Quantifying and Exploiting Speech Memory for the Improvement of Narrowband Speech Bandwidth Extension

Amr H. Nour-Eldin

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

November 2013

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

to Dina, Hana, and the hometown of Euclid

Abstract

Since its standardization in the 1960s, the bandwidth of traditional telephony speech has been limited to the 0.3–3.4 kHz range. Such *narrowband* speech exhibits not only a quality that is noticeably inferior to its wideband counterpart, but also reduced intelligibility especially for consonant sounds. Wideband speech reconstruction through artificial *bandwidth extension* (BWE) attempts to regenerate the *highband* frequency content above 3.4 kHz in the receiving end, thereby providing backward compatibility with existing networks. Although BWE has been the subject of considerable research, BWE schemes have primarily relied on *memoryless mapping* to capture the correlation between narrowband and highband spectra. In this thesis, we investigate exploiting *speech memory*—in reference to the long-term information in segments longer than the conventional 10–30 ms frames—for the purpose of improving the cross-band correlation central to BWE.

With speech durations of up to 600 ms modelled through *delta features*, we first quantify the correlation between long-term parameterizations of the narrow and high frequency bands using *information-theoretic* measures in combination with statistical modelling based on Gaussian mixture models (GMMs) and vector quantization. In addition to showing that the inclusion of memory can indeed increase *certainty* about highband spectral content in joint-band GMMs by over 100%, our information-theoretic investigation also demonstrates that the gains achievable by such acoustic-only memory inclusion saturate at, roughly, the syllabic duration of 200 ms—thereby coinciding with similar findings to the same effect in earlier works studying the long-term information content of speech.

To translate the highband certainty gains achievable by memory inclusion into tangible BWE performance improvements, we subsequently propose two distinct and novel approaches for *memory-inclusive* GMM-based BWE where highband spectra are reconstructed given narrowband input by minimum mean-square error estimation. In the first approach, we incorporate delta features into the feature vector representations whose underlying cross-band correlations are to be modelled by joint-band GMMs. Due to their non-invertibility, however, the inclusion of delta features into the parameterization frontend in lieu of some of the conventional static features imposes a *time-frequency information tradeoff*. Accordingly, we propose an empirical optimization process to determine the optimal allocation of available dimensionalities among static and delta features such that the certainty about static highband content is maximized. Requiring only minimal

modifications to our memoryless BWE baseline system, integrating *frontend-based memory inclusion* optimized as such results in performance improvements that, while modest, involve no increases in extension-stage computational cost nor in training data requirements, thereby providing an easy and convenient means for exploiting speech dynamics to improve BWE performance.

In our second approach, we focus on modelling the high-dimensional distributions underlying sequences of joint-band feature vectors as an alternative to the frontend dimensionality-reducing transform used in our first approach above. To that end, we extend the GMM framework by presenting a novel training approach where sequences of past frames are progressively used to estimate the parameters of high-dimensional *temporally-extended* GMMs in a *tree-like time-frequency-localized* fashion. By breaking down the infeasible task of modelling high-dimensional distributions into a series of localized modelling operations with considerably lower complexity and fewer degrees of freedom, our proposed tree-like extension algorithm circumvents the complexities associated with the problem of GMM parameter estimation in high-dimensional settings. Incorporating novel algorithms for *fuzzy GMM-based clustering* and *weighted Expectation-Maximization*, we also attempt to present our proposed temporal-based GMM extension approach in a manner that emphasizes its wide applicability to the general contexts of source-target conversion and high-dimensional modelling. By integrating temporally-extended GMMs into our memoryless BWE baseline system, we show that our model-based memory-inclusive BWE technique can outperform not only our first frontend-based approach, but also other comparable and oft-cited model-based techniques in the literature. Although this superior BWE performance is achieved at a significant increase in extension-stage computational costs, we nevertheless show these costs to be within the typical capabilities of modern communication devices such as tablets and smart phones.

Sommaire

Depuis sa normalisation dans les années 1960, la bande passante traditionnelle de la téléphonie de la parole a été limitée à la gamme de 0,3 à 3,4 kHz. Cette téléphonie de la parole *à bande étroite* présente non seulement une qualité évidemment inférieure à sa contrepartie large bande, mais aussi une intelligibilité réduite, en particulier pour les sons consonnes. La reconstruction de la parole à large bande à travers l'*extension* artificielle *de la bande passante* (EBP) essaye de régénérer la *bande passante à haute fréquence* au-dessus de 3,4 kHz au niveau du récepteur, ce qui permet la rétrocompatibilité avec les réseaux existants. Bien que l'EBP a fait l'objet de nombreuses recherches, les travaux proposés ont principalement utilisé une *cartographie sans mémoire* pour modéliser la corrélation entre les spectres à bande étroite et ceux à haute fréquence. Dans cette thèse, nous étudions l'exploitation de la *mémoire vocale* en référence à l'information à long terme dans des segments plus longs que les cadres conventionnels de 10–30 ms; ceci est dans le but d'améliorer la corrélation inter-bande capitale pour l'EBP.

Focalisant sur des durées de parole modélisées jusqu'à 600 ms par des *coefficients delta*, nous quantifions d'abord la corrélation entre les paramétrisations à long terme des bandes à bases et hautes fréquences en utilisant *la théorie de l'information* et la modélisation statistique basée sur des modèles de mélanges Gaussiens (GMMs) ainsi que la quantification vectorielle. En plus de montrer que l'inclusion de la mémoire peut en effet augmenter *la certitude* sur le contenu spectral de la haute bande dans des GMMs de bandes jointes de plus de 100%, notre étude démontre également que les gains réalisables par une telle inclusion sature, à peu près, à la durée syllabique de 200 ms—ce qui coïncide avec des résultats similaires réalisés avec des travaux précédentes concernant l'information à long terme de la parole.

Afin de transformer ces gains théoriques de certitude sur la bande haute à des améliorations tangibles en performance de l'EBP, nous proposons ensuite deux nouvelles approches pour l'EBP *avec mémoire* qui sont basées sur des GMMs et où les spectres à haute bande sont reconstruits, sachant ceux de la bande étroite, par l'estimation de l'erreur quadratique moyenne. Dans la première approche, nous incorporons des coefficients delta dans les représentations vectorielles dont les corrélations inter-bandes sont modélisées par des GMMs de bandes jointes. En raison de la non-inversibilité des coefficients delta, cependant, remplaçant les paramètres statiques classiques par de tels coefficients delta impose

un *compromis d'information temps-fréquence*. En conséquence, nous proposons un processus d'optimisation empirique pour déterminer l'allocation optimale des dimensionnalités disponibles parmi les paramètres statiques et coefficients delta de sorte que la certitude sur le contenu statique de la haute bande est maximisée. Ne nécessitant que des modifications minimes à notre système de base d'EBP sans mémoire, l'intégration de la mémoire optimisé de cette manière dans la paramétrisation entraîne des améliorations de performances qui, bien que modestes, n'impliquent aucune augmentation du coût de calcul associé à l'étape d'extension, ni des besoins de données de formation, offrant ainsi un moyen facile et pratique pour exploiter les caractéristiques dynamiques de la parole afin d'améliorer les performances d'EBP.

Dans notre deuxième approche, nous nous concentrons sur la modélisation des distributions de dimensionnalités élevées qui sous-tendent des séquences de vecteurs de paramètres de bandes conjointes, plutôt que d'utiliser une transformation pour la réduction de la dimensionnalité de paramétrisation comme suivie dans notre première approche. À cette fin, nous étendons le cadre de GMMs en présentant une nouvelle approche d'apprentissage où les séquences des cadres passés sont progressivement utilisées afin d'estimer les paramètres des GMMs de dimensionnalités élevées qui sont *temporellement étendus* d'une manière *arborescente* et *localisée en temps-fréquence*. En décomposant la tâche irréalisable de modélisation des distributions de dimensionnalités élevées en une série d'opérations de modélisation localisée qui exigent une complexité considérablement plus faible avec des degrés de liberté moindre, notre algorithme d'extension arborescente contourne les complexités liées aux problèmes de l'estimation des paramètres des GMMs de dimensionnalités élevées. En plus d'incorporer des nouveaux algorithmes pour le *regroupement floue basé sur des GMMs* et le *Espérance-Maximisation pondéré*, nous tentons également de présenter notre approche d'extension temporelle des GMMs en soulignant sa large applicabilité aux contextes généraux de la transformation source-cible et de la modélisation en dimensionnalités élevées. En intégrant des GMMs temporellement étendus dans notre système de base d'EBP sans mémoire, nous montrons que cette technique d'EBP avec mémoire modelisée peut surpasser non seulement notre première approche basée sur les coefficients delta, mais aussi d'autres techniques souvent citées dans la littérature. Bien que cette performance supérieure est réalisée au coût d'une augmentation significative des calculs associés à l'étape d'extension, nous démontrons néanmoins que ces coûts sont conformes aux capacités typiques des appareils de communication modernes tels que les tablettes et les téléphones intelligents.

Acknowledgments

I would like to express my gratitude to the many people without whom this thesis could not have been possible. First, I would like to thank my supervisor, Prof. Peter Kabal, to whom I am much indebted for his guidance, support, and continued mentorship throughout the many years it took to complete this work. I also thank Prof. Fabrice Labeau and Prof. Richard Rose, whose advice and comments during my research proposal were invaluable in shaping the work presented here. I would also like to express my many thanks to my colleagues, and foremost my friends, Imen Demni and Hany Kamal, for their help with the *Sommaire*; and Amr El-Keyi, Hafsa Qureshi, Joachim Thiemann, Qipeng Gong, Turaj Z. Shabestary, and the late Yasheng Qian, for their input and helpful discussions, inside the lab and outside.

Finally, special thanks go to my family for their support. To Dina, whose support, motivation, and patience were, and continue to be, unlimited and unconditional, I express my utmost gratitude.

Contents

List of Figures

List of Tables

List of Acronyms

ACR	**A**bsolute **C**ategory **R**ating
AR	**A**uto-**R**egressive
ASR	**A**utomatic **S**peech **R**ecognition
(E)BP-MGN	(**E**qualized) **B**and**P**ass-**M**odulated **G**aussian **N**oise
BWE	**B**and**W**idth **E**xtension
CCR	**C**omparison **C**ategory **R**ating
DCR	**D**egradation **C**ategory **R**ating
(I)DCT	(**I**nverse) **D**iscrete **C**osine **T**ransform
DSR	**D**istributed **S**peech **R**ecognition
DTW	**D**ynamic **T**ime **W**arping
EM	**E**xpectation-**M**aximization
(I)FFT	(**I**nverse) **F**ast **F**ourier **T**ransform
FIR	**F**inite **I**mpulse **R**esponse
FLOP(s)	**FL**oating-point **OP**eration(**s**)
GMM(s)	**G**aussian **M**ixture **M**odel(**s**)
HMM(s)	**H**idden **M**arkov **M**odel(**s**)
i.i.d.	**i**ndependent & **i**dentically **d**istributed
KLT	**K**arhunen-**L**oève **T**ransform
LDA	**L**inear **D**iscriminant **A**nalysis
LPC(s)	**L**inear **P**rediction **C**oding/**C**oefficent(**s**)
LSD	**L**og-**S**pectral **D**istortion
LSF(s)	**L**ine **S**pectral **F**requency(ie**s**)
MBE	**M**ulti-**B**and **E**xcitation
MFCC(s)	**M**el-**F**requency **C**epstral **C**oefficient(**s**)
MI	**M**utual **I**nformation

ML	**M**aximum **L**ikelihood
MLP(s)	**M**ulti-**L**ayer **P**erceptron(**s**)
(C|D)MOS	(**C**omparison|**D**egradation) **M**ean **O**pinion **S**core
MRS	**M**ean-**R**oot-**S**quare
(M)MSE	(**M**inimum) **M**ean-**S**quare **E**rror
MUSHRA	**MU**ltiple **S**timuli with **H**idden **R**eference and **A**nchor
PCM	**P**ulse **C**ode **M**odulation
pdf	**p**robability **d**ensity **f**unction
PESQ	**P**erceptual **E**valuation of **S**peech **Q**uality
PSTN	**P**ublic **S**witched **T**elephone **N**etwork
RMS	**R**oot-**M**ean-**S**quare
SNR	**S**ignal-to-**N**oise **R**atio
SPL	**S**ound **P**ressure **L**evel
SQ	**S**calar **Q**uantization
STC	**S**inusoidal **T**ransform **C**oding
VOT	**V**oice **O**nset **T**ime
VQ	**V**ector **Q**uantization

Chapter 1

Introduction

The thesis presented herein concerns the artificial extension of traditional telephony speech bandwidth for the purpose of improving quality and intelligibility.[1] In particular, we focus on quantifying and exploiting speech memory to improve bandwidth extension performance. Speech memory comprises the well-known dynamic spectral and temporal properties of speech. Such properties account for a significant portion of the information content of speech. To some extent, speech memory has been successfully exploited to improve performance in fields such as speech coding and automatic speech recognition using short-term speech memory (few tens of milliseconds). For the most part, however, bandwidth extension of telephony speech has continued to rely on the conventional memoryless static representation of speech. A few exceptions show improved extension performance but, nevertheless, only make use of short-term speech memory. In this work, we quantify and demonstrate the importance of long-term speech memory for bandwidth extension, and propose techniques to translate the benefits of memory into tangible performance improvements.

This introductory chapter lays the background necessary for our work. We first describe the effects of the bandwidth limitations of traditional telephony on speech quality and intelligibility by studying the spectral characteristics of speech sounds and their role in speech perception. We then review the extent and the nature of the spectral and temporal dynamics of speech. Such an understanding of the dynamic nature of speech is central to our work. Indeed, it is that dynamic nature that we attempt to quantify and exploit through modelling speech memory. In our experience, previous bandwidth extension work

[1] Speech quality refers to the quality of a reproduced speech signal with respect to the amount of audible distortions, while speech intelligibility refers to the probability of correctly identifying meaningful speech sounds.

lacks a review of the relationships between speech phonetics and their acoustic realizations, despite the fact that bandwidth extension attempts to improve speech perception (the interpretation of phonetic speech qualities) through enhancing speech acoustically (reconstructing spectral content). Similarly, descriptions of the dynamic characteristics of speech and their significance for perception are typically inadequate or omitted in bandwidth extension works. As such, the reviews presented in this chapter can themselves be viewed as a contribution. Finally, we conclude the chapter by introducing the concept of bandwidth extension as an alternative to wideband speech coding, and describe the scope, contributions, and organization of this thesis.

1.1 The Motivation for Bandwidth Extension

The telephone system can easily be regarded as one of man's most successful inventions. It provided the spark from which our twenty-first century intricate and vast communication networks evolved. This resounding success lies in the ability to communicate speech—the most natural and convenient means of human communication—over great distances with little to no delay. As a speech communication system, the performance of telephony over the public switched telephone network (PSTN[2]) is subjectively measured in terms of perceived speech quality and intelligibility. While the relations of quality and intelligibility to the various physical properties of a speech communication system are complex and still not fully known, acoustic frequency response and bandwidth are considered the most important among a system's physical variables [1, 2].

1.1.1 Bandwidth of traditional telephony

Since its inception in 1876 by Alexander Graham Bell [3], the telephone system has undergone many technological advances. The first telephones had no network but were in private use, connected together in pairs. Each user needed as many telephone sets as the number of different people to be connected to. Soon, however, telephones took advantage of the exchange principle already employed in telegraph networks. Each telephone was connected to

[2]While the term "PSTN" technically refers to the whole telephone network which has evolved to include many technologies with different bandwidths, "PSTN" and "POTS" (plain old telephone service) have been interchangeably used in the literature to refer to the traditional analog/copper technology. In the sequel, we exclusively use "PSTN" to refer to traditional 300–3400 Hz telephony.

a local telephone exchange, and the exchanges were connected together with trunks. Networks were connected together in a hierarchical manner until they spanned cities, countries, continents and oceans. Notable advances include the introduction of pulse dialing, followed by more sophisticated address signaling including multi-frequency signalling—later evolving to the modern dual-tone multi-frequency signalling (or Touch-Tone)—as well as the use of time-division multiplexing to increase the capacity of communication links. The most important improvement to the PSTN, however, was the digitization of telephony speech using pulse code modulation (PCM) [4].

Despite these advances, the acoustic frequency characteristics of the PSTN have remained, interestingly enough, virtually unchanged. While most automated telephone exchanges and trunks now use digital rather than analog switching, analog two-wire circuits are still used to connect the last mile from the exchange to the end-user's telephone (also called the local loop). The analog audio signal from a calling party is digitized at the exchange at a sampling rate of 8 kHz using 8-bit μ- or A-law PCM, routed and transmitted over the network to the called party after passing through a digital-to-analog converter at the destination's exchange.

In designing the frequency response characteristics of the nascent analog telephone network, telephone companies needed to balance the requirements of perceived quality and intelligibility (as understood in the early twentieth century) with the economic viability associated with building and expanding the network to cover large areas and as many subscribers as possible.[3] In the early days of the telephone network, limitations of analog circuitry and channel multiplexing techniques were the chief reasons for limiting the telephone bandwidth to as low as 2.5 kHz (or 2500 *cycles*, by that era's nomenclature). At the lower end of the spectrum, the problems of crosstalk due to AC coupling of telephone wires as well as interference from AC mains frequency were the main concerns.[4] Thus, a cutoff frequency in the lower end of the spectrum was required while ensuring a minimum level of naturalness and intelligibility.

> It was concluded in 1930 [5] that, "*based on tests showing the effect upon articulation of varying the upper and lower cutoff frequencies*", there was "*lit-*

[3] As put by Martin in 1930 [5, page 483]; "*In setting up the requirements for the various transmission characteristics of telephone message circuits, the aim is to arrive at the combination of requirements which will give the most economic telephone system for furnishing the desired grade of transmission service.*"

[4] It was already understood by 1925 that speech contains frequencies as low as 60 *cycles* [6, page 547].

tle effect on articulation of cutoffs below 400 cycles". At the higher end of the spectrum, it was concluded that, "*while there is some articulation advantage in going further than 2750 cycles, observations of the number of repetitions occurring in conversations over circuits having different cutoff frequencies have indicated but little reduction in repetitions by going beyond about 2750 cycles with commercial types of terminal sets*". Furthermore, "*the extension necessary to effect a material improvement in naturalness—largely as the result of better reproduction of the fricative consonants and some of the incidental sounds which accompany speech—is a matter of a thousand cycles or more, rather than hundreds of cycles*". Consequently, "*it has been considered that such an extension for message circuits is not now justified*", especially when bearing in mind that "*an extension of the transmission range will in general increase the amount of noise on the circuit and magnify the crosstalk problem*", while also "*increasing the difficulties of securing proper impedance balances and of equalizing amplitude and phase distortion*". Ultimately, the conclusion in [5] was that "*new designs of telephone message circuits for the Bell System should have an effective transmission band width of at least 2500 cycles, extending from about 250 to 2750 cycles*". With advances in circuitry and multi-channel carrier systems, it was concluded a few years later in 1938 that "*a 3000-cycle band properly used gives good transmission both in articulation and naturalness*" [7, page 373].

The bandwidth of the PSTN was eventually standardized in the 1960s by the CCITT[5] to the 300–3400 Hz range. The most recent ITU-T standards specifying frequency characteristics of the telephone channel are G.232 [8] (giving equipment design objectives for analog 12-channel terminal equipment), and G.712 [9] (giving equipment design objectives for digital PCM channelizing equipment). Figure 1.1, reproduced from [8], illustrates the recommended range of power level attenuation across frequency. Such illustrations, are often referred to as frequency *masks*.

[5]The Comité Consultatif International Téléphonique et Télégraphique (CCITT) is one of the three sectors of the International Telecommunication Union (ITU). CCITT was renamed in 1992 to ITU-T (ITU Telecommunication Standardization Sector).

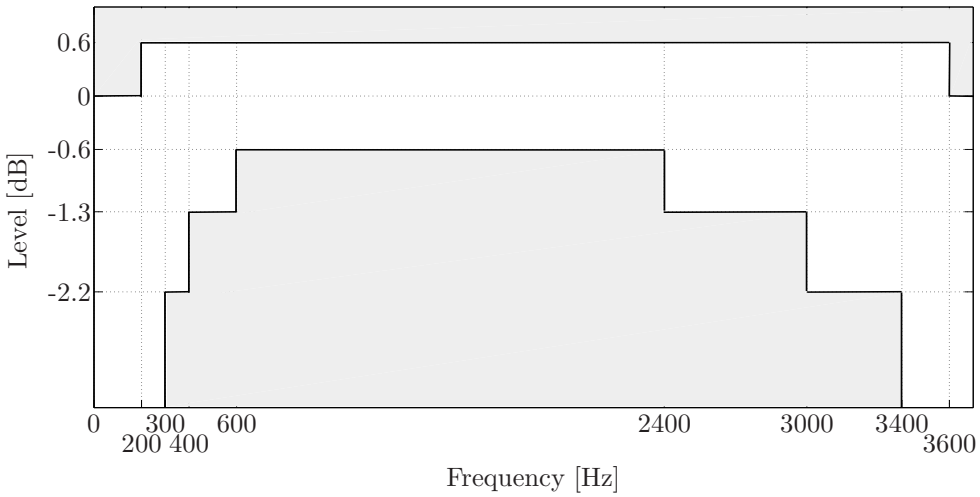


Fig. 1.1: Allowable limits for the variation, as a function of frequency, of the relative power level at the output of the sending or receiving equipment of any channel of a 12-channel (analog) terminal. Figure 2/G.232 [8]

1.1.2 Speech production

The frequency characteristics of speech sounds are a direct consequence of the physical properties of the speech production organs of the vocal tract[6]. Sounds can be acoustically classified according to two main physical aspects of sound production: (a) vibration of the vocal folds, and (b) manner and place of airflow constriction (articulation) in the vocal tract. Vibration of the vocal folds, or voicing, results in periodic signals with energy concentrated at harmonics of the fundamental frequency of vibration, F0, while unvoiced sounds are aperiodic. Constriction of the airflow at any of the vocal tract articulators results in consonants, while airflow is relatively unimpeded for vowels. The shape of the vocal tract (manner of articulation) and the place of airflow constriction, along with periodicity, determine the frequency characteristics of sounds. In general, sounds have energy peaks at formants—the resonant frequencies of the vocal tract—with the first three formants—F1, F2 and F3—generally ranging from 250–3300 Hz [10, Section 3.4]. Secondly, the degree of airflow constriction determines whether the consonant's spectrum is predominantly that of noise (as in unvoiced fricatives, plosives, and affricates), or similar to vowels (as in diphthongs, glides, liquids, and nasals), or a mixture of both (voiced fricatives, plosives, and affricates). Table 1.1 lists the properties of the English phonemes.

[6]Namely the lungs, vocal folds (or cords), tongue, lips, teeth, velum, and, indirectly, the jaw.

Table 1.1: English phonemes (using IPA—international phonetic alphabet—symbols) and corresponding features [10, Table 3.1].

Manner of articulation	Phoneme	Place of articulation	Voicing	Example word
Vowels	i	high front tense	yes	b**ea**t
	ɪ	high front lax	yes	b**i**t
	e	mid front tense	yes	b**ai**t
	ɛ	mid front lax	yes	b**e**t
	æ	low front tense	yes	b**a**t
	ɑ	low back tense	yes	c**o**t
	ɔ	mid back lax rounded	yes	c**au**ght
	o	mid back tense rounded	yes	c**oa**t
	ʊ	high back lax rounded	yes	b**oo**k
	u	high back tense rounded	yes	b**oo**t
	ʌ	mid back lax	yes	b**u**t
	ɝ	mid tense (retroflex)	yes	c**ur**t
	ə	mid lax (schwa)	yes	**a**bout
Diphthongs	ɑj (ɑɪ)	low back → high front	yes	b**i**te
	ɔj (ɔɪ)	mid back → high front	yes	b**oy**
	ɑw (ɑʊ)	low back → high back	yes	b**ou**t
Glides	j	front unrounded	yes	**y**ou
	w	back unrounded	yes	**w**ow
Liquids	l	alveolar	yes	**ll**ull
	r	retroflex	yes	**r**oa**r**
Nasals	m	labial	yes	**m**ai**m**
	n	alveolar	yes	**n**o**n**e
	ŋ	velar	yes	**ng**
Fricatives	f	labiodental	no	**ff**luff
	v	labiodental	yes	**v**alve
	θ	dental	no	**th**in
	δ	dental	yes	**th**en
	s	alveolar sibilant	no	**s**a**ss**
	z	alveolar sibilant	yes	**z**oo**s**
	ʃ	palatal sibilant	no	**sh**oe
	ʒ	palatal sibilant	yes	mea**s**ure
	h	glottal	no	**h**ow
Plosives	p	labial	no	**p**o**p**
	b	labial	yes	**b**i**b**
	t	alveolar	no	**t**o**t**
	d	alveolar	yes	**d**i**d**
	k	velar	no	**k**i**ck**
	g	velar	yes	**g**i**g**
Affricates	tʃ	alveopalatal	no	**ch**ur**ch**
	dʒ	alveopalatal	yes	**j**u**dg**e

More importantly for telephone communications, the distribution of sound energy across frequency generally depends on the excitation source generating the sound. For *sonorants*, voiced sounds where the vocal folds excite the full length of the vocal tract, energy is concentrated at the lower frequencies. Vowel energy, in particular, is primarily concentrated below 1 kHz near the low formant. Unvoiced sounds, on the other hand, are characterized by a major vocal tract constriction acting as the excitation to the shorter anterior portion of the vocal tract, thus concentrating energy at the higher frequencies. Energy in unvoiced fricatives, for example, is concentrated above 2.5 kHz. Voiced fricatives have a double acoustic source, resulting in a mixed energy distribution with features of both voiced and unvoiced sounds.

1.1.3 Effect of the telephone bandwidth on perceived quality and intelligibility

Although the long-term average speech spectrum shows speech energy to be mainly concentrated in vowels below 1 kHz [11], the full spectrum of speech sounds plays a crucial role in quality (naturalness) and intelligibility. Speech frequencies range from as low as 60 Hz (frequency of vocal fold vibration for a large man) to over 15 kHz. Consequently, *narrowband* speech—speech limited to the 300–3400 Hz PSTN band—lacks many of the distinctive frequency characteristics of some sounds.

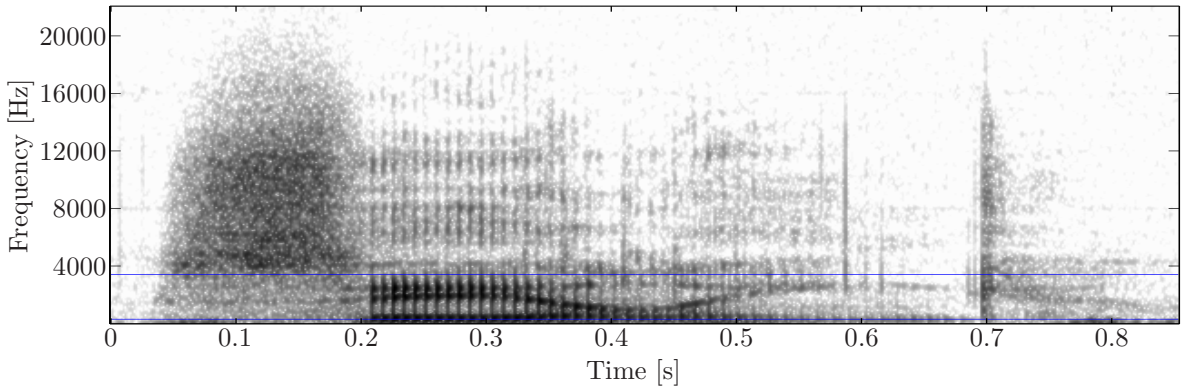
1.1.3.1 Spectral characteristics of speech sounds

Consonants, the sounds most important for intelligibility,[7] are also the sounds most negatively impacted by the bandwidth limitations of telephony. Energy for fricatives is primarily concentrated above 2.5 kHz. Labial and dental fricatives—/f/, /v/, /θ/ and /δ/ (also referred to as *nonsibilants*[8])—have relatively low energy compared to the sibilant alveolar and palatal fricatives—/s/, /z/, /$\int$/ and /ʒ/—due to a very small front cavity [13]. Sibilants are characterized by relatively steep high-frequency spectral peaks, while nonsibilants are characterized by relatively flat and wider band spectra. Alveolar sibilants, /s/ and

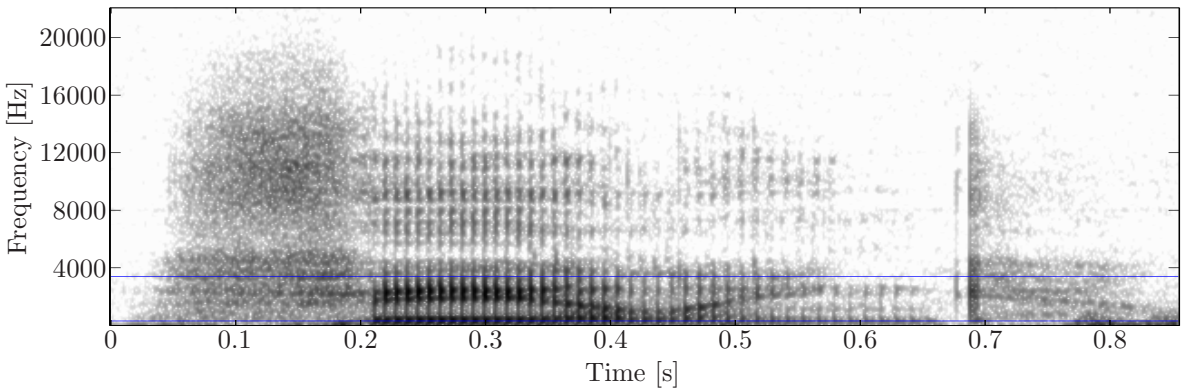
[7] The importance of consonants for intelligibility was measured as early as 1917. Crandall concluded in [12, page 75] that: "*The interesting thing, in the energy distribution in speech, is that the vowels are the determining factors of this distribution, whereas the consonants are the determining factors in the matter of importance to articulation. The importance of the consonant frequencies in speech is thus utterly out of proportion to the amount of energy associated with them.*"

[8] The alveolars /s/ and /z/, and the palatals /$\int$/ and /ʒ/, are called *sibilants* due to their hissing or shushing quality.

/z/, lack significant energy below 3.2 kHz [10, Section 3.4.6], and are distinguished from the palatal sibilants, /ʃ/ and /ʒ/, by the location of their lowest spectral peak which is around 4 kHz for the alveolars and 2.5 kHz for the palatals for a typical male speaker [13]. The PSTN bandwidth, thus, removes all spectral distinction between alveolar sibilants and nonsibilant fricatives, resulting in the well-known difficulty of distinguishing such fricatives in telephony speech (particularly the /s/ and /f/ pair). Figure 1.2 clearly illustrates this problem by comparing the spectrograms of the two words *sailing* and *failing*, showing the effect of the 300–3400 Hz PSTN bandwidth limitation in virtually removing the distinctive spectral features of /s/ and /f/—represented mostly by the higher energy above 3.4 kHz for the fricative /s/—in the 20–200 ms interval.



(a) *sailing*



(b) *failing*

Fig. 1.2: Spectrograms of the two words *sailing* and *failing* showing the effect of the PSTN bandwidth limitation on the /s/ and /f/ fricatives in the 20–200 ms interval. The boundaries of the telephone channel are marked by the two lines at 300 and 3400 Hz.

At the lower end of the spectrum, voiced fricatives are often differentiated from unvoiced ones, at least for syllable-initial fricatives, by the presence of energy at the fundamental and low harmonics (a *voice bar* on spectrograms) due to vocal cord vibration [10, Section 5.5.3]. Average F0 values for males and females, however, are 132 Hz and 223 Hz, respectively [14], i.e., below the lower 300 Hz cutoff frequency, leading to some ambiguity in perceiving voicing in fricative pairs: /s/ and /z/, /f/ and /v/, /θ/ and /δ/, and /$\int$/ and /$\jmath$/.

Plosives (or stops) are the second class of consonants adversely affected by the 300–3400 Hz bandwidth limitation. Plosives consist of a complete occlusion of the vocal tract followed by a brief (a few ms) burst of noise then longer frication at the opening constriction [10, Section 3.4.7]. For voiced stops, a voice bar of energy confined to the first few harmonics of the fundamental frequency may be present during the closure portion. As described above, since the average fundamental frequencies are below the lower 300 Hz cutoff, such voice bars separating voiced stops from unvoiced ones are usually removed or attenuated. The initial noise burst following release of the vocal tract occlusion primarily excites frequencies of fricatives having the same place of articulation. Hence, the burst release energy of alveolar stops, /t/ and /d/, usually peaks near 3.9 kHz (coinciding with the spectral peak at 4 kHz for the alveolar fricatives, /s/ and /z/). Labial stops, /p/ and /b/, also have similar burst release properties but—in a manner similar to the difference between labial/dental fricatives and alveolar ones—are distinguished from alveolar bursts by being considerably less intense (about 12 dB weaker). The loss of such plosive characteristics due to the bandwidth limitation of the PSTN, leads to significantly diminished intelligibility and naturalness for stops. The acoustics of affricates resemble those of the constituent stop+fricative sequences.

Similarly, the intelligibility of nasals is adversely affected by the lower 300 Hz cutoff frequency of the telephone bandwidth as the spectra of nasals are dominated by the first formant (the *nasal murmur*) occurring near 250 Hz.

In contrast, vowel intelligibility is largely unaffected by the higher 3400 Hz frequency cutoff as vowel energy is primarily concentrated below 1 kHz. Furthermore, the first three formants—crucial for vowel intelligibility—fall mostly within the telephone bandwidth [14]. However, while almost irrelevant for intelligibility when compared to higher frequencies, frequency content below 300 Hz is important for naturalness [10, Section 4.3.2]. As such, the lack of frequency information below 300 Hz for vowels in particular, and all sounds in general, is an important limitation distinctive of the *toll quality* of telephony speech.

1.1.3.2 Effect of bandwidth on speech intelligibility

Since the early days of the Bell Telephone Laboratories, significant efforts have been made to understand and quantify the effects of the telephone channel—particularly its bandwidth limitations—on speech intelligibility. Between 1910 and 1918, Campbell [15] and Crandall [12] were the first to use *articulation tests* and proposed the idea that speech intelligibility is based on the sum of the contributions from individual frequency bands. Building on this work, Fletcher extended the analysis in 1921 to account for the effects of filtering speech into 20 bands extending to 7 kHz [16, 17]. In particular, Fletcher first derived relations for *articulation*—the probability of correctly identifying nonsense speech sounds spoken with syllables [18]—as a function of speech frequency and SNR, then later extended the relations to include the intelligibility of words and sentences [11, 19]. Fletcher showed that while *no detectable loss in articulation results until the lower cutoff is raised to 250 cycles, or until the upper one is lowered to 7000 cycles* [20], limiting telephony speech bandwidth to the 300–3400 Hz range causes syllable articulation to drop from 98–99% to 89–92%,[9] although whole-sentence intelligibility only drops negligibly from 99.9% to 99.3% [19].[10] More recently, however, it has been shown that the effect of obstruent consonants—as described above, these are the sounds with energy concentrated mostly at the high frequencies near or above the 3400 Hz cutoff, i.e., fricatives, plosives, and affricates—on word and sentence intelligibility is quite higher than suggested by Fletcher's sentence intelligibility scores. In [22], for example, replacing obstruents in fluent speech by white noise results in 87% intelligibility for words and only 60% for sentences. The figures drop to 82% and 50% for words and sentences, respectively, when using periodic noise (sinusoids with frequencies ranging from 200 to 4000 Hz) as replacement. French's method [11] for the calculation of articulation—a simpler version of Fletcher's [19] that later became known as the *Articulation Index theory*—was standardized by the ANSI[11] in 1969 [23], then updated and renamed in 1997 to the *Speech Intelligibility Index* (or SII) [24].

[9] Using Table XII in [19] which lists values for the articulation index, A_f, as a function of the frequency *importance* function, D, the articulation index for the 300–3400 Hz band is determined as $A_f = \int_0^{3400} D\,df - \int_0^{300} D\,df \cong \int_{310}^{3390} D\,df = 0.74$. Table III is then used to arrive at the corresponding articulation values for sounds, syllables, and simple sentences.

[10] Since Fletcher's sentence intelligibility scores were based on binary right-or-wrong answers to interrogative or imperative sentences—rather than scoring sentences based on whether all words were correctly recognized—[16], the reliability of Fletcher's sentence intelligibility figures has been questioned, as in [21], for example.

[11] American National Standards Institute.

1.1.3.3 Effect of bandwidth on speech quality

With the advent of PCM in 1949 [4] based on Shannon's proof of the Sampling Theorem [25],[12] speech digitization proliferated all means of speech communication, particularly that of telephony. Digital speech transmission generally involves loss of information due to quantization and channel noise, resulting in the degradation of output speech. While quality degradation due to channel noise can be overcome by error detection and correction techniques, such techniques typically require bit protection overhead, and hence, lead to an overall bit rate increase. As more efficient transmission requires lower bit rates, understanding the importance of the different frequency bands for perceived quality is, thus, of crucial importance for speech coder design in general, and particularly for the PSTN bandwidth of 300–3400 Hz. Such an understanding allows more efficient transmission either through frequency-dependent bit allocation in frequency-domain coding, or through compromising between bandwidth (through the sampling rate) and bit protection in time-domain coding.

The subjective experiments of Voran in [27] provide an important investigation into the effects of coding bandwidth on perceived quality. In the absence of coding distortions, the perceptual quality of several passbands of varying bandwidths is compared to the 300–3400 Hz passband of narrowband speech. Most notably, the *wideband* G.722 ITU-T standard passband of 50–7000 Hz [28]—the largest bandwidth in the study—is shown to be perceptually superior to the traditional 300–3400 narrow band by 36%, relatively (1.42 points on a custom 7-point subjective scoring scale).[13] The study also shows that, while keeping bandwidth fixed, shifting passbands downwards by extending them below 300 Hz at the expense of higher frequencies results in improved quality, but only up to a certain limit that varies depending on bandwidth. In other words, up to a point that varies with bandwidth, the perception gained by additional low frequency content seems to outweigh the perception loss due to removed high frequency content. Thus, while the results of [27] confirm the importance of frequencies below the PSTN's lower 300 Hz frequency cutoff for perceptual quality, they also indirectly demonstrate the importance of different frequency subbands outside the narrowband range relative to each other. For example, the 0.8 bark subband

[12] The origins of the Sampling Theorem can be traced back to Borel as far back as 1897. Several authors have independently published essentially the same ideas between Borel in 1897 and Shannon's 1949 proof; including, Ogura, Nyquist, Whittaker, Raable, Someya, Kotelnikov, and Weston [26].

[13] In [27], listeners score a test recording against a narrowband reference by selecting one of the seven options: The second version sounds *much better than* (3), *better than* (2), *slightly better than* (1), *the same as* (0), *slightly worse than* (−1), *worse than* (−2), *much worse than* (−3), the first version.

of 3400–3889 Hz is about 5% more perceptually important than the 0.8 bark subband of 50–131 Hz, and 7% more important for quality than the 0.8 bark range of 4691–5362 Hz.[14] Finally, an interesting result of [27] is that extending the upper limit from 3400 Hz to 7000 Hz appears to be effective perceptually only when the lower 300 Hz limit is extended downwards as well, suggesting a complex nonlinear inter-band relationship between subbands and perceived quality—in contrast to the additive nature of the relationship between subbands and the articulation index. In particular, extending the upper 3400 Hz limit alone results in a maximum 4% perceptual improvement,[15] but the same highband extension, however, results in 12% improvement when applied to speech where the lower limit has already been extended down to 50 Hz.

Other works investigating the effects of coding bandwidths on perceived quality agree that wider bandwidths outperform the traditional PSTN bandwidth in terms of perceived quality, although with varying results as to the extent of differences in quality. In [29], for example, MOS[16] values for 10 and 7 kHz speech are 4 and 3.6, respectively, compared to only 2.5 for 3.6 kHz speech. In [31], the DMOS[17] values—using 15 kHz reference speech—for 10 and 7 kHz speech are 4.2 and 3.4, respectively, compared to only 1.9 for 3.6 kHz speech.

The works described above clearly demonstrate the perceptual superiority of wideband speech over narrowband telephony speech in terms of both quality and intelligibility. To conclude this section, we note, however, that intelligibility—although adversely affected by the PSTN bandwidth limitations—is still reasonable for all but the lowest-bit-rate coders [10, Section 7.4]. Moreover, while intelligibility only assesses the recognizability of speech sounds, quality is a multi-dimensional measure that encompasses many perceptual properties of sounds that are typically difficult to quantify, e.g., loudness, clarity, fullness, spaciousness, brightness, softness, nearness, and fidelity [1], but which constitute the perceived quality of speech. Thus, speech quality—rather than intelligibility—has been the criterion

[14] See Section 4.2.1 for more details on the perceptual Bark scale for frequency.

[15] Interestingly, while it is shown in [27] that extending the upper 3400 Hz limit to 5083 Hz results in 4% improvement in perceived quality, extension to 7000 Hz results in no discernable improvement.

[16] Mean opinion score (MOS) is the absolute average score obtained from absolute category rating (ACR) tests where listeners judge a test speech signal on a scale from 5 (best—imperceptible impairment) to 1 (very annoying) without referring to an original reference signal [10, Section 7.4]. MOS is the most prevalent among subjective quality measures. Guidelines for ACR testing methodology are specified in the ITU-T P.800 standard [30].

[17] A variant of MOS, degradation mean opinion score (DMOS) is obtained through degradation category rating (DCR) tests used for relative judgements where test speech is compared to a superior reference on a scale from 5 (inaudible differences) to 1 (very annoying); see [10, Section 7.4; 30].

typically used to assess speech coder performance [32]. Similarly, it has also been the criterion overwhelmingly used for the evaluation of artificial Bandwidth Extension (BWE) techniques, and is, thus, also the measure used in our work presented herein.

1.2 Dynamic and Temporal Properties of Speech and their Importance

The spectral characteristics of speech, described in Section 1.1.3.1 above, are relatively fixed or *quasi-stationary* only over short periods of time (few tens of milliseconds) as one sound is produced, whereas the signal varies substantially over intervals greater than the duration of a distinct sound (syllable duration is typically 200 ms, with stressed vowels averaging 130 ms and other phones about 70 ms in total). Typical phonetic events last more than 50 ms on average, but some, like stop bursts, are shorter [10, Section 6.10.1]; rapid spectral changes occur in stop onsets and releases and in phone boundaries involving a change in manner of articulation[18]. Hence, windows no more than 10–30 ms wide are typically used for speech analysis and processing, including BWE, such that quasi-stationarity is preserved as much as possible to allow coding and parameterization of speech. This conventional short-term analysis, however, ignores the considerable longer-term information integral to speech perception. Such information varies from the relatively short subtle temporal cues extending across and in between phonemes, such as phonemic duration and voice onset time (VOT), to the more obvious long-term effects of *coarticulation*[19] and the inherent inter- and intra-speaker variability on the spectral properties of speech. Coarticulation, in particular, effectively results in diffusing perceptually-important phonemic information across time, often across syllable and syntactic boundaries, at the expense of phonemic spectral distinctiveness. An even longer-term form of information underlying speech segments is that of *prosody*, referring to the suprasegmental and syntactic informa-

[18] See Table 1.1; *manner of articulation* refers to the classification of sounds depending on the general shape of the vocal tract and degree of airflow constriction into vowels, glides, liquids, diphthongs, fricatives, stops, and affricates, while *place of articulation* refers to the finer discrimination of sounds into phonemes depending on the point of narrowest vocal tract constriction.

[19] Coarticulation is attributed to the tendency to communicate speech with least effort; it requires less muscle effort to move an articulator gradually in anticipation toward a target over several phones than to force its motion into a short time span between phonemes; similarly, letting an articulator gradually return to a neutral position over several phones is easier than using a quick motion immediately after the phone that needed the articulator.

tion that extends beyond phone boundaries into syllables, words, phrases, and sentences. Since prosody mostly follows from language-specific rhythm, intonation, syntax, and semantics, however, the effects of such information on the acoustics of speech are much more subtle and less relevant to the acoustic-only BWE processing of speech than those of the temporal cues and coarticulation noted above.

To illustrate their importance as cues complementing—and often integral to—speech perception, we discuss these dynamic properties of speech in more detail in Appendix A. We note here, however, an important result of the analyses of such properties; as observed in [10, Section 5.4.2], the mapping from phones (with their varied acoustic correlates) to individual phonemes is likely accomplished by analyzing dynamic acoustic patterns—both spectral and temporal—over sections of speech corresponding roughly to syllables. Accordingly, a BWE system exploiting such long-term information—extending up to syllabic durations—as a means for better identification of the frequency content to be reconstructed will, thus, inherently improve perception of the extended speech.

1.3 Extending the Bandwidth of Telephony Speech

1.3.1 Wideband speech coding

Section 1.1 clearly illustrated the inferiority of narrowband telephony speech—in both quality and intelligibility—as a result of the detrimental effects of the bandwidth limitations of legacy telephone networks. Several new codecs have thus been introduced to achieve superior wideband speech communications. Such wideband codecs extend speech communication bandwidth to 50 Hz at the lower end and up to 7 kHz at the higher end of the spectrum. Super-wideband coders extend bandwidth to an even higher 10 and 15 kHz [29, 31], and further yet to 19.2 kHz [33]. Most notable among wideband codecs are G.722 [28] and G.722.2—otherwise known as Adaptive Multi-Rate Wideband (AMR-WB) [34]. As noted in [28, Section I.2], applications of the wideband G.722 codec, standardized in 1988, include: commentary quality channels for broadcasting purposes and high quality speech for audio and video conferencing applications. Indeed, the G.722 standard has become widely used in Voice over Internet Protocol (VoIP) telephony applications. More recent, the AMR-WB codec was introduced in 2000 and adopted by the ITU-T[20] as G.722.2 [34].

[20] See Footnote 5.

AMR-WB is increasingly pervading mobile phone devices and networks.

While such wideband codecs provide superior quality and intelligibility, their use in telephony is, nonetheless, limited by the traditional narrowband limitations ubiquitous in the PSTN. True wideband communication can only be possible if the call remains on an entirely wideband-capable network; the entire route must support digital wideband transmission, in addition to both transmitting and receiving parties. All benefits of wideband telephony are lost when routed through the PSTN. The growth of true wideband telephony thus requires modifying current networks. Hence, for clear economic reasons, existing telephony networks will continue to suffer—at least partially—the narrowband limitations for the foreseeable future, particularly when considering the prohibitive cost of replacing analog two-wire local loop connections still in use today. For a long transitional period, telephone networks will continue to be mixed with both narrowband and wideband capabilities.

1.3.2 Artificial bandwidth extension

Through reconstructing wideband speech rather than explicitly coding it, artificial bandwidth extension (BWE) of narrowband speech at the receiving end provides a network-independent alternative to wideband speech coding. Using only the narrowband input available at the receiver, BWE attempts to reconstruct wideband speech by estimating missing frequency content through modelling the correlation between narrowband speech and its *highband* counterpart. Alternatively, by modelling the correlation between narrowband speech and its original *wideband*—rather than highband—counterpart, the wideband signal can be estimated as a whole.

By using only narrowband speech, BWE provides backward compatibility with existing networks. Figure 1.3 illustrates how BWE can be easily integrated into the peripherals of the traditional PSTN. Natural speech, a super-wideband signal (denoted by s_{swb}) with frequencies extending up to 22 kHz (as shown, for example, in the spectrograms of Figure 1.2), is recorded at the transmitter, bandpass filtered, coded and transmitted across the telephone network. Typically, a sampling frequency of $F_{\text{s}} = 8\,\text{kHz}$ is used. At the receiving end, a wideband estimate, $\hat{s}_{\text{wb}}$, extending up to 7 or 8 kHz, is obtained through BWE having only narrowband speech, s_{nb}, as input.

In the work presented herein, we focus on improving BWE based on modelling the correlation between narrowband and highband frequency content. As described below and

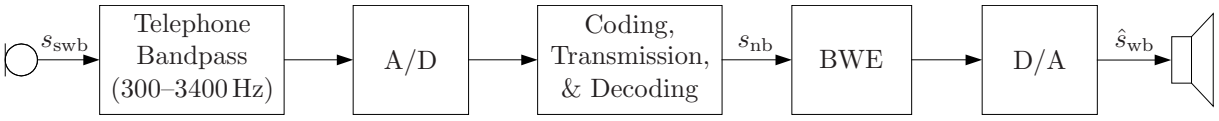


Fig. 1.3: Overall system diagram for telephone communication with bandwidth extension integrated at the receiver.

further detailed in Chapter 4, such cross-band correlation from the perspective of BWE can be quantified as the certainty about the high band given only the narrow band. As such, we use both terms—cross-band correlation and highband certainty—in the sequel synonymously.

1.4 Scope and Contributions of the Thesis

As described in Section 1.2, a significant portion of the information content in speech is carried by the dynamic spectral and temporal properties manifesting in long-term segments of speech. Indeed, exploiting these properties—instead of, or in addition to, the conventional *static* 10–30 ms parameterization of speech—has been shown to considerably improve performance in many speech processing fields, e.g., speech coding and automatic speech recognition (ASR). Examples of coding techniques exploiting speech memory include *differential coding*[21], e.g., [35], *target matching*[22] [36], and *memory vector quantization*[23]. Similarly, the use of *hidden Markov models* (HMMs) in ASR to model the temporal order of events in speech has become a de facto standard [10, Section 10.7.1].

In contrast, BWE schemes have, for the most part, primarily used *memoryless mapping* to model the correlation between narrowband and highband spectra. Exceptions to the pervasiveness of memoryless mapping in BWE are based mainly on the implementation

[21]Rather than code each frame or sample independently, differential coding makes used of short-term memory by coding only interframe differences.

[22]Target matching jointly smoothes both the residual signal and the frame-to-frame variation of linear prediction coefficients (LPCs) by matching the output of a formant predictor to a target signal constructed using smoothed pitch pulses.

[23]Memory vector quantization (VQ) incorporates knowledge of previously quantized data in the quantization process. As such, memory quantizers exploit memory between the vectors in the input process (intervector dependencies), and therefore, perform better than conventional VQ of the same dimension [37]. A common application of memory quantization methods is the quantization of spectrum parameters in linear prediction coding, e.g., [37, 38].

of highband spectrum envelope estimation using HMMs, e.g., [39], such that the dynamic properties of speech are embedded into spectrum estimation. HMM-based techniques, however, are generally marked by higher complexity and training data requirements, which increase with the number of HMM states. To mitigate the potential complexity and data insufficiency problems, first-order Markov models are assumed almost universally. This limits such HMM-based techniques to modelling the dependencies between consecutive signal frames only, effectively restricting the ability of the model to capture only 20–40 ms of memory. As described in Section 1.2, however, the information carried by speech temporally extends well beyond such 20–40 ms intra- and inter-phoneme durations. In particular, we noted that the identification of phonemes is likely accomplished by analyzing patterns with roughly syllabic durations, i.e., around 200 ms. While increasing the number of states partially alleviates the memory limitations of first-order HMMs (by modelling more longer-duration sequences of individual frames and the corresponding single-frame transitions), the inability to capture unsegmented long-term information in contiguous patterns remains. Thus, current memory-inclusive BWE techniques exploit only a fraction of the memory available in speech. Furthermore, despite the established importance of memory in speech, there have been no attempts, to the best of our knowledge, to explicitly quantify the gain of exploiting memory to improve the cross-band correlation assumption underlying the bandwidth extension of narrowband speech.

The goal of this thesis is to advance current BWE paradigms in regards to exploiting speech memory by addressing the aforementioned deficiencies. As shown in Sections 2.2 and 2.3, BWE implementations vary widely in all aspects—the properties of speech chosen for modelling in the different bands, dimensionalities and types of parameterizations used, nature of the joint-band correlation modelling employed, complexity and amounts of training data required, et cetera. As such, we strive to quantify and demonstrate the benefits of exploiting long-term speech memory in BWE conceptually and in a universal manner to imbue our theses with as much generality as possible, such that our findings can be adapted and implemented in other BWE techniques. Therefore, we focus our attention on studying the role of memory theoretically as well as the means and effects of its inclusion in practical BWE systems, rather than studying the effects of improving the various BWE implementation-specific details mentioned above. Similarly, although BWE refers, per se, to the reconstruction of lowband frequencies (< 300 Hz) as well as highband ones (> 3400 Hz), we focus on the latter in the context of studying the role of speech memory

since highband reconstruction is that which is of primary concern in bandwidth extension. Indeed, the vast majority of BWE techniques exclusively address the reconstruction of highband content, with very few works additionally addressing lowband reconstruction. Works dedicated to reconstructing only the low band are quite rare.

The contributions of our work can be summarized as follows (listed in descending order of impact in our view):

Modelling speech memory and quantifying its effects on cross-band correlation

Using parameterization-independent *delta features*, we model speech memory by explicitly parameterizing it for durations extending up to 600 ms—far greater than the indirect modelling of memory through cumulative HMM state transition probabilities of previous memory-inclusive BWE techniques. By exploiting *information-theoretic* measures to represent the correlation between narrow- and high-band speech memory thus modelled, we achieve our goal of quantifying the role of memory in increasing *certainty* about the high band. Highband certainty—the ratio of mutual information between the narrow and high bands to the discrete entropy of the high band—represents cross-band correlation normalized to the $[0, 1]$ range. By estimating highband certainty for parameterizations incorporating delta features, we are, in fact, estimating upper bounds on achievable BWE performance when memory is included in BWE. This follows from the fact that highband certainty estimation is not affected by the several components of an actual BWE system which inevitably introduce errors in reconstructing the missing high frequency content. This bounding property is demonstrated analytically by making use of a previously-derived lower bound on a common spectral distortion measure, shown to be a function of information-theoretic measures. Through highband certainty estimates, one can then determine the optimality, or lack thereof, of any BWE system incorporating memory. The ideal BWE system is that which can translate the estimated highband certainty gains into matching BWE performance improvements. Our method of modelling and quantifying memory shows that, regardless of the parameterization used, exploiting long-term memory through delta features at least doubles the cross-band correlation central to BWE, and hence, can potentially result in considerable BWE gains if efficiently made use of.

Formulation of a memory-based extension to the GMM framework

As delta features are non-invertible, they can not be directly used to reconstruct high-

band frequency content. Thus, using delta features in BWE with fixed dimensionalities results in the loss of some spectral detail as fewer invertible static parameters are available for speech reconstruction. This *time-frequency information tradeoff* provides the motivation to embed speech memory directly into the Gaussian mixture model (GMM) structure used for statistical joint-band modelling in current state-of-the-art BWE techniques. To that end, we *extend the GMM formulation to take memory into account*, presenting a novel *tree-like* training approach to estimate the parameters of *temporally-extended* GMMs. In particular, sequences of past frames are progressively used to grow high-dimensional GMMs in a tree-like fashion, effectively transforming the parameter estimation problem of such high-dimensional GMMs into a state space modelling task where the states correspond to *time-frequency-localized* regions in the full high-dimensional space underlying the modelled feature vector sequences. By breaking down the infeasible task of modelling high-dimensional distributions as such into a series of localized modelling operations with considerably lower complexity and fewer degrees of freedom, our tree-like memory-based extension of the GMM framework thus circumvents the complexities associated with the parameter estimation of GMMs in high-dimensional settings. In developing this temporal-based extension to the GMM framework, we also introduce a novel *fuzzy GMM-based clustering* algorithm, as well as a *weighted* implementation of the *Expectation-Maximization* (EM) algorithm used for GMM parameter estimation. These latter algorithms are proposed in order to maximize the information content of the aforementioned temporally-extended GMMs while ensuring that the effects of class overlap in high-dimensional spaces are reliably accounted for in our time-frequency localization approach. To emphasize their wide applicability to contexts other than that of BWE, these proposed algorithms are developed, derived, and evaluated, with the focus being on generality as feasibly possible.

Novel BWE techniques with frontend- and model-based memory inclusion

To translate the highband certainty gains achievable by the inclusion of speech temporal information into practical BWE performance improvements, we implement two GMM-based BWE techniques. The first technique employs frontend-based memory inclusion through delta features, thereby requiring minimal changes to the baseline memoryless BWE reference. As described in Section 2.3.3.4, GMMs are known for

their superior modelling of the continuous nonlinear acoustic feature space of speech compared to other techniques, albeit with increased complexity and higher computational cost that further increases with higher dimensionality. When delta features are used to replace part of the conventional static features such that overall GMM dimensionalities are unchanged, no increase in GMM complexity is involved, thereby requiring no increase in training data amounts nor in extension-stage computational resources. On the other hand, the inclusion of delta features into the parameterization frontend imposes a run-time algorithmic delay that limits our ability to exploit the full potential of memory inclusion to improve BWE performance. In addition, an empirical optimization procedure is required during training to achieve optimal allocation of the available overall dimensionalities among static and delta features. This procedure thus involves additional computations during the offline training stage. The second technique employs model-based memory inclusion implemented using the memory-based extension of the GMM framework described above. It addresses the drawbacks of the frontend-based system and improves on the BWE performance gains at the cost of higher complexity. Both techniques are compared to relevant techniques in the literature, with the latter shown to particularly outperform comparable model-based approaches, in some cases significantly. Furthermore, both proposed techniques are designed with generality in mind such that the underlying memory inclusion methodology can be adapted to other BWE implementations.

Novel MFCC-based BWE

While BWE schemes have traditionally used LP-based parameterizations, our work on quantifying cross-band correlation shows that *mel-frequency cepstral coefficient* (MFCC) parametrization results in higher certainty about the highband. We show that the superior MFCC cross-band correlation advantage extends as well to parameterizations with memory inclusion. The difficulty, however, of synthesizing speech from MFCCs—due to the non-invertibility of several steps employed in MFCC generation—has restricted their use to fields that do not require inverting MFCC vectors back into time-domain speech signals. By employing previous work on the high-resolution inverse discrete cosine transform (IDCT) of MFCCs, we achieve high-quality highband power spectra through the inversion of highband MFCCs obtained from narrowband ones by statistical estimation. Our MFCC-based highband power

spectra are comparable to conventional LP-based ones from which the time-domain speech signal can be reconstructed. Implementing this scheme for BWE thus allows capitalizing on the higher correlation advantage of MFCCs to increase the potential for memory-inclusive BWE performance improvements.

Detailed analysis of the effect of GMM covariance type on BWE performance
In order to reduce the computational complexity associated with GMM-based statistical modelling, spectral transformation techniques—including those of BWE—have, in general, relied on diagonal approximations to GMM Gaussian covariances. Indeed, employing diagonal Gaussian covariances, rather than full, reduces the computational costs associated with both the training and extension stages of a BWE GMM-based system—with the cost reduction especially significant during training. Such diagonal covariance approximations have been motivated by the argument that, since Gaussians in a GMM act in unison to model the overall probability density function of the spectral transformation in question, the effect of using a GMM with a particular number of full-covariance Gaussians can be equally obtained by a GMM with a larger set of diagonal-covariance Gaussians [40]. For BWE techniques where the computational cost of the offline maximum likelihood (ML) training stage is of increasingly less importance (particularly with the continuous advances in offline computational power), the diagonal covariance approximation has not been adequately evaluated in the literature. As GMMs are central to our work presented herein, we carefully investigate the effect of GMM covariance type on BWE performance. In particular, we compare diagonal- and full-covariance GMMs in terms of BWE performance as a function of the exact computational and memory costs associated with both covariance types during the extension stage. Emphasizing the fact that our investigation focuses on the complexities involved with only the extension stage, our analysis leads us to conclude that, to achieve similar BWE performance, using full-covariance GMMs is, in fact, more efficient than using GMMs with diagonal covariances.

1.5 Outline of the Thesis

The thesis is organized as follows. In Chapter 2, we review BWE techniques and underlying principles. We describe spectral envelope reconstruction techniques in some detail, with

particular emphasis on statistical modelling—central to our work.

In Chapter 3, we describe the details of our *dual-mode BWE implementation* used throughout the thesis for both memoryless and memory-based extension. As our BWE system coincides with current state-of-the-art techniques in employing GMMs for the statistical modelling of speech frequency bands, a review of the mathematical principle underlying GMM-based BWE is first presented, namely the *minimum mean-square error (MMSE) estimation of highband spectra using joint-density GMMs*. The details of our *memoryless BWE implementation* are then presented, providing the reference baseline for memory inclusion evaluation throughout the thesis. As part of the development of our baseline, we study the effects of varying the number of components in the BWE Gaussian mixtures, as well as the effects of using diagonal and full covariance matrices. This analysis represents one of the contributions of this thesis. Finally, we describe the measures used for BWE performance evaluation throughout our work and the motivations behind their choice. These measures are the *log-spectral distortion*; two variants of the Itakura-Saito distortion, the gain-optimized *Itakura distortion* and the gain-sensitive symmetrized *COSH* measure; and the *PESQ* measure. We conclude the chapter by evaluating these measures for the memoryless BWE baseline.

Chapters 4 and 5 represent our main contributions described in Section 1.4 above. In particular, Chapter 4 presents our work on modelling speech memory in the narrow and high frequency bands, and quantifying its effects on correlation between both bands. Two types of parameterizations are chosen for this analysis, line spectral frequencies (LSFs) as well as MFCCs. The justification for the choice of both types of parameters for BWE in general, and for the evaluation of the role of memory inclusion in particular, is provided. The most notable result of this chapter is the finding through quantifiable information-theoretic measures that speech memory can improve certainty about the high band by over 100%—quite a large figure, even for an upper bound. Another notable finding is that the effects of speech memory saturate at durations corresponding roughly to those of syllables, coinciding with similar hypotheses and measurements made in previous works in the context of speech perception and coding. Finally, our analysis shows the superiority of MFCCs over conventional LSFs in capturing the temporal information in speech, providing the motivation for MFCC-based BWE.

Chapter 5 builds on the theoretical results of Chapter 4 by first describing our implementation of speech reconstruction from MFCCs, then by integrating memory inclusion into

our GMM-based baseline BWE system. Through substituting part of the static features with delta ones, we show that BWE performance improvements can be attained through frontend-based memory inclusion. Although a computationally-demanding optimization procedure is required during model training in order to attain the best achievable improvements, such frontend-based memory inclusion involves no additional computational cost during extension relative to the memoryless baseline BWE system.

Using the aforementioned information-theoretic measures, we find, however, that the BWE performance improvements attained by frontend-based memory inclusion represent only a fraction of those theoretically achievable by memory inclusion in general. Furthermore, the inclusion of memory through the non-causal delta features imposes a run-time algorithmic delay that requires favourable network and computational latencies in order to achieve maximum BWE performance improvements while ensuring acceptable interactive real-time speech communication. As such, we continue Chapter 5 by addressing the drawbacks of frontend-based memory inclusion in BWE through transferring the task of modelling speech memory from the frontend to the modelling space. We derive an extension to the GMM formulation whereby we explicitly exploit speech memory to construct temporally-extended GMMs. Then, by integrating these temporally-extended GMMs into our MFCC-based dual-model BWE system, we show this novel technique to outperform not only our frontend-based approach, but also other comparable model-based memory-inclusive techniques, thereby demonstrating its superiority in regards to the efficiency of transferring the highband certainty gains associated with memory inclusion into tangible BWE performance improvements.

Concluding the thesis, Chapter 6 provides an extended summary of all research and work presented herein, discusses possible avenues for improving our proposed techniques, and finally, addresses the potential and applicability of our work to BWE and other related fields. The extended summary effectively encapsulates the entire thesis into a few pages for the purposes of a quick but comprehensive review.

1.6 Notation

As there is no consensus in the literature on mathematical notations, particularly for vectors, matrices, and probabilities, we herein define the notation used in this thesis. Unless otherwise indicated for exceptions, clarifications, or disambiguations, we represent:

- the *probability* of an event by $P(\cdot)$ and the *probability density function* (*pdf*) of a random variable X by $p_X(x)$.[24] Subscripts are dropped when clear from the context.

- *scalars* by italic letters, e.g., F_s for the sampling frequency, a_i for the coefficients of a prediction filter, and μ for the mean of a Gaussian density. Scalar random variables are represented by uppercase letters, e.g., X for arbitrary narrowband speech representation, and their realizations in the target space[25] by lowercase letters, e.g., x. For example, the probability distribution function of a scalar discrete random variable is defined as

$$F_X(x) \triangleq P(X \leq x) = \sum_{\xi \in (-\infty, x]} p_X(\xi). \tag{1.1}$$

- *vectors* by bold upright letters, e.g., $\mathbf{a} = [1, a_1, \ldots, a_p]^T$ for a prediction error filter. Unless otherwise stated, we always assume vectors to be column vectors. Random vectors are represented by uppercase letters, e.g., $\mathbf{X}$ for narrowband speech random feature vectors, and their realizations by lowercase letters, e.g., $\mathbf{x}$. For example, the probability distribution function of a vector random variable composed of the variables $X_1, \ldots, X_n$ is defined as

$$F_{\mathbf{x}}(\mathbf{x}) \triangleq P(X_1 \leq x_1, \ldots, X_n \leq x_n). \tag{1.2}$$

  An exception are vectors represented by Greek letters which we represent by their bold italic version for aesthetics of typography, e.g., $\boldsymbol{\mu}$ rather than $\boldsymbol{\mu}$ for the mean of a multivariate Gaussian density.

- *matrices* by uppercase bold upright letters, e.g., $\mathbf{C}$ or $\boldsymbol{\Sigma}$ for covariances of multivariate Gaussian densities,

- *sets* by uppercase upright or calligraphic letters, e.g., $\mathrm{A} = \{\alpha_i\}_{i \in \mathcal{I}}$ and $\Lambda = \{\lambda_j\}_{j \in \mathcal{J}}$.

[24] In the literature, *pdf*s are commonly denoted by f, e.g., $f_Y(y)$, to differentiate them from *probability mass functions* of discrete random variables denoted by, for example, $p_Y(y)$. However, since the overwhelming majority of random variables in our work are continuous, we prefer and use the latter form for *pdf*s . Exceptions where random variables are discrete are explicitly stated as such.

[25] Formally, a random variable $X{:}\Omega \to \Psi$, is a function that maps the events $\mathscr{F}$ with probabilities P from a sample space Ω, i.e., the probability space $(\Omega, \mathscr{F}, P)$, into a set of corresponding measurable sets $\mathscr{E}$ with the same probabilities P in the target measurable space Ψ, i.e., the probability space $(\Psi, \mathscr{E}, P)$.

Chapter 2

BWE Principles and Techniques

2.1 Introduction

As described in Section 1.1.1, traditional telephone networks limit speech bandwidth to the *narrowband* 300–3400 Hz range. As a result, narrowband speech has sound quality inferior to its *wideband* counterpart, and shows reduced intelligibility especially for consonant sounds. Such adverse effects of bandwidth limitation have been detailed in Section 1.1.3. Wideband speech reconstruction through bandwidth extension (BWE) attempts to regenerate as much as possible of the low- (< 300 Hz) and high-band (> 3.4 kHz) signals lost during the filtering processes employed in traditional networks.

Such reconstruction is based on two assumptions. The first is that narrowband speech correlates closely with the highband signals, and thus, given some a priori information about the nature of this correlation, the higher frequency speech content can be estimated. The second assumption is that even if the reconstructed highband signal does not exactly match the missing original one, it significantly enhances the perceived quality of telephony speech. Indeed, a variety of listening tests confirm this latter property of bandwidth extension [41]. The greatest advantage of BWE is that it generates enhanced wideband speech without any additional transmitted information, thereby providing backward compatibility with existing networks. It is worth noting that such *blind* BWE (i.e., where no side information is transmitted) has been applied to a very limited extent in some speech and audio coders. In AMR-WB coding [34], for example, blind BWE is used to reconstruct only the 6.4–7 kHz band (except at the highest 23.85 kbit/s mode where excitation gain information is encoded into the bitstream as side-information). This implies the daunting nature of the task of

extending speech bandwidth from 3.4 kHz up to 7 or 8 kHz.

BWE schemes have primarily used the source-filter model of speech, where narrowband and highband linear prediction (LP)-based envelopes are jointly modelled. As such, LP coefficients (LPCs[26]) of highband envelopes—estimated from the corresponding narrowband ones—can, then, be combined with a highband residual error (excitation) signal in an LP synthesis filter to regenerate the missing highband signal. This signal is, in turn, added to the available narrowband signal to generate wideband speech. Alternatively, full wideband—rather than only highband—envelopes and excitation signals can be estimated based on the narrowband input, with the advantage that lowband content is also generated in addition to that of the high band. Wideband speech generated as such is typically bandstop filtered to preserve only the lowband (< 300 Hz) and highband (> 3400 Hz) content, which can then be added to the available narrowband signal thereby avoiding introducing any distortions to the base narrowband signal. However, as argued in Section 3.3.2, this alternate approach is less efficient in modelling the cross-correlation between the available narrowband content and that which is of primary interest—the highband content.

In contrast, early BWE approaches do not make use of any particular model of speech generation neither do they make use of any a priori knowledge about speech properties. Such historical non-model-based techniques are much simpler, but typically inferior, compared to model-based methods.

Since many of the basic ideas underlying non-model-based BWE techniques are shared with model-based excitation generation methods, we present a brief overview of non-model-based techniques in the following to serve as an introduction of those ideas. We then review model-based BWE techniques in more detail due to their relevance to our work, with particular emphasis on spectral envelope generation techniques employing statistical modelling. An illustrative example comparing the properties and performance of several spectral envelope reconstruction techniques is presented.[27]

[26] The acronym LPC has been interchangeably used in the literature to refer to linear prediction coding/coefficient. When clear from the context, we will use the acronym to denote either, otherwise writing it out if disambiguation is needed.

[27] Detailed comparisons of the various techniques described below—in terms of their effect when used in our BWE implementation—is outside the scope of this thesis. As noted in Section 1.4, it is the role of speech memory—which manifests more clearly in measurable spectral envelope changes—that represents the focus of the work presented here, rather than comparing the various BWE implementation-specific techniques (particularly for excitation generation since, as discussed in Section 2.3.5, spectral envelopes are far more important for perception than excitation).

2.2 Non-model-based BWE

2.2.1 Spectral folding

Through insertion of zeros between adjacent samples (thereby increasing sampling rate), the narrowband spectrum is simply folded, or aliased, at half the original sampling frequency resulting in a mirrored highband spectrum. Examples of such a straightforward aliasing technique include the BWE schemes of [42] and [43]. While simple, this method has several problems when applied to telephony speech. First, it is unlikely that the new high frequency harmonics will reside at integer multiples of the voiced speech's fundamental frequency, F0. Secondly, as the pitch of the narrowband moves higher or lower in frequency, the corresponding high-frequency harmonics of the new wideband signal move in the opposite direction, causing speech to be somewhat garbled, especially in intervals with rapid F0 variations. Finally, the resulting wideband speech exhibits a band gap in the middle of the spectrum when half the narrowband sampling frequency (typically, F_s = 8 kHz) is higher than the telephone bandlimiting cutoff frequency, i.e., a gap corresponding to the eliminated frequency content in the 3.4–4 kHz range. While spectral folding works surprisingly well for extending the bandwidth of signals bandlimited to 8 kHz, for example, this BWE technique performs poorly for telephony speech [44, Section 5.4.1].

2.2.2 Spectral shifting

Rather than fold the narrowband spectrum into the high band, spectral shifting addresses the problems of spectral folding by shifting a weighted copy of part of the short-term narrowband spectrum in different manners into the extension regions [44, Section 5.4.2]. As such, both low (< 300 Hz) and high (> 3.4 kHz) frequency content can be generated in contrast to spectral folding which can only generate the latter. The high band is initially generated by zero-extending the narrowband signal's analysis FFT, fast Fourier transform, at π. The length of FFT zero padding depends on the desired new sampling frequency (e.g., padding an N-length FFT with N zeroes effectively doubles the sample rate). Fixed spectral shifting uses fixed values for the edge frequencies of the narrowband spectral subband to be copied into the high band. The copied subband is then weighted to mimic the average spectral decay associated with higher frequencies in speech, followed by inverse FFT to reconstruct the wideband signal. While such spectral shifting using fixed edge frequency

values eliminates the second and third problems associated with spectral folding, i.e., the problems of garbling and mid-frequency gap, it still usually results in misaligned high-frequency harmonics—and the corresponding artifacts—for voiced speech. Pitch-adaptive spectral shifting improves on the fixed scheme by incorporating pitch detection and estimation to adapt the edge frequencies of the narrow subband such that pitch structure is maintained even at the transition regions from the telephone bandpass to the extension regions.

2.2.3 Nonlinear processing

Nonlinear processing of the time-domain narrowband signal provides another means of bandwidth extension [44, Sections 5.4.3 and 5.5.1.2]. The application of nonlinear characteristics—e.g., quadratic, cubic, half- and full-wave rectification—generally broadens the band of the signal. Full-wave rectification, in particular, has been more common, e.g., [45]. When applied to a periodic signal, e.g., voiced speech, harmonics are preserved in the narrowband and are extended throughout the resulting broad band in a seamless continuous manner. Nonlinear processing thus provides the advantages of generating low frequency content (as well as high content), in addition to the benefits of pitch-adaptive spectral shifting while precluding the need for pitch detection. This latter property is quite desirable since the accuracy of pitch estimates heavily affects the performance of pitch-adaptive techniques. Furthermore, by virtue of broadening the signal—rather than flipping it, for example—no spectral gaps occur within the higher frequency extensions.

On the other hand, nonlinear processing may—depending on the effective bandwidth, the sampling rate and the kind of characteristic—require additional processing to avoid aliasing in the nonlinearly processed signal. Similarly, nonlinear processing generates strong undesired components around 0 Hz, which, in turn, have to be removed. The application of nonlinear characteristics may also result in undesired spectrum coloration (concentration of energy in one or more subbands), further requiring the use of whitening filters. Another disadvantage of nonlinear processing is that it reproduces the harmonics of any periodic noise that may be present in the narrowband signal. Furthermore, power normalization is required in the case of signals processed using quadratic and cubic characteristics due to the resulting wide dynamic range.

2.3 Model-based BWE

2.3.1 The source-filter model

The parametric source-filter speech production model, as described by Fant in [46], is by far the model most commonly used in BWE, followed by the sinusoidal model described in Section 2.3.6. The source-filter model assumes that the vocal cords are the source of a spectrally flat *excitation signal*, and that the vocal tract acts as a *spectral shaping filter* that shapes the spectra of various speech sounds. While an approximation, this model is widely used in speech analysis and coding in the form of LPC—linear prediction coding.[28] Its popularity derives from its compact yet precise representation of speech spectral properties as well as the relatively simple computation associated with LPC. As described in Section 1.1.2, phonemes can be distinguished by their excitation (source) and spectral shape (filter). Voiced sounds, e.g, vowels, have an excitation signal that is periodic and can be viewed as a uniform impulse train having a line spectrum with regularly-spaced uniform-area harmonics. Unvoiced sounds, e.g., unvoiced fricatives[29], have an excitation signal that resembles white noise. Mixed sounds, e.g., voiced fricatives, have an excitation signal consisting of harmonic and noisy components. Figure 2.1 illustrates how the source-filter model represents such excitation signals, $e(n)$, through a time-varying continuous measure of periodicity versus noisiness, $g(n)$ where $0 \leq g(n) \leq 1$, making use of the pitch frequency, F0, as well as the overall excitation signal gain, $\sigma(n)$.

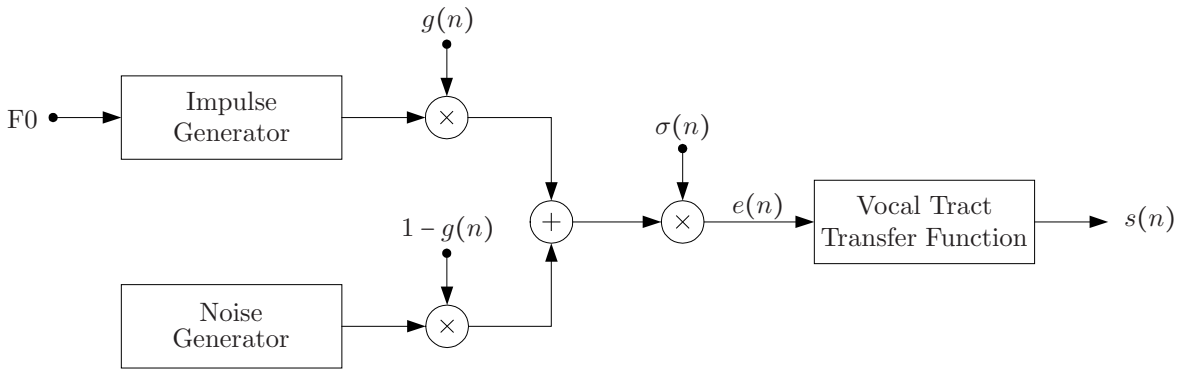


Fig. 2.1: The source-filter speech production model.

[28] See [47, Chapter 12] for a detailed analysis of LPC.
[29] See Table 1.1.

The vocal tract transfer function is predominantly assumed to be an all-pole model with fixed parameters for short segments of time (frames). In other words, speech is assumed to be an autoregressive (AR) random process with the spectrally flat excitation its corresponding innovations process. Thus, the vocal tract transfer function can be written as $H(z) = 1/A(z)$, where, for p poles,

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}, \tag{2.1}$$

and the speech signal, $S(z) = E(z)H(z)$ where $E(z)$ is the z-transform of $e(n)$, can then be written as

$$S(z) = \frac{\sigma}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{2.2}$$

When applied to the speech signal, $s(n)$, the all-zero inverse filter, $A(z)$, acts as a prediction error filter. As such, the parameters $\{a_k\}_{k\in\{1,\ldots,p\}}$ are obtained through the MMSE solution to the *normal equations* of the pth-order predictor. Since $s(n)$ is assumed to be an AR process, the normal equations also correspond to the *Yule-Walker equations*, and are commonly referred to as thus in the context of LPC. Similarly, the gain parameter, σ, represents the square root of the power density of the spectrally-white excitation innovations, and is computed as the square root of the power density of the predictor error filter output (i.e., the root-mean-square forward prediction error). Due to its AR property, the autocorrelation matrix of $s(n)$ is Töeplitz and positive definite. These two properties are exploited by the Levinson-Durbin and Schür algorithms, respectively, to solve the normal equations in a recursive manner.[30] As described in Section 1.2, speech has a quasi-stationary character only for short periods of time, and hence, an LPC model's parameters need to be estimated periodically roughly every 10 ms.

First applied for the task of BWE in 1994 by Yoshida [49], and independently by Carl [50], the source-filter model of speech, thus, reduces the problem of reconstructing highband (or wideband) speech given only the narrow band, into two tasks:

- generating a highband (or wideband) excitation signal, $e(n)$, containing the voiced and unvoiced excitation characteristics described above, and

[30] In addition to the well-known Levinson-Durbin and Schür algorithms, there are also other *fast* algorithms for solving the Yule-Walker equations—namely the Euclidean and the Berlekamp-Massey algorithms. See [48] for a comparison of these algorithms.

- generating an estimate of the highband (or wideband) spectral envelope, $H(z)$.

The excitation and spectral envelope estimates can then be combined in a synthesis filter[31] to reconstruct $s(n)$. It should be noted that since most of the signal in the higher bands of wideband speech is not harmonically structured, the spectral envelope is usually deemed sufficient for highband reconstruction, i.e., phase estimation is commonly bypassed.

2.3.2 Generation of the highband (or wideband) excitation signal

The first methods for the generation of highband excitation signals derived from the so-called baseband coders [51].[32] In baseband coders, only a low-frequency portion of the excitation (the residual at the output of the analysis filter in the transmitter), known as the baseband, is transmitted and used at the receiver to regenerate the high-frequency portion of the excitation.[33] The wideband LPCs are transmitted separately. The sum of the transmitted baseband excitation and the regenerated high-frequency excitation constitute the wideband excitation to the synthesis filter at the receiver. This technique is sometimes referred to in the literature as HFR, high-frequency regeneration, and was used in early RELP speech coders.[34]

BWE excitation generation techniques can generally be classified as follows.

2.3.2.1 Nonlinear processing

The high-frequency excitation generation techniques applied in baseband coders were mostly based on nonlinear processing of the baseband excitation through waveform rectification. To avoid aliasing potentially introduced by the nonlinearities, the baseband excitation is first interpolated. The nonlinearly processed signal is then spectrally flattened before it is

[31] The filters $A(z)$ and $H(z)$ are typically referred to as the *analysis* and *synthesis* filters, respectively.

[32] Baseband coders (also known as voice-excited coders) were originally proposed as a compromise between waveform coders—the simplest speech coders—and the relatively more complex pitch-excited coders (also known as vocoders). Vocoders, e.g., LPC, employ a speech production model, usually the source-filter model, and hence, operate on blocks of quasi-stationary speech. Waveform coders, on the other hand, analyze, code, and reconstruct speech sample-by-sample.

[33] Baseband excitation is extracted through a lowpass or bandpass filter of width B, usually determined such that the full bandwidth, W, is an integer multiple of B.

[34] Originally proposed in the 1970s, residual-excited linear prediction (RELP) coding [52] is a predecessor of code-excited linear prediction (CELP) coding [53]. However, unlike CELP where a limited set of excitation signal parameters are transmitted and used at the decoder to generate the excitation signal through an adaptive and a fixed codebook, RELP directly transmits the residual signal. To achieve lower rates, that residual signal is usually lowpass filtered and downsampled; e.g. $F_s = 1.6\,\text{kHz}$ in [52].

used as excitation to the synthesizer. In the context of BWE of telephony speech where the narrowband signal corresponds to the baseband signal of baseband coders, nonlinear processing can be applied to all or portion of either the narrowband signal itself, e.g., [54, 55], or its residual, e.g., [56]. As shown in Section 3.2.4, highband excitation generation in our BWE system employs nonlinear processing in the form of full-wave rectification of the equalized 3–4 kHz subband of the narrowband signal followed by spectral flattening through white noise modulation.

2.3.2.2 Spectral folding

Spectral folding, similar to the technique described in Section 2.2.1, can also be applied only to the narrowband/baseband excitation signal. Introduced in [51], baseband excitation spectral folding eliminates the need for the spectral flattening associated with nonlinear processing, since the baseband excitation that is mirrored into the high-frequency region is already spectrally flat. It suffers, however, from the drawbacks described earlier—namely the potential for spectral gaps and the problems associated with irregular pitch harmonics. The problem of spectral gaps is often mitigated by downsampling and upsampling the available bandpass residual, as in the BWE method of [57]. Despite its disadvantages compared to other techniques, spectral folding is frequently used primarily for its simplicity, e.g., [50, 58–60].

2.3.2.3 Modulation techniques

Similar in concept to the spectral shifting technique discussed in Section 2.2.2, modulation techniques—more common in recent BWE works—effectively shift the residual extracted by the LPC analysis of narrowband speech into the high band. Modulation is performed through the time-domain multiplication

$$e_{\mathrm{m}}(n) = \tilde{e}_{\mathrm{nb}}(n)\, 2\cos(\omega_{\mathrm{m}} n), \tag{2.3}$$

where $\tilde{e}_{\mathrm{nb}}(n)$ is the interpolated version of the narrowband excitation $e_{\mathrm{nb}}(n)$, i.e., upsampled to a sampling frequency that is sufficient to represent the extended wideband speech signal, e.g. $F_{\mathrm{s}} = 16$ kHz, and lowpass filtered. The narrowband excitation is the residual obtained by LP analysis of the narrowband telephone signal at the receiver. The modulation

frequency is $\omega_\mathrm{m} = 2\pi F_\mathrm{m}/F_\mathrm{s}$, and $e_\mathrm{m}(n)$ is the resulting modulated excitation which now extends above F_m. Spectrally, this multiplication generates generates two shifted copies of $E_\mathrm{nb}(\omega)$, the narrowband excitation spectrum:

$$E_\mathrm{m}(\omega) = \tilde{E}_\mathrm{nb}(\omega + \omega_\mathrm{m}) + \tilde{E}_\mathrm{nb}(\omega - \omega_\mathrm{m}). \tag{2.4}$$

To prevent potential spectral overlap of the shifted spectra depending on the choice of ω_m, the upsampled narrowband excitation is lowpass filtered prior to modulation (part of the interpolation process), while the modulated excitation is highpass filtered to preserve only the desired highband components, $e_\mathrm{hb}(n)$. The wideband excitation signal, $e_\mathrm{wb}(n)$, can then be formed by adding the two signals. In BWE techniques where high-frequency speech content is first reconstructed then added to the available narrowband content (in contrast to techniques which model and reconstruct wideband speech as a whole from the narrowband input), only the corresponding highband components of the excitation are technically needed. However, the computationally-trivial addition of narrowband and highband excitation signals eliminates any potential spectral gaps due to misalignments between the bandwidth edge frequencies of the highband excitation and the highband spectral envelope estimated separately.[35]

In BWE, the modulation frequency, F_m, is typically chosen around the 3.4 kHz narrowband higher cutoff frequency to ensure a seamless continuation of the excitation spectrally, thereby avoiding any spectral gaps, e.g., [39, 61]. Furthermore, pitch structure can be preserved across the wide band by incorporating pitch detection to adaptively modify F_m through floor and ceiling functions such that

$$F_\mathrm{m} = \left\lfloor \frac{3.4}{\widehat{\mathrm{F0}}} \right\rfloor \widehat{\mathrm{F0}} \qquad \text{or} \qquad F_\mathrm{m} = \left\lceil \frac{3.4}{\widehat{\mathrm{F0}}} \right\rceil \widehat{\mathrm{F0}} \qquad [\mathrm{kHz}], \tag{2.5}$$

as implemented in [61], for example. Pitch estimation must be reliable, however, since pitch-adaptive modulation reacts quite sensitively to small errors in $\widehat{\mathrm{F0}}$ estimates (errors

[35] As seen in Chapter 3, our BWE technique, for example, uses midband equalization to reconstruct content in the 3.4–4 kHz range, and statistical modelling to reconstruct highband spectral envelopes above 4 kHz. Thus, only the excitation content above 4 kHz is technically needed. Nonetheless, had we been using such an excitation signal obtained by modulation, any minor changes to the frequency ranges of midband equalization or highband statistical modelling would necessitate corresponding changes in the system components generating the highband excitation signal.

are magnified by the factor $3.4/\widehat{\text{F0}}$) [39]. Figure 2.2 depicts wideband excitation generation through pitch-adaptive modulation.

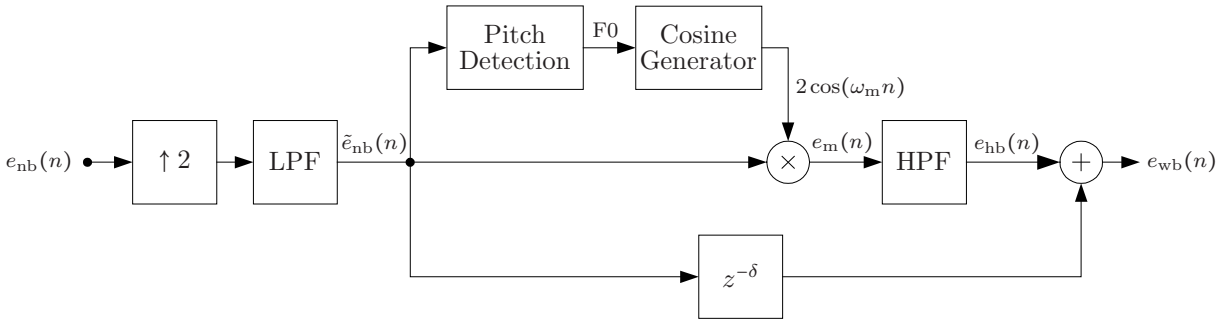


Fig. 2.2: Wideband excitation generation through pitch-adaptive modulation. The δ delay applied to $\tilde{e}_{\text{nb}}(n)$ compensates for the HPF delay.

2.3.2.4 Harmonic modelling

An attractive technique proposed in [62] generates highband excitation by parameterizing the *harmonicity* of speech such that the correlation between narrowband and wideband harmonicity can be modelled in the training stage, in a manner similar to the modelling of spectral envelopes. This approach performs such modelling using a harmonic-plus-noise model (HNM) where the degree of voicing (harmonicity) in 32 separate bands (with each band centered on a harmonic multiple of F0) is quantified by measuring the squared distance in the spectral domain between the actual wideband excitation signal in each band and a Gaussian-shaped window scaled such that its peak has the same amplitude of the harmonic of that band; the smaller the distance the higher the degree of voicing is in that band. Subbands above 32-band range are assumed to be entirely unvoiced.

A codebook is trained on such harmonicity feature vectors such that, in the extension stage, harmonicity of the wideband excitation signal, as a whole, can be estimated from narrowband harmonicity. The obtained per-band harmonicity values are then used during reconstruction to appropriately weight the Gaussian-shaped voiced components (Gaussian windows in the frequency domain centered on multiples of F0) as well as Rayleigh-distributed random unvoiced components. All excitation components, voiced and unvoiced, are then summed. Excitation amplitudes in each subband at the harmonics are assumed to be unity with the usual assumption that the LP model whitens the excitation. The gain of the frame is extracted as an LP gain value for which another codebook is trained

in conjunction with a narrowband-to-wideband spectral envelope codebook. Finally, the excitation thus reconstructed is multiplied by the wideband LP spectrum and a phase component to form the speech spectrum in each frame.

The use of the harmonicity model for reconstruction of the excitation signal is compared in [62] to the nonlinear bandpass-modulated Gaussian noise (BP-MGN) method of [54]. This latter method is an earlier implementation of the more superior technique used in our BWE system—equalized BP-MGN (EBP-MGN) [55].[36] Results show that the harmonicity-based technique outperforms the BP-MGN method particularly for highband content with more harmonically structured patterns, i.e., voiced components. However, as stated in [62], the harmonicity technique requires pitch detection whose accuracy is crucial for estimating reliable harmonicity levels. Moreover, the performance difference between the two approaches is more pronounced for voiced, rather than unvoiced, highband content. As discussed in Section 1.1.3, it is rather the noisy unvoiced content—mostly associated with fricatives, stops, and affricates, with energy concentrated in higher frequencies—that is more adversely affected by narrowband telephony bandwidth limitations.

2.3.3 Generation of the highband (or wideband) spectral envelope

BWE hinges on the assumption that narrowband speech correlates closely with the highband signal such that high-frequency content can be estimated given only the narrowband signal and learning a priori the nature of the cross-band correlation. However, due to the dynamic nature and the inherent variability of speech described in Section 1.2, such cross-band correlation is significantly more complex than to allow an ideal closed form solution for the narrowband-to-highband mapping problem, notwithstanding whether it is even sufficient to guarantee uniqueness of the solution. In fact, uniqueness of the solution is quite unlikely; there is likely no underlying one-to-one mapping between narrowband and highband features over any arbitrary duration. Thus, BWE techniques rather attempt to model cross-band correlation, as described below, in order to allow mapping that is as accurate as possible, with performance varying greatly with choice of model. In particular, it will be shown that modelling techniques allowing many-to-many mapping between narrowband and highband (or wideband) acoustic subspaces provide better BWE performance.

[36] See Section 3.2.4 for more details regarding the superior performance of the EBP-MGN method over the BP-MGN one for the generation of the highband excitation signal.

2.3.3.1 Linear mapping

In the simplest terms, narrowband-to-highband spectral envelope mapping can be modelled as a single-matrix linear transformation where a highband feature vector, $\mathbf{y}$, is obtained from that of the narrowband input, $\mathbf{x}$, through the mapping

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \tag{2.6}$$

with the transformation matrix $\mathbf{W}$ determined using least squares over all narrowband and highband feature vectors, $\mathbf{X}$ and $\mathbf{Y}$, respectively, from a large training database, as [63]

$$\mathbf{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2.7}$$

Although quite simple, such single-matrix linear mapping is, however, an unrealistic over-simplification of the highly nonlinear narrowband-to-highband space mapping problem. Hence, several variations have been proposed to improve mapping capability either by refining linear mapping itself, or by introducing some nonlinearity into the basic algorithm. These improvements involve the use of multiple matrices, rather than a single matrix, with each matrix optimized for a particular subspace of either the narrowband or highband (or wideband) spaces. The BWE technique of [58], for example, refines linear mapping by optimizing multiple-input single-output linear filters where each filter generates an estimate for one of the wideband features as a linear combination of all input narrowband features within a window of 100 ms. More common, however, are the *piecewise-linear mapping* techniques which use some form of clustering—a nonlinear operation—to partition the narrowband space into disjoint subspaces. The subspaces are defined either by the codewords of a VQ codebook (described below), as in [63], or by the regions delimited by thresholds of one or more parameters, as in [60]. In the extension stage, each input narrowband feature vector is classified in a preprocessing step prior to being linearly mapped. The desired highband (or wideband) feature vector is then obtained through the particular transformation matrix optimized for the class assigned to the input narrowband vector. Alternatively, a linear combination of the transformation matrices corresponding to the K nearest codewords can be used, as in [56], resulting in superior smoothed highband (or wideband) vectors.

As shown by the results of [63], for example, single-matrix linear mapping is inferior to most—if not all—other techniques because of its over-simplification of the BWE mapping

problem. While the refinements and piecewise-linear approaches perform somewhat better, they are still nevertheless inferior to the more common codebook approaches.

2.3.3.2 Codebook mapping

Introduced independently for BWE by both Yoshida [49] and Carl [50], codebook mapping is the first and most common model-based approach to reconstruct highband (or wideband) spectral envelopes. Codebook mapping is based on the vector quantization (VQ) of one or more spaces parameterized into feature vectors. VQ partitions a continuous feature vector space into disjoint polytope partitions, or *Voronoi*, represented by their centre *codevectors*, such that a particular distortion measure calculated over all training vectors is minimized [64, Sections 10.1 and 10.2]. Codebook VQ training is typically performed using the Linde-Buzo-Gray (LBG) iterative algorithm [65].

In the context of BWE, simpler codebook mapping approaches quantize only the wideband space and, hence, require only one codebook. Optimization in the training stage is performed on the entire wideband envelopes, e.g., [50, initial approach; 66]. In the extension stage, by calculating distortion over only the narrowband portion, the wideband codevector closest to the input narrowband vector is selected. Alternatively, more advanced approaches quantize only the narrowband space to generate a narrowband codebook, which is then *shadowed* by another highband (or wideband) codebook where codevectors are obtained by averaging the highband (or wideband) vectors corresponding to the narrowband training vectors falling in each Voronoi of the narrowband codebook, e.g., [50, 59, 63, 67]. In the extension stage, the high (or wideband) codevector with the same codebook index as that of the narrowband codevector closest to the narrowband input, is selected. This more common approach to codebook mapping is illustrated in Figure 2.3.

Since codebook mapping involves quantization of the continuous feature vector space into a limited number of codewords, discontinuities occasionally result in perceptually-annoying artifacts in the extended signal—namely highband power overestimation and overly rapid spectral envelope changes. While increasing codebook size—thereby decreasing overall VQ distortion—alleviates some of these artifacts at a higher computational cost, simpler and more effective techniques have been proposed for this purpose. Similar to the interpolation method described above for piecewise-linear techniques, *codebook mapping with interpolation* selects the K narrowband envelopes closest to that of the input nar-

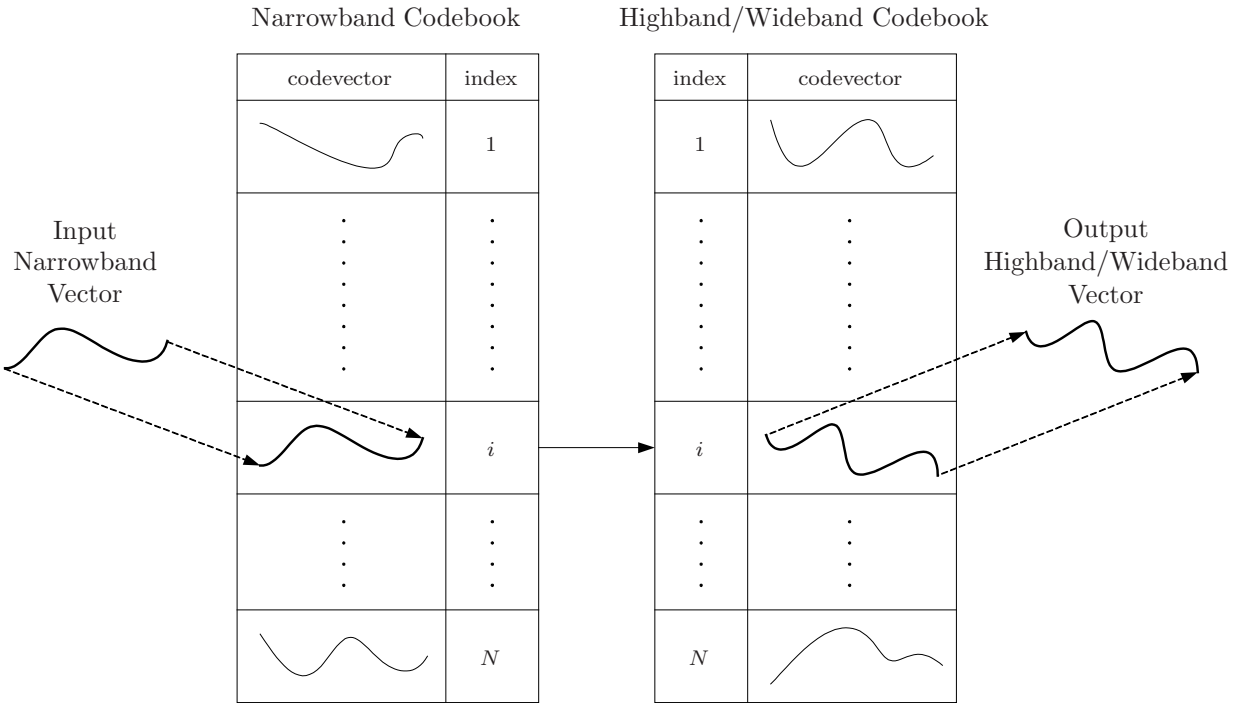


Fig. 2.3: Highband spectral envelope generation using codebook mapping.

rowband signal, combining their mapped highband codevectors. The combined envelopes can be simply averaged, as in [63], for example, or—in a manner similar to that used in [56] for piecewise-linear mapping—can be weighted depending on the proximity of each selected codevector to the input narrowband vector, e.g., [68]. Hence, codebook mapping with interpolation is also referred to as codebook mapping with *fuzzy* or *soft* VQ. As shown in [63], codebook mapping with interpolation generally outperforms conventional mapping due to its ability to predict envelope shapes not contained in the highband codebook. Other variations of the same concept involve envelope-domain smoothing, as in [59], where the wideband envelope is produced as the weighted sum of the last three chosen codewords.

Even better codebook mapping performance can be obtained by making use of measurable signal properties to directly improve the VQ partitioning of the feature space itself. Using voicing, for example, to *split* the feature space into voiced and unvoiced partitions allows building two separate smaller—but overall more accurate—codebooks, e.g., as in [63]. This particularly helps minimize artifacts due to highband overestimation. An alternate technique in [59] identifies codevectors in the trained codebook that are *dangerous* for

voiced sounds. If a marked codebook vector is chosen during a voiced sound, the power of the generated highband speech is lowered by 10 dB. Yet another attractive technique exploits voicing *periodicity* to partition the narrowband space into three separate codebooks representing voiced, unvoiced, and mixed sounds [69]. All these techniques report improved highband signal reconstruction compared to conventional mapping. They require, however, additional voicing detection.

2.3.3.3 Neural networks

Artificial neural networks are known for their superior ability to learn complex nonlinear relationships, and thus, have been widely used in pattern recognition applications including automatic speech recognition (ASR). In the context of BWE, however, neural networks have not received as much adoption as other techniques despite having been introduced in [70] for the purpose of BWE around the same time as codebook mapping. This follows mainly from the difficulty of analyzing the nonlinear processing in the *hidden* layers of a neural network, making system development mostly an empirical exercise.

Neural networks are generally composed of neurons organized in a regular structure. The type of neural network most often applied to the BWE mapping problem is the *multi-layer perceptron* (MLP) network with feed-forward operation.[37] Illustrated in Figure 2.4, perceptrons perform mapping as given by

$$y = \varphi(\tau + \sum_{i=1}^{N} w_i x_i) \tag{2.8}$$

for N inputs, x_i, where the bias, τ, and weights, w_i, are parameters to be trained, and φ is a nonlinear *activation function*, typically a sigmoid function.

In an MLP network, layers of perceptrons are arranged in cascade as shown in Figure 2.5. The output layer, generating the desired highband (or wideband) features, is preceded by one or more hidden layers, referred to as such as their outputs are inaccessible externally. As shown in Figure 2.5, a single hidden layer is typically used due to its capability to model any nonlinear continuous function. The input layer is only a pass-through layer distributing input narrowband features to the perceptrons of the hidden layer. Training is achieved in a supervised manner typically using the *back-propagation* algorithm [73], which

[37] See [71, Chapter 6; 72, Chapter 4] for detailed description and analysis of multi-layer perceptrons.

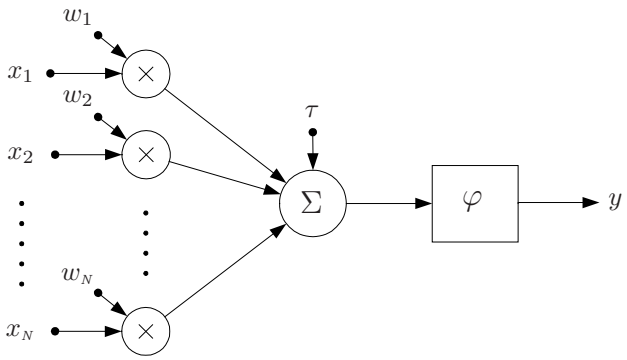


Fig. 2.4: The perceptron of a neural network.

applies gradient-descent until a stopping criterion is reached for the *training error*.

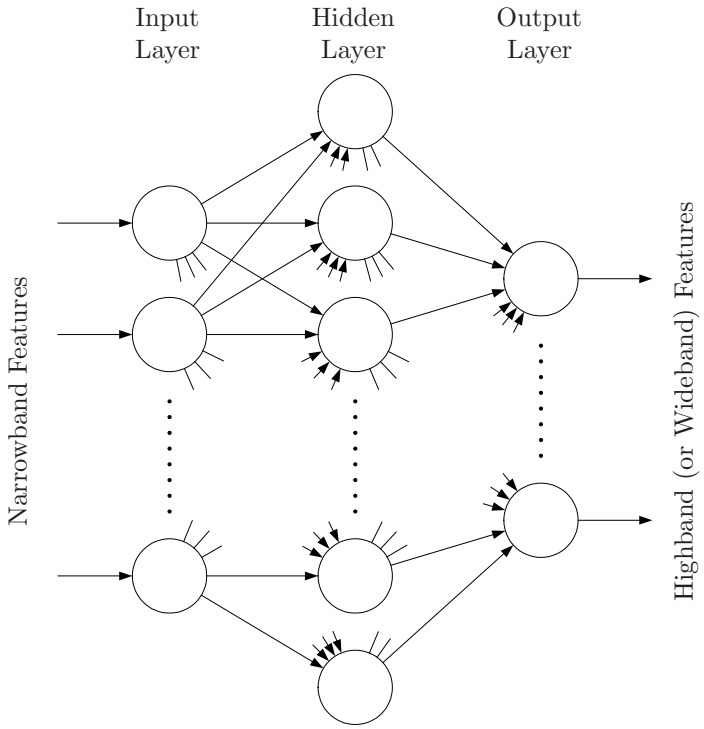


Fig. 2.5: Multi-layer perceptron neural network.

Despite the nonlinear *expressive power* of multi-layer neural networks [71, Section 6.2.2], works comparing their BWE performance to that of codebook and linear mapping report mixed results. In [56], for example, spectral envelopes generated using neural networks show less distortion than both codebook and linearly mapped envelopes in speaker-dependent training and testing conditions. In speaker-independent and noisy testing conditions, how-

ever, neural networks lag in performance, indicating that neural networks lack robustness against training-testing mismatches. Similarly, it is shown in [41] that while neural network BWE performance outperforms that of codebook mapping using four different objective measures, subjective evaluations lead to the opposite result. In particular, when compared to narrowband speech, codebook mapping is found to be approximately 1 point better than neural networks in terms of MOS. When choosing which approach produces better results, around 80% of listeners voted for the codebook-based scheme.

Because of their ability to learn complex tasks using comparatively few layers and neurons, neural networks nevertheless represent an attractive approach since they provide the potential for superior modelling of the complex nonlinear cross-band correlations in speech. Moreover, since neural networks do not require evaluating a distance measure in the extension stage, they require lower computational cost than codebook-based methods for the same input and output dimensionalities. Although not pursued in this thesis, we find these advantages particularly attractive for BWE with short-term memory inclusion where supervectors composed of current and few surrounding frames can be directly used as inputs without prohibitively increasing complexity and training data requirements, as would be the case with codebook-based BWE as well as the GMM-based BWE described in the next section. Indeed, similar ideas of modelling temporal information have been successfully applied in *dynamic* and *recurrent* neural networks for system identification and time-series prediction problems.[38] Their application to memory-inclusive BWE, however, has not been investigated to the best of our knowledge.

2.3.3.4 Statistical modelling

Despite the success of linear mapping and—to a larger extent—codebook mapping in achieving reasonable BWE performance with relatively little computational complexity, both techniques suffer a fundamental limitation in their ability to model the complex nonlinear continuous acoustic distributions of speech. As described in Section 2.3.3.1, linear mapping effectively reduces the N-dimensional distribution of the acoustic space modelled by N features, into a linear hyperplane (or multiple hyperplanes in the case of piecewise-linear mapping). Similarly, codebook mapping partitions the continuous N-dimensional acoustic space into polytopes where the continuous acoustic distribution within a poly-

[38] See [72, Chapters 13 and 15] for details on temporal processing using feed-forward and dynamically-driven recurrent networks.

tope partition is quantized into a single codevector. As mentioned in Section 2.3.3.2, this typically results in speech discontinuities in addition to imposing one-to-one mapping on narrowband and highband (or wideband) vectors. While codebook mapping with interpolation replaces such hard-classification quantization into a local continuous approximation of the distribution in the subspace around a polytope match, such interpolation is still a sub-optimal smooth fit that is based on only a few quantized points in space, thereby ignoring the true distribution within these local subspaces. These deficiencies of linear and codebook mapping are exposed through an illustrative example in the next section—Section 2.3.3.5. Given their gross approximations, the reasonable BWE performance of linear and codebook mapping techniques can, therefore, be attributed to the aforementioned second assumption underlying BWE; that *even if the reconstructed highband signal does not exactly match the missing original one, it significantly enhances the perceived quality of telephony speech.*

In contrast to the deterministic and quantizing nature of linear and codebook mapping, respectively, statistical modelling techniques employ a probabilistic framework to produce a continuous approximation of the complex nonlinear many-to-many acoustic space. During training, cross-band correlation is learned by statistically modelling the joint *pdf*, $p_{\mathbf{XY}}(\mathbf{x},\mathbf{y})$, of the narrowband and highband (or wideband) spectral envelopes (with features for both shape and gain) represented by the continuous vector variables, $\mathbf{X}$ and $\mathbf{Y}$, respectively. This probabilistic approach thus allows a better continuous many-to-many model of the underlying mapping. In the extension stage, highband (or wideband) spectral envelopes can then be obtained from input narrowband envelopes as a function of the conditional *pdf*, $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$, derived from the joint *pdf*.

I. Statistical recovery based on autoregressive Gaussian sources model[39]

Statistical modelling was first applied for spectral envelope reconstruction by Cheng [74]. In particular, the K-sample narrowband and highband speech frames—represented by $\mathbf{X}$ and $\mathbf{Y}$, respectively—are assumed to be generated by a combination of N and M random sources, $\Lambda = \{\lambda_i\}_{i\in\{1,\ldots,N\}}$ and $\Theta = \{\theta_j\}_{j\in\{1,\ldots,M\}}$, respectively, which, in turn, are assumed to be correlated by a many-to-many mapping given by $\mathrm{A} = \{\alpha_{ij}\} = \{P(\theta_j|\lambda_i)\}$[40]. Highband speech is synthesized by assigning different weights to the corresponding sources, with the

[39] Although not a spectral envelope reconstruction technique per se, the statistical recovery function technique of [74] is described here in the context of statistical modelling.

[40] $p(\theta|\lambda)$ is a probability mass function.

weights estimated based on the available narrowband speech. By modelling the sources Λ and Θ as autoregressive Gaussian sources,[41] a *statistical recovery function* can be derived to estimate $\mathbf{Y}$ as a function of the narrowband input, $\mathbf{X}$, and model parameters, $\Xi = \{\mathrm{A}, \Lambda, \Theta\}$; i.e.,

$$\mathbf{Y} = f(\mathbf{X}, \Xi). \tag{2.10}$$

By further restricting $\mathbf{Y}$ and $\mathbf{X}$ to dependence only upon their respective sources, Θ and Λ, the cross-correlation between highband and narrowband speech can be reduced into only the probabilities $\{P(\theta_j|\lambda_i)\}$, such that the joint *pdf*, $p(\mathbf{y}_t, \mathbf{x}_t, \lambda_i, \theta_j)$, at time t, is given by

$$\begin{aligned} p(\mathbf{y}_t, \mathbf{x}_t, \lambda_i, \theta_j) &= p(\mathbf{y}_t|\theta_j)P(\theta_j|\lambda_i)p(\mathbf{x}_t|\lambda_i)P(\lambda_i) \\ &= p(\mathbf{y}_t|\theta_j)\alpha_{ij}p(\mathbf{x}_t|\lambda_i)P(\lambda_i). \end{aligned} \tag{2.11}$$

Thus, the statistical mapping model can be fully represented by the autoregressive Gaussian densities $\{p(\mathbf{x}_t|\lambda_i)\}$ and $\{p(\mathbf{y}_t|\theta_j)\}$,[42] the prior probabilities $\{\alpha_{ij}\}$ and $\{P(\lambda_i)\}$, in addition to a gain parameter for each output source, β_{θ_j}, estimated as a function of the ratio of highband to narrowband signal energies weighted by the posterior *pdf*, $p(\theta_j|\mathbf{x}_t, \mathbf{y}_t)$, of the relevant source, θ_j. Using the popular Expectation-Maximization (EM) algorithm [76] to maximize the likelihood $p(\mathcal{X}, \mathcal{Y}|\Xi)$ for the training sequences $\mathcal{X} = \{\mathbf{x}_t\}_{t\in\{1,\ldots,T\}}$ and $\mathcal{Y} = \{\mathbf{y}_t\}_{t\in\{1,\ldots,T\}}$, the parameters needed for the extension stage—namely, $\{a_k^{(i)}\}$, $\{a_k^{(j)}\}$, $\{\alpha_{ij}\}$, $\{P(\lambda_i)\}$, and $\{\beta_{\theta_j}\}$, for all $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, M\}$ and $k \in \{1, \ldots, p\}$—can be iteratively estimated. In the extension stage, the MMSE solution, $\hat{\mathbf{Y}}$, is derived as a function of the quantities in Eq. (2.11) and makes use of the autoregressive model of the output sources, such that Eq. (2.10) giving the output signal is shown to be, at frame t,

$$\hat{\mathbf{Y}}_t(z) = \sum_{j=1}^{M} f_{t,j}\frac{U(z)}{A_j(z)}, \quad \text{where} \quad f_{t,j} = \sqrt{\mathcal{E}(\mathbf{x}_t)}\beta_{\theta_j}\sum_{i=1}^{N}\alpha_{ij}p(\mathbf{x}_t|\lambda_i)P(\lambda_i), \tag{2.12}$$

[41] For the p-order autoregressive signal $x(n) = \sum_{i=1}^{p} a_i x(n-i) + e(n)$ with zero-mean and σ^2-variance Gaussian innovations $e(n)$, the conditional *pdf* of the K-sample vector $\mathbf{x} = [x(1), \ldots, x(K)]^T$ given the parameter vector $\mathbf{p} = [\sigma^2, a_1, \ldots, a_p]^T$ can be shown to be, for $K \gg p$ [75],

$$p(\mathbf{x}|\mathbf{p}) = (2\pi\sigma^2)^{-K/2}\exp\left(-\frac{K}{2\sigma^2}\left[\mathbf{a}^T\mathbf{R}_{\mathbf{x}}\mathbf{a}\right]\right), \tag{2.9}$$

where $\mathbf{a} = [a_1, \ldots, a_p]^T$ and $\mathbf{R}_{\mathbf{x}}$ is the autocorrelation matrix of $\mathbf{x}$.

[42] By using unit-variance Gaussian sources, the *pdf*s $\{p(\mathbf{x}_t|\lambda_i)\}_{i\in\{1,\ldots,N\}}$ and $\{p(\mathbf{y}_t|\theta_j)\}_{j\in\{1,\ldots,M\}}$, defined as described in Footnote 41, are effectively reduced to requiring only the estimation of the predictor coefficients of the input and output sources, i.e., $\{a_k^{(i)}\}_{\forall i,k}$ and $\{a_k^{(j)}\}_{\forall j,k}$, respectively, during training.

where $U(z)$ is a zero-mean unit-variance Gaussian source, $\mathcal{E}(\mathbf{x}_t)$ is the energy of the input in frame t, and $p(\mathbf{x}_t|\lambda_i)$ is given by Eq. (2.9) with $\sigma^2 = 1$ and estimated for each frame $\mathbf{x}_t$. Figure 2.6 illustrates this BWE technique.

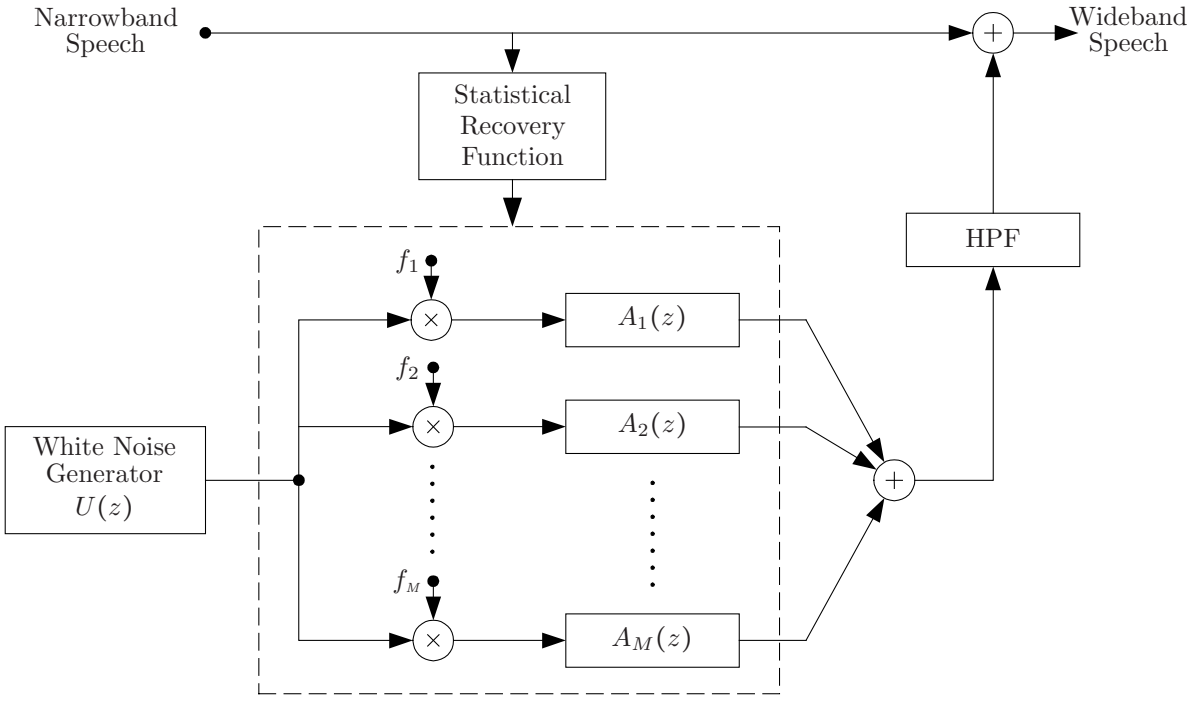


Fig. 2.6: BWE with statistical recovery using autoregressive Gaussian sources.

The performance of this initial attempt to statistically achieve BWE was not appropriately measured. By merely comparing narrowband and reconstructed wideband spectrograms to those of the original wideband signal, it is reported in [74] that wideband speech reconstructed through this technique is better than narrowband speech. The authors especially note the inaccurate reconstruction of the fricatives /f/ and /s/. No comparison of performance relative to other techniques, however, is reported. Furthermore, as can be deduced from the discussion above, the computational cost of this technique is quite high, even when only considering the extension stage. Indeed, as reported in [74], values of $N = 64$ and $M = 16$, for example, are required for reasonable performance. It is likely that such high computational requirements are behind the lack of its adoption in the literature, particularly when compared to the less computationally-expensive yet highly-performing GMM-based techniques described next.

II. Gaussian mixture models

Gaussian mixture models (GMMs) have been widely and successfully used to statistically model speech signals in a variety of fields, most notably ASR [77], speaker identification [40], and speaker—or voice—conversion [78, 79]. First proposed and detailed in [80] as an approximation to arbitrary densities, a GMM $\mathcal{G}(\mathbf{x}; M, \mathrm{A}, \Lambda)$[43] approximates the distribution of an n-dimensional random vector $\mathbf{X}: \Omega \to \mathbb{R}^n$ by a mixture of M n-variate Gaussians defined by the set of 2-tuples $\Lambda = \{\lambda_i := (\boldsymbol{\mu}_i, \mathbf{C}_i)\}_{i\in\{1,\dots,M\}}$ and weighted by the priors $\mathrm{A} = \{\alpha_i := P(\lambda_i)\}_{i\in\{1,\dots,M\}}$, i.e.,[44]

$$\begin{aligned} \mathbf{x} \sim \mathcal{G}_{\mathbf{x}} := \mathcal{G}(\mathbf{x}; M, \mathrm{A}, \Lambda) &\triangleq \sum_{i=1}^{M} \alpha_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \mathbf{C}_i) \\ &= \sum_{i=1}^{M} \frac{\alpha_i}{(2\pi)^{n/2}|\mathbf{C}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{C}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]. \end{aligned} \tag{2.13}$$

The ability of GMMs to model the complex realizations of speech is most aptly described in [40]—quoted below—which was mainly concerned with speaker identification, but whose arguments nevertheless equally apply to speaker-independent speech in general (our generalizations and notes in parenthesis).

> The first motivation (for using Gaussian mixture densities as a representation of speaker identity and speech in general) is the intuitive notion that the individual component densities of a multi-modal density, like the GMM, may model some underlying set of acoustic classes. It is reasonable to assume the acoustic space corresponding to a speaker's voice (and speaker-independent speech in general) can be characterized by a set of acoustic classes representing some broad phonetic events, such as vowels, nasals, or fricatives. The spectral shape of the ith acoustic class can in turn be represented by the mean $\boldsymbol{\mu}_i$ of the ith component density, and variations of the average spectral shape can be represented by the covariance matrix $\mathbf{C}_i$. Because all training or testing speech is (usually) unlabeled, the acoustic classes are *hidden* in that the class of an observation is unknown. Assuming independent feature vectors, the observa-

[43] Unless needed for clarity, we will often drop the variables from a distribution's notation in order to simplify expressions.

[44] The symbol ~ denotes *is drawn from the distribution.*

tion density of feature vectors drawn from these hidden acoustic classes is a Gaussian mixture.

The second motivation is the empirical observation that a linear combination of Gaussian basis functions is capable of representing a large class of sample distributions. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily-shaped densities.

Indeed, it was shown in [80] that any continuous *pdf* can be approximated arbitrarily closely by a Gaussian mixture. This important property is primarily the reason that GMMs generally outperform other mapping techniques in regards to speech modelling. We illustrate this property next in Section 2.3.3.5.

We further add a third motivation for specifically using Gaussian mixtures to model speech, as opposed to other multi-modal densities. By considering that each of the different phonetic events of speech is, in fact, a sum of the acoustic manifestations of several independent physiological variables with specific means and variances tied to that phonetic event, e.g., glottal excitation, tongue position, lip rounding, etc., then, by the Central Limit theorem[45], the sum of these random variables for each acoustic class is asymptotically a normal distribution, and the overall multi-class distribution is asymptotically a Gaussian mixture.

In the context of BWE, GMMs were first proposed for highband *and* lowband spectral envelope reconstruction by Park [82]. For spectral transformation in general, a single GMM is used to model the joint density, $p_{\mathbf{XY}}(\mathbf{x},\mathbf{y})$, of narrowband random features vectors, $\mathbf{X}$, and the target random feature vectors, $\mathbf{Y}$. The target feature space is either that of wideband speech including lowband as well as highband frequencies, as in [82], or of only highband speech, as in [54, 55]. The advantages of the two approaches are compared in Section 3.3.2. Parameters of the GMM are optimized in a training stage using the EM algorithm for maximum likelihood (ML) estimation. As derived by Kain in [78],[46] an MMSE highband (or wideband) spectral envelope estimate, $\hat{\mathbf{y}}$, is generated in the extension

[45]The Central Limit Theorem (with Lindeberg's condition) states that *the normalized sum of a large number of mutually independent random variables with zero means and finite variances tends to the normal distribution provided that the individual variances are sufficiently small.* See [81, Chapters 1 and 2] for a history of the development of the theorem.

[46]Kain's paper—[78]—was, in fact, concerned with speaker conversion rather than bandwidth extension. In the speaker conversion problem, the source speaker's speech is represented by the random feature vectors $\mathbf{X}$, and the target speaker's by $\mathbf{Y}$.

stage as a function of the input vector, $\mathbf{X} = \mathbf{x}$, and quantities derived from the joint *pdf*, and is given by

$$\hat{\mathbf{y}} = \sum_{i=1}^{M} P(\lambda_i|\mathbf{x})\, E[\mathbf{Y}|\mathbf{x}, \lambda_i]. \tag{2.14}$$

The derivation of this MMSE estimation is given in Section 3.3.1, and will be integral to our work in Chapter 5 on extending the GMM framework to exploit speech memory for BWE performance improvement.

Computationally, GMMs are more expensive in the training stage than the popular codebook mapping techniques since: (a) the EM algorithm is more expensive than the LBG algorithm, and (b) clustering during codebook training for BWE is typically performed only on the narrowband feature vectors, $\mathbf{X}$, whereas joint density parameter estimation is performed on the longer supervectors, $\mathbf{Z} = \begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix}$, thereby requiring more complex models, i.e., models with more parameters to model the additional degrees of freedom, and, in turn, higher training data and computational requirements. The earlier GMM-based speaker conversion technique of [79] is akin to codebook mapping in that it considers only the narrow band during GMM training, and hence, is computationally less expensive than joint density modelling.[47] In the context of BWE, however, this earlier technique discards the superior ability of GMMs to capture the cross-band correlations central to BWE since it only models narrowband—rather than wideband—speech. Generally, concerns regarding training computational requirements should not be overstated. With the ongoing increase in computational power of signal processing hardware and and the fact that model training is almost always performed offline, the computational cost associated with offline training is increasingly becoming a secondary concern much less important than modelling capability and BWE performance.

Confirming the validity of the motivations described above, the performance of GMM-based BWE techniques has been shown to be superior to that of codebook-based ones, subjectively as well as objectively. In [82], for example, wideband speech reconstructed through GMMs as described above, is judged preferable to codebook-based wideband speech 65% of the time, in both speaker-independent and -dependent implementations. Objectively, the spectral distortion—calculated over the full wideband, i.e., including distortions in both

[47] The target data in [79] is obtained from source data using a piecewise-linear mapping function of quantities derived from the source data GMM. Parameters of the mapping are computed by solving normal equations for a least squares problem, based on the correspondence between the source and target data.

lowband and highband frequencies—of GMM-based extended wideband speech relative to the original wideband reference is 0.56 dB and 0.42 dB lower than the distortion in wideband speech extended using codebook mapping, in the speaker-independent and -dependent implementations, respectively. An even higher spectral distortion reduction of 0.96 dB is reported in [54], although calculated only for the highband frequencies.

III. Hidden Markov models

Ubiquitous in ASR [10, 77], hidden Markov models (HMMs) can be viewed as an extension to the statistical modelling achieved by GMMs.[48] Rather than using a single GMM to model the whole acoustic space as described above, HMMs employ multiple GMMs by dedicating a GMM to each individual HMM state. These states—the characteristic feature of HMMs distinguishing them from single GMMs—exploit interframe dependencies as an integral factor in the statistical modelling of speech (by generating a probabilistic model of state transitions). Thus, HMMs can be thought of as providing a finer resolution of the acoustic space along a temporal axis in addition to the spectral axes of GMMs. Due to the additional complexity associated with such a temporal axis, however, they are limited to first-order modelling (where the probability of being in a particular state depends only on the immediately preceding state) in the vast majority of implementations, in ASR and elsewhere.

There have been two distinct approaches to using HMMs for BWE statistical modelling. The first approach, proposed in [84], employs conventional first-order left-to-right HMMs typical in ASR, where models correspond to phonemes. HMM states with diagonal-covariance GMMs model wideband speech represented by the concatenation of subband feature vectors, i.e., GMMs model the joint narrowband-highband feature vector *pdf*, $p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})$, thereby learning cross-band correlations. Conventional HMM training to estimate transition probabilities and GMM parameters is performed using the Baum-Welch algorithm [85].[49] By simply splitting the mean and covariance diagonals, the trained wideband HMMs, $\{\Xi\}$, are split into separate subband HMMs, $\{\Xi_{\mathbf{x}}\}$ and $\{\Xi_{\mathbf{y}}\}$ for the narrowband and highband subband HMMs, respectively. These subband HMMs share the same HMM structure

[48] The basic theory of HMMs was published in a series of classic papers by Baum and his colleagues in the late 1960's and early 1970's, and was implemented for speech processing applications by Baker at CMU and by Jelinek and his colleagues at IBM in the late 1970's. See [83, Section 2.2].

[49] The Baum-Welch algorithm is an example of a forward-backward algorithm, and is a special case of the EM algorithm.

and transition probabilities but differ in GMM parameters. In the reconstruction phase, observation sequences of narrowband feature vectors are decoded by the Viterbi algorithm [86] using $\{\Xi_{\mathbf{x}}\}$; for each observation sequence $\mathbf{X}(m) = [\mathbf{x}(1), \ldots, \mathbf{x}(m)]$, the overall state sequence $\mathbf{S}(m)$—stretching across narrowband phoneme models—maximizing the likelihood $P(\mathbf{X}(m)|\{\Xi_{\mathbf{x}}\})$ is found. Since $\{\Xi_{\mathbf{x}}\}$ and $\{\Xi_{\mathbf{y}}\}$ models share the same state sequences and transition probabilities, the highband models corresponding to the sequence of phonemes obtained by Viterbi decoding are simply connected. This narrowband-to-highband state sequence mirroring is illustrated in Figure 2.7. Finally, the optimal sequence of highband envelope feature vectors is calculated through the highband models and state sequence as that which maximizes the likelihood $p(\mathbf{Y}(m)|\mathbf{S}(m), \{\Xi_{\mathbf{y}}\})$. This technique has the advantage of jointly modelling narrowband and highband content through GMMs. However, it requires large amounts of labelled training data such that phoneme HMMs can be adequately trained. Despite the potential of this HMM-based BWE approach, its performance has not been compared to that of others, statistical or otherwise, and has not received much adoption beyond [84], likely due to its high complexity and training data requirements. Furthermore, no objective or subjective performance evaluations, other than visual spectrogram comparisons, are reported in [84].

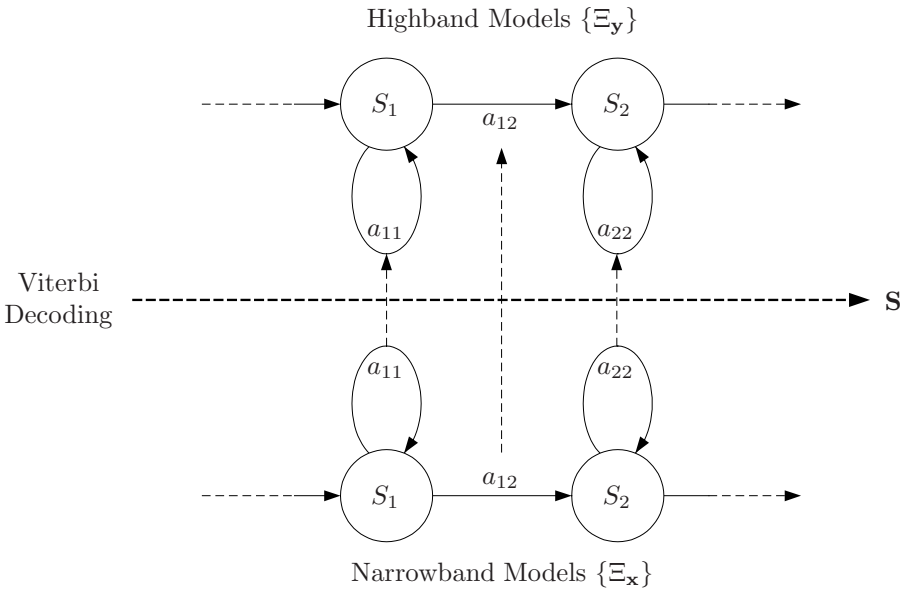


Fig. 2.7: Narrowband-to-highband state sequence mirroring in BWE using subband HMMs.

The second approach, proposed in [39] and, with a slight variation, in [87], uses a single

HMM where the left-to-right transitional constraint is relaxed, i.e., in addition to self-transitions, transitions are allowed back and forth between all N_s states of the model. In contrast to the first approach described above, only narrowband spectral envelopes are modelled by the state-specific GMMs. Thus, cross-band correlations are not modelled through joint-density Gaussian mixture modelling as in the first approach. Rather, cross-band correlations are learned indirectly by associating a VQ codebook of highband spectral envelopes with the HMM states modelling the corresponding narrowband envelopes. In [39], the highband codebook is trained first in a preprocessing step. Each of the highband codewords is then assigned to a particular HMM state. HMM parameters—namely GMM parameters and state prior and transition probabilities—can then be easily estimated given the true highband feature vector sequences and their narrowband counterparts in the training data set. Alternatively, as shown in [87], the HMM can be trained using the Baum-Welch algorithm on the narrowband training data independently of the highband data. The highband codebook can then be built in a postprocessing step by associating each of the HMM states to a particular codebook centre codevector based on the available correspondence between narrowband and highband training data.

In the extension stage, a continuous MMSE estimate of the highband spectral envelope at frame m, $\hat{\mathbf{y}}(m)$, is derived and estimated as a function of the highband codebook centres, $\{c_i^{\mathbf{y}}\}_{i\in\{1,\ldots,N_s\}}$, and the posterior probabilities $\{P[S_i(m)|\mathbf{X}(m)]\}_{i\in\{1,\ldots,N_s\}}$—the probabilities of being in each of the states $\{S_i\}_{i\in\{1,\ldots,N_s\}}$ at frame m given the narrowband observation sequence up to frame m, $\mathbf{X}(m) = [\mathbf{x}(1),\ldots,\mathbf{x}(m)]$. The MMSE estimate is given by

$$\hat{\mathbf{y}}(m) = \sum_{i=1}^{N_s} c_i^{\mathbf{y}} P[S_i(m)|\mathbf{X}(m)], \tag{2.15}$$

where the probabilities $\{P[S_i(m)|\mathbf{X}(m)]\}_{i\in\{1,\ldots,N_s\}}$ are estimated through a recursive technique similar to the forward pass of the forward-backward algorithm, making use of the first-order Markov assumption as well as Bayes' rule to estimate $\{P[S_i(m)|\mathbf{X}(m)]\}$ as a function of the state GMM *pdf*s, $\{p[\mathbf{x}(m)|S_i(m)]\}$.

The BWE performance gains achieved by this second HMM-based approach increase with the number of states/codevectors as well as the number of components in state GMMs. Performance in both [39] and [87] seems to saturate at $N_s = 64$. No performance comparison relative to other techniques (even those using a single large GMM as in [55, 82]) is reported in [39]. In [87], performance was compared only to the piecewise-linear mapping

approach of [60] (where narrowband space is clustered using thresholds of reflection coefficients), rather than GMMs, showing an average PESQ[50] improvement of roughly 0.28 (from 3.72 for piecewise-linear mapping per [60] to 4.0 using an HMM with $N_s = 64$), a modest figure considering that the reference is that of piecewise-linear mapping. Computationally, however, this single-HMM approach is much less expensive than the first approach of [84], particularly in training (since neither labelled data nor Baum-Welch training are required) and to a lesser extent in extension, although more expensive than single-GMM approaches nonetheless.

2.3.3.5 Comparing mapping performance: An illustrative example

To illustrate the performance of the spectral envelope mapping methods described above in regards to their ability to model the true narrowband-to-highband mapping, we use a simple one-to-one 1-dimensional mapping problem as follows. Let $X{:}\,\Omega \to \mathbb{R}^1$ and $Y{:}\,\Omega \to \mathbb{R}^1$ represent continuous random variables on the input and and output sample spaces, respectively. We assume that the input features, x, have an underlying 4-component GMM distribution with equal weights, unit variances and means drawn randomly from the uniform distribution $\mathcal{U}(1,9)$;[51] i.e.,

$$x \sim \sum_{i=1}^{M^x} \alpha_i \mathcal{N}(x;\mu_i,\sigma_i^2) = \sum_{i=1}^{M^x} \alpha_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2}\left[\frac{x-\mu_i}{\sigma_i}\right]^2\right), \tag{2.17}$$

with

$$M^x = 4, \quad \text{and} \quad \forall i \in \{1,\ldots,M^x\}: \;\; \alpha_i = \frac{1}{M^x}, \;\; \sigma_i = 1, \;\; \mu_i \sim \mathcal{U}(1,9). \tag{2.18}$$

We also assume that the output target space, $\Omega_Y \subseteq \mathbb{R}^1$, is a nonlinear one-to-one mapping of the input target space, $\Omega_X \subseteq \mathbb{R}^1$, given by the Gaussian transformation:

$$Y = \mathcal{T}(X) \triangleq b \sum_{j=1}^{M^y} \alpha_j \mathcal{N}(x;\mu_j,\sigma_j^2), \tag{2.19}$$

[50] The PESQ—perceptual evaluation of speech quality—measure was developed to model subjective tests commonly used in telecommunications, particularly MOS. See Section 3.4 for details.

[51] The distribution $\mathcal{U}(a,b)$ denotes the uniform *pdf* of a random variable $X{:}\,\Omega \to \mathbb{R}^1$; i.e.,

$$\mathcal{U}(a,b) \coloneqq p_X(x) = \begin{cases} \dfrac{1}{b-a}, & \text{for } a < x < b, \\ 0, & \text{elsewhere.} \end{cases} \tag{2.16}$$

where

$$M^y = 100, \quad b = 100, \quad \text{and}$$
$$\forall j \in \{1,\ldots,M^y\}: \ \alpha_j = \frac{|w_j|}{\sum_{k=1}^{M^y}|w_k|}, \ w_j \sim \mathcal{N}(5,1), \ \sigma_j \sim \mathcal{U}\left(\tfrac{1}{4},\tfrac{1}{2}\right), \ \mu_j \sim \mathcal{U}(0,10). \tag{2.20}$$

Using this true model of the $\Omega_{XY} \subseteq \mathbb{R}^2$ space with a fixed realization of the parameters μ_i, w_j, σ_j, and μ_j in Eqs. (2.18) and (2.20), we generate 10^5 2-dimensional data points for the training of the various mapping techniques to be compared. Figure 2.8 illustrates the true $\Omega_X \to \Omega_Y$ mapping as well as the mapping modelled by each of the following techniques:

Figure 2.8(a) **Linear mapping** The $\Omega_X \to \Omega_Y$ mapping is modelled as $y = a_1 x + a_0$ where the slope, a_1, and scale, a_0, are obtained using a least-squares fit of the training data.

Figure 2.8(b) **Codebook mapping** A 4-codevector[52] input space codebook, C^x, is trained using VQ of the input features, x, of the training data. A shadow output space codebook, C^y, is then generated with the y codevectors, $\{c_i^y\}_{i\in[1,4]}$, obtained by averaging the y features corresponding to the x features classified into each of the C^x Voronoi.

Figure 2.8(c) **Piecewise-linear mapping** Similar to the piecewise-linear technique of [63], the C^x codebook trained above is used to cluster the training (x, y) pairs into 4 separate clusters for each of which a linear model is estimated.

Figure 2.8(d) **Codebook mapping with interpolation** The shadow codebook output described above is smoothed using weighted interpolation of the K-nearest c^y codevectors in a manner similar to that of [68], where $K = 3$ and the weights are determined based on the squared Euclidean distance between the input features x and the c^x codevectors. Interpolation at the outer halves of edge cells increases distortion, and hence, is omitted in these regions. Thus, output feature estimates, $\hat{y}$, are given by

$$\hat{y} = \begin{cases} \sum_{k=1}^{K} w_k c_k^y \ \text{ where } \ w_k = \dfrac{\|x - c_k^x\|^{-2}}{\sum_{i=1}^{K}\|x - c_i^x\|^{-2}}, & \text{for } \ \min\limits_i c_i^x \le x \le \max\limits_i c_i^x, \\ c_i^y \ \text{ where } \ i = \arg\min\limits_i c_i^x, & \text{for } \ x < \min\limits_i c_i^x, \\ c_i^y \ \text{ where } \ i = \arg\max\limits_i c_i^x, & \text{for } \ x > \max\limits_i c_i^x. \end{cases} \tag{2.21}$$

[52] Since we are using scalar features in this example, referring to codebook centres as codevectors is technically a misnomer. To avoid confusion, however, we continue to refer to codebook centres as such in conformity with convention.

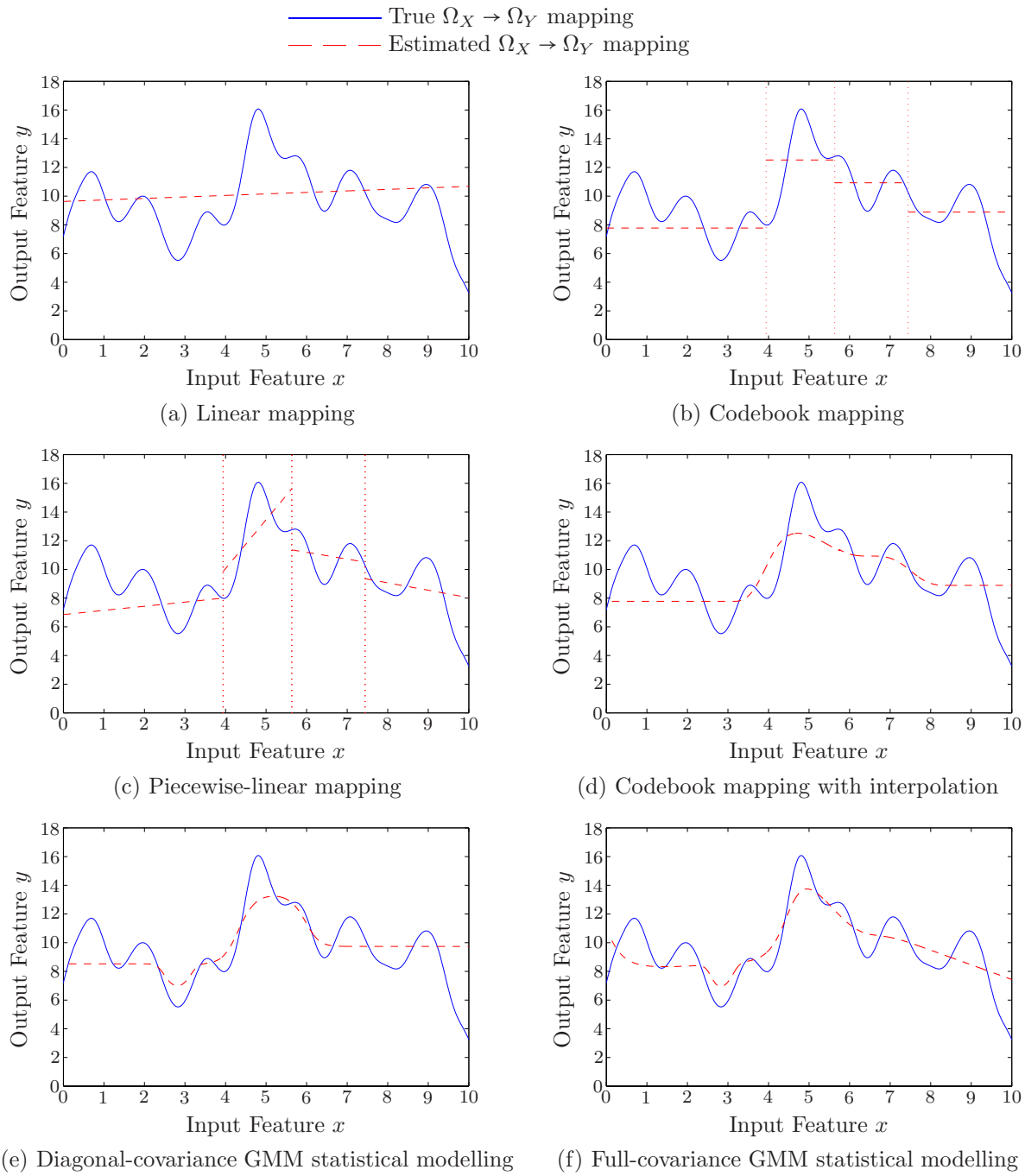
Figure 2.8(e) **Statistical modelling using diagonal-covariance GMMs** A GMM with 4 diagonal-covariance component densities is trained on the 10^5 training (x, y) pairs using the EM algorithm. Output feature estimates, $\hat{y}$, are obtained using MMSE estimation as described in Section 3.3.1. Figure 2.8(e) shows the $\hat{y}$ estimates corresponding to the training x features.

Figure 2.8(f) **Statistical modelling using full-covariance GMMs** As above but using full-covariance component densities.

Figure 2.8 clearly illustrates the continuity properties (or lack thereof) of these mapping techniques as well as their ability to model a nonlinear mapping relationship. Table 2.1 further compares the MSE performance and the complexity of these techniques in terms of the number of model parameters requiring estimation. It is clear that, at comparable or slightly higher model complexity, statistical modelling through GMMs outperforms all other techniques in its ability to closely model nonlinear relationships. As described in Section 2.3.3.4, GMMs are characterized, however, by higher computational cost in the offline training stage compared to other techniques. Nonetheless, GMMs are increasingly becoming the method of choice for BWE spectral envelope mapping due to their superior modelling ability, particularly with the computational concerns associated with offline training further becoming a distant second to that of BWE performance. Indeed, as described in Section 3.3.3, it is the superior modelling ability of GMMs with full covariances (where cross-band correlations can be explicitly captured in cross-band covariance terms) that make them the best tool to study the role of cross-band correlation on BWE performance in general, and the role of speech memory in increasing such cross-band correlations in particular.

Table 2.1: MSE performance and model complexity of the mapping methods used in Figure 2.8.

Mapping method		MSE	Number of model parameters
Linear mapping	[Figure 2.8(a)]	6.56	2
Codebook mapping	[Figure 2.8(b)]	3.13	8
Piecewise-linear mapping	[Figure 2.8(c)]	2.29	12
Codebook mapping with interpolation	[Figure 2.8(d)]	2.55	9
Diagonal-covariance GMM statistical modelling	[Figure 2.8(e)]	1.77	12
Full-covariance GMM statistical modelling	[Figure 2.8(f)]	1.40	20



(a) Linear mapping

(b) Codebook mapping

(c) Piecewise-linear mapping

(d) Codebook mapping with interpolation

(e) Diagonal-covariance GMM statistical modelling

(f) Full-covariance GMM statistical modelling

Fig. 2.8: Comparing the performance of spectral envelope mapping techniques using a simple one-to-one 1-dimensional $\Omega_X \rightarrow \Omega_Y$ mapping problem. See Table 2.1 for a comparison of MSE performance and model complexity.

2.3.4 Highband energy estimation

As highband (and, optionally, lowband) content generated by BWE is combined with the original narrowband signal to generate wideband speech, it is important that highband energy is adjusted to suitable levels relative to narrowband signal energy. Highband energy overestimation introduces audible artifacts in the extended region that can make the extended wideband signal often sound more annoying than the original narrowband signal. In contrast, underestimation of highband energy undermines the value of bandwidth extension itself, particularly for sounds with high-frequency energies, e.g., fricatives.

For BWE techniques where the entire wideband spectrum is reconstructed then bandstop filtered before being added to the narrowband signal, wideband energy adjustments can easily be performed by scaling the reconstructed signal prior to bandstop filtering such that the reconstructed and original input signals have the same energy in the narrowband region, e.g., [61]. Alternatively, appropriate highband energies can be estimated based on the narrowband input, in a manner similar to the mapping or statistical estimation of highband spectral envelopes. This latter approach is, in fact, required for BWE techniques where highband content is directly estimated or mapped from the narrowband input. During training, such techniques typically model the cross-correlation between the usual narrowband feature vectors and an energy ratio σ^2_{rel} (which is more robust than modelling absolute energy values). The ratio is either that of highband to narrowband energy calculated from the wideband training data, or, the ratio of the original highband energies of the training data to those of the corresponding highband signals reconstructed during training specifically for that purpose. In the extension phase, the energy ratio is estimated given the available narrowband input then multiplied by narrowband energy (or the energy of the reconstructed highband signal), thereby generating adequate scaling values for the highband extension.

Thus, any of the aforementioned spectral envelope mapping techniques can be used for energy-ratio modelling. Codebook mapping is used in [82], for example, whereas a dedicated GMM is used in [55]. In [84], a dedicated energy-ratio subband HMM is extracted from the wideband HMM, and is used to estimate highband energy in a manner identical to that used for highband feature vector estimation as described in Section 2.3.3.4 and illustrated in Figure 2.7; i.e., energy-ratio HMMs are connected according to the optimal narrowband HMM state sequence obtained by Viterbi decoding. An elaborate scheme is

further proposed in [57] for the purpose of reducing highband energy overestimations in particular. An asymmetric cost function is introduced such that highband energy overestimations are penalized more than underestimations during MMSE energy-ratio estimation via a highband-to-narrowband energy-ratio GMM. As shown in [57], such an asymmetric cost function results in MMSE energy-ratio estimates as functions of the GMM posterior distributions, $\{p(\sigma^2_{\text{rel}}|\lambda_i)\}_{i\in\{1,\ldots,M\}}$,[53] such that broad distributions are penalized more than narrow distributions. This results in energy-ratio estimates that take into account the confidence of the estimate (the narrower the posterior probability of the GMM, the higher the confidence in the derived energy-ratio estimate), where frames with unreliable highband energy-ratio estimates are attenuated. Listening tests of GMM-based extended speech employing this technique in [57] show a significant reduction of severe and moderate highband artifacts.

2.3.5 Relative importance of accuracies in spectral envelope and excitation generation

Many BWE works have observed and reported that accuracy and quality in highband spectral envelope reconstruction is far more important for the subjective quality of extended speech, than in excitation signal generation. For example, informal listening tests in [39]—where modulation is used for highband excitation signal generation—show that, assuming that BWE of the spectral envelope works well, the human ear is amazingly insensitive to distortions of the excitation signal at frequencies above 3.4 kHz. Spectral gaps of moderate width resulting from choosing a modulation frequency above 3.4 kHz are almost inaudible. Furthermore, misalignments of the harmonic structure of speech at high frequencies does not significantly degrade the subjective quality of the extended speech signal. Similarly, in [58] where spectral folding is used, the authors conclude that as long as the spectral envelope shape is similar to the original, the excitation used made almost no difference for the recovery of high frequencies. A similar conclusion is also noted in [88] where the effect of replacing an original wideband excitation signal by another reconstructed using full-wave rectification is very small.

Thus, for the focus of this thesis—studying the effect of speech memory inclusion on BWE performance—we only consider speech memory in spectral envelopes.

[53] See GMM definition in Section 2.3.3.4.

2.3.6 Sinusoidal modelling

BWE techniques synthesizing highband speech through sinusoidal modelling are a less common class of BWE techniques that do not employ LP synthesis but, nevertheless, employ the source-filter model. These techniques make use of the sinusoidal transform coding (STC) [89] and multi-band excitation (MBE) [90] models of speech. Both models make use of the fact that high-quality speech can generally be synthesized as a sum of sinusoids with appropriate frequencies, amplitudes and phases. Rather than estimate a highband excitation signal to excite an LP-synthesis filter defined by the highband LP-based spectral envelopes estimated separately, sinusoidal-based BWE generates highband speech by using the estimated highband spectral envelopes themselves to determine the amplitudes of sinusoids representing the voiced components of speech as well as the spectral shape of white noise representing unvoiced components. Other sinusoid parameters, i.e., frequency and phase, as well as the degree of mixing voiced and unvoiced components, are determined from the narrowband signal. Both components are then added to generate the highband signal. Unlike conventional source-filter model-based BWE, spectral flatness of the excitation is, thus, not an issue in sinusoidal-based BWE since sinusoid amplitudes are determined directly by the spectral envelope. However, pitch estimation is required.

In the context of BWE, highband speech synthesis through STC—proposed in [91]—makes use of the mixed excitation of the source-filter model—as described in Section 2.3.1—where the weights of the periodic (voiced) and random (unvoiced) components are determined based on degree of voicing over the entire speech bandwidth. The periodic component is synthesized using the STC model as harmonically-spaced sinusoids. The narrowband signal is analyzed to estimate the model's parameters of phase, pitch and degree of voicing, while the highband spectral envelope is used to determine sinusoid amplitudes. The random component is generated as a highband random sequence spectrally shaped by the estimated highband spectral envelope and scaled according to the estimated degree of voicing.

In the MBE model, on the other hand, the speech spectrum is divided into a number of bands centered on the pitch harmonics where each band can be individually declared as voiced or unvoiced. The MBE model parameters consist of a set of band magnitudes and phases, a set of binary voiced/unvoiced (V/UV) decisions, and a pitch frequency. Proposed by [66], MBE-based BWE is implemented by applying various codebooks to narrowband speech in order to estimate the required per-band high-frequency V/UV decisions as well as

magnitudes for the voiced and unvoiced bands. The highband voiced signal is then obtained in the time domain by applying the estimated parameters to harmonic oscillators. To ensure signal continuity across frames, band magnitudes are linearly interpolated between frames. Unvoiced speech is synthesized in the frequency domain by shaping a unity-variance white noise spectrum with the estimated highband unvoiced spectrum.

While mean opinion scores and informal listening tests reported in [66] and [91], respectively, indicate clear preference for the sinusoidally-extended speech over narrowband speech, it is difficult to quantify the performance of sinusoidal-based BWE since very limited comparisons were made with conventional source-filter model-based techniques. Moreover, the additional complexity associated with estimating the parameters required for sinusoidal-based BWE (namely, pitch, phases, and degree of voicing), compared to conventional techniques, has most likely hindered wider adoption and improvements.

2.4 Summary

BWE relies on the assumption that highband speech closely correlates with its narrowband counterpart. Thus, by learning the cross-band relationships a priori, highband frequency content can be reconstructed given only narrowband input. By using the source-filter model, the BWE problem is reduced to two separate tasks—generating a highband excitation signal and a highband spectral envelope. Several works have shown the latter to be of more importance for the subjective quality of extended speech. Extensive work has been dedicated to investigating and proposing techniques by which to learn the spectral envelope cross-band correlations. Through our analysis of speech and its dynamics presented in Chapter 1, we have shown these cross-band correlations to be rather complex and nonlinear. As such, the ability of the surveyed techniques to model such complex correlations varies greatly depending on their continuity and nonlinearity properties, or lack thereof. We find GMMs, in particular, the tool most suited to our purpose—investigating the role of speech memory in improving BWE performance through apt modelling of cross-band correlations. They outperform codebook-based techniques—the most common of spectral envelope mapping techniques—at comparable or slightly higher model complexity. With offline training concerns being secondary to those of BWE performance, GMMs become especially attractive. Finally, we note that while HMMs provide the additional advantage of exploiting interframe dependencies, their use of speech memory is rather limited.

Chapter 3

Memoryless Dual-Mode GMM-Based Bandwidth Extension

3.1 Introduction

In this chapter, we describe the details of our BWE implementation that will be used as the basis for all developments and evaluations thenceforth in the thesis. We employ a dual-mode BWE system based on that of Qian and Kabal in [55]. Per our comparative analysis of model-based BWE techniques in Section 5.4.3.3, the dual-mode technique of [55] is shown to outperform nearly all comparable techniques, in some cases by a rather wide margin. Furthermore, in addition to using GMM-based statistical modelling—the approach we concluded in Section 2.3.3.5 to be the most suited for our purpose of studying the role of memory in improving the cross-bad correlations central to BWE—for the reconstruction of highband spectral envelopes as well as highband energy ratios, the dual-mode technique exploits equalization to extend the apparent bandwidth of narrowband speech to 100 Hz at the low end and to near 4 kHz at the high end. The *dual-mode* designation thus refers to the use of both equalization and statistical modelling. The complementary highband spectrum up to 8 kHz is statistically estimated using a GMM given parameters of the narrowband signal enhanced by midband equalization in the 3.4–4 kHz range. In parallel, the midband-equalized narrowband signal is also processed to generate an enhanced excitation signal—the equalized bandpass-modulated Gaussian noise (EBP-MGN). Following [55], our baseline BWE system uses line spectral frequencies (LSFs) to parameterize both narrowband and highband spectral envelopes. The motivation for choosing LSFs—briefly mentioned in [55]—is provided in more detail in this chapter. The estimated

highband LSF features, converted to LPCs, are then used together with the estimated excitation signal to reconstruct highband speech through LP synthesis, followed by level adjustment using the statistically estimated energy ratios. Particular details of our BWE implementation—namely parameterization, dimensionality, training and test data, and filter response characteristics—are described.

Since all our BWE systems—memoryless as well as memory-inclusive—presented in this thesis employ GMMs for statistical modelling, the derivation of the MMSE estimation of target features using joint-density GMMs is presented in detail. We also discuss the choice of jointly modelling highband—rather than wideband—spectra with their narrowband counterparts. We then introduce the measures used for BWE performance evaluation throughout our work and discuss the motivations for their choice. Finally, we evaluate BWE performance in memoryless conditions, i.e., without making use of the information in speech dynamics, studying in the process the effects of varying the number of components in the Gaussian mixture, as well as the effects of using diagonal and full covariance matrices. Based on these results, we conclude by establishing the memoryless performance baseline for the future MFCC- and memory-inclusive BWE evaluations in Chapter 5.

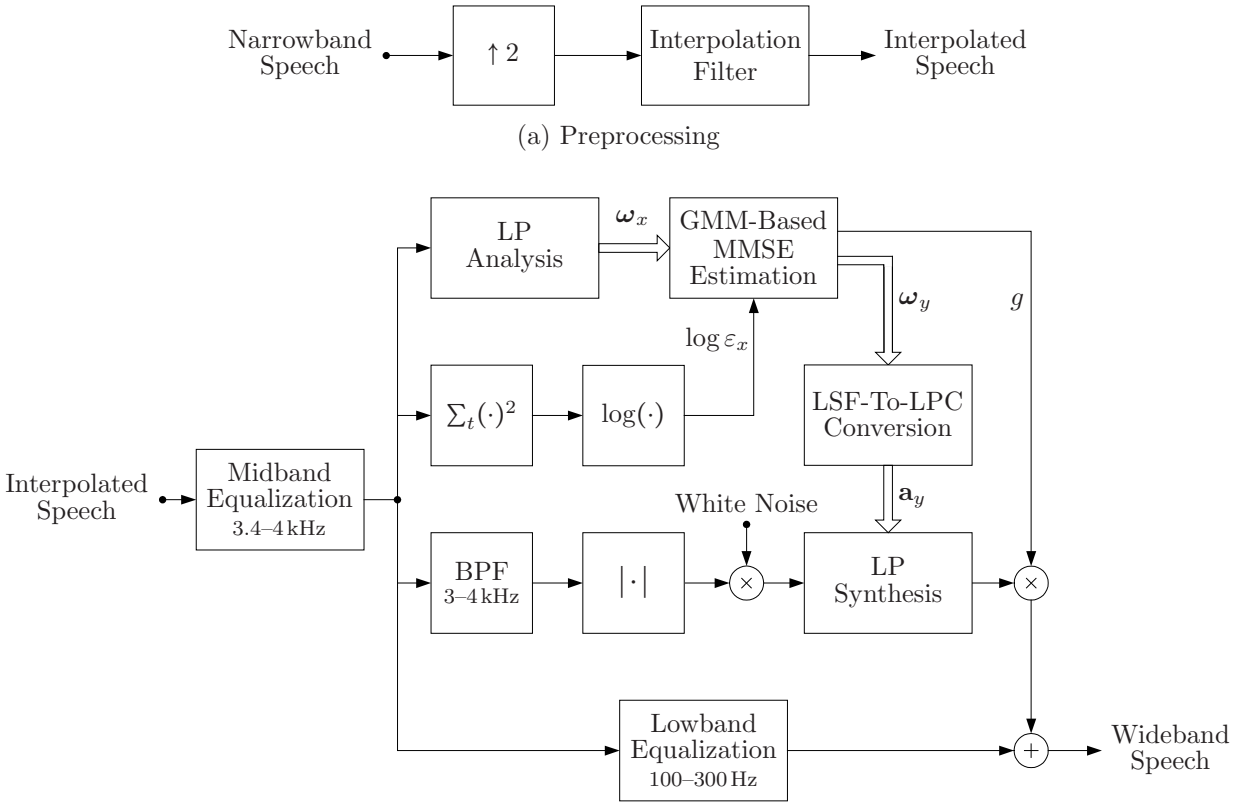
3.2 Dual-Mode Bandwidth Extension

3.2.1 System block diagram and input preprocessing

Figure 3.1 shows the overall system block diagram. As shown in Figure 3.1(a), the input narrowband signal sampled at $F_s = 8\,\text{kHz}$ is preprocessed by first upsampling to $F_s = 16\,\text{kHz}$. All subsequent processing is performed at $F_s = 16\,\text{kHz}$. A lowpass interpolation filter is then used for anti-aliasing, with its frequency response shown in Figure 3.2(a).[54] All filters described in this chapter are equiripple linear-phase finite impulse response (FIR) filters designed using the filter design tool of Kabal [92].[55]

[54] To better illustrate response in transition regions, some of the filter frequency responses illustrated in this chapter are shown only for part of the full 0–8 kHz frequency range of the filter.

[55] Filters are specified in terms of the desired response in multiple passbands and stopbands. Band specifications include desired value, relative weighting and limits on the allowed response values in the band. The resulting filters are weighted minimax approximations (with constraints) to the given specifications. Filter coefficients have an even symmetry around the middle of the filter. See [92] for more details on the design procedure and constraint definitions.



(a) Preprocessing

(b) Main processing

(c) GMM-based MMSE estimation

Fig. 3.1: The dual-mode bandwidth extension system.

3.2.2 LSF parameterization

Originally developed by Itakura in [93] as an alternative representation of LPCs, LSFs have become ubiquitous in speech processing for their quantization error resilience and perceptual significance properties. It is well known that LPCs are not suited for speech coding and quantization due to their large dynamic range and, more importantly, due to

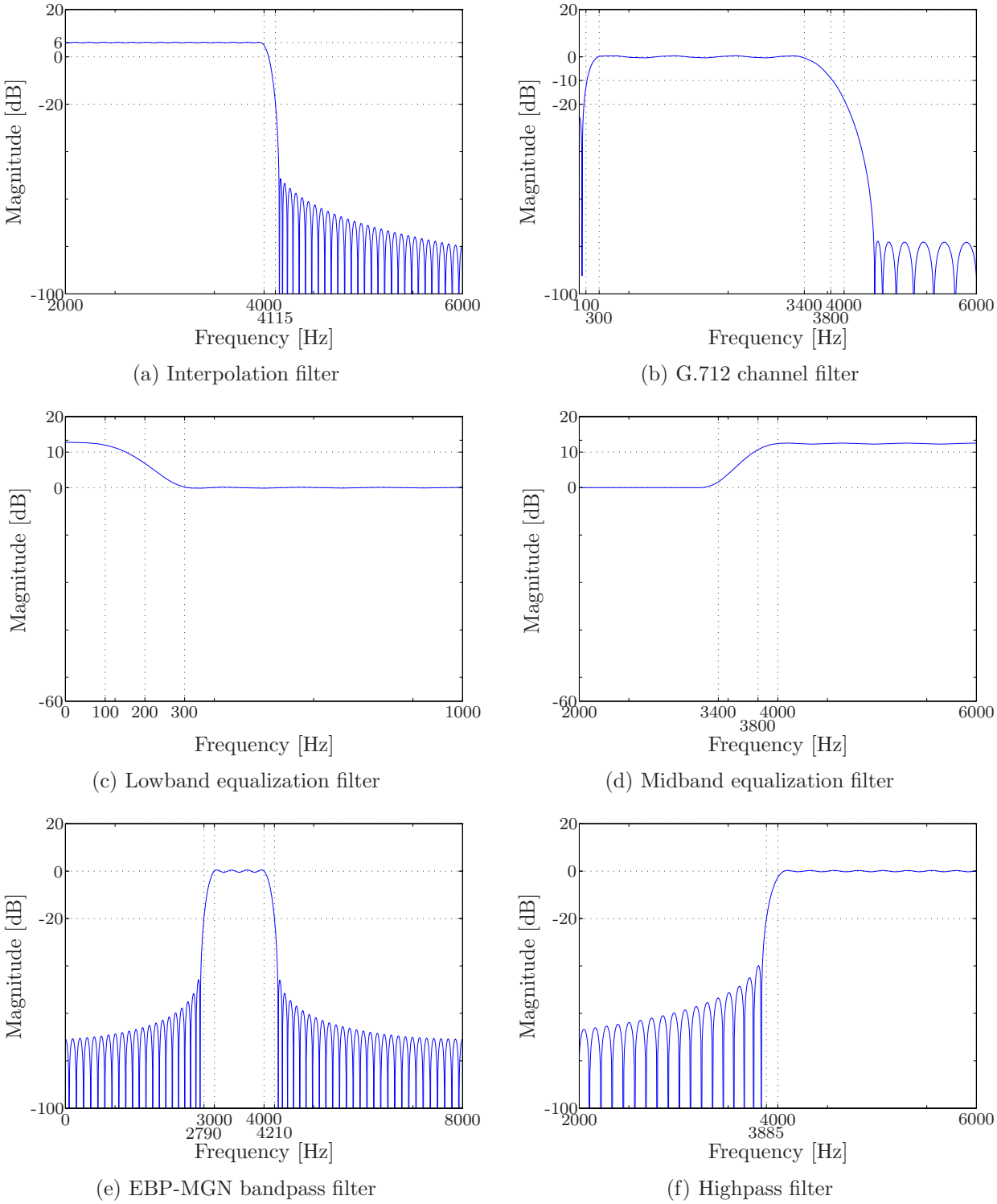


Fig. 3.2: Dual-model BWE system filter responses.

the potential instability of synthesis filters based on quantized and/or interpolated LPCs. In contrast, LSFs are quite robust to estimation and quantization errors, and furthermore, easily guarantee synthesis filter stability—by ensuring appropriate LSF ordering.

LSFs are an artificial mathematical representation generated from LPCs by finding the roots of the two z-polynomials, $P(z)$ and $Q(z)$, corresponding to the p-order LP analysis filter, $A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$, with additional reflection coefficients of 1 and −1, respectively. In other words, $P(z)$ corresponds to the vocal tract represented by $A(z)$ but with the glottis completely closed while $Q(z)$ corresponds to that with an open glottis; i.e.,

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A(z^{-1}), \\ Q(z) &= A(z) - z^{-(p+1)} A(z^{-1}), \end{aligned} \tag{3.1}$$

and

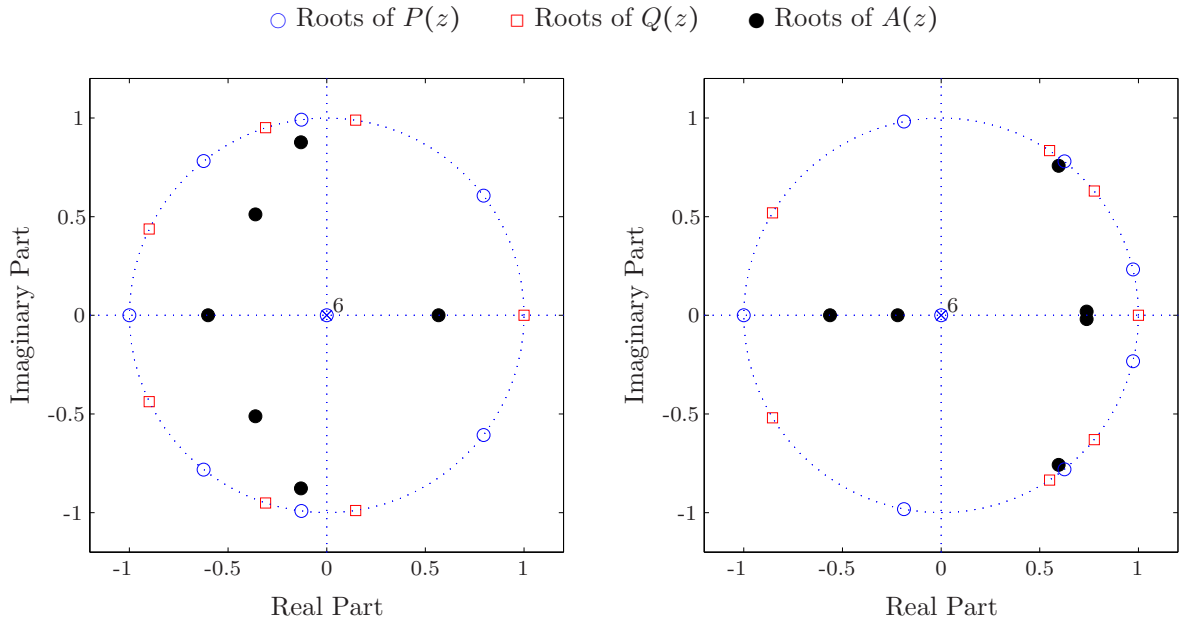
$$A(z) = \frac{P(z) + Q(z)}{2}. \tag{3.2}$$

Due to the symmetry and anti-symmetry properties of $P(z)$ and $Q(z)$, respectively, it can be shown that their roots exist in conjugate pairs, representing interlaced zeroes existing only on the unit circle. The phases of these zeroes in the z-plane represent frequencies, and hence, are referred to as line spectral frequencies. Since the zeroes occur in conjugate pairs, only those within the open $(0, \pi)$ range are needed to fully represent the original LPCs. Furthermore, the interlaced order of LSFs allows the minimum-phase property of $A(z)$ to be easily preserved with LSF quantization, thus ensuring stability of the corresponding LP synthesis filters. These properties have been proven by Soong and Juang [94] for LSFs in particular, and independently proven earlier by Schüssler [95] in the more general context of the stability of discrete systems. In [96], Bäckström provides rigourous and up-to-date proofs and extensions of the properties of line spectrum pair polynomials in general.

By representing the vocal tract transfer function $H(z) = 1/A(z)$ in terms of $P(z)$ and $Q(z)$ as in Eq. (3.2), LSFs are shown to demonstrate a direct correspondence to the shape of the spectral envelope. The closed $[0, \pi]$ range corresponds to the whole frequency range of the spectrum. Dense distributions of LSFs represent high magnitude regions of the spectrum, while scattered distributions represent low magnitude ones. Hence, in contrast to LPCs, local errors in LSF values only tend to cause local spectral distortions.

Figure 3.3 illustrates these properties for two 20 ms windows from the *sailing* waveform of Figure 1.2(a), after it has been lowpass filtered and downsampled to $F_s = 16$ kHz. The

interlaced ordering of LSFs is clear. Figure 3.3(a), corresponding to the fricative /s/, shows a dense distribution of LSFs for phases greater than $\pi/2$, i.e., frequencies above 4 kHz, indicating mostly highband energy. In contrast, Figure 3.3(b), corresponding to the vowel /e/, shows the opposite scenario. These observations agree with the energy distributions for the same intervals in the spectrogram of Figure 1.2(a).



(a) /s/ (b) /e/

Fig. 3.3: Illustrating the properties of LPCs and LSFs in the z-plane; roots of the 6-order LP analysis filter $A(z)$ are represented by ●, while roots of the symmetric $P(z)$ and anti-symmetric $Q(z)$ LSF polynomials are represented by ○ and □, respectively. Subfigure (a) represents the zeroes of the fricative /s/ in the 100–120 ms window of Figure 1.2(a) (after the waveform was lowpass filtered and downsampled to F_s = 16 kHz), whereas Subfigure (b) represents the zeroes of the vowel /e/ in the 240–260 ms window of the same waveform.

These properties make LSFs especially attractive for BWE, and as such, have been used to varying extents for BWE spectral envelope parameterization in [55, 59, 60, 66, 87], among others. In particular:

- any linear combination of LSF vectors (as in the case of GMM-based highband MMSE estimates) will always preserve the interlaced ordering property, thus guaranteeing the minimum-phase and LP synthesis filter stability properties,
- unlike LPCs, the perceptual significance of LSFs (where the properties of formants

and valleys can be related to LSF pairs) improves the ability of GMMs to capture perceptually significant characteristics of the acoustic space of speech,

- by virtue of their correspondence to the spectral envelope, BWE using LSFs is more robust to estimation errors as individual errors do not degrade the whole envelope.

Conversion of LSFs back to LPCs is rather straightforward; the symmetric $P(z)$ and anti-symmetric $Q(z)$ polynomials of Eq. (3.1) are generated using the interlaced LSFs as the phases of the polynomial unit-circle roots, followed by averaging per Eq. (3.2) to obtain the analysis filter, $A(z)$.

In this work, we denote LSF feature vectors by $\boldsymbol{\omega}$, where an n-LSF vector $\boldsymbol{\omega}$ is interpreted as a realization of the continuous LSF random vector $\boldsymbol{\Omega} \in \{\omega \in \mathbb{R} : 0 < \omega < \pi\}^n$. Thus, $\boldsymbol{\omega}_x$ and $\boldsymbol{\omega}_y$ in Figure 3.1 denote narrowband and highband LSF feature vectors, respectively.

3.2.3 Equalization

In reality, typical telephone channel attenuation in the 100–300 and 3400–4000 Hz bands is not abrupt; rather, it is somewhat smooth. Thus, provided filtering response characteristics are known, the speech signal in those ranges can be reconstructed by equalization more accurately than by estimation algorithms. Indeed, the ITU-T G.712 Recommendation [9] provides attenuation/frequency requirements for both ranges in the form of frequency masks similar to that of Figure 1.1, e.g., [9, Figure 3/G.712] for channels between two-wire analog ports, in addition to an out-of-band attenuation filter characteristic for $f > 3400$ Hz [9, Figure 10/G.712]. Using these specifications, telephony speech signals can be characterized as follows [55]:

- The channel filter attenuates the speech signal by 0–18 dB in the 3400–4000 Hz range, and by 0–10 dB from 300 to 100 Hz. Figure 3.2(b) shows our implementation of the G.712 channel based on these characteristics. Given the relatively low attenuation in these two bands, speech content therein can be accurately recovered by equalization. The value of equalization over estimation for these ranges becomes even greater when considering their perceptual importance.[56] As discussed in Section 1.1.3.3, the 0.8 bark subband of 3400–3889 Hz was found in [27] to be particularly important. Similarly, it was concluded that highband extension is most effective perceptually

[56] Due to the particular perceptual importance of the low and mid bands, exploiting GMM-based statistical estimation as a corrective post-equalization step to further improve the reconstruction in these bands is discussed in Section 6.2 as potential future work.

when accompanied by lowband extension. Indeed, we showed in Section 1.1.3.1 that the content below 300 Hz provides important cues that help distinguish nasals as well as distinguish between voiced and unvoiced fricatives, stops, and affricates.

- Frequency content above 4000 Hz is missing due to the 8 kHz sampling rate. These lost components can only be reconstructed using any of the spectral envelope reconstruction methods described in Section 2.3.3. Our method of choice is that of statistical GMM estimation.
- To suppress AC coupling interference, current telephone networks provide at least 22 dB attenuation in the 50–60 Hz band using a highpass filter at the transmission side. Hence, these components can not be recovered by equalization. Furthermore, since average fundamental frequencies—whose first few harmonics are important for naturalness—are above 100 Hz, and the finding in [27] that the 0.8 bark 50–131 Hz subband is the least important perceptually below 300 Hz, we do not attempt to reconstruct signals below 100 Hz by statistical estimation.

After [55], two equalizers are designed to recover the attenuated components. The first, shown in Figure 3.2(c), provides a gain of 10 dB at 100 Hz, while the second, shown in Figure 3.2(d), provides a similar gain of 10 dB in the 3800–4000 Hz range. The frequency response of the equalized channel is, thus, almost flat from 100 to 3850 Hz. Although the equalized signal extends only to 4 kHz, it was observed in [55] that its quality is noticeably better than that of narrowband speech, thus confirming the aforementioned perceptual importance of the equalized ranges.[57] The narrowband signal enhanced by midband equalization is used for the generation of the enhanced excitation signal—in the next-to-lowermost path of Figure 3.1(b)—as well as the spectrum envelope and the excitation gain for content above 4 kHz (in the two upper paths of Figure 3.1(b)).

3.2.4 EBP-MGN excitation generation

The basis for the generation of a wideband excitation in [54] and later in [55], is the application of a nonlinearity—full-wave rectification—to a subband of the narrowband signal. As argued in [88], the absolute value function is a good candidate since, unlike the square

[57] Worthy of note is also the confirmation in [55] that equalization, in both the lowband and midband regions, does not unduly emphasize quantization noise for PCM encoded speech, thus allaying the authors' early concern about the potential negative impact of equalizing quantized speech in regions where the signal has been already attenuated prior to quantization—i.e., regions with low signal-to-quantization-noise ratio.

value, it does not require energy normalization. A wideband excitation generated in this fashion will be phase-coherent with the original narrowband signal and further preserves the harmonic structure without any spectrum discontinuities.

As discussed in Section 1.1.3, the average long-term speech energy is mainly concentrated below 1 kHz [11], falling off with a long-term average of 6 dB/octave [10].[58,59] Indeed, as confirmed by the observations in [54], the LP residual of voiced phonemes contains weak pitch harmonics over 4 kHz (in addition to noise-like components in the case of voiced fricatives, stops and affricates) compared to strong harmonics below 3.5 kHz. The unvoiced residuals are noisy in the high band as well as in the low band. As such, the narrowband speech signal in the 2–3 kHz range was initially chosen in [54] as the basis for highband excitation generation. This frequency range, however, is inappropriate since many phonemes, including voiced ones, have weak responses in that region. As described in Section 1.1.3.1, unvoiced fricatives, e.g., /s/ and /f/, have almost no energy below 2.5 kHz. More importantly, however, the nasal /n/ exhibits a spectral null in the 1450–2200 Hz region [10, Section 3.4.5], and the liquid /l/ in English is also often characterized by a deep antiresonance near 2 kHz [10, Section 3.4.4]. In comparison, the 3–4 kHz band is superior; it contains distinctive spectral cues for many fricatives, stops, and affricates, while still containing enough harmonic structure to reproduce high-quality voiced sounds. Hence, since content in this region has already been enhanced by midband equalization, the 2–3 kHz bandpass filter of [54] was replaced by a 3–4 kHz bandpass filter in [55]. Figure 3.2(e) shows the frequency response of this filter.

The midband-equalized bandpass (EBP) signal is then spectrally broadened through straightforward full-wave rectification. The spectrum of the resulting wideband signal exhibits pitch harmonics (for vowel-like voiced sounds), noise (for unvoiced sounds), or both (for mixed sounds), without the discontinuities often associated with the spectral folding and modulation techniques of Section 2.3.2. Finally, the EBP-MGN excitation is obtained by using the bandpass-envelope signal to modulate white Gaussian noise. For voiced sounds, this corresponds in the frequency domain to superimposing the fine harmonic structure of

[58] While the 6 dB/octave rolloff applies only to vowel-like voiced phonemes, unvoiced phonemes—which tend to have a flat spectrum at high frequencies—are typically weaker than voiced ones (compare, for example, spectrogram peak energies in Figure 1.2 for the leading fricatives, /s/ and /f/, versus those of the ensuing vowel /e/).

[59] Pre-emphasis is typically applied to compensate for the 6 dB/octave roll off such that high frequency content is emphasized.

the wideband bandpass-envelope signal on the flat spectrum of the Gaussian noise; see, for example, Figure 3 in [54]. The next-to-lowermost path of the main processing block in Figure 3.1(b) illustrates the generation of the EBP-MGN excitation signal.

3.2.5 Reconstruction of highband spectral envelopes and excitation gain

As described in Section 1.1.3.1, band energies are an important perceptual cue, particularly for manner of articulation. Thus, in addition to the midband-equalized narrowband LSF feature vectors, $\boldsymbol{\omega}_x$, we explicitly include midband-equalized narrowband frame log-energy, $\log \varepsilon_x$, in the narrowband random feature vector representation $\mathbf{X}$; i.e., $\mathbf{X} \triangleq [\boldsymbol{\Omega}_x^T, \log \mathcal{E}_x]^T$. Similarly, in addition to the highband $\boldsymbol{\omega}_y$ LSF feature vectors, we use a highband excitation gain, g, representing the gain required to scale the reconstructed highband signal such that the energy of the reconstructed highband components is equal to the energy of the corresponding frequency band in wideband speech. The excitation gain is calculated as the square root of the energy ratio of the original highband signal, $y(n)$, to the reconstructed one, $\hat{y}(n)$, in each frame; i.e.,

$$g = \sqrt{\frac{\|y(n)\|^2}{\|\hat{y}(n)\|^2}}. \tag{3.3}$$

The true values of the excitation gain g are determined during training by artificially synthesizing the highband signal using: (a) the EBP-MGN excitation generated as described above, and (b) the true highband LPCs.

To model cross-band spectral envelope shape and gain correlations, the dual-model BWE scheme uses two GMMs:

1. $\mathcal{G}_{\mathbf{X}\boldsymbol{\Omega}_y} \coloneqq \mathcal{G}(\mathbf{x}, \boldsymbol{\omega}_y; M^{\mathbf{x}\boldsymbol{\omega}_y}, \mathrm{A}^{\mathbf{x}\boldsymbol{\omega}_y}, \Lambda^{\mathbf{x}\boldsymbol{\omega}_y})$, to statistically model the joint density of midband-equalized narrowband feature vectors, $\{\mathbf{x}\}$, and highband LSF feature vectors $\{\boldsymbol{\omega}_y\}$; and
2. $\mathcal{G}_{\mathbf{X}G} \coloneqq \mathcal{G}(\mathbf{x}, g; M^{\mathbf{x}g}, \mathrm{A}^{\mathbf{x}g}, \Lambda^{\mathbf{x}g})$, to statistically model the joint density of narrowband feature vectors, $\{\mathbf{x}\}$, and the excitation gains, $\{g\}$.

To simplify notation in the sequel, we will often drop the subscript y in GMM and parameter notation when clear from the context; e.g., $\mathcal{G}_{\mathbf{X}\boldsymbol{\Omega}} \coloneqq \mathcal{G}_{\mathbf{X}\boldsymbol{\Omega}_y}$, as well as denote a dual-mode BWE system's $(\mathcal{G}_{\mathbf{X}\boldsymbol{\Omega}}, \mathcal{G}_{\mathbf{X}G})$ GMM tuple by $\boldsymbol{\mathcal{G}}$; i.e., $\boldsymbol{\mathcal{G}} \coloneqq (\mathcal{G}_{\mathbf{X}\boldsymbol{\Omega}}, \mathcal{G}_{\mathbf{X}G})$. Details of the training procedure are discussed in Section 3.2.6 below. In the extension stage, MMSE estimation of $\boldsymbol{\omega}_y$ and g—illustrated in Figure 3.1(c)—is performed are described in Section 3.3.1

3.2.6 System training

Starting with wideband speech sampled at F_s = 16 kHz, the training stage proceeds in a speaker-independent manner as follows:

1. Wideband speech is first filtered by the G.712 channel bandpass filter of Figure 3.2(b) and the highpass filter of Figure 3.2(f), resulting in narrowband and highband signals in the 0.3–3.4 and 4–8 kHz ranges, respectively.
2. Mimicking extension stage processing, the narrowband signal is then equalized in the 3.4–4 kHz range using the midband equalization filter of Figure 3.2(d).
3. The midband-equalized narrowband signal and that of the high band are LP-analyzed to obtain LPCs representing narrowband and highband spectra.
4. The midband-equalized narrowband signal is bandpass filtered in the 3–4 kHz range using the EBP-MGN filter of Figure 3.2(e). The resulting bandpass signal is then full-wave rectified and used to modulate unit-variance Gaussian noise, providing the EBP-MGN excitation signal.
5. Excitation gain data is calculated per Eq. (3.3), using the true highband signal and its artificial counterpart obtained by LP-synthesis with the EBP-MGN excitation and the true highband LPCs obtained in Steps 4 and 3, respectively.
6. Midband-equalized narrowband and highband LPCs are then converted to LSFs.
7. Midband-equalized narrowband log-energies are calculated and appended to narrowband LSFs.
8. Finally, the two GMMs, $\mathcal{G}_{\mathbf{x}\boldsymbol{\Omega}}$ and $\mathcal{G}_{\mathbf{x}G}$, are trained using the EM algorithm [76], for which we calculate initial estimates through Lloyd's K-means clustering algorithm [97].[60]

3.2.7 Dimensionality

The choice for the dimensionality of a spectral representation is a compromise between spectral accuracy, complexity/computation cost, and bandwidth. For an LP-based representation, as in the dual-mode BWE system, poles are needed to represent all formants—two poles per resonance—in the signal bandwidth plus an additional 2–4 poles to approximate

[60] EM training is iteratively performed until a stopping criterion—typically the change in the log-likelihood of the training data given the estimated model parameters—is reached. Similarly, we perform K-means clustering iterations until the relative changes of either: (a) the total squared-error over the training data, or (b) cluster centres, fall below particular thresholds.

possible zeroes in the spectrum and general spectral shaping (e.g., 8 kHz sampled speech is typically represented by 10 poles) [10, Section 6.5.5]. In our implementation of the memoryless dual-model BWE system, we represent midband-equalized narrowband content in the 0.3–4 kHz range by 9 LSFs[61,62] in addition to frame log-energy as mentioned in Section 3.2.5 above, resulting in a total dimensionality of $\mathrm{Dim}(\mathbf{X}) = p = 10$ for the narrowband random feature vector $\mathbf{X}\colon\Omega \to \mathbb{R}^p$.

Since the highband 4–8 kHz frequency range is generally dominated by unvoiced sounds with flat spectra in addition to the fact that high-frequency formants of voiced speech often have wide bandwidths, e.g., in nasals, and low energy compared to unvoiced speech, fewer poles can be used for the high band in comparison to the narrow band. Due to the dominance of unvoiced sounds in the high band, however, the accurate modelling of highband energy becomes particularly important, especially since the usual all-pole autoregressive (AR) LP model results in higher prediction errors for unvoiced sounds relative to voiced ones [10, Section 6.5.5]. As such, we represent highband content by 6-LSF feature vectors $\mathbf{\Omega}$ in $\mathcal{G}_{\mathbf{X}\mathbf{\Omega}}$, as well as separately modelling its energy in $\mathcal{G}_{\mathbf{X}G}$.[63] Thus, the total dimensionalities for $\mathcal{G}_{\mathbf{X}\mathbf{\Omega}}$ and $\mathcal{G}_{\mathbf{X}G}$ are $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{X}\\ \mathbf{\Omega}\end{bmatrix}\right) = 16$ and, $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{X}\\ G\end{bmatrix}\right) = 11$, respectively.

3.2.8 Windowing

We process wideband training data as well as narrowband test data in the time-domain using 20 ms frames with 50% overlap. For windowing, we employ the modified Hann window as defined in [98];

$$w[n] = \begin{cases} \dfrac{1}{2} - \dfrac{1}{2}\cos\left(\dfrac{\pi(2n+1)}{N}\right), & \text{for } 0 \le n \le N-1, \\ 0, & \text{elsewhere.} \end{cases} \tag{3.4}$$

This N-sample window is the sampled version of the continuous-time Hann window of length W where the N samples are uniformly spaced between the end points given by—assuming the continuous-time window is symmetric about zero, i.e., defined over the interval

[61] As described in Section 3.2.2, a set of m poles is fully represented by the m LSFs in the $(0, \pi)$ range.

[62] Our experiments on the effect of narrowband LP order—for a fixed highband LP order—showed that BWE performance nearly saturates above 8 poles.

[63] As in Footnote 62, our experiments on the effect of highband LP order show negligible performance improvements above 6 poles. Using 12 poles, for example, results in log-spectral distortion improvement of < 0.01 dB; see Section 3.4 for details on performance evaluation.

$[-\frac{W}{2}, \frac{W}{2}]$—$t_1 = -t_0 = \frac{W}{2} - \frac{\Delta t}{2}$ with $\Delta t = \frac{W}{N}$. As shown by Kabal in [98], this modified sampling pattern gives the smallest value for the sampling interval Δt for particular values of W and N while still covering the continuous-time window symmetrically. Small values of the sampling interval generally reduce aliasing due to sampling of the continuous-time window.

The study presented in [98] provides a further important motivation for choosing a Hann window for time-domain windowing, particularly for the dual-model BWE system using LSF parameterization; the smoothness of the LSF tracks resulting directly from the fact that the Hann window is, in fact, a raised-cosine window with no *rectangular pedestal*, i.e., where the cosine is raised (and weighted) such that its range extends from zero to the peak with no discontinuities at the edges. In contrast, the cosine in the more common Hamming window is raised such that it effectively *sits* on a rectangular pedestal (with a relative height of 0.08), thereby resulting in discontinuities at the edges. These discontinuities can cause substantial changes in the estimated LP parameters even when the window moves ahead by a single sample. The result is that LSF tracks will often exhibit spurious variations, potentially leading to undesirable LSF outliers when such tracks are sub-sampled at the actual frame rate. The simple expedient of using a window with no pedestal removes the spurious variations in LSF tracks and ensures smooth LSF evolution.

3.2.9 Formant bandwidth expansion

A well-known problem with LP-based spectral envelope modelling is that LP envelopes often exhibit unnaturally sharp peaks [99]. For high-pitched voiced speech in particular, LP envelope estimation often fails to separate the vocal tract's transfer function effect (the envelope) from the glottal excitation source (the pitch). The result is, due to bias towards pitch harmonics, LP spectra overestimate and overemphasize spectral powers at formants, providing a sharper contour compared to the original vocal tract response. Contrary to good design methodology, increasing the LP model order does not necessarily lead to better results and often exacerbates the problem. Instead, formant bandwidth expansion is employed whereby the bandwidths of peaks in LP spectra are broadened. Such expansion can be implemented through one or more of the following approaches:

- before LP analysis using *time-domain windowing* and/or *lag windowing* of the autocorrelation sequences [100, 101];
- after LP analysis through *scaling the radii* of the poles of the AR model [102];

- during LP analysis itself through *regularization smoothing* where a penalty measure representing the *peakiness* of the spectral envelope is included into the estimation of the AR model parameters [103]. Such regularization introduces a trade-off between the fit to data (i.e., the conventional minimum prediction error variance) and the smoothness of the envelope.

Since time-domain windowing of the input signal prior to estimating the correlation values corresponds in the frequency domain to convolution of the window's frequency response with that of the input signal, such time-domain windowing of the training and testing data constitutes, in itself, a form of implicit formant bandwidth expansion since the window response has a non-zero main lobe width [104, 105]. For the modified Hann window of Eq. (3.4), the 6 dB main lobe bandwidth—the double-sided bandwidth measured at the half-amplitude point—is $\frac{4\pi}{N}$ (where N is the window length in samples) [98, Table 1]. Thus, for 20 ms windows at $F_s = 16$ kHz (after the 8 to 16 kHz sample rate conversion applied during preprocessing as described in Section 3.2.1), $N = 320$ and the 6 dB main lobe bandwidth is 100 Hz (resulting in expanding peak bandwidths by 100 Hz at the half-amplitude point).

For explicit formant bandwidth expansion, we apply lag windowing using a Gaussian-shaped window as well as through radial scaling, both as developed in [104] and previously implemented in the dual-mode BWE system of [55]. Since the autocorrelation sequence has as its Fourier transform the power spectrum, lag windowing of the correlation corresponds to a periodic convolution of the frequency response of the window with the power spectrum of the signal. For the continuous-time Gaussian window

$$w(t) = \exp\left(-\frac{1}{2}[at]^2\right), \tag{3.5}$$

the frequency response also has a Gaussian shape;

$$W(\Omega) = \frac{\sqrt{2\pi}}{a}\exp\left(-\frac{1}{2}\left[\frac{\Omega}{a}\right]^2\right), \tag{3.6}$$

i.e., having a single lobe, with a two-sided 1-σ bandwidth—the bandwidth measured between the 1 standard deviation points—of $\omega_\sigma = 2a$ radians, and a two-sided 3 dB bandwidth of $\omega_{3\text{dB}} = \sqrt{8\log(2)}a$ radians. The discrete-time window is

$$w[k] = \exp\left(-\frac{1}{2}\left[\frac{ak}{F_s}\right]^2\right), \tag{3.7}$$

where F_s is the sampling rate. The parameter a can be expressed in terms of F_σ or F_{3dB} as

$$a = \pi \frac{F_\sigma}{F_s} = \frac{\pi}{\sqrt{2\log(2)}} \frac{F_{3dB}}{F_s}. \tag{3.8}$$

In our implementation, we use $F_\sigma = 120\,\text{Hz}$, resulting in a 3 dB bandwidth expansion of $F_{3dB} \cong 141\,\text{Hz}$ (also the double-sided expansion value at the 6 dB half-amplitude point).

Finally, we apply formant bandwidth expansion after LP analysis using radial scaling, where LPCs are windowed using an exponential sequence. Radial scaling involves moving the poles of the AR model inwards in the z-domain through replacing z by z/α [102]. Choosing $\alpha < 1$ has the effect of expanding resonance bandwidths. For a causal filter $H(z)$, the effect of replacing z with z/α is such that the impulse response of the filter is modified to become

$$h'[n] = \alpha^n h[n], \tag{3.9}$$

i.e., the impulse response coefficients are multiplied by an exponential (infinite length) time window. In the frequency domain, the frequency response of the filter is convolved with the frequency response of the window. As shown in [104, 105], the expanded 3 dB bandwidth obtained through this frequency-domain convolution can be well approximated by the first two terms of the corresponding Taylor series such that, for a given 3 dB bandwidth, α can be estimated by

$$\alpha = 2 - \sqrt{1 + 2\pi F_{3dB}/F_s}. \tag{3.10}$$

For the AR LP model, since $H(z) = 1/A(z)$, then $H(z/\alpha) = 1/A(z/\alpha)$. In other words, the radial scaling of the all-pole $H(z)$ can be implemented by multiplying the LPCs by the exponential time window. In our implementation of radial scaling in the dual-model BWE system, we use $\alpha = 0.994$ corresponding to $F_{3dB} \cong 31\,\text{Hz}$.

3.2.10 Training and testing data

We use the popular TIMIT speech corpus [106] to supply the wideband speech used for system training as well as for testing throughout our work. Training and testing are both performed in a speaker-independent manner. TIMIT contains phonetically diverse speech sampled at $F_s = 16\,\text{kHz}$ from a total of 630 male and female speakers from 8 major dialect regions of the United States. As described in the database distribution, the texts and

speakers in TIMIT have been subdivided into suggested training and test sets such that:

1. No speaker appears in both the training and testing portions.
2. All dialect regions are represented in both subsets, with at least 1 male and 1 female speaker from each dialect.
3. Text material in the two subsets do not overlap, i.e., no sentence text appears in both the training and test material.
4. All phonemes occur at least once in the test material.

For BWE system training, we use all 3696 waveforms from the TIMIT training set as suggested by the distribution and determined per the criteria above. Extracting 20 ms frames with 50% overlap from all 3696 training files results in $\approx 1.125 \times 10^6$ training frames.

For testing, we use the core test set also suggested by the distribution. The set consists of 24 speakers, 2 male and 1 female from each dialect region, with 8 distinct sentences per speaker for a total of 192 unique test sentences, corresponding to $\approx 58 \times 10^3$ test frames. Wideband test data is filtered using the G.712 filter of Figure 3.2(b) to simulate the bandwidth limiting effects of the telephone channel. The narrowband versions of the test data obtained as such are used as inputs to our dual-model BWE system implementation, depicted in Figure 3.1, while the original wideband material is used as the reference for performance evaluation as described in Section 3.4.

3.3 Gaussian Mixture Modelling

3.3.1 Joint density MMSE estimation

The minimum mean square error estimation of a target random variable (or vector) given another dependent source random variable (or vector) with their joint *pdf* modelled by a GMM is a special case of MMSE estimation based on an arbitrary multi-modal joint density whose set of parameters represent a third random variable (or vector) evaluated independently in a training stage.

For simplicity, let $X{:}\,\Omega \to \mathbb{R}^1$ and $Y{:}\,\Omega \to \mathbb{R}^1$ be the continuous source and target random variables. The function $\hat{y} = f(x)$ that minimizes the mean square error $\varepsilon = E\left[\|y - f(x)\|^2\right]$, is the Expectation $E[Y|X = x] = \int_{\Omega_y} y\, p(y|x)\, dy$. We introduce a third random variable Λ taking the discrete values $\{\lambda_i\}_{i\in\{1,\ldots,M\}}$, which in turn represent the parameters of the M-modal joint *pdf* $p(y, x)$, i.e.,

$$p(y,x) = \sum_{i=1}^{M} p(y,x,\lambda_i). \tag{3.11}$$

Given a source realization $X = x$ and the known parameter set Λ evaluated in the training stage, the objective is now to represent $E[Y|x]$ as a function of both x and Λ, i.e., $E[Y|x] = f(x,\Lambda)$. Thus, we rewrite $E[Y|x]$ as

$$\begin{aligned}
E[Y|x] &= \int_{\Omega_y} y\, p(y|x)\, dy \\
&= \int_{\Omega_y} y\, \frac{p(y,x)}{P(x)}\, dy \\
&= \int_{\Omega_y} y\, \frac{\sum_{i=1}^{M} p(y,x,\lambda_i)}{\int_{\Omega_y} \sum_{j=1}^{M} p(y,x,\lambda_j)\, dy}\, dy \\
&= \int_{\Omega_y} y\, \frac{\sum_{i=1}^{M} p(y|x,\lambda_i)\, P(x|\lambda_i)\, P(\lambda_i)}{\sum_{j=1}^{M} \int_{\Omega_y} p(y,x|\lambda_j)\, P(\lambda_j)\, dy}\, dy \\
&= \sum_{i=1}^{M} \frac{P(\lambda_i)\, P(x|\lambda_i) \int_{\Omega_y} y\, p(y|x,\lambda_i)\, dy}{\sum_{j=1}^{M} P(\lambda_j)\, P(x|\lambda_j)} \\
&= \sum_{i=1}^{M} P(\lambda_i|x)\, E[Y|x,\lambda_i],
\end{aligned} \tag{3.12}$$

where, by Bayes' rule,

$$P(\lambda_i|x) = \frac{P(\lambda_i)\, P(x|\lambda_i)}{\sum_{j=1}^{M} P(\lambda_j)\, P(x|\lambda_j)}. \tag{3.13}$$

In other words, the expectation $E[Y|x,\lambda_i]$ is weighted by the posterior probability of the mode $\Lambda = \lambda_i$ given $X = x$.

The application of the MMSE estimation rule to GMMs, e.g., the $\mathcal{G}_{\mathbf{X}\Omega_y}$ and $\mathcal{G}_{\mathbf{X}G}$ GMMs described above for the dual-model BWE system, easily follows by substituting the terms

in the last equality of Eq. (3.12) by their GMM counterparts. Let $\mathcal{G}_{\mathbf{XY}}$ represent a GMM jointly modelling feature vectors $\mathbf{X}$ and $\mathbf{Y}$, then, we have from Eq. (2.13) (rewriting joint vectors as supervectors for notational purposes)

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{G}_{\mathbf{z}} := \mathcal{G}(\mathbf{z}; M^{\mathbf{z}}, \mathrm{A}^{\mathbf{z}}, \Lambda^{\mathbf{z}}) = \sum_{i=1}^{M^{\mathbf{z}}} \alpha_i^{\mathbf{z}} \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i^{\mathbf{z}}, \mathbf{C}_i^{\mathbf{z}}), \tag{3.14}$$

with

$$\alpha_i^{\mathbf{z}} = \alpha_i^{\mathbf{x}} = \alpha_i^{\mathbf{y}}, \quad \boldsymbol{\mu}_i^{\mathbf{z}} = \begin{bmatrix} \boldsymbol{\mu}_i^{\mathbf{x}} \\ \boldsymbol{\mu}_i^{\mathbf{y}} \end{bmatrix}, \quad \text{and} \quad \mathbf{C}_i^{\mathbf{z}} = \begin{bmatrix} \mathbf{C}_i^{\mathbf{xx}} & \mathbf{C}_i^{\mathbf{xy}} \\ \mathbf{C}_i^{\mathbf{yx}} & \mathbf{C}_i^{\mathbf{yy}} \end{bmatrix}. \tag{3.15}$$

Then, by the properties of multivariate normal distributions[64]

$$P(\lambda_i|\mathbf{x}) = \frac{\alpha_i^{\mathbf{x}} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i^{\mathbf{x}}, \mathbf{C}_i^{\mathbf{xx}})}{\sum_{j=1}^{M} \alpha_j^{\mathbf{x}} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_j^{\mathbf{x}}, \mathbf{C}_j^{\mathbf{xx}})}, \tag{3.16}$$

and

$$E[\mathbf{Y}|\mathbf{x}, \lambda_i] = \boldsymbol{\mu}_i^{\mathbf{y}} + \mathbf{C}_i^{\mathbf{yx}} \mathbf{C}_i^{\mathbf{xx}^{-1}} [\mathbf{x} - \boldsymbol{\mu}_i^{\mathbf{x}}]. \tag{3.17}$$

3.3.2 Wideband versus highband spectral envelope modelling

As mentioned in Section 2.3.3.4, the target feature vectors $\mathbf{Y}$ can represent the spectra of either the high band, as in our dual-mode BWE system (where, for the GMMs $\mathcal{G}_{\mathbf{x\Omega}}$ and $\mathcal{G}_{\mathbf{x}G}$, $\mathbf{Y} = \boldsymbol{\Omega}$ and $\mathbf{Y} = G$ represent highband envelope shape and gain, respectively), or the full wide band, as in the GMM-based scheme of [82]. Modelling the full wide band as the target space provides the advantage that MMSE-estimated extensions contain lowband content ($< 300\,\mathrm{Hz}$) in addition to that of the high band. Thus, the need for further processing in order to estimate lowband content is eliminated, in contrast to the first approach where the target space is exclusively that of the high band. However, as described in Section 3.2.3, knowledge of the general attenuation characteristics of the G.712 channel allows reconstruction of the lowband frequencies more accurately than can be obtained by GMM-based estimation. Moreover, by focusing only on the narrower highband frequency range as the target space, the superior ability of the GMM to learn complex cross-

[64] If the random vector $\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}$ has a multivariate normal distribution, then the marginal $p(\mathbf{x})$ and conditional $p(\mathbf{y}|\mathbf{x})$ distributions are also normal. See [71, Section A.5.2] for a proof in the simpler bivariate case, and [107, Section 1.2.1] for the proof in the more general multivariate case.

correlations—as illustrated in the example of Section 2.3.3.5—is fully dedicated to those correlations between the non-overlapping frequency ranges of the narrow band and that in which we are primarily concerned, i.e., the high band, rather than the full wide band. A further motivation is that of reconstructed highband signal quality for similar model complexity; assuming fixed dimensionalities for $\mathbf{Y}$, devoting the target parameters fully to modelling highband spectral envelopes results in better spectral fidelity in the high band, as opposed to spreading out envelope modelling capability across the wide band only to discard the narrowband portions later through bandstop filtering.

3.3.3 Diagonal versus full covariances

The question of whether to use diagonal or full covariance GMM matrices in spectral transformation techniques, in general, depends on compromising between two factors: (a) computational complexity in both training and transformation stages, and (b) the ability of the model to provide a better fit for the underlying distribution. For GMM-based BWE, however, the computational cost associated with offline ML training is of increasingly secondary concern (as described in Section 2.3.3.4). As such, we focus only on the complexity associated with the extension stage as performed through MMSE estimation.

By reviewing Eqs. (3.16) and (3.17), it can be seen that MMSE estimation using diagonal-covariance GMMs should, indeed, be much simpler than that using similarly-sized (i.e., with the same number of Gaussian components) full-covariance GMMs since: (a) the cross-covariance terms $\{\mathbf{C}_i^{\mathbf{yx}}\}_{i\epsilon\{1,\ldots,M\}}$ in Eq. (3.17) are zero for diagonal covariances, and hence, the second term can be discarded altogether (thereby reducing computations), and (b) full matrix inversion is not required for the estimation of the probabilities $\{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_i^{\mathbf{x}},\mathbf{C}_i^{\mathbf{xx}})\}_{i\epsilon\{1,\ldots,M\}}$ in Eq. (3.16). Moreover, a GMM with diagonal covariances involves significantly fewer parameters compared to a similarly-sized full-covariance GMM. However, while diagonal-covariance GMMs are clearly less costly computationally compared to full-covariance ones when the number of Gaussian components is comparable in both types of the GMM, they are essentially an approximation the extent of which depends on the statistical dependence between the two feature vector spaces being jointly modelled. Using diagonal covariances thus, generally, requires an increase in the number of components in the Gaussian mixture in order to achieve the same performance obtained with full covariances. Nevertheless, it has typically been assumed that the additional com-

putational cost incurred by such an increase is quite low compared to the cost savings associated with using diagonal covariances. Indeed, it has been argued in [40] that "*because the component Gaussians are acting together to model the overall pdf, full covariance matrices are not necessary even if the features are not statistically independent. The linear combination of diagonal-covariance Gaussians is capable of modelling the correlations between feature vector elements. The effect of using a set of M full-covariance Gaussians can be equally obtained by using a larger set of diagonal-covariance Gaussians*". While the diagonal covariance cost-saving assumption underlying this statement is true when the computational complexities of ML training with full covariances are taken into account, it requires re-evaluation if such offline training costs become secondary to spectral transformation performance as in the case of BWE.

For LSF parameterization with practical dimensionalities, we show in Section 3.5.1 that using a GMM with a larger set of diagonal-covariance Gaussian components does *not*, in fact, lead to the same effect as that of a GMM with fewer full-covariance Gaussians unless the number of Gaussians is increased to the extent that diagonal covariances no longer correspond to lower computational costs. In particular, we compare BWE performance of full-covariance GMM tuples, $\{\boldsymbol{\mathcal{G}}^{\text{full}}\}$, with varying number of Gaussian components, M^{full}, to that of diagonal-covariance GMM tuples, $\{\boldsymbol{\mathcal{G}}^{\text{diag}}\}$, with M^{diag} Gaussians, in two scenarios where memory and computational cost during the extension stage are taken into account:

- In the first scenario, we compare BWE performance with M^{diag} set to a sufficiently large value; $M^{\text{diag}} > M^{\text{full}}$, calculated such that the total number of GMM parameters is the same for a particular $\boldsymbol{\mathcal{G}}^{\text{diag}}$-$\boldsymbol{\mathcal{G}}^{\text{full}}$ pair. We find that BWE performance of $\boldsymbol{\mathcal{G}}^{\text{diag}}$ is still inferior to that of the corresponding $\boldsymbol{\mathcal{G}}^{\text{full}}$ with $M^{\text{full}} < M^{\text{diag}}$.
- In the second scenario, we compare the performance of $\boldsymbol{\mathcal{G}}^{\text{full}}$-$\boldsymbol{\mathcal{G}}^{\text{diag}}$ pairs where the values of M^{diag} and M^{full} are calculated such that the total number of operations, or FLOPs (floating-point operations), needed to perform highband MMSE estimation per Eq. (3.12), is identical for both covariance types. Again, we find that BWE performance of $\boldsymbol{\mathcal{G}}^{\text{diag}}$ is inferior to that of the corresponding $\boldsymbol{\mathcal{G}}^{\text{full}}$.

In other words, even when the number of Gaussians in the diagonal-covariance GMM is increased such that both memory and computational cost are identical to those of the full-covariance GMM being compared to, performance remains inferior. In order to achieve similar performance, M^{diag} has to be increased by more than an order of magnitude compared to M^{full}, resulting in an overall increase in the number of GMM parameters to be

estimated during training as well as in the number of operations required during extension, compared to a full-covariance GMM. Thus, we conclude that diagonal-covariance GMMs are, in fact, more computationally expensive compared to full-covariance GMMs if equivalent BWE performance is desired.

To better understand these findings, we examine the MMSE estimation of Eq. (3.12) more closely. While the source-target feature vector correlations (or cross-band correlations in the case of BWE) are indirectly captured by the various GMM parameters—A and Λ, i.e., the sets of Gaussian component priors and their means and covariances[65]—during training on joint vectors (as suggested in [40]), the inter-band cross-covariance terms, $\{\mathbf{C}_i^{\mathbf{yx}}\}_{i\in\{1,\dots,M\}}$, directly reflect these correlations. As the second term in Eq. (3.17) shows, the influence of the difference terms, $\{\mathbf{x}-\boldsymbol{\mu}_i^{\mathbf{x}}\}_{i\in\{1,\dots,M\}}$, on the MMSE estimate, $\hat{\mathbf{y}}$, is greater for higher inter-band to intra-band cross-covariance *ratios*, $\{\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}\}_{i\in\{1,\dots,M\}}$.[66] By eliminating such cross-covariances, diagonal covariances effectively result in discarding an important parameter of the cross-band correlations underlying BWE. We confirm this observation in Section 3.5.1 by evaluating the average matrix Frobenius and p-, or, L_p-norms (for $p=1,2,\infty$) of the multiplicative $\{\mathbf{C}_i^{\boldsymbol{\omega}\mathbf{x}}\mathbf{C}_i^{\mathbf{xx}^{-1}}\}_{i\in\{1,\dots,M\}}$ factors for full-covariance $\mathcal{G}_{\mathbf{x\Omega}}$ GMMs with increasing number of components, M.[67] We find that these norms—representing the weight of the multiplicative term otherwise discarded by diagonal covariances—are almost consistently increasing for higher M. In other words, model accuracy and, consequently, BWE performance, directly correlate with higher ratios of inter-band to intra-band cross-covariances. In fact, as discussed in Section 5.4.2.1 in the context of high-dimensional GMM-based modelling, these multiplicative $\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}$ factors—representing the weights on the contributions of the source data to the MMSE estimates of the target—will result in oversmoothed target data, and hence, an unclear low-quality highband speech signal, when their norms are too low. In essence, these *ratios* partially represent a joint-band GMM's ability to model information mutual to the disjoint frequency bands rather than band-specific information.

[65] See Eq. (2.13).

[66] While the quantity $\mathbf{C}^{\mathbf{yx}}\mathbf{C}^{\mathbf{xx}^{-1}}$ is, strictly speaking, not a ratio, but rather the product of the matrix $\mathbf{C}^{\mathbf{yx}}$ and the inverse matrix $\mathbf{C}^{\mathbf{xx}^{-1}}$, conceptually this product is equivalent to a ratio of $\mathbf{C}^{\mathbf{yx}}$ to $\mathbf{C}^{\mathbf{xx}}$.

[67] Matrix norms represent measures of distance or weight in the space of matrices [108]. For a matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$, the Frobenius norm is given by $\|\mathbf{A}\|_{\mathrm{F}}=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$. The L_p-norms are given by $\|\mathbf{A}\|_1=\max_{1\le j\le n}\sum_{i=1}^{m}|a_{ij}|$, $\|\mathbf{A}\|_\infty=\max_{1\le i\le m}\sum_{j=1}^{n}|a_{ij}|$, and $\|\mathbf{A}\|_2=\sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}$, where $\lambda_{\max}(\mathbf{A}^T\mathbf{A})$ is the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$.

3.3.4 On the number of Gaussian components

As described in Section 2.3.3.4, a GMM—as a multi-modal density—is intuitively a good means for the modelling of a multi-class distribution. Given sufficient training data and a training procedure optimized to maximize their likelihood given the model parameters, it can be reasonably assumed that individual Gaussian component densities of the GMM will correspond to the individual classes underlying the distribution being modelled. With that perspective, the choice for the number M of Gaussians in a GMM ultimately depends on two factors: (a) the nature of the distribution being modelled as determined by the choice of its feature vector representation, and (b) the amount of data available for training.

The first factor relates to the *true* number of underlying classes and the complexity of their distributions. An overall distribution of C well-decorrelated normally-distributed classes, for example, requires roughly the same number of Gaussians if reliable modelling is to be achieved, i.e., $M \cong C$. On the other hand, a complex distribution comprising highly-overlapping non-normally-distributed classes will require a larger number of Gaussians, i.e., $M > C$. The nature of the feature vector space also determines the number of underlying classes. The acoustic space corresponding to a scalar parameter representing the degree of voicing, for example, rather consists of only a few underlying classes—a voiced class, an unvoiced one, and one or few more classes representing sounds with mixed voicing. In contrast, the acoustic space to be modelled by GMMs, in the case of BWE, is much more complex. It spans a large number of overlapping and non-linearly related acoustic properties. The realizations of these acoustic properties—parameterized into spectral envelope feature vectors—exhibit several levels of prominent trends of joint behaviour, comprising the acoustic classes to be modelled by the multi-modal GMM. In other words, there is no single *true* number of underlying acoustic classes. Rather, there are several levels of acoustic resolution where the number of underlying classes increases with higher resolution, thereby guiding the choice for the number of Gaussians:

- At the lowest level of resolution, a memoryless spectral envelope space consists of roughly 8 classes corresponding to manners of articulation as listed in Table 1.1.
- As spectral and modelling resolutions increase, finer classification tends towards densities corresponding to places of articulation.
- A resolution of approximately 40 classes corresponds roughly to the spectral characteristics of the phonemes of Table 1.1 for English.

- With increasing resolution, more classes will correspond to finer phonemic spectral detail, e.g., separate classes for the onset, steady-state, and trailing portions of phoneme spectra. With yet further resolution, the acoustic space consists of finer underlying classes representing spectral characteristics of the > 100 allophonic variations of phonemes (resulting from coarticulation; see Section A.2).
- Whereas the number of classes in a memoryless acoustic space saturates for C in the order of 100–200 (corresponding to the total number of allophonic variations), a dynamic space where spectral envelopes can be further classified along a temporal axis potentially introduces several more levels of finer resolution depending on the extent and representation of temporal information.
- As shown in Chapter 5, extending spectral envelope representations temporally considerably increases the number of underlying classes, requiring a proportional increase in the number of Gaussian components. Results show that the number of classes saturates at values roughly corresponding to 100–200 ms of temporal information.

We note that the number of classes and their covariances in this increasingly-fine categorization further increase as a result of inter-speaker as well as intra-speaker variability.

While increasing the number of Gaussian components in a GMM generally translates into a better fit to an underlying distribution with finer spectral—and optionally, temporal—resolution, and hence, improved BWE performance, such increases are constrained by the amount of training data (frames) available. For a full-covariance GMM with M components modelling a D-dimensional distribution, the relation between M and the amount N_f of available training data points can be written as

$$M^{\mathrm{full}} = \left\lfloor \frac{N_\mathrm{f}}{N_{\mathrm{f/p}}\left(1 + D + \frac{D(D+1)}{2}\right)} \right\rfloor, \tag{3.18}$$

where $N_{\mathrm{f/p}}$ represents the number of data points (frames) required per GMM parameter such that the parameter is reliably estimated. For a diagonal GMM, the relation is

$$M^{\mathrm{diag}} = \left\lfloor \frac{N_\mathrm{f}}{N_{\mathrm{f/p}}(1 + 2D)} \right\rfloor. \tag{3.19}$$

In [109], it was empirically found that $N_{\mathrm{f/p}} \approxeq 100$. However, through our GMM experiments described in Section 3.5.2, we show that $N_{\mathrm{f/p}}$ can be as low as 10 with negligible loss in

BWE performance ($\sim 0.1\%$ relative degradation in log-spectral distortion, for example).

3.4 Performance Evaluation

As discussed in Section 1.1.3.3, the intelligibility of telephony speech, although adversely affected by the PSTN bandwidth limitations, is nevertheless still reasonable for all but the lowest-bit-rate coders. Moreover, since intelligibility only assesses the recognizability of speech while quality encompasses many more perceptual properties of sounds, quality has been the criterion typically used for the evaluation of BWE performance. Generally, the perceived quality, or naturalness, of speech, is comprised of several factors that are difficult to quantify, e.g., loudness, clarity, fullness, spaciousness, brightness, softness, nearness, and fidelity [1]. As such, the optimal means by which to evaluate BWE performance—in reference to the perceived quality of extended speech—is that of subjective listening tests. Formal listening tests are known, however, for being time-consuming, labour-intensive, and potentially expensive. They also suffer from inherent variability caused by any changes in testing conditions or listener pool. This variability is typically addressed by diversifying and increasing test data and listener pool as much as possible, thus further adding to the difficulty associated with formal subjective testing. Informal listening tests, on the other hand, are ill-equipped for finely quantifying distortions over multiple simulations or for evaluating the quality of isolated speech aspects such as envelopes or excitation.

In contrast, objective quality measures attempt to analytically measure physical characteristics of the speech signal that closely correlate with quality. They thus provide a considerable advantage over subjective evaluations by being cheaper and easier to implement. Such objective measures are, however, clearly suboptimal to subjective ones since:

- Different objective measures tend to focus on different types of distortions. Thus, in contrast to subjective measures which inherently encompass several perceptual aspects of speech, no single objective measure can completely replace subjective evaluations.
- Objective measures vary considerably in terms of making use of the knowledge about the human auditory system and speech perception. Among the well known properties of speech perception, for example, is that sensitivity to smaller differences in time, amplitude and frequency of sounds generally increases as the frequency of the sound decreases; i.e., *difference limens*, or *just-noticeable differences*, increase with frequency

[10, Section 4.3.3]. Another important property quantified by the so-called hearing threshold is that sounds outside the 1–5 kHz frequency range require significantly more energy to be heard than those inside this range [10, Section 4.3.2]. Objective measures that weight distortions in a manner that takes account of such perceptual properties provide a better measure of quality than those treating distortions equally over the wideband frequency range.

In the context of our work on the evaluation of the effects of speech memory—in reference to speech dynamics—on BWE performance, we study incorporating such memory with various durations and by different means. Considering the numerous combinations by which such memory inclusion is investigated in Chapter 5, the difficulties of performing subjective listening tests become quite apparent. Instead, we evaluate BWE performance in this work by quantifying distortions in spectral envelopes using a few objective measures chosen such that the two following requirements are satisfied by the ensemble of measures as a whole:

Popularity Log-spectral distance/distortion (LSD), or some variant thereof, is the most common objective measure used in the BWE literature for performance evaluation, e.g., [39, 41, 55, 56, 62, 63, 67, 68, 82]. For the benefit of allowing performance comparisons, we conform with these works in using LSD.

Perceptual relevance Despite their relatively higher correlation with subjective measures compared to LSD, the Itakura-Saito distortion and related measures have seldom been used for BWE performance evaluation. In our work, we use two variants of such measures. Furthermore, we use the superior PESQ—perceptual evaluation of speech quality—measure, which has been specifically designed using a psychoacoustic model to map objective scores to subjective MOS—mean opinion scores. In contrast to the LSD and Itakura-based measures where distortions are evaluated between smoothed LP-derived versions of spectral envelopes, no such LP smoothing is performed by the PESQ model.

These distance measures are described in detail below. Finally, we note that, by using distance relative to the optimal original wideband signal as the measure of BWE performance, objective measures do not reflect the aforementioned observation whereby an extended signal may sound natural despite having mismatches relative to the true wideband signal.

3.4.1 Log-spectral distortion

Since its formulation in the mid 1970s, log-spectral distortion [110] has been the de facto measure for the evaluation of LP-based speech coders and quantization techniques. LSD is a measure of the distance between smooth test LP-based or quantized spectra and their original reference counterparts. Since the objective of BWE is the reconstruction of spectra foremost in the missing highband frequency range, LSD is a natural and popular choice for BWE performance evaluation. For a particular frame, LSD, expressed in decibels, is generally given by

$$d^2_{\text{LSD}} = \int_{-\pi}^{\pi} \left(20 \log_{10} \frac{\sigma}{|A(e^{j\omega})|} - 20 \log_{10} \frac{\hat{\sigma}}{|\hat{A}(e^{j\omega})|} \right)^2 \frac{d\omega}{2\pi}, \tag{3.20}$$

where ω is the normalized frequency, σ and $A(e^{j\omega})$ are the LP gain and inverse filter of the reference signal frame's auto-regressive (AR) model, respectively, while $\hat{\sigma}$ and $\hat{A}(e^{j\omega})$ are those of the test signal frame's. Since our focus is evaluating highband reconstruction only in the 4–8 kHz range without the effects of other system processing, e.g., lowband and midband equalization, we isolate this range by limiting the range of the integration in Eq. (3.20) to the 4–8 kHz band. Thus, for the dual-model BWE system, Eq. (3.20) can be rewritten using the true and MMSE estimates of the highband signal excitation gain, g,[68] and the spectral envelope inverse filter, $A_y(e^{j\omega})$, obtained through the GMMs $\mathcal{G}_{\mathbf{x}G}$ and $\mathcal{G}_{\mathbf{x}\boldsymbol{\Omega}}$ (as defined in Section 3.2.5), respectively, i.e.,

$$d^2_{\text{LSD}} = 2 \int_{\omega_l}^{\omega_h} \left(20 \log_{10} \frac{g}{|A_y(e^{j\omega})|} - 20 \log_{10} \frac{\hat{g}}{|\hat{A}_y(e^{j\omega})|} \right)^2 \frac{d\omega}{2\pi}, \tag{3.21}$$

where ω_l and ω_h correspond to 4 and 8 kHz, respectively.

Performance over a set of N test frames is evaluated either as the mean-root-square (MRS) average of the set of $\{d^2_{\text{LSD}_n}\}_{n\in\{1,\ldots,N\}}$; i.e.,

[68] As described in Section 3.2.4, the EBP-MGN excitation signal $e(n)$ is a spectrally white signal whose variance depends on the energy in the equalized 3–4 kHz range, i.e., $e(n) \cong \beta u(n)$. Since β is the same for both true and reconstructed highband signals, the LP prediction gains, σ and $\hat{\sigma}$, of the true and reconstructed highband signals, respectively, are related to the true and estimated excitation signal gains, g and $\hat{g}$, respectively, by the same multiplicative constant, i.e., $\sigma \cong \beta g$ and $\hat{\sigma} \cong \beta \hat{g}$. Then, by the logarithm subtraction in Eq. (3.20), the common factor β can be omitted.

$$\overline{d}_{\mathrm{LSD(MRS)}} = \frac{1}{N}\sum_{n=1}^{N}\left[d^2_{\mathrm{LSD}_n}\right]^{\frac{1}{2}} \qquad [\mathrm{dB}], \tag{3.22}$$

or as the root-mean-square (RMS) average,

$$\overline{d}_{\mathrm{LSD(RMS)}} = \left[\frac{1}{N}\sum_{n=1}^{N} d^2_{\mathrm{LSD}_n}\right]^{\frac{1}{2}} \qquad [\mathrm{dB}]. \tag{3.23}$$

Generally, the MRS average is lower than the corresponding RMS one, and has typically been more popular [111]. As such, it is the average used primarily in our work, for BWE performance evaluation as well as for discrete highband entropy estimation—in Chapters 4 and 5—for the purpose of quantifying certainty about the high band given the narrow band. In Sections 4.3.5 and 4.4.3.2, we also use an RMS-LSD lower bound to demonstrate the effects of memory inclusion in improving potential BWE performance. Thus, we will also report relevant BWE $\overline{d}_{\mathrm{LSD(RMS)}}$ results when needed in the context of determining a BWE system's optimality. In the sequel, unless otherwise indicated, we refer to the typical LP-based MRS average LSD simply as *average* LSD, denoting it by $\overline{d}_{\mathrm{LSD}}$. In contrast, reported RMS averages are explicitly denoted by $\overline{d}_{\mathrm{LSD(RMS)}}$.

Although LSD does not make use of any perceptually-related knowledge in measuring distances between spectra, it correlates reasonably well with subjective speech quality. A correlation of 0.63 with the diagnostic acceptability measure (DAM) [112], for example, was measured in [113]. In the early perceptual studies of Flanagan in [114] on difference limens, varying intensity alone resulted in a barely perceptible difference of about 1.5 dB for vowels and 0.4 dB for synthesized unvoiced sounds with entirely flat spectra. These intensity figures were related to similar LSD numbers in [110].

Through informal subjective testing on LPC quantization in [115], Paliwal and Atal later found the following three conditions to jointly represent the threshold for spectral *transparency* in the 0–3 kHz band (i.e., the threshold below which quantization errors are inaudible): (a) an average LSD of 1 dB, (b) no outlier frames with LSD greater than 4 dB, and (c) less than 2% of frames with LSD in the 2–4 dB range. As noted in [109], however, since level discrimination decreases for higher frequencies (i.e., higher difference limens), the average LSD threshold for spectral transparency for frequencies above 3 kHz is, in fact, higher than 1 dB. Nevertheless, the 1 dB average LSD threshold can still be applied to the highband frequency range but as a rather conservative estimate. Similarly, the LSD values

for the spectral transparency conditions on outlier frames are rather higher for frequencies above 3 kHz than those of [115].

3.4.2 Itakura-Saito distortion variants

While LSD is widely used for evaluating spectral envelope degradation due to its tractability and historic value, it does not take into account the differences in perceptual importance of some aspects of the AR LP speech spectrum representation. For example, errors in spectral peaks and valleys are weighted equally. While some attempts have been made to improve the perceptual relevance of LSD resulting in variants that show higher correlation with subjective measures, BWE performance evaluations have, for the most part, continued to use the conventional form of Eq. (3.20). An example of such perceptually-weighted LSD variants is the frequency-weighted LSD [113].

In contrast, the Itakura-Saito distortion measure [116] has some perceptual relevance. Arising from the formulation of LP as an approximate maximum likelihood estimation, the Itakura-Saito distortion is an asymmetric gain-sensitive measure that weights spectral density underestimation more heavily than overestimation (specifically, positive log spectral differences similar to the integrand of Eq. (3.20) are weighted more heavily than negative errors) [110]. Since underestimation in LP spectra typically occurs at spectral peaks corresponding to formants while overestimation occurs at spectral valleys, the Itakura-Saito distortion has been argued to be a subjectively meaningful distortion measure for the spectral shape of speech [117]. This follows from the fact that—as described in Section 1.1.3.1—the amplitudes and frequency locations of high-energy regions of spectra play a central role in the perception of sounds. Due to its sensitivity to LP gain, however, a gain-optimized variant—the Itakura distance or log-likelihood ratio distance—was derived in [75] by finding the AR model gains that minimize the Itakura-Saito distortion, rendering it gain-independent. This variant was shown in [113] to have a correlation of 0.73 with the subjective DAM (versus 0.63 for LSD), reaching up to a correlation of 0.89 with MOS for an enhanced version that exploits auditory masking properties [118].

Despite the higher correlation of the Itakura-Saito distortion variants with subjective measures compared to LSD, they have rarely been used in BWE. A notable use is the early work in [50] using codebook mapping where codebook search is performed using the gain-normalized Itakura-Saito distortion—also referred to as the likelihood ratio distortion. This

measure is similar to the log-likelihood ratio distortion but without the logarithm operation.

With the same definitions used in Eq. (3.21) where the original and reconstructed highband spectral envelopes (in the 4–8 kHz band) are represented by the AR LP models $\frac{g}{|A_y(e^{j\omega})|}$ and $\frac{\hat{g}}{|\hat{A}_y(e^{j\omega})|}$, respectively, the Itakura-Saito distortion can be written as (dropping the arguments in $A_y(e^{j\omega})$ and $\hat{A}_y(e^{j\omega})$ to simplify notation)

$$d_{\text{IS}}\left(\tfrac{g}{|A_y|}, \tfrac{\hat{g}}{|\hat{A}_y|}\right) = 2\int_{\omega_l}^{\omega_h}\left[\frac{g^2/|A_y|^2}{\hat{g}^2/|\hat{A}_y|^2} - \log\frac{g^2/|A_y|^2}{\hat{g}^2/|\hat{A}_y|^2} - 1\right]\frac{d\omega}{2\pi} \qquad [\text{dB}], \tag{3.24}$$

where the notation $d_{\text{IS}}(R,T)$ indicates R is the reference spectrum, and T is the test spectrum under evaluation.

The Itakura-Saito distortion, d_{IS}, does not fulfill the symmetry condition for distance metrics [110]; i.e., $d_{\text{IS}}(R,T) \neq d_{\text{IS}}(T,R)$. A symmetrized version of it, however, can be constructed by the arithmetic mean;

$$d^*_{\text{IS}}\left(\tfrac{g}{|A_y|}, \tfrac{\hat{g}}{|\hat{A}_y|}\right) = \frac{1}{2}\left[d_{\text{IS}}\left(\tfrac{g}{|Y|}, \tfrac{\hat{g}}{|\hat{Y}|}\right) + d_{\text{IS}}\left(\tfrac{\hat{g}}{|\hat{Y}|}, \tfrac{g}{|Y|}\right)\right] \qquad [\text{dB}], \tag{3.25}$$

which, by substitutions from Eq. (3.24), can be written as

$$d^*_{\text{IS}}\left(\tfrac{g}{|A_y|}, \tfrac{\hat{g}}{|\hat{A}_y|}\right) = 2\int_{\omega_l}^{\omega_h}\left\{\cosh\left[\frac{g^2/|A_y|^2}{\hat{g}^2/|\hat{A}_y|^2}\right] - 1\right\}\frac{d\omega}{2\pi}, \tag{3.26}$$

and is, hence, called the COSH measure [110]. The effect of symmetrizing d_{IS} is to weight large differences in log spectra (regardless of error sign, i.e., regardless of whether the error is due to under- or over-estimation) more heavily than the LSD measure [110]. Since larger deviations generally correspond to the regions of changing formant frequencies, d^*_{IS} can be viewed as a symmetric distance measure that emphasizes more perceptually-important errors in spectra.

Similarly, the gain-optimized Itakura distortion is also asymmetric; it is given by

$$\begin{aligned} d_{\text{I}}\left(\tfrac{g}{|A_y|}, \tfrac{\hat{g}}{|\hat{A}_y|}\right) &\triangleq \min_{\hat{g}>0} d_{\text{IS}}\left(\tfrac{g}{|A_y|}, \tfrac{\hat{g}}{|\hat{A}_y|}\right) \\ &= \log\left(\frac{\hat{\mathbf{a}}_y^T\mathbf{R}_y\hat{\mathbf{a}}_y}{g^2}\right), \end{aligned} \tag{3.27}$$

where $\hat{\mathbf{a}}_y^T$ is the reconstructed LP coefficient vector and $\mathbf{R}_y$ is the Töeplitz autocorrelation matrix of the original signal LP model. In the same manner described above for the asymmetric d_{IS}, we symmetrize d_{I} by the arithmetic mean; i.e.,

$$d_{\text{I}}^* = \frac{1}{2}\left[d_{\text{I}}\left(\frac{g}{|A_y|}, \frac{\hat{g}}{|\hat{A}_y|}\right) + d_{\text{I}}\left(\frac{\hat{g}}{|\hat{A}_y|}, \frac{g}{|A_y|}\right)\right] \qquad [\text{dB}]. \tag{3.28}$$

As described for LSD, we evaluate performance over a test set of N frames by the simple averages $\overline{d}_{\text{IS}}^* = \frac{1}{N}\sum_{n=1}^{N} d_{\text{IS}_n}^*$ and $\overline{d}_{\text{I}}^* = \frac{1}{N}\sum_{n=1}^{N} d_{\text{I}_n}^*$. By employing both $\overline{d}_{\text{IS}}^*$ and $\overline{d}_{\text{I}}^*$ to evaluate the effect of memory inclusion on BWE performance, not only do we obtain relatively higher subjectively-correlated measures of performance improvement, but more importantly, we also obtain an implicit breakdown of that improvement into separate gain-related and spectral shape-related improvements (by exploiting the gain-sensitivity of $\overline{d}_{\text{IS}}^*$ and the lack of it for $\overline{d}_{\text{I}}^*$).

3.4.3 Perceptual evaluation of speech quality

As described in the Section 3.4 preamble, it is our view that in order for an objective evaluation to provide a complete picture of BWE performance, such an evaluation should make use of an ensemble of measures to collectively ensure that results: (a) can be compared to those of other BWE techniques in the literature where objective measures are commonly employed, and (b) are perceptually relevant in the sense that such objective results correlate with subjective ones as much as possible. In our work, we satisfy the first requirement by employing the most popular objective evaluation measure, LSD. While Itakura-based distortion variants provide a more subjectively-correlated measure of BWE performance, their value lies rather in the finer detail they provide about the quality of reconstructed highband spectra in terms of their shapes and gains. To satisfy the perceptual-relevance requirement, we make use of the superior PESQ—perceptual evaluation of speech quality—ITU-T P.862.2 standard [119] for the objective assessment of the perceived quality of wideband telephony speech—a wideband extension to the earlier PESQ ITU-T P.862 standard [120] intended for narrowband telephony speech.

Starting in the early 1990s, researchers have been attempting to improve the objective assessment of perceived telephony speech quality. The motivation for this research arose from the increase in the number of transmission and coding technologies for digital telephony services, thereby introducing new types of spectral and temporal distortions

affecting the subjective quality of telephony speech (e.g., packet loss, variable delay, front-end clipping, etc.) for which classical quality measurement techniques—using concepts like signal to noise ratio, frequency response functions, etc.—have become grossly inaccurate [121]. To account for the perceptual effects of such distortions, perception-based approaches rather attempt to quantify these distortions in time and frequency with weighting derived from psychoacoustic models such that distortions are effectively translated into subjectively-correlated scores. During a training stage, the system's parameters for the detection and quantification of the distortions under consideration are optimized under various testing conditions such that final objective scores are maximally correlated with a particular subjective measure, typically MOS scores from an ACR (absolute category rating) evaluation.[69]

Several perception-based assessment techniques have been proposed and evaluated by the ITU-T in a series of benchmark tests, culminating in the adoption of the PESQ technique in ITU-T Recommendation P.862.[70] For a variety of 22 benchmark ITU experiments covering mobile, fixed and VoIP telephony network and codec conditions, PESQ achieves objective scores with an average correlation of 0.935 with subjective MOS scores [120, 121]. Consistency of the PESQ measure was further confirmed by achieving a similar correlation of 0.935 on a set of 8 independent experiments—unknown during the development of PESQ—used in ITU-T's final validation [120, 121]. Such a superior subjective correlation with MOS scores over a wide range of telephony distortions makes the PESQ measure quite attractive for our purposes of evaluating BWE performance, especially when considering the aforementioned difficulties associated with performing a large number of subjective listening tests for the numerous combinations by which we investigate speech memory inclusion. Compared to the LSD and Itakura-based measures where distortions can be easily represented mathematically on a per-frame basis independently of surrounding frames, PESQ, however, is rather a quite complex measure whose calculation involves many time- and frequency-domain processing steps over the length of a test speech signal. As such, we detail the construction and optimization of the PESQ algorithm separately in Appendix B.

In our context of BWE performance evaluation, the PESQ algorithm's reference signal is the original wideband test speech while the test signal is that extended through BWE. For

[69] See Footnote 16.

[70] The PESQ P.862 standard, in fact, replaced the earlier perceptual speech quality measure (PSQM) and measuring normalizing blocks (MNB) approaches standardized in Recommendations P.861 and P.861.1, respectively.

a test material of M speech files, we evaluate the perceived quality of the extended speech using the simple average of the per-file PESQ scores, i.e, $\overline{Q}_{\text{PESQ}} = \frac{1}{M}\sum_{m=1}^{M} Q_{\text{PESQ}_m}$, where the MOS-like Q_{PESQ} score typically ranges from 1.0 (bad) to 4.5 (no distortion) [120–122].

Finally, we note that, unlike LSD and the Itakura-based measures where we limit distortion calculation to the 4–8 kHz range (by limiting the integrations in Eqs. (3.21) and (3.24)), the PESQ algorithm compares and, in fact, requires the original and extended signals over the wideband 50–7000 Hz range.[71] As such, PESQ scores reported in the sequel not only assess highband GMM-based extension in the smaller 4–7 kHz range, but they also take into account the distortions associated with imperfect lowband (< 300 Hz) and midband (3400–4000 Hz) equalization-based extensions. However, since in all experiments:

(a) we compare speech with highband extensions obtained using some means of memory inclusion to speech with extensions obtained by the conventional static GMM-based approach, and

(b) the content below 4 kHz is identical for any particular test file regardless of the method used for highband extension (since the lowband and midband equalization-based extensions are independent of extension above 4 kHz),

any improvements obtained in $\overline{Q}_{\text{PESQ}}$ will directly correspond to improved highband extension above 4 kHz.

3.5 Memoryless BWE Baseline

In order to arrive at a well-performing memoryless baseline given the amount of training data described in Section 3.2.10 and our parameterization and dimensionality choices described in Section 3.2.7, we study the role of the remaining variables on BWE performance. Specifically, we investigate the effects of the number and covariance type of the Gaussian kernels in the model's $\mathcal{G}_{\mathbf{x\Omega}}$ and $\mathcal{G}_{\mathbf{x}G}$ GMMs, as well as the effect of the amount of data available for training.

3.5.1 Effect of number and covariance type of Gaussian components

To compare BWE performances using full- and diagonal-covariance GMMs while simultaneously investigating the effect of the number of Gaussian components, we train two

[71] Level alignment, for example, for both reference and test signals is performed based on the narrowband content in the 300–3000 Hz range [121].

separate sets of $(\mathcal{G}_{\mathbf{x}\boldsymbol{\Omega}}, \mathcal{G}_{\mathbf{x}G})$ GMM tuples. For the full- and diagonal-covariance tuples given by

$$\boldsymbol{\mathcal{G}}^{\text{full}} := \left(\mathcal{G}^{\text{full}}_{\mathbf{x}\boldsymbol{\Omega}} := \mathcal{G}(\mathbf{x}, \boldsymbol{\omega}; M^{\text{full}}, \cdot), \mathcal{G}^{\text{full}}_{\mathbf{x}G} := \mathcal{G}(\mathbf{x}, g; M^{\text{full}}, \cdot)\right), \tag{3.29}$$

and

$$\boldsymbol{\mathcal{G}}^{\text{diag}} := \left(\mathcal{G}^{\text{diag}}_{\mathbf{x}\boldsymbol{\Omega}} := \mathcal{G}(\mathbf{x}, \boldsymbol{\omega}; M^{\text{diag}}, \cdot), \mathcal{G}^{\text{diag}}_{\mathbf{x}G} := \mathcal{G}(\mathbf{x}, g; M^{\text{diag}}, \cdot)\right), \tag{3.30}$$

respectively, the two sets are $\{\boldsymbol{\mathcal{G}}^{\text{full}}_i\}$ and $\{\boldsymbol{\mathcal{G}}^{\text{diag}}_j\}$, where $i, j \in \{1, \ldots, 8\}$, $M^{\text{full}}_i = 2^i$, and $M^{\text{diag}}_j = 2^j$.

Figure 3.4 illustrates LSD performance for $\boldsymbol{\mathcal{G}}^{\text{full}}$ and $\boldsymbol{\mathcal{G}}^{\text{diag}}$ as a function of M. As expected, performance consistently improves with higher M values regardless of covariance type. Secondly, at a particular $M = M^{\text{diag}} = M^{\text{full}}$, i.e., $i = j$, $\boldsymbol{\mathcal{G}}^{\text{diag}}$ has fewer parameters compared to $\boldsymbol{\mathcal{G}}^{\text{full}}$, translating into fewer degrees of freedom for acoustic space modelling and, hence, expectedly poorer BWE performance.

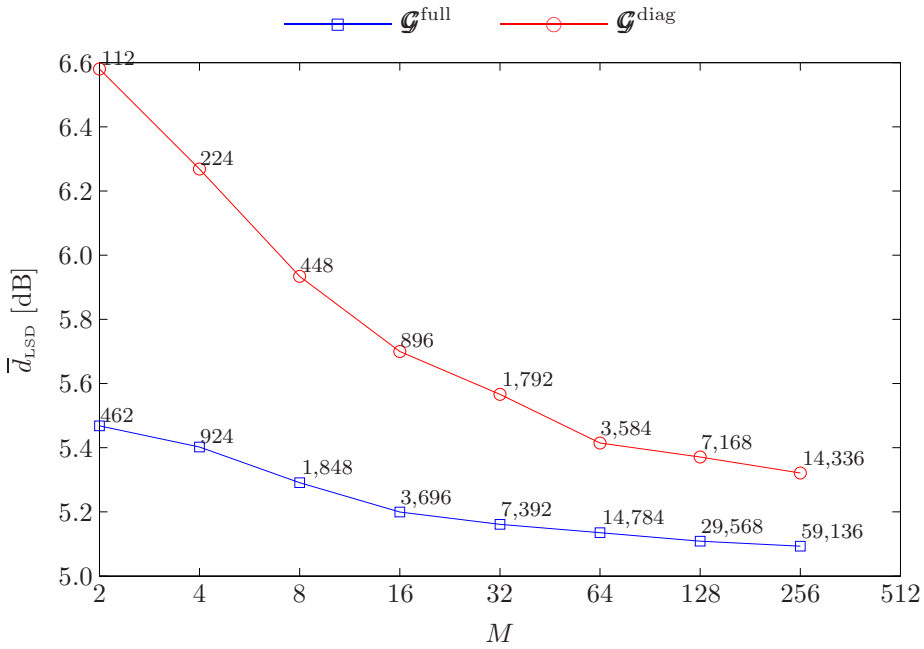


Fig. 3.4: BWE $\overline{d}_{\text{LSD}}$ performance as a function of the number of Gaussian components, M, for the GMM tuples $\boldsymbol{\mathcal{G}}^{\text{full}}$ and $\boldsymbol{\mathcal{G}}^{\text{diag}}$, defined in Eqs. (3.29) and (3.30), respectively. Data labels represent the numbers of GMM parameters, N_{p}.

While Figure 3.4 illustrates the performance gap between $\mathcal{G}^{\text{diag}}$ and $\mathcal{G}^{\text{full}}$ GMMs for a fixed number of Gaussians, it is rather the performance as a function of both:

1. the total degrees of freedom available for modelling, represented by the total number of GMM parameters, N_{p}, in a $\boldsymbol{\mathcal{G}}^{\mathrm{full}}$ or $\boldsymbol{\mathcal{G}}^{\mathrm{diag}}$ tuple; and
2. the computational complexity of performing extension through MMSE estimation per Eq. (3.12), represented by the total number of operations per input frame, $N_{\mathrm{FLOPs/f}}$;

that can determine the superiority of one type of covariances over the other.

Given that $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{x}\\ \boldsymbol{\Omega}\end{bmatrix}\right) = 16$ and $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{x}\\ G\end{bmatrix}\right) = 11$, the total number of GMM parameters available in the tuple $\left(\mathcal{G}^{\mathrm{full}}_{\mathbf{x}\boldsymbol{\Omega}} := \mathcal{G}(\mathbf{x}, \boldsymbol{\omega}; M^{\mathrm{full}}, \cdot), \mathcal{G}^{\mathrm{full}}_{\mathbf{x}G} := \mathcal{G}(\mathbf{x}, g; M^{\mathrm{full}}, \cdot)\right)$ for the modelling of the $\begin{bmatrix}\mathbf{x}\\ \boldsymbol{\Omega}\end{bmatrix}$ and $\begin{bmatrix}\mathbf{x}\\ G\end{bmatrix}$ spaces is given by

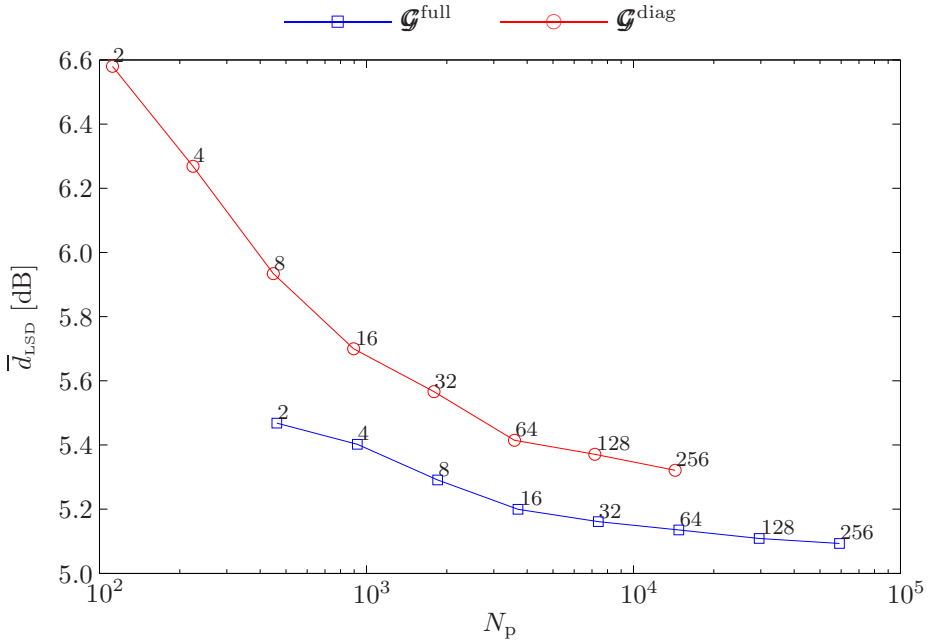
$$N^{\mathrm{full}}_{\mathrm{p}} = M^{\mathrm{full}}[(1 + 16 + 0.5(16 \cdot 17)) + (1 + 11 + 0.5(11 \cdot 12))] = M^{\mathrm{full}}[231], \tag{3.31}$$

while that for $\left(\mathcal{G}^{\mathrm{diag}}_{\mathbf{x}\boldsymbol{\Omega}} := \mathcal{G}(\mathbf{x}, \boldsymbol{\omega}; M^{\mathrm{diag}}, \cdot), \mathcal{G}^{\mathrm{diag}}_{\mathbf{x}G} := \mathcal{G}(\mathbf{x}, g; M^{\mathrm{diag}}, \cdot)\right)$ is

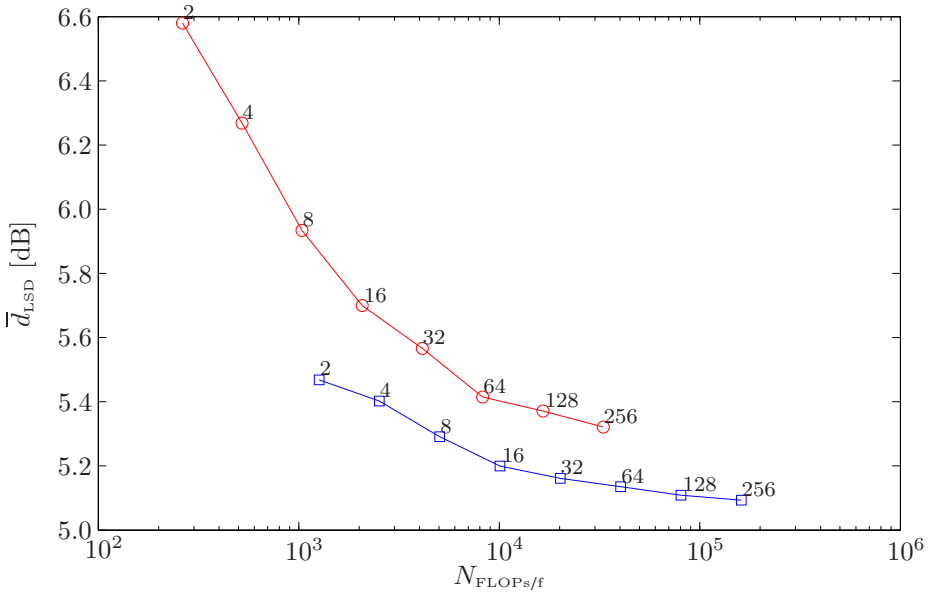
$$N^{\mathrm{diag}}_{\mathrm{p}} = M^{\mathrm{diag}}[(1 + 16 + 16) + (1 + 11 + 11)] = M^{\mathrm{diag}}[56]. \tag{3.32}$$

Using Eqs. (3.31) and (3.32) to calculate the LSD performance obtained in Figure 3.4 as a function of N_{p} results in Figure 3.5(a). In effect, we are comparing the performance of $\{\boldsymbol{\mathcal{G}}^{\mathrm{diag}}\}$ to that of $\{\boldsymbol{\mathcal{G}}^{\mathrm{full}}\}$ at those particular values of $M^{\mathrm{diag}} = kM^{\mathrm{full}}$ where $k > 1$ is determined such that the number of GMM parameters is the same for both $\boldsymbol{\mathcal{G}}^{\mathrm{diag}}$ and $\boldsymbol{\mathcal{G}}^{\mathrm{full}}$, i.e., $N^{\mathrm{diag}}_{\mathrm{p}} = N^{\mathrm{full}}_{\mathrm{p}}$. It is clear from Figure 3.5(a) that even when the number of Gaussians in the diagonal-covariance GMM tuple is increased such that the overall number of parameters is the same as that of the full-covariance GMM tuple being compared to, performance remains inferior. In order to achieve similar performance, M^{diag} has to be increased by more than an order of magnitude compared to M^{full} (e.g., $\overline{d}_{\mathrm{LSD}}$ performance is roughly the same at $M^{\mathrm{full}} = 4$ and $M^{\mathrm{diag}} = 64$), resulting in an overall increase—rather than a decrease—in the number of GMM parameters to be estimated during training compared to a full-covariance GMM ($N^{\mathrm{diag}}_{\mathrm{p}} = 3,584$ compared to $N^{\mathrm{full}}_{\mathrm{p}} = 924$ for $M^{\mathrm{diag}} = 64$ and $M^{\mathrm{full}} = 4$).

To perform a similar analysis of BWE performance as a function of per-frame extension-stage computational complexity, $N_{\mathrm{FLOPs/f}}$, we examine MMSE estimation more closely. It is clear from Eqs. (3.16) and (3.17) that the computational cost associated with MMSE estimation is dominated by the matrix inversion and determinant operations—the most expensive in those formulae; evaluating $\{E[\mathbf{Y}|\mathbf{x}, \lambda_i]\}_{i\in\{1,\ldots,M\}}$ requires calculating $\{\mathbf{C}_i^{\mathbf{xx}^{-1}}\}_{i\in\{1,\ldots,M\}}$



(a) BWE $\overline{d}_{\text{LSD}}$ performance as a function of the number of GMM parameters, N_{p}.



(b) BWE $\overline{d}_{\text{LSD}}$ performance as a function of the number of extension-stage computations per frame, $N_{\text{FLOPs/f}}$.

Fig. 3.5: BWE $\overline{d}_{\text{LSD}}$ performance as a function of memory (represented by N_{p}, the number of GMM parameters) and computational complexity (represented by $N_{\text{FLOPs/f}}$, the number of per-frame computations) required during extension for the $\mathcal{G}^{\text{full}}$ and $\mathcal{G}^{\text{diag}}$ GMM tuples defined in Eqs. (3.29) and (3.30), respectively. Data labels represent M, the number of Gaussian components.

while evaluating $\{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_i^{\mathbf{x}},\mathbf{C}_i^{\mathbf{xx}})\}_{i\in\{1,\ldots,M\}}$ requires calculating both $\{\mathbf{C}_i^{\mathbf{xx}^{-1}}\}_{i\in\{1,\ldots,M\}}$ and $\{|\mathbf{C}_i^{\mathbf{xx}}|\}_{i\in\{1,\ldots,M\}}$. Representing narrowband and highband dimensionalities in the two joint GMMs ($\mathcal{G}_{\mathbf{x\Omega}}$ and $\mathcal{G}_{\mathbf{x}G}$), by p and q, respectively, we have: (a) $p \coloneqq \mathrm{Dim}(\mathbf{X}) = 10$ and $q \coloneqq \mathrm{Dim}(\mathbf{\Omega}) = 6$ for $\mathcal{G}_{\mathbf{x\Omega}}$, and (b) $p \coloneqq \mathrm{Dim}(\mathbf{X}) = 10$ and $q \coloneqq \mathrm{Dim}(G) = 1$ for $\mathcal{G}_{\mathbf{x}G}$. Thus, the $\mathbf{C}_i^{\mathbf{xx}}$ matrix inversion and determinant operations result in an overall extension-stage complexity of $O(p^3)$ for MMSE estimation using full-covariance GMMs, compared to only $O(p)$ when using diagonal covariances.[72] While these orders of complexity favour diagonal-covariance GMMs over those with full covariances, they do not, however, account for two important factors:

1. As described in Section 3.3.3, a higher number of components in the Gaussian mixture is required when using diagonal covariances compared to using full covariances since such diagonal covariances are essentially an approximation.
2. In practice, the rather costly operations in Eqs. (3.16) and (3.17) associated with the full $\mathbf{C}_i^{\mathbf{xx}}$ and $\mathbf{C}_i^{\mathbf{yx}}$ matrices—namely, matrix multiplication, inversion, and the determinant—can be eliminated from online MMSE estimation altogether. This follows from the fact that these matrices—determined during the training stage—are already known and fixed beforehand.

By performing the following matrix operations offline for all $i \in \{1,\ldots,M\}$ prior to extension: (a) $-\frac{1}{2}\mathbf{C}_i^{\mathbf{xx}^{-1}}$, (b) $\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}$, and (c) $\frac{\alpha_i}{(2\pi)^{p/2}|\mathbf{C}_i^{\mathbf{xx}}|^{1/2}}$, the total number of FLOPs required to perform MMSE estimation per Eq. (3.12) for each input narrowband frame reduces to:[73]

$$N_{\mathrm{FLOPs/f}}^{\mathrm{diag}} = M^{\mathrm{diag}}(4p+q+21)+q-1, \quad \text{for } \mathcal{G}^{\mathrm{diag}}, \tag{3.33}$$

and

$$N_{\mathrm{FLOPs/f}}^{\mathrm{full}} = M^{\mathrm{full}}(2p^2+2pq+2p+q+21)+q-1, \quad \text{for } \mathcal{G}^{\mathrm{full}}. \tag{3.34}$$

For $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{X}\\ \mathbf{\Omega}\end{bmatrix}\right) = \begin{bmatrix}10\\6\end{bmatrix}$ and $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{X}\\ G\end{bmatrix}\right) = \begin{bmatrix}10\\1\end{bmatrix}$, Eq. (3.33) gives $N_{\mathrm{FLOPs/f}}^{\mathrm{diag}} = M^{\mathrm{diag}}[129]+5$ for the diagonal-covariance GMM tuple $\boldsymbol{\mathcal{G}}^{\mathrm{diag}}$, while Eq. (3.34) gives $N_{\mathrm{FLOPs/f}}^{\mathrm{full}} = M^{\mathrm{full}}[629]+5$ for $\boldsymbol{\mathcal{G}}^{\mathrm{full}}$. Using these relations, we obtain the LSD performance illustrated in Figure 3.5(b) as a function of $N_{\mathrm{FLOPs/f}}$ complexity, for both $\boldsymbol{\mathcal{G}}^{\mathrm{diag}}$ and $\boldsymbol{\mathcal{G}}^{\mathrm{full}}$. Similar to the findings of

[72] Most algorithms for matrix inverse or determinant calculation involve $O(n^3)$ complexity. Among those algorithms, Gaussian elimination [108, Section 3.2] is the most common. It requires $\approx 2n^3/3$ operations.

[73] Following [123], we assume that the exponential operation requires 20 FLOPs for x86 (32-bit) architectures.

Figure 3.5(a), we find that, even with M^{diag} increased relative to M^{full} such that overall extension-stage computational cost is identical in both GMM implementations, diagonal-covariance GMMs remain inferior to those with full covariances. Thus, we conclude that diagonal-covariance GMMs are, in fact, more computationally expensive compared to full-covariance GMMs if equivalent BWE performance is desired.

The lower LSD performance of diagonal-covariance GMMs compared to those with full covariances, even at equivalent complexity as measured in both scenarios above, indicates an inferior ability of diagonal-covariance GMMs to model the cross-band correlations fundamental to bandwidth extension. Indeed, by using diagonal covariances, cross-covariance terms—which explicitly capture cross-band correlations—are eliminated. Instead, it is assumed that such cross-band information will indirectly be captured by other parameters of the model, i.e., component priors, means, and variances, through the joint modelling action of the Gaussian components, provided that the number of components is sufficiently increased. We empirically showed above that this assumption is invalid; simply substituting cross-covariance terms by an equal number of additional diagonal-covariance Gaussian parameters is insufficient. The cross-band information modelled by cross-covariance terms requires, in fact, an exponentially higher number of such diagonal-covariance Gaussian parameters.

As Eq. (3.17) demonstrates, cross-covariance terms explicitly influence MMSE estimation through the inter-band to intra-band cross-covariance *ratios*, $\{\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}\}_{i\in\{1,\dots,M\}}$. A joint-band GMM's ability to model information mutual to the disjoint frequency bands, rather than band-specific information, is explicitly represented by these ratios, in contrast to the indirect and equally shared modelling through other model parameters. The higher these ratios are—on average—for a GMM, the more superior this full-covariance GMM is for BWE through MMSE, and the more difficult it is to achieve comparable performance through a diagonal-covariance GMM. Figure 3.6 illustrates, for example, the average Frobenius and L_p-norms[74] (for $p = 1, 2, \infty$) for $\{\mathbf{C}_i^{\boldsymbol{\omega}\mathbf{x}}\mathbf{C}_i^{\mathbf{xx}^{-1}}\}_{i\in\{1,\dots,M\}}$ as a function of M. Comparing Figure 3.6 to Figure 3.4 confirms the strong correlation between LSD performance and a full-covariance GMM's efficiency in modelling cross-band correlations represented by inter-band to intra-band cross-covariance ratios. The increased inefficiency of diagonal-covariance GMMs compared to full-covariance ones with higher average norms for the $\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}$ ratios is indirectly illustrated in Figure 3.4; while using a diagonal-covariance

[74] See Footnote 67.

GMM (with $M^{\text{diag}} = 64$) requires a relative increase of 389% in model parameters relative to a full-covariance GMM (with $M^{\text{full}} = 4$) to achieve an $\overline{d}_{\text{LSD}} \approxeq 5.4\,\text{dB}$, the relative increase required to achieve a similar $\overline{d}_{\text{LSD}} \approxeq 5.3\,\text{dB}$ is 776% (at $M^{\text{diag}} = 256$ and $M^{\text{full}} = 8$).

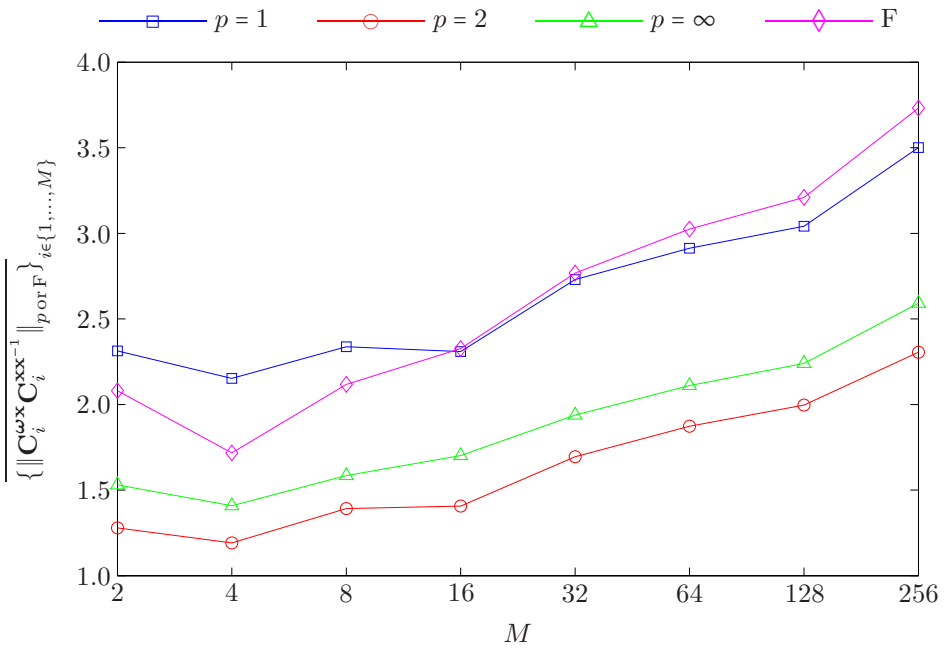


Fig. 3.6: Average norms of inter-band to intra-band Gaussian component cross-covariance ratios, $\{\|\mathbf{C}_i^{\boldsymbol{\omega}\mathbf{x}}\mathbf{C}_i^{\mathbf{x}\mathbf{x}^{-1}}\|_{p\,\text{or}\,\text{F}}\}_{i\in\{1,\ldots,M\}}$, for $\{\mathcal{G}^{\text{full}}_{\mathbf{x}\boldsymbol{\Omega}_j} := \mathcal{G}(\mathbf{x},\boldsymbol{\omega}; M_j^{\text{full}},\cdot)\}_{j\in\{1,\ldots,8\}}$ where $M_j^{\text{full}} = 2^j$.

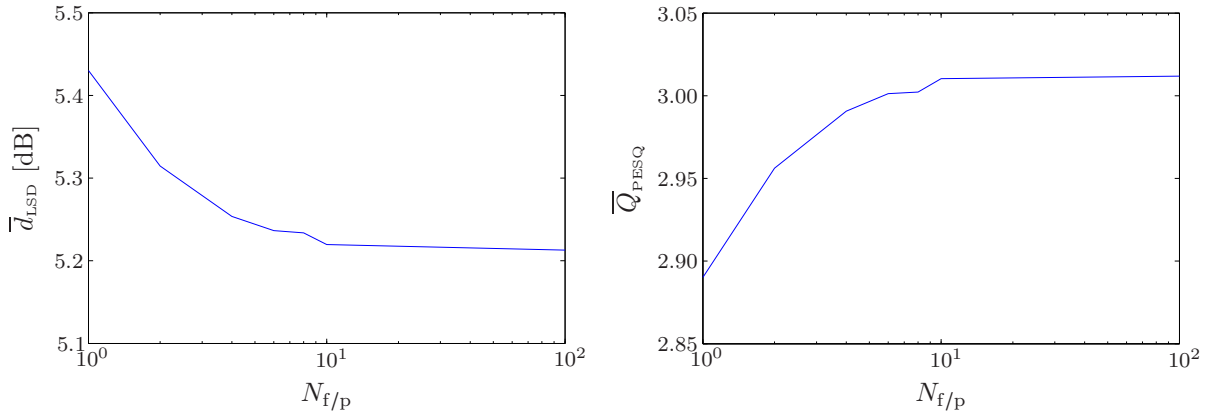
As a result of these findings, we only use full-covariance GMMs for our memoryless BWE baseline, as well as elsewhere in the sequel.

3.5.2 Effect of amount of training data

As discussed in Section 3.3.4 and showed by the results in Figure 3.4, increasing the number of Gaussian components, M, in a GMM translates into a better fit to the underlying distribution with the ability to capture finer details of the modelled space, and consequently, improved BWE performance. Such increases are generally constrained, however, by the amount of training data available. To assess the reliability of the EM-trained GMMs used in our work given the training data available through TIMIT, we perform a series of experiments where BWE performance is evaluated for varying amounts of GMM training data. Representing the relation between training data amount and GMM complexity by

the average number of data points (frames) per GMM parameter, i.e., $N_{\mathrm{f/p}} = N_{\mathrm{f}}/N_{\mathrm{p}}$, allows our results to be independent of GMM dimensionalities.

Using full-covariance GMM tuples, $\{(\mathcal{G}^{\mathrm{full}}_{\mathbf{x\Omega}}, \mathcal{G}^{\mathrm{full}}_{\mathbf{x}G})\}$, with $M^{\mathrm{full}} = 16$ and $N_{\mathrm{p}}^{\mathrm{full}}$ given by Eq. (3.31), we obtain the BWE performance—evaluated on the TIMIT test data described in Section 3.2.10—illustrated in Figure 3.7 as a function of $N_{\mathrm{f/p}}$. Figure 3.7 shows that BWE performance is virtually unaffected, i.e., the estimated GMM parameters are reliable, for $N_{\mathrm{f/p}} \geq 10$. Compared to $N_{\mathrm{f/p}} = 100$, the value suggested in [109] for reliable GMM parameter estimation, degradation in BWE performance with $N_{\mathrm{f/p}} = 10$ is quite negligible (relative degradations in $\overline{d}_{\mathrm{LSD}}$ and $\overline{Q}_{\mathrm{PESQ}}$ are $\approx 0.13\%$ and $\approx 0.05\%$, respectively). Thus, for the maximum dimensionality of $D = \mathrm{Dim}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{\Omega}\end{bmatrix}\right) = 16$, $N_{\mathrm{f/p}} = 10$ frames/parameter, and $N_{\mathrm{f}} \approxeq 1.125 \times 10^6$ frames[75], Eq. (3.18) gives $M_{\max} = 735$ for the maximum reliable number of Gaussian components in $\mathcal{G}^{\mathrm{full}}_{\mathbf{x\Omega}}$, thereby confirming the reliability of the results in Figure 3.4.



(a) $\overline{d}_{\mathrm{LSD}}$ performance (b) $\overline{Q}_{\mathrm{PESQ}}$ performance

Fig. 3.7: Average BWE $\overline{d}_{\mathrm{LSD}}$ and $\overline{Q}_{\mathrm{PESQ}}$ performance as a function of the amount of training data represented by $N_{\mathrm{f/p}}$ using full-covariance GMM tuples, $\{(\mathcal{G}^{\mathrm{full}}_{\mathbf{x\Omega}}, \mathcal{G}^{\mathrm{full}}_{\mathbf{x}G})\}$, with $M^{\mathrm{full}} = 16$. Performance shown is the average over 10 GMM tuple training instances at each $N_{\mathrm{f/p}}$ value.

3.5.3 Baseline performance

Based on the results presented above, we select the full-covariance GMM tuple $(\mathcal{G}^{\mathrm{full}}_{\mathbf{x\Omega}}, \mathcal{G}^{\mathrm{full}}_{\mathbf{x}G})$ with $M^{\mathrm{full}} = 128$ for our memoryless BWE baseline. Table 3.1 lists the baseline BWE performance—evaluated for the TIMIT core test set with $N_{\mathrm{f}} \approxeq 58 \times 10^3$ frames[76]—using the

[75] See Section 3.2.10.
[76] See Section 3.2.10.

measures detailed in Section 3.4.[77]

Table 3.1: Speaker-independent memoryless BWE baseline performance using full-covariance GMMs with $M = 128$, and LSF parameterization with $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{x}\\ \boldsymbol{\Omega}_y\end{bmatrix}\right) = 16$ and $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{x}\\ G\end{bmatrix}\right) = 11$.

$\overline{d}_{\mathrm{LSD}}$ [dB]	$\overline{d}_{\mathrm{LSD(RMS)}}$ [dB]	$\overline{Q}_{\mathrm{PESQ}}$	$\overline{d}^{\star}_{\mathrm{IS}}$ [dB]	$\overline{d}^{\star}_{\mathrm{I}}$ [dB]
5.11	5.82	3.06	10.53	0.5835

3.6 Summary

A thorough description of the dual-mode system used as the basis for BWE throughout our work was presented. Most relevant to our later investigations of the effect of memory inclusion on BWE performance is the GMM-based statistical modelling employed in order to reconstruct highband spectra in the 4–8 kHz range. As such, particular attention was given to the GMM framework. A general derivation was presented for joint density MMSE estimation using multi-modal densities, which was then applied to the GMM special case. In addition, the role of the number and covariance type of Gaussian components as well as the relation between the amount of training data available and GMM complexity were carefully examined. This analysis, quite important to establish and confirm the reliability of GMM-based BWE in general, is especially lacking in the literature. Based on our findings, we concluded that full-covariance GMMs are, in fact, more computationally efficient compared to diagonal-covariance GMMs with equivalent performance, and hence, are used as the means for statistical modelling in our work.

For BWE performance evaluation, an ensemble of objective measures was selected such that results obtained in our work are: (a) comparable to those of previous works (LSD), (b) quite highly correlated with subjective measures (PESQ), and (c) sufficiently detailed to allow separately studying gain-related and spectral shape-related BWE performance improvements (symmetrized Itakura-Saito and Itakura distortion measures).

Finally, based on the analysis described above, a well-performing memoryless BWE baseline for the work to follow was selected and its performance presented using the chosen ensemble of objective measures.

[77] Since GMM training is sensitive to initialization conditions, all GMM-derived results listed here and in the sequel, including BWE performance figures such as those of Table 3.1, are based on averages of at least 4 realizations with random initializations.

Chapter 4

Modelling Speech Memory and Quantifying its Effect

4.1 Introduction

In contrast to the considerable research published on BWE techniques, only a few researchers have actually investigated the correlation assumption between narrowband and highband spectral envelopes. In [124], an approximate lower bound on the mutual information (MI) between narrow- and high-frequency bands was derived. This initial attempt was extended in [109] to quantify the *certainty* about the high band given the narrow band by determining the ratio of the MI between the two bands to the discrete entropy of the high band. The authors show that this ratio (representing the dependence between the two bands) is quite low. The relation of this ratio to BWE performance was further confirmed in [125] by deriving an upper bound on achievable BWE performance—represented by log-spectral distortion (LSD)—given a certain amount of MI and highband entropy.

Despite the low dependence, BWE schemes have, for the most part, continued to use *memoryless mapping* between spectra of both bands. It was thus concluded in [109] that these schemes "*perform reasonably, not because they accurately predict the true high band, but rather by extending the narrow band such that the overall wideband signal sounds pleasant*". Accordingly, BWE methods should make use of perceptually-relevant properties to improve the subjective quality of extended speech. This implies that, for the vast majority of BWE schemes employing linear prediction for the representation of spectral envelopes, characteristics of the excitation of input speech, e.g., gain or voicing, should be included in the feature vector mapping in addition to the well-tried spectral envelope parameters [126].

As described in Sections 1.4 and 2.3.3.4, a few works, based primarily on hidden Markov models (HMMs), have been proposed for the purpose of exploiting the benefits of speech memory to improve BWE performance, most notably [39, 84, 87]. Due to their high complexity and training data requirements, however, these HMM-based approaches are limited to first-order Markov modelling—effectively restricting the memory modelled to only 20–40 ms. It has been shown, however, that speech temporal information extends up to 1000 ms [127], with energies of modulation spectra (spectra of the temporal envelopes of the signal) peaking around 4–5 Hz—corresponding to 200–250 ms [128]. This latter finding coincides with the aforementioned conclusion in [10, Section 5.4.2] that the perception of phonemes utilizes dynamic acoustic patterns over sections of speech corresponding roughly to syllables.

In addition to these HMM-based approaches, a handful of other works have also been proposed to make use of speech dynamics to improve BWE performance. However, these works, discussed in Sections 5.3.1 and 5.4.1, are also characterized either by their limitations on the extent of memory used, e.g., [129–132], by their excessive computational requirements, e.g., [133], and/or by using a speech production model other than the source-filter model (thereby making performance comparisons to source-filter model-based techniques nearly impossible without subjective evaluations), e.g., [132].

While all approaches exploiting memory are reported to show superior performance compared to memoryless ones, it is notable that none has explicitly quantified the gain of exploiting the considerable information in the dynamic temporal and spectral patterns of speech. In our work presented in this chapter, first introduced in [134] and continued in [135], we explicitly account for speech memory through *delta features* [136]—widely used in speech recognition. Delta features incorporate the considerable temporal correlation properties in long-term speech, otherwise neglected by conventional *static* parametrization. They can be applied to almost any form of parametrization, thus partially transferring the task of capturing temporal information from the modelling space (through GMMs or HMMs) to the frontend (i.e., parameterization). By substituting higher-order static feature vectors by *dynamic* vectors comprising lower-order static parameters as well as their delta features, speech dynamics are modelled while overall feature vector dimensionalities can be preserved, thereby requiring no increase in statistical modelling complexity or training data requirements. More importantly for our work, delta features are obtained through linearly weighted differences between neighbouring static feature vectors. Thus, they also

provide a significant advantage over first-order Markov chains; the extent of embedded temporal information for a signal frame is controlled by varying the span of neighbouring static feature vectors involved in the calculation of the delta features for that specific frame. This property eliminates the need for complex HMM structures (with high-order Markov chains), and hence, also eliminates the associated increases in computational resources and data required for statistical training. Through this *frontend-based memory inclusion*, we study the effects of including up to 600 ms of memory (300 ms on each side of a signal frame) in speech parametrization.

To examine the effect of memory inclusion on highband certainty, we consider mel-frequency cepstral coefficients (MFCCs) [137] as well as line spectral frequencies (LSFs) [93] for the parameterization of the same signals representing the two speech frequency bands as described in Chapter 3, i.e., the midband-equalized narrowband (0.3–4 kHz) and highband (4–8 kHz) signals. MFCCs were shown in [126] to have the highest class separability and second highest MI content among several speech parameterizations, while LSFs are widely used in speech coding for their quantization error resilience and perceptual significance properties. Similar to [109] and [125], we estimate MI using the numerical method of stochastic integration, where the marginal and joint distributions of the narrow and high band parameterizations are modelled by Gaussian mixture models (GMMs) for both static and dynamic (static+delta) acoustic spaces. Rather than estimate the discrete highband entropy indirectly from the differential one through scalar quantization (SQ) of the highband space as in [109] (where stochastic integration is also used to estimate differential entropy), we estimate discrete entropy directly by vector quantizing (VQ) the highband space such that the average LSD corresponding to all quantized highband feature vectors is equal to 1 dB; the first spectral transparency threshold of [115].[78] Our VQ approach results in more realistic and accurate discrete entropy estimates than those of [109] and, more importantly, allows entropy estimation for LSFs as well as MFCCs (unlike the indirect SQ approach of [109] applicable only to MFCCs).

By varying the number of static feature vectors involved in the estimation of the delta features, we show that frontend-based memory inclusion can increase certainty about the highband by over 100% for both LSFs and MFCCs. Expressed alternatively, the relative decrease in *uncertainty* about the highband—corresponding to a potential decrease in BWE distortion—is shown to be, approximately, 20% and 38%, for LSFs and MFCCs,

[78] See Section 3.4.1.

respectively. Furthermore, our results show that certainty gains due to memory inclusion saturate at durations corresponding roughly to inter-phoneme (or syllabic) temporal information. This latter result coincides with earlier findings about the contribution of memory to phoneme identification. Phonemes with mostly highband energy, e.g., fricatives, stand to have the most benefit of such short-term syllabic memory inclusion. Since BWE schemes generally perform poorly when reconstructing such phonemes, we expect BWE performance to be generally improved by memory inclusion.

4.2 Speech Parameterization

4.2.1 On the perceptual properties of speech

As described in Section 3.2.2, LP-derived LSFs have the desirable properties of synthesis filter stability, error resilience and localization, and correspondence to properties of formants and valleys. Since the vast majority of BWE techniques employ the source-filter model of Section 2.3.1, these properties make LSFs particularly attractive for such LP-based BWE schemes, especially so for those employing GMM-based statistical estimation. LSFs, however, do not incorporate some of the most important aspects of speech perception—the nonlinear relation between a sound's perceived pitch and the sound's frequency [138], and the critical-band nature of perception [139]. The first aspect relates to the psychoacoustic property whereby the perceived pitch is essentially linear with frequency up to 1 kHz and logarithmic at higher frequencies, resulting in the perceptual *mel* scale for pitch [138].[79] The mel scale, thus, gives higher resolution to lower frequencies. The most popular linear-to-mel-scale frequency mapping is that of [10, Section 4.3.6];

$$f_{\text{mel}} = 2595 \log_{10}\left(1 + \frac{f_{\text{Hz}}}{700}\right). \tag{4.1}$$

The second aspect relates to another important psychoacoustic property whereby the perception of sound stimuli is defined by ranges of sound frequencies known as *critical bands* [139]. The loudness of a band of noise at constant sound pressure remains constant as the

[79] The mel scale is a perceptual scale of the pitch of pure tones where tone frequencies in Hz are mapped to subjective pitch values in mels as judged by listeners. As a reference point, the pitch of of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Other subjective pitch values in mels are obtained by adjusting the frequency of a stimulus tone such that its perceived pitch is half or twice that of a reference tone.

noise bandwidth increases up to the width of the critical band, beyond which increased loudness is perceived. Similarly, a sub-critical bandwidth multi-tone sound of constant intensity is perceived as loud as an equally intense pure tone at the centre frequency of the band, regardless of the overall frequency separation of the multiple tones. When the separation exceeds the critical bandwidth, the complex multi-tone sound is perceived as becoming louder. Below 500 Hz, critical bandwidth is roughly constant at ≈ 100 Hz, increasing roughly logarithmically with higher frequencies above 1 kHz [10, Section 4.3.6]. Closely related to the mel scale, the Bark scale—proposed in [140]—relates acoustical frequency to perceptual frequency resolution where one Bark covers one critical bandwidth.

The subjective importance of these two perceptual properties is demonstrated by the superior subjective correlation of PESQ scores with MOS relative to other distortion measures (as described in Section 3.4.3, the PESQ perceptual model explicitly employs binning of FFT coefficients on a modified Bark scale). Given their importance, the lack of accounting for these properties in LSF parameterization motivates us to seek a more perceptually-inspired parameterization to be used—in addition to LSFs—for the investigation of cross-band correlations described in this chapter. As described below, the properties of mel-frequency cepstral coefficients (MFCCs) make them a means of parameterization well suited for the task. While such a parameterization may not be as amenable to actual highband speech reconstruction as LSFs are, our focus in this chapter is to rather quantify the role of memory in improving cross-band correlations, represented by certainty about the highband. As such, using the more subjectively-correlated MFCCs, in addition to LSFs as reference, makes our findings more relevant perceptually.

4.2.2 MFCCs

In contrast to the conventional cepstrum defined as the Fourier transform of the logarithm of the signal spectrum, MFCCs—attributed to Mermelstein [137]—parameterize a short-time spectrum perceptually through filterbank analysis—simulating critical bands—on the mel scale, thereby modelling the two perceptual properties described above. In addition, MFCC parameterization employs the discrete cosine transform (DCT) rather than the Fourier transform. We apply MFCC parametrization—MFCCs are denoted below by $\{c_n\}_{n\in\{0,\ldots,K-1\}}$ for K mel-scale filters—of the midband-equalized narrowband (0.3–4 kHz) and highband (4–8 kHz) signals as follows:

1. **No pre-emphasis:** Typically, a high-pass filter with a single pole (at $z = -0.97$, for example) is used to compensate for the long-term average speech energy roll-off of 6 dB/octave and to generally emphasize high-frequency content. For our implementation, however, we do not apply such pre-emphasis.[80]

2. **Windowing:** The modified Hann window described in Section 3.2.8 is used to mitigate the edge effect of discontinuities due to framing. As in Section 3.2.8, we use 20 ms frames with 50% overlap.

3. **Power spectrum:** FFT (Fast Fourier transform) is applied followed by a magnitude and squaring operation (thereby discarding phase).

4. **Mel-scale filterbank binning:** Mel-scale triangular filters (based on the conversion formula of Eq. (4.1)) are applied to the power spectrum in each of the two frequency bands such that the squared absolute values of FFT coefficients within each filter are summed resulting in mel-scale filterbank energies. Corresponding to the perceptual measurements of Zwicker in [139] where approximately 21–22 critical bands span the 0–8 kHz frequency range, we use 15 filters for the 0–4 kHz narrow band and 7 for the 4–8 kHz high band with the filters being equally-spaced within each band. Similar to [109], we ensure there is no overlap between the two sets of filters in order to avoid introducing artificial dependencies between the two disjoint frequency bands. Figure 4.1 illustrates the two filter banks.

5. **Log operation:** Filterbank log-energies are obtained.

6. **DCT:** The binned mel-scale log spectrum is converted to the *cepstral* domain through

[80] As described in Section 4.3.3, Euclidean distances between MFCC vectors directly correspond to a perceptually-weighted LSD measure provided that MFCCs are not *liftered*—i.e., filtered in the cepstral domain—and c_0 is scaled appropriately to ensure unitary DCT. Pre-emphasizing speech through time-domain filtering corresponds to additive liftering in the cepstral domain that would unevenly bias the LSD measure towards higher frequencies, and hence, requires undoing the liftering by subtracting the MFCC vector corresponding to the pre-emphasis filter from MFCC feature vectors prior to LSD calculation.

Applying pre-emphasis, however, resulted in no tangible gains in our MFCC-based certainty evaluations described in this chapter, as well as in the BWE performance evaluations described in Chapter 5. As such, we concluded that the additional computational costs associated with pre-emphasis filtering and unliftering—albeit minor—were unjustified.

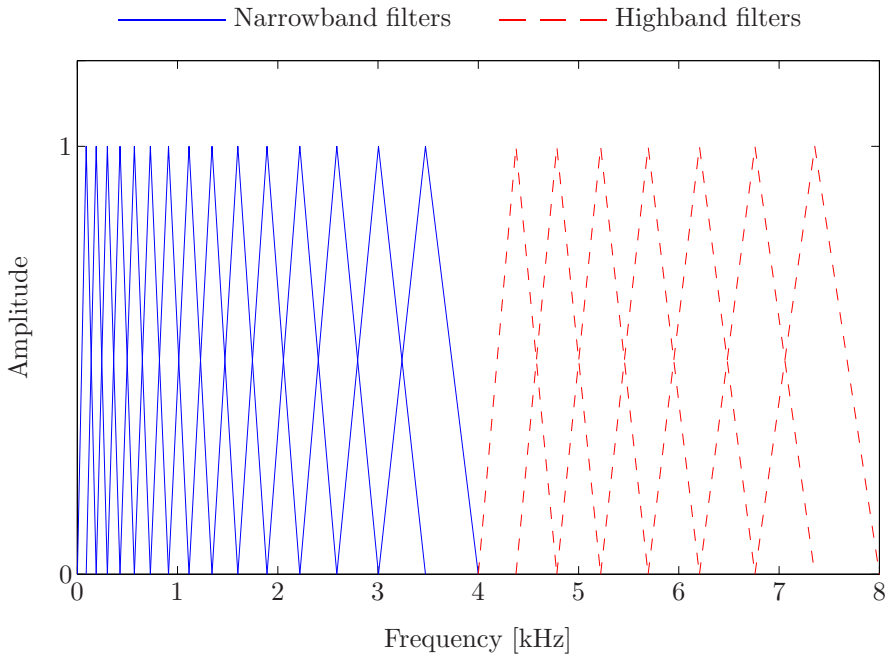


Fig. 4.1: Mel-scale equally-spaced filter bank used for MFCC parameterization. Frequency scale conversion is based on Eq. (4.1).

a discrete cosine transform (DCT) [141]. In particular, we use Type-II DCT per

$$c_n = a \sum_{k=0}^{K-1} (\log_e \varepsilon_k) \cos\left(n(k+\frac{1}{2})\frac{\pi}{K}\right), \text{ where } a = \begin{cases} \sqrt{\frac{1}{K}}, & \text{for } n = 0, \\ \sqrt{\frac{2}{K}}, & \text{for } n = 1, \ldots, K-1, \end{cases} \tag{4.2}$$

c_n is the nth MFCC, K is the number of mel-scale filters of the pertaining frequency band, and ε_k is the kth mel-scale filter energy. Using $K = 7$ filters for the high band results in 6 MFCCs, $\{c_n\}_{n\epsilon\{1,\ldots,6\}}$, representing highband spectral envelope shape (thereby corresponding exactly to the 6 highband LSFs used in our memoryless baseline BWE system) and 1 coefficient, c_0, representing highband energy.

A well-known property of MFCCs is that the terms $\{c_n\}$ are well-decorrelated; this follows directly from the decorrelating effect of the DCT. The magnitudes of the off-diagonal covariance terms for an arbitrary set of MFCC vectors are considerably lower than those of the diagonal terms. As such, the DCT can be viewed as a unitary rotation of principal axes which, in effect, *orthogonalizes* and reduces the scatter of data points around their K-dimensional mean. Assuming that feature vectors follow an underlying distribution of

overlapping classes, the decorrelating/orthogonalizing rotation performed by the DCT thus improves class *separability*. Separability is a measure of the quality of a particular feature set in terms of classification [71, Section 3.8.3]. For a set of classes defined over a feature vector space, the separability of feature vectors is given by the ratio of between-class scatter to within-class scatter. Consequently, and as shown in [126], MFCCs exhibit the highest class separability among the common parameterizations of LPCs, LSFs, ACF (auto-correlation function) features, and conventional as well as LP-based cepstral coefficients (where cepstral coefficients are calculated from smooth LP-based spectra rather than the signal spectra). The improved class separability associated with a particular parameterization translates into acoustic-space modelling that is more discriminative of these classes, with a better rate-distortion curve compared to other parameterizations with lower separability; i.e., fewer bits are required to achieve the same classification performance of a different feature set with lower separability. As described below in Section 4.3.2, this implies lower entropy for the quantization of MFCCs compared to LSFs for the same LSD performance. Given sufficient MI between MFCC-parameterized narrowband and highband spectral envelopes, the lower MFCC highband entropy results in higher cross-band correlation as quantified by highband certainty.

To conclude this motivation and analysis of our use of MFCCs, it is worth noting that the superior decorrelation properties described above are frequency band-specific, i.e., they do not extend across the wideband space underlying joint-band feature vectors. In fact, it is this very property that leads to the superior multiplicative $\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}$ factors—which, as discussed in Section 3.3.3, represent the weights on the contributions of the source data to the MMSE estimates of the target—for MFCCs, relative to LSFs. By being frequency band-specific, the DCT decorrelation effects reduce the norms of the within-band $\mathbf{C}_i^{\mathbf{xx}^{-1}}$ covariances, but not those of the cross-band $\mathbf{C}_i^{\mathbf{yx}}$ terms, thereby resulting in higher overall weights for the MMSE multiplicative $\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}$ factors.

4.3 Highband Certainty Estimation

To verify and quantify the cross-band correlation assumption underlying BWE in both memoryless and memory-inclusive conditions, we exploit the information-theoretic measure of highband certainty—the ratio of mutual information (MI) between the narrow and high frequency band representations to the discrete entropy of the highband representation—

proposed in [109]. The motivation for using MI arises from the fact that it measures all statistical dependence between two random variables, linear as well as non-linear. In contrast, the common correlation coefficient, often used as a measure of dependence between random variables, only measures linear dependence or second order statistics between the variables. We have shown in Chapter 1 that the relationship between the narrow and high frequency bands is a complex and nonlinear one. Accordingly, the cross-band dependencies of interest can only be measured through MI.

MI—denoted by $I(\mathbf{X};\mathbf{Y})$—quantifies the information mutual to the particular parameterizations of both bands; i.e., it measures the information available in narrowband feature vectors, $\mathbf{X}$, about those of the highband, $\mathbf{Y}$. For the purpose of highband reconstruction, however, it is not the quantity of such shared information that matters per se, but rather, it is the relevance of that quantity in relation to the total information in the highband representation—i.e., highband entropy, $H(\mathbf{Y})$. Thus, in the context of BWE, MI alone is not sufficient; a more relevant measure of cross-band dependence is rather the ratio of MI to highband entropy, $\frac{I(\mathbf{X};\mathbf{Y})}{H(\mathbf{Y})}$. This ratio, quantifying certainty about the highband parameterization given the narrowband's, is, in fact, a normalized measure of cross-band dependence; the minimum highband certainty value of 0 indicates statistical independence between the two bands, while a maximum certainty of 1 indicates complete knowledge about highband content given that of the narrow band. Given this interpretation, we denote highband certainty, given the narrow band, by the more representative

$$C(\mathbf{Y}|\mathbf{X}) := \frac{I(\mathbf{X};\mathbf{Y})}{H(\mathbf{Y})}, \tag{4.3}$$

with the uncertainty remaining in the high band given by $1 - C(\mathbf{Y}|\mathbf{X})$. Similar normalizations have previously been proposed in other contexts; e.g., the *relative information transmitted* of [142]—given by $\frac{I(\mathbf{X};\mathbf{Y})}{\min[H(\mathbf{X}),H(\mathbf{Y})]}$—normalizes MI relative to the maximum amount of information that can be shared, regardless of whether that information corresponds to the source or target.

4.3.1 Mutual information

Given the narrow and high frequency bands represented by the continuous vector variables $\mathbf{X}$ and $\mathbf{Y}$, respectively, with the marginal and joint *pdf*s: $p_{\mathbf{X}}(\mathbf{x})$, $p_{\mathbf{Y}}(\mathbf{y})$, and $p_{\mathbf{XY}}(\mathbf{x},\mathbf{y})$,

the mutual information $I(\mathbf{X};\mathbf{Y})$ between the two bands is equal to the Kullback-Leibler divergence; i.e., it can be written in terms of the marginal and joint *pdf*s as [64, Section 8.5]

$$I(\mathbf{X};\mathbf{Y}) = \int_{\Omega_{\mathbf{y}}}\int_{\Omega_{\mathbf{x}}} p_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \log_2\left(\frac{p_{\mathbf{XY}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})p_{\mathbf{Y}}(\mathbf{y})}\right) d\mathbf{x}\, d\mathbf{y} \qquad \text{[bits]}. \tag{4.4}$$

Rewriting Eq. (4.4) as

$$I(\mathbf{X};\mathbf{Y}) = E\left[\log_2\left(\frac{p_{\mathbf{XY}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})p_{\mathbf{Y}}(\mathbf{y})}\right)\right], \tag{4.5}$$

and replacing the expectation operator by the sample mean yields (by the law of large numbers with the number of samples, N, sufficiently large)

$$I(\mathbf{X};\mathbf{Y}) \approxeq \frac{1}{N}\sum_{n=1}^{N} \log_2\left(\frac{P_{\mathbf{XY}}(\mathbf{x}_n,\mathbf{y}_n)}{P_{\mathbf{X}}(\mathbf{x}_n)P_{\mathbf{Y}}(\mathbf{y}_n)}\right). \tag{4.6}$$

As discussed in Section 2.3.3.4, GMMs provide a superior means for the modelling of arbitrary densities in general, and of speech-derived ones in particular. Thus, similar to [109] and [125], we approximate the marginal and joint densities of Eq. (4.6) using GMMs[81], thereby allowing the estimation of MI (in bits) using numerical integration per[82]

$$\hat{I}(\mathbf{X};\mathbf{Y}) = \frac{1}{N}\sum_{n=1}^{N} \log_2\left(\frac{\mathcal{G}_{\mathbf{XY}}(\mathbf{x}_n,\mathbf{y}_n)}{\mathcal{G}_{\mathbf{X}}(\mathbf{x}_n)\mathcal{G}_{\mathbf{Y}}(\mathbf{y}_n)}\right). \tag{4.7}$$

4.3.2 Discrete highband entropy

Given the continuous nature of the acoustic space, either the differential entropy or the discrete entropy—obtained through quantization of the continuous acoustic space—of the highband feature vector space, $\mathbf{Y}$, can be used to quantify highband self-information. The differential entropy of the highband feature vector space, given by,

$$h(\mathbf{Y}) = -\int_{\Omega_{\mathbf{y}}} p_{\mathbf{Y}}(\mathbf{y}) \log_2 p_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y} \qquad \text{[bits]}, \tag{4.8}$$

[81] See Eq. (2.13).

[82] As noted in [109], the technique of replacing an integration with a sample mean average has been successfully used in [111] to obtain rate-distortion curves in the context of high-rate vector quantization.

can be estimated via stochastic integration in the same manner used to estimate mutual information;[83] i.e.,

$$\hat{h}(\mathbf{Y}) = -\frac{1}{N}\sum_{n=1}^{N}\log_2 \mathcal{G}_{\mathbf{Y}}(\mathbf{y}_n). \tag{4.9}$$

However, since $h(\mathbf{Y})$—and differential entropy in general—is susceptible to any scaling of $\mathbf{Y}$ [64, Theorem 8.6.4], the discrete entropy provides a more consistent estimate of highband self-information. Representing highband self-information by discrete entropy implies quantization of the continuous random feature vectors $\mathbf{Y}$ into discrete vectors represented by the mapping $\mathcal{Q}(\mathbf{Y})$. For $q := \mathrm{Dim}(\mathbf{Y})$, a straightforward method to estimate $H(\mathcal{Q}(\mathbf{Y}))$ from $h(\mathbf{Y})$ is by entropy-constrained q-dimensional scalar quantization of the continuous feature vectors $\mathbf{Y}$—provided that $p_{\mathbf{Y}}(\mathbf{y})\log_2 p_{\mathbf{Y}}(\mathbf{y})$ is Riemann integrable [64, Theorem 8.3.1]—resulting in the approximation (dropping the hat in $\hat{h}(\mathbf{Y})$ and the mapping in $H(\mathcal{Q}(\mathbf{Y}))$ to simplify notation)

$$H(\mathbf{Y}) \approxeq h(\mathbf{Y}) - \log_2(\Delta^q), \tag{4.10}$$

where Δ is the quantization step-size.[84] The MSE distortion resulting from such scalar quantization (SQ) is given by

$$D = q\frac{\Delta^2}{12}. \tag{4.11}$$

As described in Section 4.3.3 below, Euclidean distances between MFCC vectors correspond directly to a more perceptually-relevant form of LSD. Thus, by using MFCCs as highband feature vectors $\mathbf{Y}$, the SQ distortion of Eq. (4.11) will, in fact, be equal to square LSD. This, in turn, allows estimating the discrete entropy $H(\mathbf{Y})$ corresponding to a particular LSD, e.g., the 1 dB spectral transparency threshold of [115], using Eq. (4.10) and a differential entropy estimate $h(\mathbf{Y})$ obtained via the GMM-based numerical approximation of Eq. (4.9).

Estimating discrete entropy through SQ per the approach above was proposed in [109], and applied for memoryless highband certainty estimation for the different sound classes

[83] Estimating entropy through modelling the underlying probability density or mass function is often referred to as *plug-in* estimation. These methods include histogram and mixture modelling (with the latter being the method employed here). A different class of entropy estimators uses data directly for entropy estimation without density estimation. See [143] for an overview of entropy estimators.

[84] For entropy-constrained quantization, distortion is minimized under the constraint that the average codeword length is fixed. This results in a fixed centroid density, i.e., fixed quantization step-size. Resolution-constrained quantization, on the other hand, minimizes distortion under the constraint that all codewords have a fixed length, resulting in variable centroid density and quantization step-size. See [144, Chapter 7] for details.

of Table 1.1, i.e., vowels, fricatives, et cetera. This approach, however, is an approximation that is only valid under the high-rate assumption, i.e., if the quantization step-size Δ is small enough such that the q-dimensional *pdf* of $\mathbf{Y}$ can be considered flat along each dimension in each quantization bin [145]. Furthermore, since entropy-constrained SQ partitions the multi-dimensional feature vector space into hypercubes; i.e., using the same step-size Δ for all dimensions of $\mathbf{Y}$, marginal densities along all dimensions are assumed to have similar variances. This assumption is invalid for many speech parameterizations. As a result of the energy-packing characteristics of the DCT, for example, MFCCs exhibit a large dynamic range; numerical MFCC values decrease as the order of the cepstral coefficient increases, leading to a non-uniform distribution of MFCC variances. The uniform variance assumption of SQ thus results in further distortion due to the inefficient equal allocation of available bits to dimensions with differing variances. Finally, we note that the distortion resulting from inefficiently partitioning the highband feature space $\mathbf{Y}$ into hypercubes increases with the dimensionality $q = \mathrm{Dim}(\mathbf{Y})$.

Rather than estimate discrete highband entropy, $H(\mathbf{Y})$, indirectly via GMM-based *pdf* estimation to first obtain the differential entropy—via Eq. (4.9)—followed by entropy-constrained SQ—via Eqs. (4.10) and (4.11)—as described above, we estimate $H(\mathbf{Y})$ directly by performing resolution-constrained VQ of the highband space such that the average quantization distortion corresponds to an average LSD of 1 dB—the first spectral transparency threshold of [115]. In particular, we perform VQ using the generalized Lloyd algorithm [97] in steps of increasing resolution. At each step, quantization distortion is calculated as the average LSD of all training feature vectors given their quantized VQ codevectors. The VQ codebook size is increased until average LSD falls below the 1 dB spectral transparency threshold. As noted in Section 3.4.1, the 1 dB spectral transparency threshold of [115] was determined empirically for the 0–3 kHz band. Since level discrimination decreases for higher frequencies (i.e., higher difference limens), the average LSD threshold for spectral transparency for frequencies above 3 kHz is, in fact, higher than 1 dB. Nevertheless, the 1 dB average LSD threshold can still be applied to the highband frequency range but as a rather conservative estimate. Calculating average LSD for LSF and MFCC quantized data is described in Section 4.3.3.

VQ applied as such effectively results in a q-dimensional histogram-based estimator of the *pdf* of $\mathbf{Y}$, $p_{\mathbf{Y}}(\mathbf{y})$, with $p_{\mathbf{Y}}(\mathbf{y})$ approximated by the probability mass function of $\mathcal{Q}(\mathbf{Y})$, $p_{\mathcal{Q}(\mathbf{Y})}\big(\mathcal{Q}(\mathbf{y})\big)$, estimated directly from a training data set. In other words, we apply a

mapping, $\mathcal{Q}$, of the q-dimensional feature vector Euclidean space $\mathbb{R}^q$, onto a countable set of codevectors, $\mathcal{C} = \{\mathbf{c}_i\}_{i\in\mathcal{I}}$, where $\mathcal{I}$ is a countable set of indices; i.e.,

$$\mathcal{Q}\colon \mathbb{R}^q \to \mathcal{C}, \quad \text{where } \mathbf{Y} \subseteq \mathbb{R}^q \text{ and } \mathcal{Q}(\mathbf{Y}) = \mathcal{C}. \tag{4.12}$$

Thus, for $|\mathcal{I}|$ Voronoi with the ith Voronoi defined by

$$\mathcal{V}_i = \{\mathbf{y} \in \mathbb{R}^q \colon \mathcal{Q}(\mathbf{y}) = \mathbf{c}_i\}, \tag{4.13}$$

the discrete highband entropy can be estimated by

$$H(\mathbf{Y}) \equiv H(\mathcal{Q}(\mathbf{Y})) = -\sum_{i\in\mathcal{I}} P_{\mathcal{Q}(\mathbf{Y})}(\mathbf{c}_i)\, \log_2 P_{\mathcal{Q}(\mathbf{Y})}(\mathbf{c}_i), \tag{4.14}$$

where, for a data set $\mathcal{V} = \{\mathbf{y}_n\}_{n\in\{1,\ldots,|\mathcal{V}|\}}$ with the total number $|\mathcal{V}|$ of VQ training frames,

$$\begin{aligned} P_{\mathcal{Q}(\mathbf{Y})}(\mathbf{c}_i) &\approxeq P(\{\mathbf{y}_n\colon \mathcal{Q}(\mathbf{y}_n) = \mathbf{c}_i\}) = P(\{\mathbf{y}_n\colon \mathbf{y}_n \in \mathcal{V}_i\}) \\ &= \frac{|\{\mathbf{y}_n\colon \mathcal{Q}(\mathbf{y}_n) = \mathbf{c}_i\}|}{|\mathcal{V}|} = \frac{|\{\mathbf{y}_n\colon \mathbf{y}_n \in \mathcal{V}_i\}|}{|\mathcal{V}|}. \end{aligned} \tag{4.15}$$

With the codebook cardinality constrained to powers of 2; i.e., $|\mathcal{I}| = 2^n$ where $n \in \mathbb{Z}$, we perform VQ in steps of increasing resolution until $\overline{d}_{\text{LSD}}(n)$—LSD expressed as a function of n—falls below 1 dB. The discrete entropy corresponding to an average LSD of 1 dB can then be obtained using Eqs. (4.14) and (4.15) together with linear interpolation as follows. Let $H(\mathbf{Y})\big|_{n_1}$ and $H(\mathbf{Y})\big|_{n_2}$ be the discrete entropy values at the stopping resolution and the immediately preceding resolution, respectively; i.e.,

$$H(\mathbf{Y})\big|_{n_1} \triangleq H(\mathbf{Y}) \quad \text{at } n_1 = \min_{n\in\mathbb{Z}}(|\mathcal{I}|) \text{ s.t. } \overline{d}_{\text{LSD}}(n) \le 1\,\text{dB},\ |\mathcal{I}| = 2^n, \tag{4.16}$$

and

$$H(\mathbf{Y})\big|_{n_2} \triangleq H(\mathbf{Y}) \quad \text{at } n_2 = \max_{n\in\mathbb{Z}}(|\mathcal{I}|) \text{ s.t. } \overline{d}_{\text{LSD}}(n) > 1\,\text{dB},\ |\mathcal{I}| = 2^n. \tag{4.17}$$

Then, $H(\mathbf{Y})\big|_{\overline{d}_{\text{LSD}}=1\text{dB}}$ can be estimated as

$$H(\mathbf{Y})\big|_{\overline{d}_{\text{LSD}}=1\text{dB}} \approxeq \frac{1-b}{a}, \tag{4.18}$$

where

$$a = \frac{\overline{d}_{\mathrm{LSD}}(n_1) - \overline{d}_{\mathrm{LSD}}(n_2)}{H(\mathbf{Y})\big|_{n_1} - H(\mathbf{Y})\big|_{n_2}} \quad \text{and} \quad b = \overline{d}_{\mathrm{LSD}}(n_1) - aH(\mathbf{Y})\big|_{n_1} = \overline{d}_{\mathrm{LSD}}(n_2) - aH(\mathbf{Y})\big|_{n_2}. \tag{4.19}$$

By employing VQ as such, we exploit its advantages over SQ—namely those of space filling, shape, and memory [146]. Our approach consequently results in discrete entropy estimates more realistic[85] and superior to those of [109]. Quantization error is higher with SQ than for VQ for the same bit rate, resulting in SQ-based entropy estimates for highband feature vectors that are inaccurately higher than their true values. This, in turn, results in highband certainty estimates that are lower than their true values. More importantly, in contrast to the indirect approach of [109] where the estimation of the discrete highband entropy from differential entropy through SQ requires a direct equivalence between quantization mean-square error and LSD (making this approach only applicable to cepstral parameters), our approach for estimating discrete entropies directly from the quantized highband space makes no assumptions about the relation between the two types of distances. As long as LSD can be calculated for quantized features vectors, our VQ approach can be applied to any form of parameterization.

4.3.3 Calculating the average quantization log-spectral distortion

For an $|\mathcal{I}|$-sized codebook and a distortion measure $d(\mathbf{y}_n, \mathcal{Q}(\mathbf{y}_n))$, the generalized Lloyd algorithm partitions a data set $\mathcal{V} = \{\mathbf{y}_n\}$ into the sets $\{\mathcal{V}_i\}_{i\in\mathcal{I}}$ such that

$$\mathcal{V}_i = \left\{ \mathbf{y}_n \in \mathcal{V}\colon \begin{array}{ll} d(\mathbf{y}_n, \mathbf{c}_i) \le d(\mathbf{y}_n, \mathbf{c}_m), & \forall m, i \in \mathcal{I}, m < i, \\ d(\mathbf{y}_n, \mathbf{c}_i) < d(\mathbf{y}_n, \mathbf{c}_m), & \forall m, i \in \mathcal{I}, m > i \end{array} \right\}, \tag{4.20}$$

with a total quantization distortion given by

$$D = \sum_{i\in\mathcal{I}} \sum_{\mathbf{y}_n\in\mathcal{V}_i} d(\mathbf{y}_n, \mathbf{c}_i). \tag{4.21}$$

Typically, the squared Euclidean distance is the distortion measure used, resulting in optimal codevectors $\{\mathbf{c}_i\}$ estimated simply as the means of the sets $\{\mathcal{V}_i\}$. Codebook training is carried out in iterations until a stopping criterion is satisfied, e.g., a threshold for the

[85] Scalar quantization is rarely used in speech coding.

absolute and/or relative change in total distortion. We apply VQ of the highband feature vectors, $\mathbf{Y}$, using this algorithm with squared Euclidean distance as the distortion measure and with a stopping threshold of 1×10^{-3} for the relative change in total distortion.

Given a highband feature vector VQ codebook trained as above, we calculate quantization distortion in terms of average (MRS) LSD via

$$\overline{d}_{\text{LSD}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{I}} \sum_{\mathbf{y}_n \in \mathcal{V}_i} d_{\text{LSD}}(\mathbf{y}_n, \mathbf{c}_i). \tag{4.22}$$

For LSF-parameterized highband feature vectors, $d_{\text{LSD}}(\mathbf{y}_n, \mathbf{c}_i)$ is calculated using Eq. (3.21). As described in Section 4.3.4 below, we add highband frame log-energy to highband LSF feature vectors—i.e., $\mathbf{Y} = \begin{bmatrix} \mathbf{\Omega}_y \\ \log \mathcal{E}_y \end{bmatrix}$—in order to include cross-band spectral envelope gain correlations in our highband certainty estimates (while also ensuring consistency with the highband parameterization used in our baseline BWE system, where both the shape and gain of highband spectral envelopes are jointly modelled with the narrow band via $\mathcal{G}_{\mathbf{x}\mathbf{\Omega}}$ and $\mathcal{G}_{\mathbf{x}G}$, respectively). With the addition of the highband log-energy parameter, applying Eqs. (4.22) and (3.21) for LSF-based highband feature vectors becomes rather straightforward. LSFs are converted back to LPCs to obtain the analysis $A(z)$ filters as described in Section 3.2.2. The prediction gains necessary to complete the estimation of $d_{\text{LSD}}(\mathbf{y}_n, \mathbf{c}_i)$ per Eq. (3.21)[86] can then be calculated as the scale factors required such the total energy of each frame's LP-based spectrum corresponds exactly to the frame's log-energy parameter [99, Section II.B.3]. The use of frame log-energy in our LSF parameterization—rather than LP gain or dual-mode BWE excitation gain—is motivated in Section 4.3.4 below.

To calculate average quantization LSD for MFCC highband parameterization, we exploit the equivalence of Euclidean distances between MFCC feature vectors and their quantized counterparts to LSD. Since the Type-II DCT of Eq. (4.2) is unitary, it only results in a rotation of the space over which the log mel-scale filter energy vectors—consisting of the elements $\{\log_e \varepsilon_k\}_{k \in \{0,\ldots,K-1\}}$ with K the number of mel-scale filters—are defined. As such, Euclidean distances between MFCC feature vectors are the same as those between the corresponding log mel-scale filter energy vectors; i.e., for an MFCC vector $\mathbf{y}$ and its VQ

[86] See Footnote 68 for the equivalence between prediction gains and the dual-mode BWE system excitation signal gains used in Eq. (3.21).

estimate $\hat{\mathbf{y}} \coloneqq \mathcal{Q}(\mathbf{y})$,

$$d^2_{\text{MFCC}}(\mathbf{y}, \hat{\mathbf{y}}) \triangleq \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \sum_{k=0}^{K-1} |\log_e \varepsilon_k - \log_e \hat{\varepsilon}_k|^2. \tag{4.23}$$

By comparing Eq. (4.23) to the LSD between a short-time FFT power spectrum, $P(\omega)$, and its estimate, $\hat{P}(\omega)$ (rather than the smoothed all-pole model-based LSD of Eq. (3.20)),

$$d^2_{\text{LSD}} = \int_{-\pi}^{\pi} |10\log_{10} P(\omega) - 10\log_{10} \hat{P}(\omega)|^2 \, \frac{d\omega}{2\pi}, \tag{4.24}$$

where d_{LSD} is expressed in decibels, it can be seen that d_{MFCC} is, in fact, a frequency-warped LSD that further takes the critical band structure of speech into account. By considering only the highband frequency range of $f_{\text{Hz}_l} = 4$ to $f_{\text{Hz}_h} = 8\,\text{kHz}$ with K mel-scale filters as shown in Figure 4.1, the exact relation between d_{LSD} and d_{MFCC} can be derived as

$$d^2_{\text{LSD}} = \left(\frac{10}{\log_e 10}\right)^2 \left(\frac{f_{\text{mel}_h} - f_{\text{mel}_l}}{K+1}\right) \frac{1}{f_{\text{mel}_h}} \, d^2_{\text{MFCC}}, \tag{4.25}$$

thereby allowing the estimation of the average quantization LSD—per Eq. (4.22)—for MFCC-parameterized highband feature vectors directly from the Euclidean distances between training vectors (including the 0th cepstral coefficient representing frame log-energy) and their vector-quantized counterparts.

4.3.4 Memoryless highband certainty baselines

In establishing the highband certainty memoryless baseline corresponding to our LSF-based dual-mode BWE system of Chapter 3, we should ensure consistency in terms of the resolution—i.e., dimensionality—used for spectral envelope shape and gain parameterizations in both contexts, i.e., in dual-mode BWE and highband certainty estimation. We showed in Section 1.1.3.1 that band energies play a central role in the identification of many sounds. The importance of this characteristic for BWE was discussed in Section 2.3.4, and was the basis for incorporating frame log-energy into the narrowband feature vectors of our memoryless BWE system, as well as for modelling highband excitation gains through a dedicated $\mathcal{G}_{\mathbf{x}_G}$ GMM. Thus, in contrast to highband envelopes where the shape and gain are modelled in the dual-mode BWE system via separate $\mathcal{G}_{\mathbf{x}\boldsymbol{\Omega}_y}$ and $\mathcal{G}_{\mathbf{x}_G}$ GMMs (with

$\mathrm{Dim}(\mathbf{\Omega}_y) = 6$ and $\mathrm{Dim}(\log \mathcal{E}_y) = 1$), respectively, the LSF-based narrowband feature vector space of both $\mathcal{G}_{\mathbf{X}\mathbf{\Omega}_y}$ and $\mathcal{G}_{\mathbf{X}G}$ represents both the shape and gain of narrowband envelopes conjointly (with $\mathbf{X} = \begin{bmatrix} \mathbf{\Omega}_x \\ \log \mathcal{E}_x \end{bmatrix}$ and $\mathrm{Dim}\left(\begin{bmatrix} \mathbf{\Omega}_x \\ \log \mathcal{E}_x \end{bmatrix}\right) = 9 + 1 = 10$). Accordingly, reusing the same narrowband vectors for LSF-based highband certainty estimation—specifically in the GMM training and numerical evaluation of Eq. (4.7)—ensures consistency with the dual-mode BWE system's narrowband parameterization. To be able to apply MI and discrete highband entropy estimation—via Eqs. (4.7), (4.14), and (4.18)—using a single highband feature vector $\mathbf{Y}$ while also preserving consistency with the high band's representation in dual-mode BWE, we append highband frame log-energy, $\log \mathcal{E}_y$, to the highband LSF feature vector, $\mathbf{\Omega}_y$—i.e., for highband certainty estimation, we represent highband envelopes by $\mathbf{Y} = \begin{bmatrix} \mathbf{\Omega}_y \\ \log \mathcal{E}_y \end{bmatrix}$ with $\mathrm{Dim}\left(\begin{bmatrix} \mathbf{\Omega}_y \\ \log \mathcal{E}_y \end{bmatrix}\right) = 6 + 1 = 7$.

In a similar manner, we model the band-specific spectral envelope shapes and gains for MFCCs using each band's $[c_1, \ldots, c_L]^T$ and c_0 parameters, respectively, with $L = 9$ and 6 for the narrow and high bands, respectively.

In addition to allowing the calculation of average quantization distortion in terms of LSD (thereby allowing the estimation of discrete highband entropies via VQ as described in Sections 4.3.2 and 4.3.3), these parameters, i.e., band log-energies for LSF vectors and c_0 for MFCCs, are more suitable for highband certainty estimation compared to LP and EBP-MGN excitation gains since: (a) LP gains depend on the energy as well as the predictability of the speech signal, rather than on only its energy; and (b) EBP-MGN excitation gains are derived from the 3–4 kHz midband-equalized signal and, thus, involve the inherent error associated with equalization in the 3.4–4 kHz range.

We note that our narrowband dimensionality of 10 coincides with that used in [125] for the evaluation of an LSD lower bound given MI, highband dimensionality, and differential highband entropy. While our overall joint-space dimensionality, $\mathrm{Dim}\left(\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}\right) = 17$, is slightly lower than that used in [109],[87] we employ full-covariance GMMs for MI estimation in Eq. (4.7) as opposed to the diagonal-covariance GMMs of [109]—thereby allowing us to use lower feature vector dimensionalities to obtain MI measurements that are equally or more reliable compared to those obtained using diagonal GMMs at higher dimensionalities. By using full-covariance GMMs for MI estimation, we further ensure correspondence between the highband certainty results of our reference $\mathrm{Dim}(\mathbf{X}, \mathbf{Y}) = (10, 7)$ space and our

[87] In [109], 14 MFCCs (not including c_0) were used to model the narrow band while 4 MFCCs and a highband-to-narrowband log-energy ratio were used as the components of highband feature vectors.

memoryless BWE results of Section 3.5.3.

Table 4.1 shows the memoryless cross-band correlation baseline using highband certainty for both $\mathrm{Dim}(\mathbf{X},\mathbf{Y}) = (10,7)$ LSF and MFCC parameterizations, with the $H(\mathbf{Y})|_{\overline{d}_{\mathrm{LSD}}=1\mathrm{dB}}$ discrete highband entropies obtained as illustrated in Figure 4.2. The GMMs of Eq. (4.7) and the highband VQ codebook of Eq. (4.12) are trained using the TIMIT training set described in Section 3.2.10, while the estimation of highband certainty—via Eqs. (4.7), (4.14), (4.18), (4.22), (3.21), and (4.25)—is performed using the TIMIT core test set.

Table 4.1: Memoryless baseline—information-theoretic measures (in bits) and highband certainty results for the reference $\mathrm{Dim}(\mathbf{X},\mathbf{Y}) = (10,7)$ LSF and MFCC static spaces.

	$\mathrm{Dim}(\mathbf{X},\mathbf{Y})$	$I(\mathbf{X};\mathbf{Y})$	$H(\mathbf{Y})\|_{\overline{d}_{\mathrm{LSD}}=1\mathrm{dB}}$	$C(\mathbf{Y}\|\mathbf{X})$
LSFs	(10,7)	2.24	14.11	15.9%
MFCCs	(10,7)	1.78	8.64	20.5%

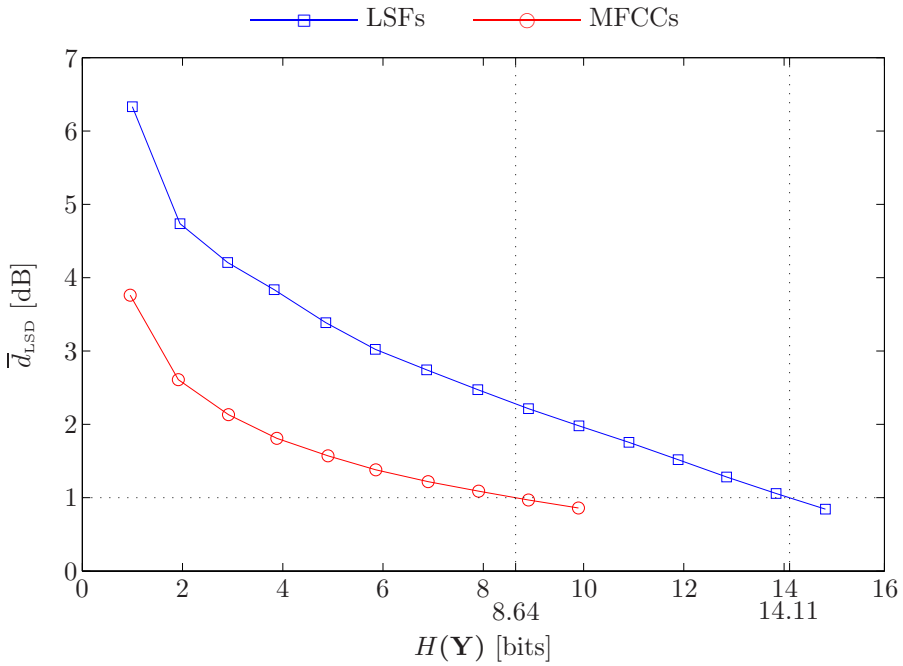


Fig. 4.2: Estimating memoryless discrete highband entropy, $H(\mathbf{Y})$, through VQ, for the memoryless reference dimensionality of $\mathrm{Dim}(\mathbf{Y}) = 7$ (including a highband energy term). Through Eqs. (4.14), (4.18), (4.22), (3.21), and (4.25), quantization error—expressed in $\overline{d}_{\mathrm{LSD}}$—is used to find the discrete entropy values corresponding to the 1 dB spectral transparency threshold of [115] for both LSFs and MFCCs.

As Figure 4.2 shows, the improved class separability of the MFCC-parameterized acoustic space—compared to the LSF-parameterized space—consistently results in lower uncertainty about highband spectral envelopes at any particular spectral distortion level, even at identical LSF and MFCC spectral resolutions, i.e., same dimensionality used for envelope shapes and gains in both types of parameterizations. In other words, MFCC-based highband entropy is always lower than that based on LSFs for the same spectral quality. In fact, Table 4.1 shows that the decrease in $H(\mathbf{Y})|_{\bar{d}_{\text{LSD}}=1\text{dB}}$ highband entropy is sufficiently large, $\approx 39\%$, to result in an overall increase of $\approx 29\%$ in certainty about the highband given the narrowband, $C(\mathbf{Y}|\mathbf{X})$, despite the relatively lower cross-band mutual information of MFCC-parameterized spectral envelopes compared to LSF-parameterized ones.

In Section 4.4 below, we investigate the role of speech dynamics in increasing cross-band correlation by explicitly incorporating memory, in the form of delta features, into frequency bands' feature vector representations. As shown in Section 4.4 and further detailed in Chapter 5, while such delta features increase cross-band correlation by exploiting mutual information on a temporal axis, they represent a dimensionality reduction transform, and, as such, can not be used for the reconstruction of static highband spectral envelopes. Accordingly, the value of frontend-based memory inclusion through delta features varies in relation to the highband dimensionalities of the reference memoryless baseline against which memory inclusion is compared. In particular, *dynamic* feature vectors, comprising both static and delta features, can be viewed as being the result of either:

(a) appending delta features to the existing vectors of static parameters of either or both frequency bands, thereby increasing feature vector dimensionalities, and consequently, increasing the complexities of associated GMMs and/or VQ used for statistical modelling; or

(b) substituting a higher-order subset of the static parameters of existing feature vectors by the delta features of the remaining low-order static parameters, thus preserving feature vector dimensionalities as well as associated GMM and/or VQ complexities.

While appending delta features per Context (a) increases dimensionalities and complexities, the static spectral resolution of the resulting dynamic feature vectors is not adversely affected compared to reference static vectors (since the number of static parameters that can be used for spectral envelope reconstruction is the same with or without memory inclusion). Thus, cross-band correlation can only improve in this context as a result of memory inclusion. In contrast, the substitution of spectral information (consisting in static param-

eters) by temporal information (consisting in delta features) per Context (b) represents a time-frequency information tradeoff. This tradeoff and its effect on BWE is investigated in Chapter 5. For properly assessing the value of frontend-based memory inclusion on highband certainty, however, we establish here two additional memoryless highband certainty baselines with $\text{Dim}(\mathbf{X},\mathbf{Y}) = (10,4)$ and $(5,4)$. The three memoryless baselines—including that established in Table 4.1 with $\text{Dim}(\mathbf{X},\mathbf{Y}) = (10,7)$—will be used as references to investigate memory inclusion in Section 4.4 in the two contexts listed above.

In parameterizing highband envelopes for the $(10,4)$ and $(5,4)$ spaces, we follow the same process used for the $(10,7)$ space. For LSF-based parameters, we use 3 LSFs (rather than 6) for the 4–8 kHz band with one log-energy parameter. For MFCCs, we use $K = 4$ mel-scale filters (rather than 7) resulting in 3 MFCCs representing envelope shape (rather than 6) and one MFCC representing envelope log-energy. Highband spectral envelope shapes are, thus, represented in the $(10,4)$ and $(5,4)$ reference spaces by half the number of parameters used for the $(10,7)$ space. $I(\mathbf{X},\mathbf{Y})$ and $H(\mathbf{Y})\big|_{\overline{d}_{\text{LSD}}=1\text{dB}}$ are estimated as described previously. In Section 4.4, the $C(\mathbf{Y}|\mathbf{X})$ certainty estimates obtained as such for the $(5,4)$ space will represent the references for memory inclusion per Context (a), while those of the $(10,7)$ and $(10,4)$ spaces the references per Context (b).

Since highband envelopes are parameterized using different resolutions in the $(\cdot,4)$ baselines relative to the $(10,7)$ baseline, the $\overline{d}_{\text{LSD}}$ measures—calculated using Eqs. (3.21), (4.23) and (4.25)—used to estimate $H(\mathbf{Y})\big|_{\overline{d}_{\text{LSD}}=1\text{dB}}$ for the $(\cdot,4)$ baselines are not comparable with that of the $(10,7)$ baseline. Estimates for the $(\cdot,4)$ spaces do not account for the lower spectral resolution relative to the $(10,7)$ space. Accordingly, the corresponding $C(\mathbf{Y}|\mathbf{X})$ estimates can not also be directly compared. To account for this difference in spectral resolution when comparing cross-band correlations using different highband dimensionalities (and their potential effect on highband envelopes reconstructed through BWE), we define $\mathbf{Y}_{\text{ref}}$, representing the reference unquantized highband feature vectors used in the calculation of $\overline{d}_{\text{LSD}}$ for highband VQ codebooks, as follows:

LSFs Using the $\text{Dim}(\mathbf{Y}) = 4$ LSF-based highband feature vectors obtained from the TIMIT training set, the highband VQ codebook needed for estimating $H(\mathbf{Y})\big|_{\overline{d}_{\text{LSD}}=1\text{dB}}$ is trained in iterations of increasing codebook cardinality as previously described in Section 4.3.3. To calculate the average quantization LSD via Eq. (3.21) at the end of each iteration, however, we use a parallel set of $\text{Dim}(\mathbf{Y}) = 7$ LSF-based highband

feature vectors, $\mathbf{Y}_{\text{ref}}$, as the reference unquantized vectors, obtained from the TIMIT core test set. Each of these $\mathbf{Y}_{\text{ref}}$ *shadow* vectors is the higher-dimensionality parameterization of the test frame represented by the lower-dimensionality $\mathbf{Y}$ vector. Finally, we use lower-dimensionality $\mathcal{Q}(Y)$ VQ codevectors as the quantized test vectors to be used in Eq. (3.21). As such, we effectively use the low-dimensionality $\text{Dim}(\mathbf{Y}) = 4$ VQ codebook to estimate $\overline{d}_{\text{LSD}}$ (and $H(\mathbf{Y})\big|_{\overline{d}_{\text{LSD}}=1\text{dB}}$, in turn) for the higher-dimensionality $\text{Dim}(\mathbf{Y}_{\text{ref}}) = 7$ reference LSF-based highband feature vectors.

MFCCs In a manner similar to that of LSFs, the MFCC-parameterized highband VQ codebook is trained using the $\text{Dim}(\mathbf{Y}) = 4$ training MFCC highband feature vectors. Rather than use $K = 4$ mel-scale filters as described previously for the $(\cdot, 4)$ MFCC spaces, we use $K = 7$. In effect, this translates into a truncated MFCC highband representation where the truncated higher-order coefficients are assumed to be zero. To estimate $\overline{d}_{\text{LSD}}$ at each VQ training iteration, we perform inverse DCT on: (a) the shadow $\text{Dim}(\mathbf{Y}_{\text{ref}}) = 7$ MFCC vectors (i.e., with no truncation) corresponding to the lower-dimensionality Voronoi; and (b) the truncated $\text{Dim}(\mathbf{Y}) = 4$ VQ codevectors; resulting in mel-scale filter log-energy vectors to be used as the unquantized reference and quantized test vectors, respectively. Since Type-II DCT, as well as its inverse, are unitary transforms, $\overline{d}_{\text{LSD}}$ can be equally calculated through Eq. (4.25) using the squared Euclidean distances between mel-scale log-energies rather than between MFCCs, as shown in Eq. (4.23).

Extending the reference baseline representation as $\text{Dim}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_{\text{ref}})$, Table 4.2 lists, in the top three rows for each parameterization type, the information-theoretic measures estimated for the three memoryless baseline $(10, 7, 7)$, $(10, 4, 4)$, and $(5, 4, 4)$ spaces used in the sequel. The $(10, 4, 7)$, and $(5, 4, 7)$ spaces, used exclusively in this section for the purpose of allowing comparisons at identical spectral resolution, are in the rows below.

Similar to the observations concluded from the results of Table 4.1, Table 4.2 shows that MFCCs outperform LSFs in terms of the relevant information shared between the midband-equalized narrow band and the high band. The cross-band correlation of MFCC-parameterized envelopes is consistently higher than that of LSF-parameterized envelopes, with the relative difference ranging from $\approx 29\%$ for $\text{Dim}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_{\text{ref}}) = (10, 7, 7)$, to $\approx 89\%$ for the $(5, 4, 4)$ baseline. Finally, we also note that increasing the dimensionality of the parameterizations of either or both bands, consistently results in higher mutual information,

Table 4.2: Memoryless highband certainty baselines and RMS-LSD lower bounds—defined in Section 4.3.5 below—at varying $\mathrm{Dim}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_{\mathrm{ref}})$ dimensionalities. $I(\cdot;\cdot)$ and $H(\cdot)$ are in bits, while $\downarrow\overline{d}_{\mathrm{LSD(RMS)}}$ is in dB.

	$\mathrm{Dim}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_{\mathrm{ref}})$	$I(\mathbf{X};\mathbf{Y})$	$H(\mathbf{Y})\big\vert_{\overline{d}_{\mathrm{LSD}}=1\mathrm{dB}}$	$C(\mathbf{Y}\vert\mathbf{X})$	$\downarrow\overline{d}_{\mathrm{LSD(RMS)}}$
LSFs	(10,7,7)	2.24	14.11	15.9%	—
	(10,4,4)	1.68	10.60	15.9%	—
	(5,4,4)	1.55	10.60	14.6%	—
	(10,4,7)	1.68	18.69	9.0%	—
	(5,4,7)	1.55	18.69	8.3%	—
MFCCs	(10,7,7)	1.78	8.64	20.5%	4.62
	(10,4,4)	1.73	5.89	29.3%	4.88
	(5,4,4)	1.62	5.89	27.6%	5.01
	(10,4,7)	1.76	9.07	19.4%	4.68
	(5,4,7)	1.69	9.07	18.7%	4.73

thereby indicating that higher spectral resolutions translate into higher shared information.

4.3.5 Highband certainty as an upper bound on achievable BWE performance

By quantifying cross-band correlation through $C(\mathbf{Y}|\mathbf{X})$, highband certainty given the narrow band at an average highband quantization LSD of 1 dB, we are, in fact, estimating upper bounds on achievable BWE performance. The memoryless $\mathrm{Dim}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_{\mathrm{ref}}) = (10, 7, 7)$ MFCC highband certainty value of $C(\mathbf{Y}|\mathbf{X}) = 20.5\%$, for example, suggests that an average BWE performance of $\overline{d}_{\mathrm{LSD}} = 1\,\mathrm{dB}$ can theoretically be achieved for approximately one fifth of the highband spectra reconstructed through BWE (assuming high-quality highband spectra can be reconstructed from MFCC vectors). This theoretical BWE performance is, however, only an upper bound since:

(a) highband certainty estimation does not account for the spectral envelope distortions inevitably introduced by components in an actual BWE system other than GMMs, e.g., imperfect midband equalization in the 3.4–4 kHz range and the subsequent errors in highband excitation signal generation, and,

(b) the remaining uncertainty about the highband implies an average error of $\overline{d}_{\mathrm{LSD}} > 1\,\mathrm{dB}$ for $1 - C(\mathbf{Y}|\mathbf{X})$ of the reconstructed highband envelopes.

This bounding relation between information-theoretic measures and achievable BWE performance was confirmed in [125]. In particular, given estimates of mutual information and

differential highband entropy, a memoryless lower bound is derived for the $\overline{d}_{\text{LSD(RMS)}}$ distortion of highband spectra that can be reconstructed by BWE, using conventional cepstral parameterization for the high band. By exploiting the correspondence we have shown in Section 4.3.3 between LSD and MFCC distances, we can easily adapt the lower bound of [125] to the case where MFCCs are used to parameterize highband spectral envelopes. This provides us with the means to map highband certainty estimates into concrete BWE performance bounds, and, more importantly, allows us to determine the potential BWE performance value of any highband certainty gains achieved as a result of memory inclusion. To provide the necessary context for our MFCC modification, we describe below the relevant outlines of the $\overline{d}_{\text{LSD(RMS)}}$ lower bound derivation of [125].

The complex cepstrum of a signal is defined as the Fourier transform of the natural logarithm of the signal spectrum. For a power spectrum (magnitude-squared Fourier transform) $P(\omega)$, which is symmetric around $\omega = 0$ and periodic for a sampled data sequence, the Fourier series representation of $\log_e P(\omega)$ is given by $\log_e P(\omega) = \sum_{i=-\infty}^{\infty} c_i e^{-j\omega i}$, where $c_i = c_{-i}$ are real and referred to as the cepstral coefficients of $P(\omega)$. Thus, for a pair of spectra, $P(\omega)$ and its estimate $\hat{P}(\omega)$, Parseval's theorem allows us to rewrite d^2_{LSD} of Eq. (4.24) using cepstral distances;[88] i.e.,

$$d^2_{\text{LSD}} = \left(\frac{10}{\log_e 10}\right)^2 \sum_{i=-\infty}^{\infty} (c_i - \hat{c}_i)^2. \tag{4.26}$$

With the per-frame LSD given by Eq. (4.26), the root-mean-square (RMS) LSD average for a set of speech frames can then be written as

$$\overline{d}_{\text{LSD(RMS)}} = \frac{10\sqrt{2}}{\log_e 10} \sqrt{E\left[\frac{1}{2}(c_0 - \hat{c}_0)^2 + \sum_{i=1}^{\infty} (c_i - \hat{c}_i)^2\right]}. \tag{4.27}$$

[88] Alternatively to this development based on [147, Section 4.5.2], the correspondence represented by Eq. (4.26) between LSD and cepstral distances can also be derived using the complex cepstrum of the signal's LP spectrum—i.e., $H(e^{j\omega})$—as shown in [148] and referenced by [125]. This provides a recursive formula by which cepstral coefficients can be calculated from a set of LPCs, and is used in [125] to parameterize highband envelopes for the evaluation of the derived $\overline{d}_{\text{LSD(RMS)}}$ lower bound for test data.

Then, by using q cepstral coefficients—noting the infinite number of coefficients—to represent highband spectral envelopes; i.e.,

$$y_i = \begin{cases} \frac{1}{\sqrt{2}} c_i, & \text{for } i = 0, \\ c_i, & \text{for } i = 1, \ldots, q-1, \end{cases} \tag{4.28}$$

and writing the BWE system's estimates of highband feature vectors given those of the narrow band as $\hat{\mathbf{y}} = f(\mathbf{x})$, with the estimation error $\mathbf{n} = \mathbf{y} - \hat{\mathbf{y}}$, Eq. (4.27) can be rewritten as

$$\overline{d}_{\text{LSD(RMS)}} \geq \frac{10\sqrt{2}}{\log_e 10} \sqrt{E\left[|\mathbf{n}|^2\right]}. \tag{4.29}$$

Using properties of mutual information and differential entropies, the authors in [125] then proceed to show that

$$E\left[|\mathbf{n}|^2\right] \geq \frac{q}{2\pi e} \exp\left[\frac{2}{q}\Big(h(\mathbf{Y}) - I(\mathbf{X};\mathbf{Y})\Big)\right]; \tag{4.30}$$

a lower bound that is independent of the type of parameterizations used for $\mathbf{X}$ and $\mathbf{Y}$, as well as independent of the BWE method used to achieve the mapping $\hat{\mathbf{y}} = f(\mathbf{x})$. Substituting Eq. (4.30) into Eq. (4.29) results in the memoryless lower bound

$$\overline{d}_{\text{LSD(RMS)}} \geq \frac{10}{\log_e 10} \sqrt{\frac{q}{\pi e}} \exp\left[\frac{1}{q}\Big(h(\mathbf{Y}) - I(\mathbf{X};\mathbf{Y})\Big)\right]. \tag{4.31}$$

To rewrite this lower bound based on MFCC Euclidean distances rather than conventional cepstral coefficient distances, we substitute d^2_{LSD} of Eq. (4.26) above by that of Eq. (4.25) from Section 4.3.3, where d^2_{LSD} is written in terms of d^2_{MFCC} given by Eq. (4.23). Repeating the derivation above with this modification results in

$$\begin{aligned} \overline{d}_{\text{LSD(RMS)}} &\geq \frac{10}{\log_e 10} \sqrt{\frac{q\left(f_{\text{mel}_h} - f_{\text{mel}_l}\right)}{\pi e (K+1) f_{\text{mel}_h}}} \exp\left[\frac{h(\mathbf{Y}) - H(\mathbf{Y})\, C(\mathbf{Y}|\mathbf{X})}{q}\right] \\ &= \frac{10}{\log_e 10} \sqrt{\frac{\text{Dim}(\mathbf{Y}_{\text{ref}})\left(f_{\text{mel}_h} - f_{\text{mel}_l}\right)}{\pi e\Big(\text{Dim}(\mathbf{Y}_{\text{ref}}) + 1\Big) f_{\text{mel}_h}}} \exp\left[\frac{h(\mathbf{Y}_{\text{ref}}) - H(\mathbf{Y}_{\text{ref}})\, C(\mathbf{Y}|\mathbf{X})}{\text{Dim}(\mathbf{Y}_{\text{ref}})}\right], \end{aligned} \tag{4.32}$$

where we have rewritten $\mathbf{Y}$ and $K = q = \text{Dim}(\mathbf{Y})$ as the more explicit $\mathbf{Y}_{\text{ref}}$ and $\text{Dim}(\mathbf{Y}_{\text{ref}})$, respectively, as well as reorganized the exponential's arguments in Eq. (4.31)—dropping

the evaluation-point qualifier in $H(\mathbf{Y}_{\text{ref}})\big|_{\overline{d}_{\text{LSD}}=1\text{dB}}$ to simplify notation—such that the lower bound is an explicit function of highband certainty rather than mutual information. By aligning notations as such with our earlier $\text{Dim}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_{\text{ref}})$ acoustic space notation, these modifications facilitate evaluation of the lower bound for the reference MFCC memoryless spaces of Table 4.2 as well as for the memory-inclusive spaces discussed in Section 4.4.3.2 below, particularly for cases where $\text{Dim}(\mathbf{Y}) \neq \text{Dim}(\mathbf{Y}_{\text{ref}})$. For these cases, the effect of using lower spectral resolutions for highband envelopes on the certainty estimated for the high band, is thus accounted for by using the reference $\mathbf{Y}_{\text{ref}}$ as the argument for $h(\cdot)$, $H(\cdot)$, and $\text{Dim}(\cdot)$, while continuing to use the lower-dimensionality $\mathbf{Y}$ in $C(\mathbf{Y}|\mathbf{X})$ since it already takes the higher reference dimensionality into account. It is also worth noting that the MFCC-based lower bound of Eq. (4.32) is, in fact, tighter than that of Eq. (4.31); in contrast to Eq. (4.29) where the inequality results from the truncation of the non-negative highband cepstral coefficients to q, Eq. (4.23) involves no MFCC truncation, and hence, the equality holds (without the $\sqrt{2}$ term).

Table 4.2 shows our estimates of the lower bound of Eq. (4.32), denoted by $\downarrow\overline{d}_{\text{LSD(RMS)}}$, obtained for the dimensionalities and the estimates of the information-theoretic measures for the memoryless MFCC-based spaces of Table 4.2, and using $h(\mathbf{Y}_{\text{ref}})$ estimates obtained by stochastic integration per Eq. (4.8). Despite identical dimensionalities, the $\downarrow\overline{d}_{\text{LSD(RMS)}}$ estimate for the MFCC $(10, 7, 7)$ baseline, in particular, is not comparable to the LSF-based dual-mode BWE $\overline{d}_{\text{LSD(RMS)}}$ result of Table 3.1 due to the difference in parameterization. Nevertheless, it indicates that we can not reduce GMM-based BWE distortion to less than $\overline{d}_{\text{LSD(RMS)}} = 4.62\,\text{dB}$ when using MFCCs with $\text{Dim}(\mathbf{X}, \mathbf{Y}) = (10, 7)$. More importantly, the $\downarrow\overline{d}_{\text{LSD(RMS)}}$ estimates for the MFCC $(10, 7, 7)$ and $(5, 5, 4)$ baselines, provide the memoryless references to the memory-inclusive bounds in Section 4.4.3.2 below, allowing us to gain important insights into the potential effect of memory inclusion on practical BWE performance.

To conclude, we note that since any highband certainty gains achieved through memory inclusion correspond to upper bounds on BWE performance gains, an ideal BWE system is that which can translate the measured certainty gains into matching BWE performance improvements—i.e., where reductions in measured $\overline{d}_{\text{LSD(RMS)}}$ performance are equivalent to the decreases in $\downarrow\overline{d}_{\text{LSD(RMS)}}$. Thus, highband certainty estimates provide the reference point against which the optimality (or lack of it) of any BWE system can be determined, as well as provide the theoretically ideal frame of reference for performance using which competing

BWE systems can be compared against each other. Indeed, in this chapter as well as in Chapter 5, we use highband certainty estimates as the basis for evaluating the role of memory inclusion in BWE, in general, as well as for comparing different BWE systems where the extent to which memory is incorporated is varied.

4.4 Memory Inclusion through Delta Features

Despite the well-known dynamic and temporal properties of speech discussed in Section 1.2, and referred to herein simply as speech memory, investigating the theoretical basis underlying the assumption that exploiting speech memory will automatically improve BWE performance has received little, if any, attention. Indeed, to our knowledge, all works showing the superiority of BWE with such memory inclusion make no attempt to determine how competent these memory-inclusive techniques actually are in making use of the temporal information available in the narrow band to improve highband reconstruction. Our objective in this section is, thus, to quantify the role of memory in improving cross-band correlations as represented by $C(\mathbf{Y}|\mathbf{X})$, certainty about the high band given the narrow band. To achieve this objective, one can follow either of two approaches:

(a) Assume temporal statistical dependence between the conventional static feature vectors representing each of the two frequency bands, consequently modelling cross-band correlations through the joint *pdf*s of sequences of narrowband and highband feature vectors. Highband certainty can then be derived accordingly. Since this approach applies no dimensionality reduction, it would fully preserve all spectral information present in the sequences of static frames. The resulting increase in complexity, however, would, in fact, be prohibitive for practical purposes. To demonstrate, the mutual information for *only first-order sequences* would be given by:

$$I(\mathbf{X}_t, \mathbf{X}_{t-1}; \mathbf{Y}_t, \mathbf{Y}_{t-1}) = \int_{\Omega_{\mathbf{y}_{t-1}}} \int_{\Omega_{\mathbf{y}_t}} \int_{\Omega_{\mathbf{x}_{t-1}}} \int_{\Omega_{\mathbf{x}_t}} p_{\mathbf{X}_t \mathbf{X}_{t-1} \mathbf{Y}_t \mathbf{Y}_{t-1}}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{y}_t, \mathbf{y}_{t-1}) \cdot \log_2 \left(\frac{p_{\mathbf{X}_t \mathbf{X}_{t-1} \mathbf{Y}_t \mathbf{Y}_{t-1}}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{y}_t, \mathbf{y}_{t-1})}{p_{\mathbf{X}_t \mathbf{X}_{t-1}}(\mathbf{x}_t, \mathbf{x}_{t-1}) p_{\mathbf{Y}_t \mathbf{Y}_{t-1}}(\mathbf{y}_t, \mathbf{y}_{t-1})} \right) d\mathbf{x}_t \, d\mathbf{x}_{t-1} \, \mathbf{y}_t \, d\mathbf{y}_{t-1}, \quad (4.33)$$

which shows that, to estimate MI merely for the first-order case, we need to double the dimensionalities of our GMMs (noting our reference memoryless dimensionality

of Dim $\left(\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix}\right)$ = 17), which in turn requires a multiple-fold increase in training data and complexity. To model higher-order dependence, dimensionality and complexity will further multiply, making this technique impractical.

(b) Transform sequences of conventional static feature vectors into dynamic lower-dimensionality vectors in which speech dynamics are directly embedded in addition to the static envelope parameters. Through such dimensionality-reducing transforms, the new memory-inclusive vectors can be assumed to be statistically independent across time, thereby allowing highband certainty estimation in the manner described above—in Section 4.3—for conventional static features, while also allowing the inclusion of temporal information from sequences of varying lengths. As described below, delta features represent a linear form of such dimensionality-reducing transforms.

As the second approach is clearly better in terms of both efficiency and the extent of memory that can be modelled, we select it for our memory-inclusive highband certainty estimation.

4.4.1 Delta features

Rather than indirectly capture speech temporal information through first-order HMM state transition probabilities or increasing the amount of overlap of speech frames, we include memory directly in spectral envelope parametrization in the form of delta coefficients appended to the static LSF/MFCC feature vectors. Initially formulated by Furui [136] in the context of speaker verification, delta coefficients (or features) are obtained from static vectors by a first-order regression (time-derivative) implemented through linearly weighted differences between neighbouring static vectors. A consequence of the time derivative is that the differences weights used in delta coefficient calculations increase in proportion to the distance (in frames) between the two static vectors whose difference is being evaluated. This translates acoustically into emphasizing long-term spectral transitions over fine short-term differences. Indeed, since immediately successive frames show only minor differences between their static features, the underlying long-term trajectory of parameter variation with time can be more accurately and easily identified as the time separation between the

static frames involved increases.[89] Delta coefficients are calculated via:

$$\boldsymbol{\delta}_t = \frac{\sum_{l=1}^{L} l \cdot (\mathbf{s}_{t+l} - \mathbf{s}_{t-l})}{2\sum_{l=1}^{L} l^2}, \tag{4.34}$$

where $\boldsymbol{\delta}_t$ is the delta coefficient vector corresponding to the signal frame at time t computed in terms of the corresponding static feature vectors $\{\mathbf{s}_{t+l}\}_{l\in[-L,L]}$, with L specifying the number of neighbouring static frames (on each side of the tth frame) to consider. Eq. (4.34) shows that the delta coefficient transfer function is a non-causal linear time-invariant function, with the impulse response illustrated in Figure 4.3 for $L = 5$, for example. As mentioned in Section 4.3.4 above and described in more detail below, the calculated delta coefficients can either replace part of the static LSF/MFCC coefficients, or be appended to them, to produce the dynamic (static+delta) spaces.

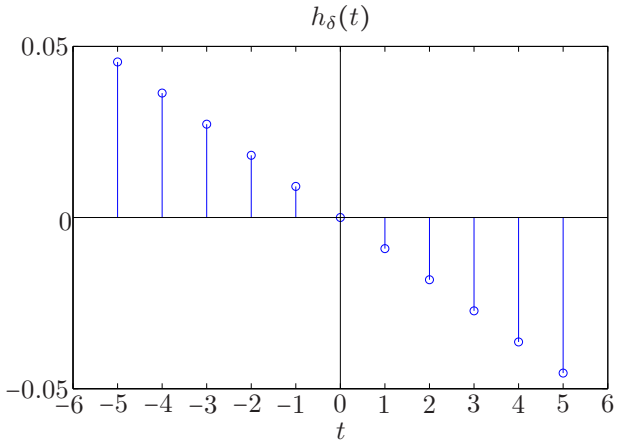


Fig. 4.3: Impulse response of delta coefficient transfer function for $L = 5$. See Eq. (4.34).

4.4.2 Comparing delta features to other dimensionality reduction transforms

Since they employ a many-to-one transform on sequences of time-indexed frames, delta features are a lossy form of compression that effectively compacts memory into a relatively small number of features. As such, delta features can be viewed as a special case of

[89] In his experiments in [136] on the effects of the speech segment length used in delta coefficient calculation on speaker verification error rates, Furui found that the minimum error rate was achieved at a length of 170 ms.

dimensionality-reducing transforms where the higher-dimensional source supervectors are simply an extension—along the temporal axis—of low-dimensional information in a space of memoryless spectrally-derived axes (where the axes are not necessarily orthonormal). When applied to sequences of time-indexed static narrowband—and optionally highband—feature vectors, other dimensionality-reducing transforms can similarly be viewed as *memory inclusion transforms*. Most notable of such transforms are linear discriminant analysis (LDA) [71, Chapter 5] and the Karhunen-Loève transform (KLT)—also referred to as principal component analysis (PCA) [71, Section 3.8]. LDA attempts to obtain a feature vector with maximal compactness by reducing the dimensionality of the source supervectors while retaining their discriminating power as much as possible. Such a reduction is performed by means of a linear transformation optimized during offline training by maximizing the class separability—the ratio of between-class to within-class scatter—of the target vectors (projections of the source supervectors unto a lower-dimensional hyperplane). The KLT, on the other hand, reduces source supervectors to a set of uncorrelated features. Worthy of note in this context is the work of [149], where several transforms, including differential transforms (delta and higher-order delta), LDA, and the KLT, were compared in the context of memory inclusion—by viewing such transforms as the application of a temporal matrix transform on a matrix comprised of stacked time-indexed cepstral vectors—for improving speech recognition performance. Results in [149] show that recognition performance is generally improved by memory inclusion. In particular, while the best performance is achieved using the KLT, the most notable among the results of [149] is that representing cepstra by delta features alone gives 13.5% higher digit recognition accuracy than achieved by static cepstra, thus confirming the ability of delta features to capture relevant information in speech memory.

As described in Sections 3.3.3 and 3.5.1, cross-band covariances play an important role for BWE since it is that cross-band correlation information that ultimately enables BWE. Thus, the superior class discrimination of LDA should intuitively improve the ability of GMM statistical modelling to discriminate non-overlapping frequency content based on temporal information. In other words, since BWE assumes that narrowband and highband content share the same underlying classes, performing LDA on temporal-based supervectors of either band (or both) would improve discrimination among such classes using the information in speech memory mutual to the two frequency bands. The KLT, on the other hand, would only diagonalize within-band covariances, i.e., it does not necessarily improve

cross-band covariances. However, as described in Section 4.2.2 and confirmed by the memoryless results of Section 4.3.4 for MFCCs, decorrelation through DCT generally results in improved highband certainty. Since the KLT completely decorrelates source features, it can be expected to result in cross-band correlation increases equal to or greater than those of MFCCs.[90]

By virtue of being dimensionality reduction transforms, however, LDA and the KLT are similar to delta features in that they can not be used for the reconstruction of highband spectral envelopes. Since any BWE system requires a conventional static representation of highband spectra, frontend-based memory inclusion through non-invertible transforms, in general, imposes a time-frequency information tradeoff for fixed overall narrowband and highband dimensionalities. This tradeoff, briefly described for delta features in Section 4.3.4, is investigated later in the thesis in more detail. We conclude that this tradeoff requires optimizing the allocation of available dimensionalities among memoryless spectral features and temporal ones, such that estimated highband certainties are maximized—taking into account the effect of static parameter dimensionality when estimating highband entropies as demonstrated for the memoryless $\mathrm{Dim}(\mathbf{X},\mathbf{Y}) = (\cdot, 4)$ baselines in Section 4.3.4. This optimization and the application of delta features for incorporating memory into BWE are the subject of Section 5.3. We note here, however, that LDA and the KLT suffer the same information tradeoff imposed by delta features. Moreover, since the estimation of transform matrices for both LDA and the KLT—involving eigen-value decomposition—is computationally more complex than the rather simple calculation of delta features, we focus on the latter for our investigation of frontend-based memory inclusion.

4.4.3 Effect of memory inclusion on highband certainty

Corresponding to the random narrowband and highband static feature vectors represented by $\mathbf{X}$ and $\mathbf{Y}$, respectively, let $\mathbf{\Delta_X}$ and $\mathbf{\Delta_Y}$ represent their random delta coefficient vector counterparts, with $\overset{\triangle}{\mathbf{X}} \triangleq \begin{bmatrix}\mathbf{X}\\ \mathbf{\Delta_X}\end{bmatrix}$ and $\overset{\triangle}{\mathbf{Y}} \triangleq \begin{bmatrix}\mathbf{Y}\\ \mathbf{\Delta_Y}\end{bmatrix}$ further representing their joint—or dynamic, i.e., static+delta—versions.

[90] In the context of the similarities between the KLT and the DCT, it is worth noting that, as shown in [149], the KLT basis functions are, in fact, almost identical to those of the DCT when estimated for feature vectors consisting of sequences of the same cepstral coefficient.

4.4.3.1 The Contexts and Scenarios of incorporating delta features

As described in Section 4.3.4, incorporating delta features into existing static feature vectors can be performed in one of two contexts;

Context A *appending* delta features to the existing vectors of static parameters of either or both frequency bands, or,

Context S *substituting* a higher-order subset of the static parameters of existing feature vectors by the delta features of the remaining low-order static parameters, preceded by recalculating the low-order static parameters if needed (e.g., when using lower-order LSFs).

Simultaneously with, but independently of these two contexts, memory inclusion through delta features can also be performed in either of the two following scenarios:

Scenario 1 Incorporating memory into the representation of *one* of the two bands only. We consider narrowband-only memory inclusion, with the reasonable assumption that—since both bands share the same underlying acoustic classes, and hence, also share their dynamic properties—the effects of single-band memory inclusion on cross-band correlation are independent of the particular band into which memory is incorporated. With narrowband-only memory inclusion, the change in certainty about the high band is given by

$$\begin{aligned}
\Delta_{C_1} &\triangleq C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}}) - C(\mathbf{Y}|\mathbf{X}) \\
&= \frac{1}{H(\mathbf{Y})\big|_{\overline{d}_{\mathrm{LSD}}=1\mathrm{dB}}} \left[I(\overset{\triangle}{\mathbf{X}};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}) \right] \\
&= \frac{1}{H(\mathbf{Y})\big|_{\overline{d}_{\mathrm{LSD}}=1\mathrm{dB}}} \left[I(\mathbf{X},\boldsymbol{\Delta}_{\mathbf{X}};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}) \right] \\
&= \frac{1}{H(\mathbf{Y})\big|_{\overline{d}_{\mathrm{LSD}}=1\mathrm{dB}}} \Delta_{I_1}; \qquad (4.35)
\end{aligned}$$

i.e., Δ_{C_1} depends only on Δ_{I_1}—the change in MI—as the static highband representation, and consequently its entropy, are unchanged. Assuming static narrowband dimensionality is preserved with memory inclusion (Context A above), the relations between the information content of the $\mathbf{X}$, $\mathbf{Y}$ and $\boldsymbol{\Delta}_{\mathbf{X}}$ feature vector spaces can be

easily visualized through the *Venn-like* diagram of Figure 4.4,[91] using which, Δ_{I_1} can be written as

$$\Delta_{I_1} \equiv (\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_4) - (\mathcal{R}_1 \cup \mathcal{R}_2) = \mathcal{R}_4, \tag{4.36}$$

representing the additional gain in MI between the two bands as a result of exploiting narrowband temporal information.

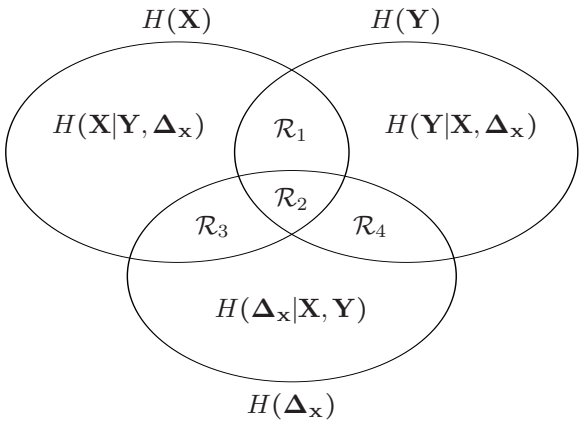


Fig. 4.4: Venn-like diagram representing the relations between the information content of the $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{\Delta_x}$ spaces.

Scenario 2 Incorporating memory into the representation of *both* bands, with the result that the entropy of the now-dynamic highband representation is changed. In this scenario, the change in certainty about the high band is given by

$$\begin{aligned} \Delta_{C_2} &\triangleq C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}}) - C(\mathbf{Y}|\mathbf{X}) \\ &= \frac{I(\overset{\triangle}{\mathbf{X}};\overset{\triangle}{\mathbf{Y}})}{H(\overset{\triangle}{\mathbf{Y}})|_{\overline{d}_{\text{LSD}}=1\text{dB}}} - \frac{I(\mathbf{X};\mathbf{Y})}{H(\mathbf{Y})|_{\overline{d}_{\text{LSD}}=1\text{dB}}}. \end{aligned} \tag{4.37}$$

Thus, in contrast to Scenario 1, the change in highband certainty, i.e., Δ_{C_2}, is now more complex as it does not depend only on the change in mutual information between representations of both bands, but also on the change in the entropy of the high band itself. Without further information about the interactions between the

[91] Although Figures 4.4 and 5.3 illustrate relationships in a manner resembling that of Venn diagrams, the relationships illustrated are those between information-theoretic quantities, rather than between sets as is the case with formal Venn diagrams. Hence, in contrast to the conventional *Venn* diagram nomenclature used in [64, Figure 2.2], for example, we refer to our illustrations of Figures 4.4 and 5.3 as *Venn-like*.

$\overset{\triangle}{\mathbf{X}}$ and $\overset{\triangle}{\mathbf{Y}}$ spaces, a general visualization similar to that of Figure 4.4 is, thus, more complex.[92] As described below, this change in highband entropy is closely tied to the aforementioned time-frequency information tradeoff.

Combining these contexts and scenarios results in four possible cases for memory inclusion where delta features:

Case A-1 are *appended* to existing static features in only *one* band—the narrow band,
Case A-2 are *appended* to existing static features in the *two* bands,
Case S-1 *substitute* higher-order static features in *one* band—the narrow band, or,
Case S-2 *substitute* higher-order static features in the *two* bands.

Extending our earlier $\text{Dim}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_{\text{ref}})$ representation of acoustic spaces introduced in Section 4.3.4 to $\text{Dim}(\mathbf{X}, \mathbf{\Delta_X}, \mathbf{Y}, \mathbf{\Delta_Y}, \mathbf{Y}_{\text{ref}})$—with the three memoryless baseline spaces now represented by $(10, 0, 7, 0, 7)$, $(10, 0, 4, 0, 4)$, and $(5, 0, 4, 0, 4)$—and representing the process of memory inclusion by $\xrightarrow{\triangle}$, we investigate the effect of memory inclusion on highband certainty in these four cases as outlined in Table 4.3 below.

Table 4.3: Breakdown of approaches to memory inclusion through delta features by context (incorporating memory by *appending* to, or *substituting*, existing static features) and scenario (incorporating memory into *one* or *two* bands), using $\text{Dim}(\mathbf{X}, \mathbf{\Delta_X}, \mathbf{Y}, \mathbf{\Delta_Y}, \mathbf{Y}_{\text{ref}})$ to represent acoustic space dimensionalities.

	Context A	Context S
Scenario 1	A-1: $(5,0,4,0,4) \xrightarrow{\triangle} (5,5,4,0,4)$	S-1: $(10,0,4,0,4) \xrightarrow{\triangle} (5,5,4,0,4)$
Scenario 2	A-2: $(5,0,4,0,4) \xrightarrow{\triangle} (5,5,4,4,4)$	S-2: $(10,0,7,0,7) \xrightarrow{\triangle} (5,5,4,4,7)$

We note that, due their importance, we always include log-energy parameters in both bands' static and delta representations for all spaces represented in Table 4.3. For example, the narrowband feature vectors $\overset{\triangle}{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \mathbf{\Delta_X} \end{bmatrix}$ of the $(5, 5, 4, 4, 4)$ LSF space consist of the static features $\mathbf{X} = \begin{bmatrix} \mathbf{\Omega}_x \\ \log \mathcal{E}_x \end{bmatrix}$ with $\text{Dim}\left(\begin{bmatrix} \mathbf{\Omega}_x \\ \log \mathcal{E}_x \end{bmatrix}\right) = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$, as well as the delta features $\mathbf{\Delta_X} = \begin{bmatrix} \delta(\mathbf{\Omega}_x) \\ \delta(\log \mathcal{E}_x) \end{bmatrix}$, similarly with $\text{Dim}\left(\begin{bmatrix} \delta(\mathbf{\Omega}_x) \\ \delta(\log \mathcal{E}_x) \end{bmatrix}\right) = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$, resulting in an overall dimensionality of 10 for the dynamic narrowband representation—the same dimensionality of the static representation. As such, in substituting static feature vectors by dynamic ones under Context S, only the

[92] Based on the findings that follow in this section in addition to certain assumptions discussed in Section 5.3.3, a simplified Venn-like diagram for Scenario 2 is presented in Figure 5.3.

resolution of the static spectral envelope *shape* representation is affected by substitution—resulting in the time-frequency information tradeoff.[93]

4.4.3.2 Implementation, results, and analysis

To estimate the information mutual to the representations of both bands in the two scenarios of memory inclusion; i.e., $I(\overset{\triangle}{\mathbf{X}};\mathbf{Y})$ and $I(\overset{\triangle}{\mathbf{X}};\overset{\triangle}{\mathbf{Y}})$, we follow the numerical integration approach described in Section 4.3.1, adapting Eq. (4.7) to the now-dynamic narrowband features vectors $\overset{\triangle}{\mathbf{X}} = \begin{bmatrix}\mathbf{X}\\ \mathbf{\Delta_X}\end{bmatrix}$—as well as to the dynamic highband vectors $\overset{\triangle}{\mathbf{Y}} = \begin{bmatrix}\mathbf{Y}\\ \mathbf{\Delta_Y}\end{bmatrix}$ in the case of Scenario 2—by replacing the static GMMs of Eq. (4.7) with their dynamic counterparts (e.g., replacing $\mathcal{G}_{\mathbf{XY}}$, $\mathcal{G}_{\mathbf{X}}$, and $\mathcal{G}_{\mathbf{Y}}$, by $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}$, $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}}$, and $\mathcal{G}_{\overset{\triangle}{\mathbf{Y}}}$, respectively, in Scenario 2).

Similarly, in order to estimate $H(\overset{\triangle}{\mathbf{Y}})\big|_{\overline{d}_{\text{LSD}}=1\text{dB}}$, the self-information in the dynamic $\overset{\triangle}{\mathbf{Y}}$ highband representation at the $\overline{d}_{\text{LSD}} = 1\,\text{dB}$ threshold of average quantization distortion, we adapt our VQ-based estimation of discrete highband entropy—described in Sections 4.3.2 and 4.3.3—by: (a) performing VQ on the now-dynamic representation of the high band, $\overset{\triangle}{\mathbf{Y}}$, while (b) estimating average $\overline{d}_{\text{LSD}}$ quantization error after each cardinality iteration of VQ codebook training using—for all cases of Table 4.3 except Case S-2—only the static $\mathbf{Y}$ subvectors of the unquantized $\begin{bmatrix}\mathbf{Y}\\ \mathbf{\Delta_Y}\end{bmatrix}$ testing data as the reference vectors, with the corresponding static $\mathcal{Q}(\mathbf{Y})$ subvectors of the quantized $\mathcal{Q}(\overset{\triangle}{\mathbf{Y}}) \triangleq \mathcal{Q}\left(\begin{bmatrix}\mathbf{Y}\\ \mathbf{\Delta_Y}\end{bmatrix}\right)$ codevectors as the LSD test vectors. In the case of memory inclusion per Context S and Scenario 2, i.e., case S-2: $(10,0,7,0,7) \xrightarrow{\triangle} (5,5,4,4,7)$, we account for the decrease in reference static highband dimensionality when calculating $\overline{d}_{\text{LSD}}$ in the manner described in Section 4.3.4 for both LSFs and MFCCs. In particular, we calculate $\overline{d}_{\text{LSD}}$ using higher-dimensionality shadow LSF and MFCC $\mathbf{Y}_{\text{ref}}$ vectors—with $\text{Dim}(\mathbf{Y}_{\text{ref}}) = 7$—as LSD reference vectors, rather than the $\mathbf{Y}$ subvectors of $\overset{\triangle}{\mathbf{Y}}$, where $\text{Dim}(\mathbf{Y}) = 4$.

Estimating mutual information and highband certainty as such allows us to quantify the effect of memory inclusion using delta features on highband certainty, as a function

[93] As discussed in detail in Section 5.3 in the context of BWE with frontend-based memory inclusion, we impose a *fixed-dimensionality constraint* in reference to the maximum joint-band dimensionality modelled by the dual-mode BWE system's GMMs. As such, while the total joint-band dimensionality for Case S-2 in Table 4.3 above increases from 17 for the static $(10,0,7,0,7)$ space to 18 for the dynamic $(5,5,4,4,7)$ space, the maximum joint-band dimensionality of the corresponding dual-mode BWE feature vectors is, in fact, fixed at 16 when considering only the parameters corresponding to the dual-mode BWE system's GMM with maximum dimensionality—i.e., $\mathcal{G}_{\mathbf{X}\mathbf{\Omega}_y}$ for the LSF-based dual-mode BWE system, for example, where $\text{Dim}\left(\begin{bmatrix}\mathbf{X}\\ \mathbf{\Omega}_y\end{bmatrix}\right) = \begin{bmatrix}10\\ 6\end{bmatrix}$.

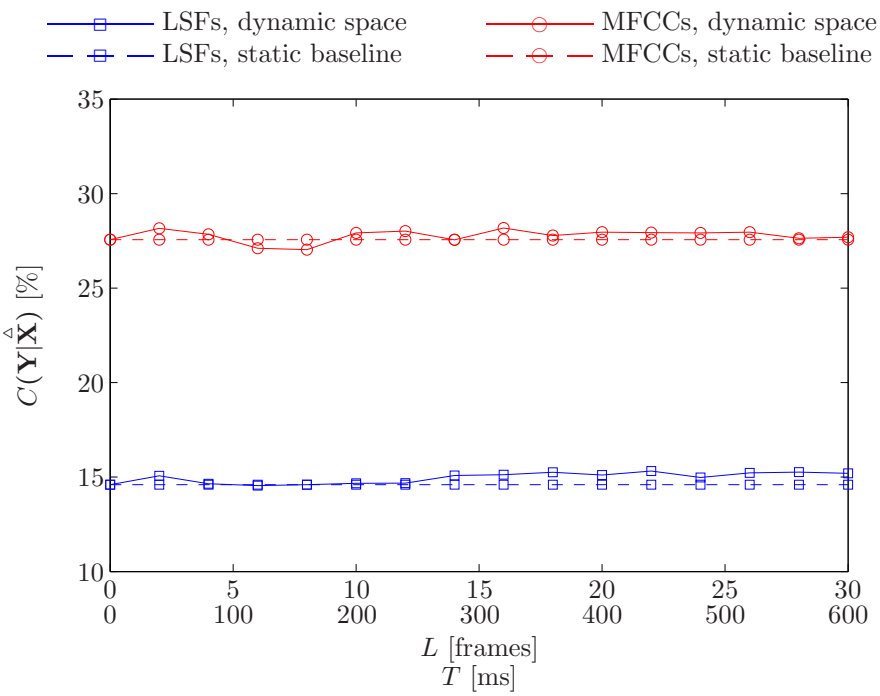
of the amount of speech memory incorporated into the dynamic frequency band representations. Figures 4.5 and 4.6 illustrate this effect, with the inclusion of memory applied per Contexts A and S of Table 4.3, respectively.[94] Highband certainty is measured as a function of L, the number of neighbouring static frames—on each side of a static signal frame—used to calculate the delta features.[95] Given our 20 ms frame length and 10 ms frame advance described in Section 3.2.8, the amount of the non-causal—i.e., two-sided—memory represented by delta features is given by $T = 10 \cdot 2 \cdot L$ ms. As the effect of memory inclusion on cross-band correlation is measured in Case S-2 relative to our memoryless $\mathrm{Dim}(\mathbf{X}, \mathbf{Y}) = (10, 7)$ baseline, which, in turn, corresponds to our dual-mode BWE baseline of Chapter 3, the information-theoretic results of such inclusion are particularly relevant to the implementation of memory-inclusive BWE in the following chapter. Thus, the effect of memory inclusion per Case S-2 on mutual information, $I(\overset{\triangle}{\mathbf{X}}; \overset{\triangle}{\mathbf{Y}})$, and highband entropy, $H(\overset{\triangle}{\mathbf{Y}})\big|_{\overline{d}_{\mathrm{LSD}}=1\mathrm{dB}}$, are examined in more detail through Figure 4.7. From the results of Figures 4.5, 4.6, and 4.7, we observe the following:

A. Narrowband spectral dynamics provide minimal additional information about the static properties of highband spectra

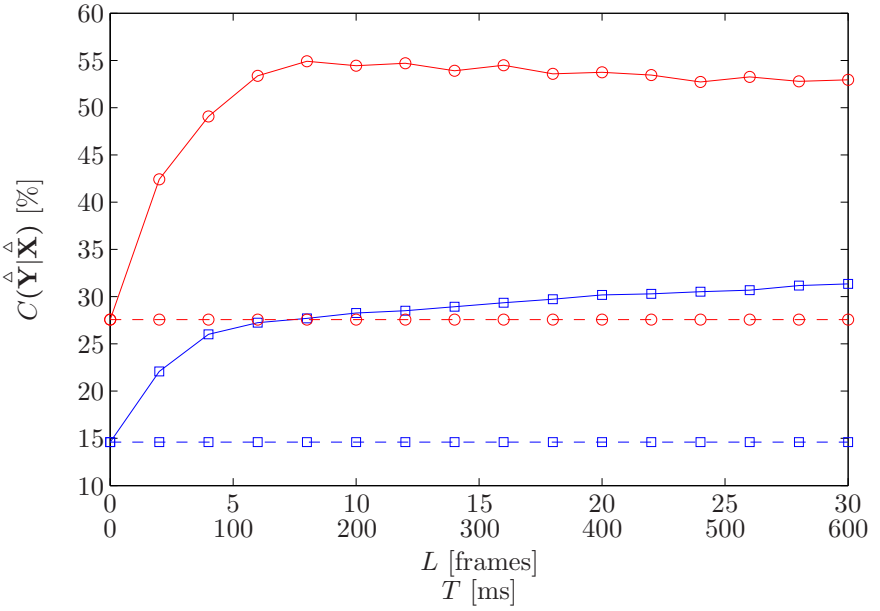
Memory inclusion per Scenario 1 can only result in modest highband certainty gains—given by Δ_{C_1} of Eq. (4.35). As shown for Case A-1 in Figure 4.5(a), extending static narrowband features, $\mathbf{X}$, by appending their $\boldsymbol{\Delta}_{\mathbf{X}}$ delta counterparts—thereby preserving the existing information mutual to the static representations of both bands—results, at best, in a mere $\frac{\Delta_{C_1}}{C(\mathbf{Y}|\mathbf{X})} \simeq 2.3\%$ relative increase in static highband certainty when using MFCCs (at $T = 320$ ms), and $\sim 5.0\%$ when using LSFs (at $T = 440$ ms). In other words, narrowband spectral dynamics and temporal information provide minimal additional information about the static properties of highband spectra, $\mathbf{Y}$. For fixed-dimensionality constraints, Figure 4.6(a), depicting Case S-1, shows that, exploiting the available narrowband dimensionality to improve the spectral representation of static narrowband spectra—rather than to include long-term narrowband information—provides, in fact, more information about the high band; i.e., narrowband delta features contain less information about the static high band than do the higher-order narrowband static features they replace.

Since knowledge about speech properties suggests that the correlation between the static

[94] See Footnote 77 regarding GMM-derived results.
[95] See Eq. (4.34).

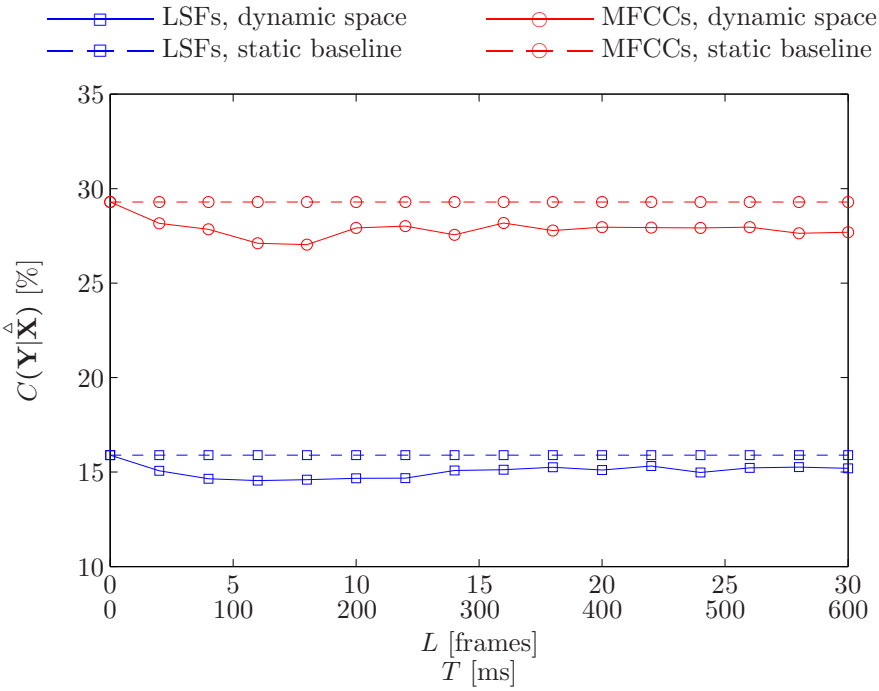


(a) Case A-1: $(5,0,4,0,4) \xrightarrow{\triangle} (5,5,4,0,4)$

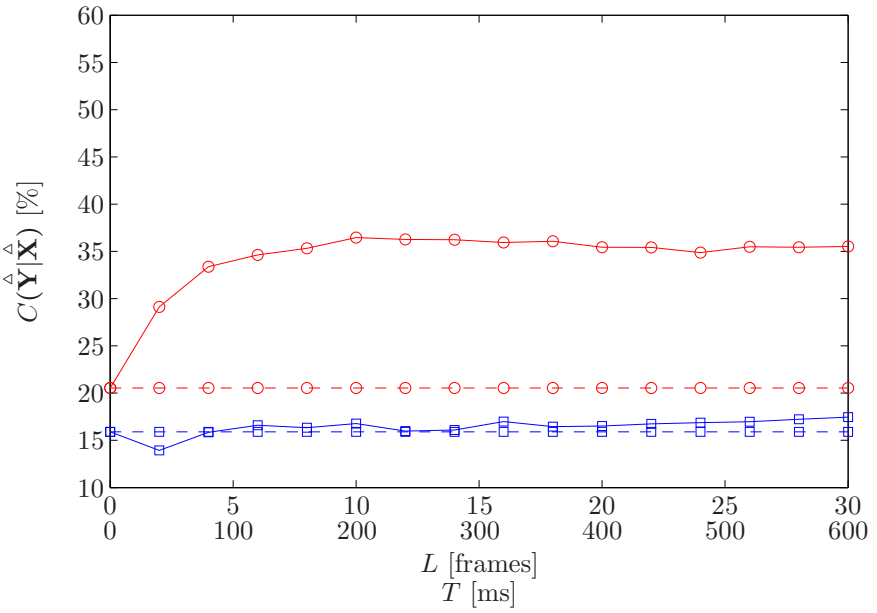


(b) Case A-2: $(5,0,4,0,4) \xrightarrow{\triangle} (5,5,4,4,4)$

Fig. 4.5: Effect of memory inclusion per Context A where LSF- and MFCC-based static features vectors are extended by appending delta features. Highband certainty is illustrated as a function of L, the number of neighbouring static frames—on each side of a static signal frame—used to calculate the delta features, per Eq. (4.34), with T representing the total two-sided memory.

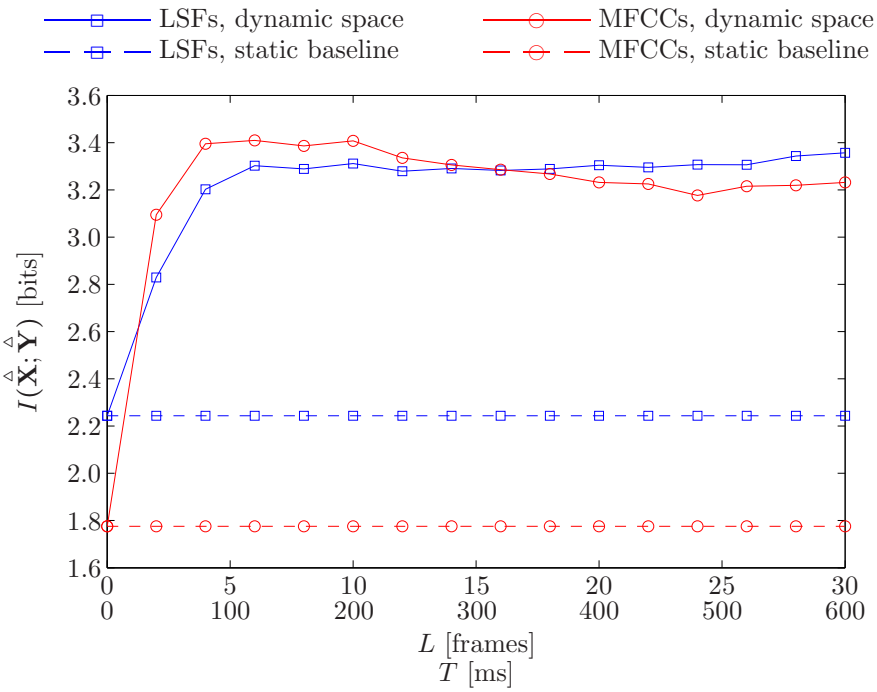


(a) Case S-1: $(10, 0, 4, 0, 4) \xrightarrow{\triangle} (5, 5, 4, 0, 4)$

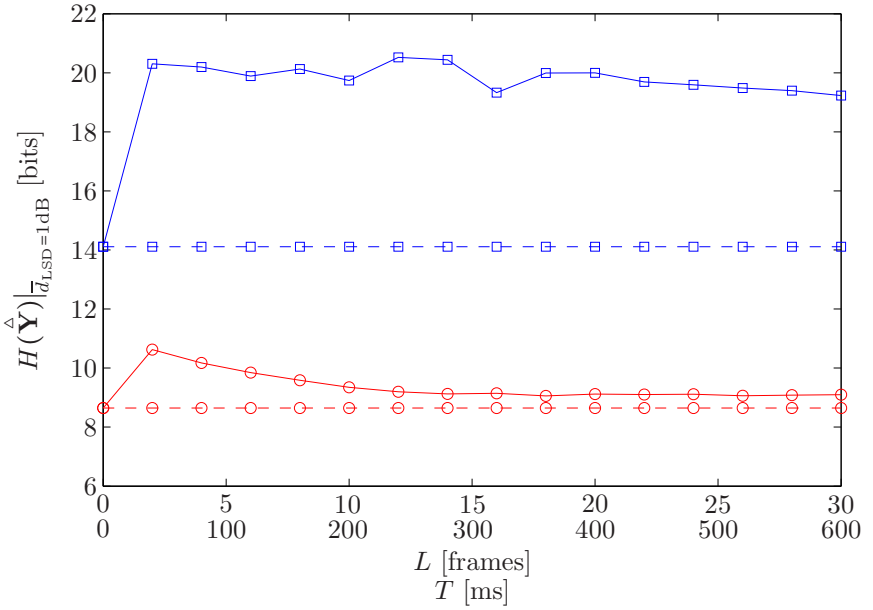


(b) Case S-2: $(10, 0, 7, 0, 7) \xrightarrow{\triangle} (5, 5, 4, 4, 7)$

Fig. 4.6: Effect of memory inclusion per Context S where a high-order subset of the LSFs and MFCCs of the static vectors are replaced by the delta features of the remaining lower-order static features. The lower-order static features of the dynamic vectors are recalculated only in the case of LSFs (lower-order static MFCCs are obtained by simply truncating the high-order static vectors).



(a) Effect of memory inclusion via $(10, 0, 7, 0, 7) \xrightarrow{\triangle} (5, 5, 4, 4, 7)$ on mutual information



(b) Effect of memory inclusion via $(10, 0, 7, 0, 7) \xrightarrow{\triangle} (5, 5, 4, 4, 7)$ on highband entropy

Fig. 4.7: Effect of memory inclusion using delta features per Context S and Scenario 2—i.e., Case S-2: $(10, 0, 7, 0, 7) \xrightarrow{\triangle} (5, 5, 4, 4, 7)$—on mutual information and discrete highband entropy.

properties of one band and the dynamic properties of the other should be higher than that shown above, these findings are revisited in Section 5.3.3 in the context of BWE.

B. Spectral dynamics in both bands are highly correlated

In contrast to Scenario 1, including memory in both narrowband and highband spectral envelope representations can result in significant highband certainty gains—given by Δ_{C_2} of Eq. (4.37)—as shown by the results of Figures 4.5(b) and 4.6(b) pertaining to Scenario 2. In particular, comparing Figure 4.5(b) to Figure 4.5(a), depicting Cases A-2 and A-1, respectively, shows that, while the narrowband dynamics represented by $\mathbf{\Delta_X}$ contain minimal additional information about static highband spectra $\mathbf{Y}$, the spectral dynamics in both bands are highly correlated, translating into considerable gains in certainty about the dynamic $\overset{\triangle}{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{\Delta_Y} \end{bmatrix}$ representation of the high band given a similarly dynamic $\overset{\triangle}{\mathbf{X}}$ narrowband representation. At the cost of increased dimensionality and complexity (but same static spectral fidelity), Figure 4.5(b), depicting Case A-2, shows relative certainty gains—$\frac{\Delta_{C_2}}{C(\mathbf{Y}|\mathbf{X})}$—reaching 99% for MFCCs (at T = 180 ms), and 115% for LSFs (at T = 600 ms), indicating that the information shared by the $\mathbf{\Delta_X}$ and $\mathbf{\Delta_Y}$ delta representations can be equal to or higher than that shared by the static $\mathbf{X}$ and $\mathbf{Y}$ representations. These certainty gains correspond to ~ 20% and ~ 38% relative decreases in the uncertainty remaining in the high band for LSFs and MFCCs, respectively.

C. Effects of time-frequency information tradeoff

More relevant to our memoryless BWE baseline, the effect of the aforementioned time-frequency information tradeoff for the high band manifests in the lower certainty results for Case S-2 relative to those of Case A-2, depicted in Figures 4.6(b) and 4.5(b), respectively, and is further detailed in Figure 4.7. In contrast to memory inclusion via Context A—represented by Figure 4.5(b)—where static feature dimensionality is preserved, replacing higher-order static highband features by delta ones per Context S—represented by Figure 4.6(b)—adversely affects highband certainty. This follows as a result of using fewer features to represent static highband spectra, thereby increasing the average quantization LSD associated with VQ when using the original high-order static feature vectors as the reference unquantized spectra. The accompanying increase in highband entropy—much smaller with MFCCs than with LSFs as described below—is illustrated in Figure 4.7(b). While reducing the number of features used to represent static highband spectra also results

in lower information about these spectra, this decrease in information is compensated by the inclusion of temporal information instead via delta features. In fact, as Figure 4.7(a) shows, this time-frequency information substitution results in significant relative mutual information gains, reaching 92% for MFCCs in particular. Based on the results of Case S-2 in Figure 4.6(b), the net effect of the time-frequency information tradeoff on highband certainty is a maximum increase of $\frac{\Delta_{C_2}}{C(\mathbf{Y}|\mathbf{X})} \simeq 78\%$ for MFCCs (at $T = 200\,\text{ms}$), but only a modest $\sim 10\%$ for LSFs (at $T = 600\,\text{ms}$), relative to the $\text{Dim}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_{\text{ref}}) = (10, 7, 7)$ memoryless baseline. These certainty gains correspond to a $\sim 20\%$ relative decrease in the uncertainty remaining in the high band for MFCCs, but only a mere $\sim 2\%$ for LSFs.

D. Effects of memory inclusion on the MFCC-based RMS-LSD lower bound

To assess the significance of the highband certainty gains shown above for Scenario 2 in terms of potential improvements in BWE performance, we make use of the MFCC-based RMS-LSD lower bound, $\downarrow\overline{d}_{\text{LSD(RMS)}}$, of Eq. (4.32). For memory inclusion per Scenario 2, we use static MFCC vectors with $\text{Dim}(\mathbf{Y}_{\text{ref}}) = 4$ and 7 for Cases A-2 and S-2, respectively, as the reference highband representation against which $\overline{d}_{\text{LSD(RMS)}}$ is calculated. Simultaneously, however, we use the dynamic $\overset{\triangle}{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{\Delta_Y} \end{bmatrix}$ MFCC vectors—with $\text{Dim}(\mathbf{Y}, \mathbf{\Delta_Y}) = (4, 4)$—to represent the high band for the purpose of cross-band correlation modelling. From the findings discussed above, it is clear that, for both Cases A-2 and S-2, the certainty $C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}})$ about the dynamic highband MFCC vectors is considerably higher than the certainty $C(\mathbf{Y}_{\text{ref}}|\overset{\triangle}{\mathbf{X}})$ about the reference static vectors given the same dynamic narrowband representation. To elaborate, let $\mathbf{X}_{\text{ref}}$ represent the reference static narrowband MFCC representation such that $\text{Dim}(\mathbf{X}_{\text{ref}}, \mathbf{Y}_{\text{ref}}) = (5, 4)$ and $(10, 7)$ for Contexts A and S, respectively. Then, for Context S where $\text{Dim}(\mathbf{X}, \mathbf{\Delta_X}, \mathbf{X}_{\text{ref}}, \mathbf{Y}, \mathbf{\Delta_Y}, \mathbf{Y}_{\text{ref}}) = (5, 5, 10, 4, 4, 7)$, the findings of memory inclusion per Case S-2 in Figure 4.6(b) showed that $C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}}) \gg C(\mathbf{Y}_{\text{ref}}|\mathbf{X}_{\text{ref}})$. In addition, Case S-1 in Figure 4.6(a) also showed that, for the same dimensionalities, $C(\mathbf{Y}|\mathbf{X}_{\text{ref}}) \geq C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}})$, and hence, $C(\mathbf{Y}_{\text{ref}}|\mathbf{X}_{\text{ref}}) \geq C(\mathbf{Y}_{\text{ref}}|\overset{\triangle}{\mathbf{X}})$. Thus, by combining the inequalities from both cases, $C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}}) \gg C(\mathbf{Y}_{\text{ref}}|\overset{\triangle}{\mathbf{X}})$. In a similar manner, Cases A-2 and A-1 of Figure 4.5 show that $C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}}) \gg C(\mathbf{Y}_{\text{ref}}|\mathbf{X}_{\text{ref}})$ and $C(\mathbf{Y}_{\text{ref}}|\mathbf{X}_{\text{ref}}) \approxeq C(\mathbf{Y}_{\text{ref}}|\overset{\triangle}{\mathbf{X}})$, respectively, and hence, $C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}}) \gg C(\mathbf{Y}_{\text{ref}}|\overset{\triangle}{\mathbf{X}})$. These observations show that a BWE system that estimates highband content in the dynamic $\overset{\triangle}{\mathbf{Y}}$ form—given a dynamic $\overset{\triangle}{\mathbf{X}}$ narrowband representation—is considered optimal if it fully translates the certainty $C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}})$ about the dynamic high band into certainty $C(\mathbf{Y}_{\text{ref}}|\overset{\triangle}{\mathbf{X}})$ about the reference static representation. Ac-

cordingly, for memory inclusion per Scenario 2, the $\downarrow\overline{d}_{\mathrm{LSD(RMS)}}$ lower bound of Eq. (4.32) can then be rewritten in terms of $C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}})$ rather than $C(\mathbf{Y}|\mathbf{X})$, while preserving other variables as functions of $\mathbf{Y}_{\mathrm{ref}}$; i.e.,

$$\overline{d}_{\mathrm{LSD(RMS)}} \geq \frac{10}{\log_e 10}\sqrt{\frac{\mathrm{Dim}(\mathbf{Y}_{\mathrm{ref}})\left(f_{\mathrm{mel}_h}-f_{\mathrm{mel}_l}\right)}{\pi e\left(\mathrm{Dim}(\mathbf{Y}_{\mathrm{ref}})+1\right)f_{\mathrm{mel}_h}}}\exp\left[\frac{h(\mathbf{Y}_{\mathrm{ref}})-H(\mathbf{Y}_{\mathrm{ref}})\,C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}})}{\mathrm{Dim}(\mathbf{Y}_{\mathrm{ref}})}\right]. \quad (4.38)$$

Figure 4.8 illustrates the effect of memory inclusion per Scenario 2 on potential BWE performance as represented by $\downarrow\overline{d}_{\mathrm{LSD(RMS)}}$. For Case A-2, the higher highband certainty reduces $\downarrow\overline{d}_{\mathrm{LSD(RMS)}}$ by up to 1.66 dB (at T = 160 ms), while for Case S-2, the decrease reaches 0.82 dB (at T = 200 ms). As expected, potential $\overline{d}_{\mathrm{LSD(RMS)}}$ performance improvements are greater when memory inclusion does not involve a reduction in static feature dimensionality.

To put these potential BWE performance gains into perspective, we compare them to the measured BWE performance gains reported in two earlier works representative of the effects of improved cross-band correlation modelling. Noting that reductions in the RMS average of LSD are, in general, only slightly higher than the corresponding MRS average reductions, the earlier version of the dual-mode BWE system in [54] achieves an average highband MRS-LSD reduction of 0.96 dB in the 3.5–7 kHz by employing GMM-based statistical mapping rather than VQ codebook mapping as in [69].[96] In the more complex speaker-independent HMM-based approach of [39], an average RMS-LSD reduction of ~ 1.1 dB is achieved in the 3.4–7 kHz range by using 64 HMM states—with 16 Gaussian components in the narrowband GMM of each state—rather than 2 states.[97] By performing HMM-based BWE in a speaker-dependent manner rather than speaker-independently, an additional average RMS-LSD advantage of ~ 1 dB is shown in [39]. From these examples, we can conclude that, with reference highband dimensionality being preserved (as in Case A-2), the potential benefit of exploiting cross-band dynamic information on BWE per-

[96] The dual-model BWE system of [54] uses 14 LSFs and a pitch gain parameter to represent narrowband envelopes while using 10 LSFs to represent those of the high band.

[97] The HMM-based BWE system of [39] uses 15-dimensional composite narrowband feature vectors (composed of 10 auto-correlation coefficients, zero-crossing rate, a time-smoothed estimate of frame energy, gradient index, local kurtosis, and spectral centroid) and 9-dimensional highband cepstral coefficient vectors. As described in Section 2.3.3.4, this approach divides highband vectors into several speech classes using VQ, with each class mapped to a dedicated HMM state consisting of a GMM trained on the corresponding narrowband vectors. Each HMM state has an associated probability and a first-order transition probability that are estimated from training wideband sequences.

formance is greater than that resulting from any of those individual cross-band correlation modelling improvements discussed above. With the time-frequency information tradeoff associated with reducing static highband dimensionality in favour of incorporating dynamic information (as in Case S-2), the potential gains of exploiting memory become lower but, nevertheless, remain comparable to those improvements of the techniques discussed above.

To conclude, we note that, in addition to the fact the BWE highband frequency range in the works cited above ($\subseteq$ 3.4–7 kHz) is, in fact, smaller than that used in our modelling of the high band (4–8 kHz), the performance gains shown in our investigation (as well as all certainty figures discussed in this chapter) are quite dependent on the dimensionalities we chose for the static and dynamic representations. For a particular total dimensionality constraint, it is unknown whether the apportionments we chose for the allocation of available dimensionality among static and delta features are optimal; i.e., the optimal allocation for maximum certainty about the high band may very well be different than those discussed in this chapter. This is, partly, the subject of Chapter 5.

E. Certainty gains due to memory inclusion saturate at the syllabic rate

By examining the certainty results of Figures 4.5(b) and 4.6(b) (depicting Cases A-2 and S-2, respectively), as well as those of the $\overline{d}_{\mathrm{LSD(RMS)}}$ lower bound in Figure 4.8, as a function of the temporal span used for memory inclusion, we observe that highband certainty reaches saturation for windows of, roughly, 200 ms. Incorporating spans of memory beyond this range has little (in the case of LSFs) or no effect (in the case of MFCCs) on certainty. Based on the duration properties of various sound units discussed in Section 1.2, we can conclude that this duration corresponds to multi-phones (phonemes with left and right contexts). Thus, the effect of memory inclusion is greatest when inter- or multi-phone (syllabic) temporal information is employed to better identify individual phonemes (by exploiting intra-syllable inter-phoneme dependencies). Indeed, as noted earlier in Section 1.2, the mapping from phones to individual phonemes is likely accomplished by analyzing dynamic acoustic patterns—both spectral and temporal—over sections of speech corresponding roughly to syllables [10, Section 5.4.2]. Acoustic-only memory inclusion provides no further information about inter-syllable dependencies. This is expected since such dependencies are determined by language-specific prosody and semantic construction rather than by phonetic speech signal characteristics. These conclusions coincide with the findings of [128] in which modulation spectra show that the acoustic information content of speech is

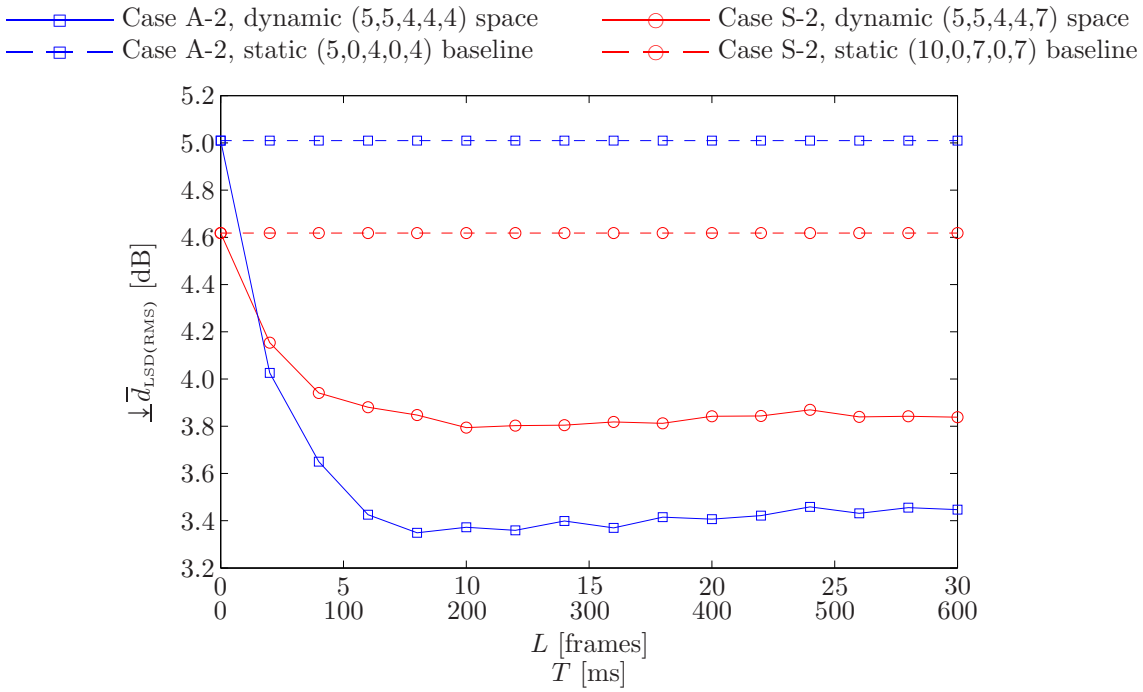


Fig. 4.8: Effect of memory inclusion using delta features per Scenario 2—i.e., per both Cases A-2: $(5,0,4,0,4) \xrightarrow{\Delta} (5,5,4,4,4)$, and S-2: $(10,0,7,0,7) \xrightarrow{\Delta} (5,5,4,4,7)$—on the MFCC-based BWE RMS-LSD lower bound, $\underline{\overline{d}}_{\mathrm{LSD(RMS)}}$, with the assumption that the certainty $C(\overset{\Delta}{\mathbf{Y}}|\overset{\Delta}{\mathbf{X}})$ about the dynamic highband MFCC vectors with $\mathrm{Dim}(\mathbf{Y}, \mathbf{\Delta}_{\mathbf{Y}}) = (4,4)$ can be fully translated into certainty $C(\mathbf{Y}_{\mathrm{ref}}|\overset{\Delta}{\mathbf{X}})$ about static vectors with $\mathrm{Dim}(\mathbf{Y}_{\mathrm{ref}}) = 4$ and 7, for Cases A-2 and S-2, respectively.

highest at the syllabic rate of 4–5 Hz, corresponding to 200–250 ms of memory.

F. The superiority of MFCCs over LSFs

Comparing the certainty results using MFCCs to those of LSFs—for the static baselines of Table 4.2 as well as for the dynamic spaces of Figures 4.5 and 4.6—shows that MFCCs consistently outperform LSFs in capturing cross-band information relevant to the high band. The superiority of MFCCs for memory inclusion per Scenario 2 and Context S, in particular, is most relevant to the implementation of memory-inclusive BWE in the sequel. While Figure 4.7(a) shows that the mutual information between dynamic MFCC-based representations of both bands is slightly superior to that of dynamic LSF-based representations only up to ~ 300 ms of memory inclusion, Figure 4.7(b) shows a consistent difference between dynamic MFCC- and LSF-based highband entropies. The considerably

lower MFCC-based entropy—resulting in the overall superior MFCC-based certainty performance of Figure 4.6(b)—is attributed to: (a) the improved class separability associated with using MFCCs, described in Section 4.2.2, and (b) the lower spectral error associated with vector-quantizing truncated MFCC vectors where $\mathrm{Dim}(\mathbf{Y}, \mathbf{Y}_{\mathrm{ref}}) = (4, 7)$, compared to that associated with vector-quantizing lower-order LSF vectors. In particular, performing IDCT on a truncated highband MFCC vector with $\mathrm{Dim}(\mathbf{Y}) = 4$ but based on $K = 7$ mel-scale filters still generates a highband spectral representation with higher resolution—albeit with error due to the truncation—than a spectrum estimated from a highband LSF vector with $\mathrm{Dim}(\mathbf{Y}) = 4$. This observation is confirmed by comparing the increases in highband entropy estimates for the $\mathrm{Dim}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_{\mathrm{ref}}) = (\cdot, 4, 7)$ baselines in Table 4.2 relative to the estimates for the $(10, 7, 7)$ baseline, for both LSFs and MFCCs; while the relative increase in highband entropy is $\approx 32\%$ for LSFs, it is only $\approx 5\%$ for MFCCs. This advantage for MFCCs makes them less susceptible than LSFs to the adverse effects associated with the time-frequency information tradeoff; while potential relative certainty gains decrease from $\frac{\Delta_{C_2}}{C(\mathbf{Y}|\mathbf{X})} \simeq 115\%$ to $\sim 10\%$ for LSFs when including delta features per Case S-2 rather than A-2, corresponding gains for MFCCs decrease from $\sim 99\%$ to only $\sim 78\%$.

For convenience of reference, Table 4.4 summarizes the highband certainty and BWE performance upper bound figures mentioned above for Scenario 2.

Table 4.4: Effect of memory inclusion per Scenario 2—where delta features are incorporated into the parameterizations of both bands—on highband certainty and RMS-LSD lower bound. Representing acoustic space dimensionalities by $\mathrm{Dim}(\mathbf{X}, \mathbf{\Delta_X}, \mathbf{Y}, \mathbf{\Delta_Y}, \mathbf{Y}_{\mathrm{ref}})$, Cases A-2 and S-2 of Scenario 2 are given by A-2: $(5, 0, 4, 0, 4) \xrightarrow{\Delta} (5, 5, 4, 4, 4)$ and S-2: $(10, 0, 7, 0, 7) \xrightarrow{\Delta} (5, 5, 4, 4, 7)$.

	Case	$\max\left[C(\overset{\Delta}{\mathbf{Y}}\|\overset{\Delta}{\mathbf{X}})\right]$	$\max\left[\frac{\Delta_{C_2}}{C(\mathbf{Y}\|\mathbf{X})}\right]$	$\min\left[\underline{\downarrow}\overline{d}_{\mathrm{LSD(RMS)}}\right]$	$\max\left[\Delta_{\underline{\downarrow}\overline{d}_{\mathrm{LSD(RMS)}}}\right]$
LSFs	A-2	31.3%	114.7%	—	—
	S-2	17.5%	9.8%	—	—
MFCCs	A-2	54.9%	99.2%	3.35 dB	1.66 dB
	S-2	36.5%	77.5%	3.79 dB	0.82 dB

4.5 Summary and Conclusions

Although the spectral dynamics and temporal properties of speech—referred to herein as speech memory—account for a significant portion of its information content, these prop-

erties have mostly been discarded by BWE schemes employing memoryless mapping. A few approaches exploiting speech memory have, however, been proposed to improve BWE performance. Nonetheless, the effect of memory on cross-band correlation—the basis underlying BWE—has not been adequately quantified in the context of BWE.

In this chapter, we presented a detailed investigation of the effect of memory inclusion on cross-band correlation, quantifying such correlation using information-theoretic measures combined with conventional GMM-based statistical modelling and vector quantization, with speech dynamics modelled through delta features. Simple yet efficient, delta features provided a means with which to represent memory extending up to 600 ms. The results of our investigation, while providing upper bounds on achievable BWE performance with the inclusion of memory, also led to several observations, most notable of which are that:

(a) the spectral dynamics of both bands are highly correlated, to the extent that—as summarized in Table 4.4—dynamic representations based on MFCCs can increase certainty about the high band given the narrow band up to 55% at the cost of doubling feature vector dimensionalities, and up to 37% with no increase in dimensionality, potentially reducing BWE RMS-LSD distortion by 1.66 and 0.82 dB, respectively;

(b) the effects of acoustic-only memory inclusion in increasing cross-band correlation saturate at, roughly, the syllabic rate of 5 Hz, and;

(c) MFCC parameters outperform LSFs in retaining mutual cross-band information content relevant to the reconstruction of the high band.

An optimal memory-inclusive BWE system is that which can translate these highband certainty and performance upper bound figures into matching improvements in reconstructed signal quality. In practice, highband content is reconstructed on a frame-by-frame basis. Thus, we can conclude from the observations above that, in order for a BWE system to efficiently make use of the considerable cross-band correlation between dynamic representations, such a system must be able to convert—partially at least—information about spectral envelope dynamics extending up to 200 ms into higher-quality static highband envelope extensions. Secondly, notwithstanding the advantages of LSFs over MFCCs, namely quantization noise robustness and straightforward speech reconstruction, we also conclude that MFCC-based BWE is potentially superior, particularly under constraints of fixed dimensionality where memory inclusion may require replacing high-order static feature vectors

by dynamic vectors consisting of delta features in addition to lower-order static features; a substitution resulting in a time-frequency information tradeoff.

Chapter 5

BWE with Memory Inclusion

5.1 Introduction

We showed in Chapter 4 that, for similar dimensionalities, parameterizing spectral envelopes using MFCCs results in consistently higher certainties about the high band than those obtained using LSFs. As shown in Tables 4.2 and 4.4, these higher MFCC-based certainties can, in fact, reach more than twice those based on LSFs, in both memoryless and memory-inclusive conditions. Thus, we concluded that, notwithstanding the LSF advantage of straightforward speech reconstruction, MFCC-based BWE is inherently better.

Accordingly, we begin this chapter by presenting our work—introduced in [150]—to exploit the superiority of MFCCs over LSFs in terms of cross-band correlation by using MFCCs to represent both narrowband and highband spectral envelopes for BWE. To reconstruct highband speech from MFCCs (obtained by GMM statistical estimation from input narrowband MFCCs), we employ *high-resolution* inverse DCT (IDCT) similar to that of [151] resulting in fine mel-scale log-energies, from which the linear power spectra can be recreated. The high-resolution IDCT effectively uses cosine functions to interpolate between mel-scale filterbank log-energies to reconstruct the spectrum with finer detail (otherwise lost due to the mel-scale filterbank binning). As in [152], we use a source-filter model to reconstruct speech from the estimated power spectra through inverse Fourier transform to obtain auto-correlation coefficients, to which the Levinson-Durbin recursion can then be applied. The LPCs thus obtained represent the synthesis filter parameters which, when combined with the enhanced EBP-MGN excitation signal of Section 3.2.4, can then be used to reconstruct highband speech through a modified MFCC-based dual-mode BWE system.

This MFCC inversion scheme thus eliminates the requirements of pitch estimation and voicing decisions of the more complex sinusoidal model-based techniques (employed in the field of distributed speech recognition) as in, e.g., [151, 153]. Using the BWE performance measures described in Section 3.4, we show that our proposed MFCC-based dual-mode technique achieves high-quality highband speech reconstruction equivalent to that of the LSF-based dual-mode system, thereby allowing us to potentially exploit the superior certainty advantages of memory inclusion associated with MFCCs in comparison to LSFs—the certainty advantages summarized in Table 4.4.

With our dual-mode MFCC-based BWE system in place, we then turn our focus to translating the considerable highband certainty gains obtained and quantified in Chapter 4 into practical and measurable BWE performance improvements. Achieved by accounting for the cross-band correlation advantages of speech memory—i.e., the temporal and dynamic spectral properties in long-term speech—through explicit delta feature inclusion into the parameterization of the narrow and high bands, we present two distinct approaches to empirically realize these theoretical certainty gains.

In the first approach, we attempt to replicate the information-theoretic effects of incorporating memory exclusively into the parameterization frontend, by integrating delta features directly into our dual-model MFCC-based BWE system. Notwithstanding the algorithmic delay entailed by the run-time calculation of non-causal delta features, the primary advantage of such frontend-based memory inclusion is the minimal modifications it requires for integration into the memoryless BWE baseline system. By re-examining the information-theoretic findings of Section 4.4.3 in the context of practical real-time BWE operating on frame-by-frame basis, we gain a better understanding of the mutual information relationships among the static and delta feature vector spaces of both bands—with $\mathbf{X}$ and $\mathbf{Y}$ representing the static narrowband and highband feature vectors spaces, respectively, and $\mathbf{\Delta_X}$ and $\mathbf{\Delta_Y}$ representing their delta counterparts. This, in turn, leads us to investigate the effect of exploiting the information in $\mathbf{\Delta_Y}$ jointly with that in $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{\Delta_X}$, in improving our GMM-based modelling of the underlying time-frequency classes shared between the two bands. Indeed, despite the fact that, in practice, only the static $\mathbf{Y}$ features can be used for the LP-based reconstruction of highband spectral envelopes since delta features are non-invertible, results show a slightly improved performance for the static highband certainty, $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}})$, when $\mathbf{\Delta_Y}$ features are included in joint-band GMM training.

By imposing a fixed-dimensionality constraint on the dual-mode system's joint-band

GMM with maximum dimensionality in order to guarantee the fairness as well as the practicality of any BWE performance improvements achieved, the inclusion of delta features in lieu of static features results in the time-frequency information tradeoff discussed in Chapter 4. Consequently, we perform empirical optimization over the frontend-based memory inclusion's dimensionalities in order to determine the optimal allocation of available dimensions among the static and delta features in both bands, such that static highband certainty is maximized. Using the optimal joint-band dimensionalities obtained as such, we then proceed to integrate frontend-based memory into our MFCC-based BWE system, followed by performance evaluations using the objective measures described in Section 3.4. Results show that the BWE performance improvements achieved as a result of frontend-based memory inclusion generally coincide with the information-theoretic certainty results. This, however, includes the modest nature of the attained performance improvements—ranging from 2.1% relative $\overline{Q}_{\text{PESQ}}$ improvement to 15.9% for $\overline{d}^{*}_{\text{IS}}$—since only a portion of the considerable gains previously shown in Section 4.4.3.2 for the dynamic highband certainty, $C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}})$, was achieved for $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}})$ using the GMM modelling improvement and optimization technique described above. Nevertheless, we also show that, in fact, these BWE performance improvements involve no additional run-time computational cost. In addition to the minimal modifications needed to the memoryless BWE baseline system and that fact that our fixed-dimensionality constraint precludes increases in requirements on training data amounts, this makes our proposed technique for frontend-based memory inclusion an easy and convenient means for translating the cross-band correlation advantages of speech memory into tangible BWE performance improvements, albeit only partially.

In analyzing the performance of our first approach described above, we conclude that such delta feature-based memory inclusion succeeds in achieving only modest improvements primarily as a result of the lossiness and non-invertibility discussed in Section 4.4.2 for dimensionality-reducing transforms in general. As such, rather than incorporate long-term spectral information through reducing dimensionalities, we focus instead in our second approach on the problem of modelling the high-dimensional distributions underlying long-term sequences of static joint-band feature vectors. With the problem of high-dimensional modelling in general having been the subject of much research in the fields of machine learning and speaker conversion, e.g., [154–158] and [159–161], respectively, we take inspiration from solutions proposed in these fields in order to devise an algorithm suited to our GMM-based approach to joint-band modelling. In particular, we use prior knowledge about

the properties of GMM speech models as well as the predictability in speech in order to constrain, or *regularize*, the degrees of freedom associated with our modelling problem in a localized manner, effectively transforming the high-dimensional GMM-based *pdf* modelling problem into a time-frequency state space modelling task. Using prior knowledge as such allows us to break down the infeasible task of estimating high-dimensional *pdf*s into a series of *incremental tree-like time-frequency-localized pdf* estimation operations with considerably lower complexity and fewer degrees of freedom. Global *temporally-extended* GMMs can then be obtained by consolidating such time-frequency-localized *pdf*s.

To maximize the information content of the temporally-extended GMMs obtained as such while ensuring their robustness to the potential oversmoothing and overfitting risks associated with the aforementioned localization, we propose a novel *fuzzy GMM-based clustering* technique as well as a *weighted* implementation of the conventional *Expectation-Maximization* (EM) algorithm used for GMM parameter estimation. The fuzzy clustering technique accounts for the effects of class overlap in high-dimensional spaces, while the second incorporates the soft weights associated to time-frequency-localized training data by fuzzy clustering into the maximum-likelihood estimation of GMM parameters.

To emphasize the wide applicability of our tree-like GMM training algorithm to the general problem of high-dimensional GMM-based modelling rather than focusing only on our BWE context, the various operations and novel techniques comprising our proposed algorithm are detailed, illustrated, and derived in as a general BWE-independent manner as possible. This is followed by an evaluation of the reliability of the obtained temporally-extended GMMs in the BWE context in terms of robustness to both oversmoothing and overfitting, with *novel proposed measures* that are equally applicable to other source-target conversion contexts.

Through a detailed analysis, we then conclude this chapter by showing that our proposed temporally-extended GMM-based dual-mode BWE technique outperforms not only our first frontend-based technique discussed above, but also other comparable BWE techniques incorporating model-based memory inclusion—most notably the oft-cited HMM-based techniques discussed in Section 2.3.3.4. In addition to achieving performance improvements of up to 9.1% and 56.1% in terms of $\overline{Q}_{\text{PESQ}}$ and $\overline{d}^{*}_{\text{IS}}$, respectively, relative to our memoryless MFCC-based dual-mode baseline, our model-based memory inclusion approach to BWE also precludes the run-time algorithmic delay associated with our non-causal delta feature-based technique, as well as requires no increases in training data requirements.

These advantages of performance and real-time practicality are achieved, however, at a run-time computational cost increase of nearly four orders of magnitude in terms of number of operations per input speech frame, relative to the memoryless baseline as well as to the computationally equally-inexpensive frontend-based approach. Nevertheless, we show that such computational costs are within the typical capabilities of modern communication devices, such as tablets and smart phones.

5.2 MFCC-Based Dual-Mode Bandwidth Extension

5.2.1 Background

Despite MFCCs' advantages in terms of speech class separability over LSFs and LP-based parameters in general, the difficulty of synthesizing speech from MFCCs has restricted their use to fields that do not require inverting MFCC vectors back into the original speech spectra or time-domain signals, e.g., automatic speech recognition, speaker verification, and speaker identification. This difficulty arises from the non-invertibility of several steps employed in MFCC generation—namely, using the magnitude of the complex spectrum, the mel-scale filterbank binning, and the possible higher-order cepstral coefficient truncation, in Steps 3, 4 and 6 of Section 4.2.2, respectively. Consequently, the vast majority of BWE techniques encountered in the literature are based on LP representations of the highband signals from which the highband frequency content is reconstructed and added to the narrowband signal.

The availability of the narrowband signal, however, has allowed researchers to investigate the effect of several types of narrowband parameterizations on increasing the correlation between narrowband feature vectors and LP-based highband (or wideband) feature vectors. Examples include [39] whose narrowband feature vectors consist of a mixture of auto-correlation coefficients, zero-crossing rate, normalized frame-energy, gradient index, local kurtosis, and the spectral centroid. A rare use of MFCCs in BWE is that of [59] which employs a VQ codebook to map MFCC-parameterized narrowband signals to LSF wideband signals. Informal listening tests in [59] show clear preference for wideband speech reconstructed using the narrowband MFCC representation compared to that of the conventional LP-based representation, despite the reported increase in LSD.

Despite the BWE performance improvements resulting from such alternative narrow-

band parameterizations, these improvements are limited by the highband (or wideband) LP-based representation. This limitation arises from the lower correlation between the alternative narrowband features and the LP-based highband ones; narrowband MFCCs, for example, correlate less with highband LSFs than with highband MFCCs.

There have been a few attempts, however, to achieve speech reconstruction from MFCCs. These attempts arose from the desire to generate speech for playback at the backend of distributed speech recognition (DSR) systems, where frontend processing—i.e., MFCC generation—takes place on the mobile device while recognition itself takes place at a central server. As fewer bits are needed to transmit MFCCs compared to the coded speech of conventional low bit-rate speech codecs employed in mobile devices, DSR thus reduces the information to be transmitted over the usually bandwidth-limited client-server channel. These attempts primarily use a sinusoidal model for speech generation, and require a pitch estimate for each speech frame to be sent as side-information in addition to the MFCC vectors. Frequencies of the sinusoids are determined from the pitch estimate, while sinusoid amplitudes are obtained from smoothed spectral envelopes inferred by applying inverse DCT and exponentiation to MFCC vectors. Sinusoid phases are also typically generated through voicing-based phase models. The works of [151] and [153] represent two notable examples employing this technique. In essence, this sinusoidal model-based technique is similar to that described in Section 2.3.6 and used in [63] and [91] for BWE, except that sinusoid amplitudes are obtained from LP-based LSFs and log envelope samples in [63] and [91], respectively, rather than MFCCs.

To overcome the aforementioned limitation of using LP-based representations for highband envelopes in BWE while also allowing us to potentially exploit the superior highband certainties associated with MFCC-based memory inclusion, we use MFCCs to parameterize both narrowband and highband envelopes—rather than limiting their use to the narrow band only as in [59]—in a manner similar to that we used for estimating MFCC-based highband certainties in Chapter 4. Using GMM-based statistical estimation as in the LSF-based dual-mode BWE system of Chapter 3, we obtain MFCCs representing highband envelope shapes given narrowband MFCC-parameterized envelopes. Then, rather than employ a sinusoidal model-based reconstruction scheme as described above which requires pitch estimation, we convert highband MFCCs into approximate LPCs through interpolation of the filterbank log-energies on the mel frequency scale through a high-resolution inverse DCT [151], followed by exponentiation, mel-to-linear frequency conversion, inverse Fourier

transform, and Levinson-Durbin recursion. Details of our proposed MFCC-based BWE technique follow below.

5.2.2 System block diagram

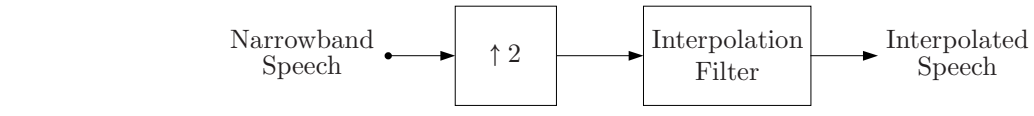
Figure 5.1 illustrates our MFCC-based modification to the dual-mode BWE system previously detailed in Section 3.2 and shown in Figure 3.1. While signal preprocessing, midband and lowband equalization, and EBP-MGN excitation signal generation, are unchanged, the parameterization of the midband-equalized narrowband signal, the subsequent GMM-based MMSE estimation, and the conversion of the estimated highband parameters to LPCs, have now been adapted to the MFCC case. We describe these modified components next.

5.2.3 Parameterization and GMMs

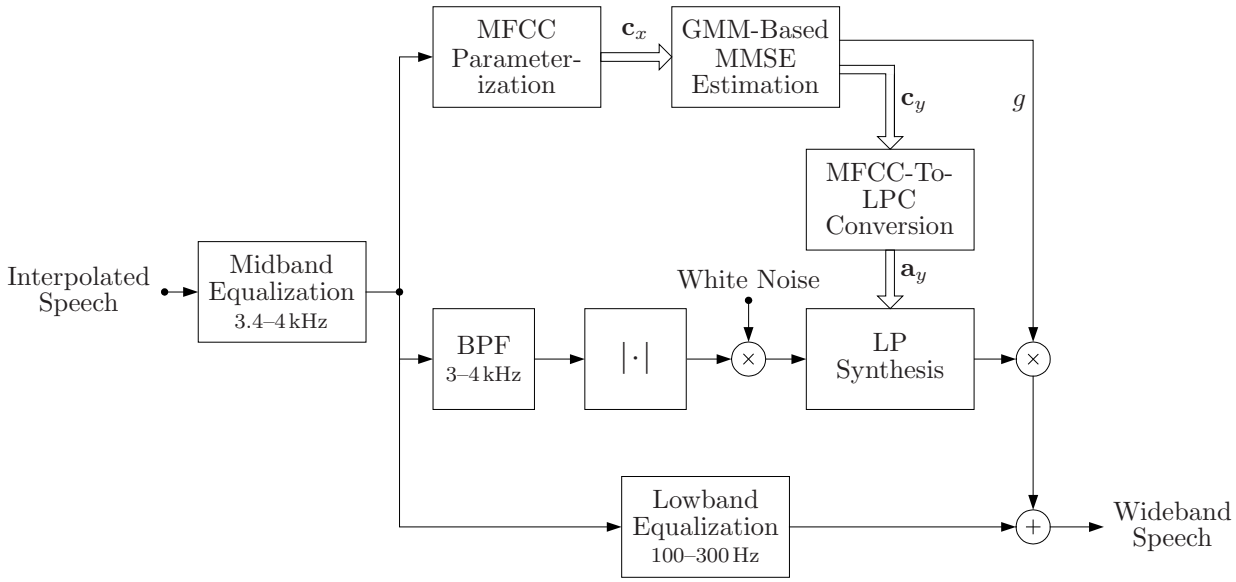
By performing MFCC parameterization as described in Section 4.2.2, and in Section 4.3.4 for the MFCC-based $\mathrm{Dim}(\mathbf{X}, \mathbf{Y}) = (10, 7)$ baseline, we ensure consistency with our LSF-based dual-mode BWE system in terms of the feature vector dimensionalities used to represent both envelopes shapes and gains. In particular, we parameterize the midband-equalized narrowband signal in the 0–4 kHz range using the 9 MFCCs, $[c_{x_1}, \ldots, c_{x_9}]^T$, and the 0th coefficient, c_{x_0}, representing narrowband envelope shape and gain, respectively.[98] As such, the MFCC-based narrowband random vector representation, $\mathbf{X} \coloneqq \mathbf{C}_x$, where the feature vector realizations corresponding to signal frames are given by $\mathbf{x} \coloneqq \mathbf{c}_x \triangleq [c_{x_1}, \ldots, c_{x_9}, c_{x_0}]^T$, coincides exactly with our LSF-based narrowband representation, $\mathbf{X} \triangleq \begin{bmatrix} \boldsymbol{\Omega}_x \\ \log \mathcal{E}_x \end{bmatrix}$, with the dimensionality $\mathrm{Dim}\left(\begin{bmatrix} \boldsymbol{\Omega}_x \\ \log \mathcal{E}_x \end{bmatrix}\right) = 9 + 1 = 10$, as detailed in Sections 3.2.5 and 3.2.7 in the context of the dual-mode BWE system, as well as in Section 4.3.4 in the context of highband certainty estimation.

As described in Sections 3.2.5 and 3.2.7, highband envelope shapes in the 4–8 kHz range were represented by 6-LSF feature vectors, $\boldsymbol{\omega}_y$, while envelope gains were modelled indirectly through the excitation gain, g, estimated such that the energy of the reconstructed highband components is equal to that of the corresponding frequency band in wideband speech. The correlation of these representations of highband envelope shapes and gains

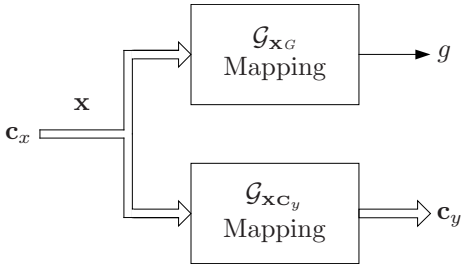
[98] In defining narrowband feature vectors as consisting of the MFCCs c_{x_n}, where n is the order of the coefficient, the subscript x was used for clarity. To simplify notation, however, we will often drop the subscripts x and y from a cepstral coefficient's symbol, e.g., c_n, when clear from the context. In contrast, we always use the subscripts in denoting MFCC feature vectors, e.g., $\mathbf{c}_x$.



(a) Preprocessing



(b) Main processing



(c) GMM-based MMSE estimation

(d) MFCC-to-LPC conversion

Fig. 5.1: The MFCC-based dual-mode bandwidth extension system.

with those of the narrow band were modelled separately through the full-covariance GMM tuple, $\boldsymbol{\mathcal{G}} = (\mathcal{G}_{\mathbf{x}\boldsymbol{\Omega}_y}, \mathcal{G}_{\mathbf{x}G})$, where $\text{Dim}\left(\begin{bmatrix}\mathbf{x}\\ \boldsymbol{\Omega}_y\end{bmatrix}\right) = \begin{bmatrix}10\\6\end{bmatrix}$ and $\text{Dim}\left(\begin{bmatrix}\mathbf{x}\\ G\end{bmatrix}\right) = \begin{bmatrix}10\\1\end{bmatrix}$. Consequently, to also ensure consistency of our MFCC-based highband parameterization with that of our LSF-based dual-mode BWE system, we use the higher-order 6 MFCCs, $\mathbf{c}_y \triangleq [c_{y_1}, \ldots, c_{y6}]^T$, and the excitation gain, g, to represent highband envelope shapes and gains, respectively. Given our MFCC-based parameterizations, the GMM tuple—which we now rewrite as $\boldsymbol{\mathcal{G}} = (\mathcal{G}_{\mathbf{x}\mathbf{C}_y}, \mathcal{G}_{\mathbf{x}G})$ with $\text{Dim}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{C}_y\end{bmatrix}\right) = \begin{bmatrix}10\\6\end{bmatrix}$ and $\text{Dim}\left(\begin{bmatrix}\mathbf{x}\\ G\end{bmatrix}\right) = \begin{bmatrix}10\\1\end{bmatrix}$—jointly modelling the feature vector spaces of both bands, is trained in the manner described in Section 3.2.6. We note, however, that the training values of the excitation gain g—used to train the $\mathcal{G}_{\mathbf{x}G}$ GMM—are calculated differently. In our LSF-based BWE system, the true values of g were determined during training by artificially synthesizing the highband signal using: (a) the EBP-MGN excitation signal described in Section 3.2.4, and (b) the true highband LPCs. With the MFCC-based highband representation, we calculate the true values of g during $\mathcal{G}_{\mathbf{x}G}$ training using the LPCs obtained, rather, through the inversion—described below—of the true highband MFCC feature vectors, $\mathbf{c}_y$.

5.2.4 High-resolution inverse DCT

As mentioned above, two of the six MFCC parameterization steps of Section 4.2.2 involve non-invertible loss of information. Phase information is discarded in Step 3 as a result of retaining only the magnitude of the spectrum. More important, however, is the partial loss of information about spectral envelopes due to the many-to-one mapping of the mel-scale filterbank binning in Step 4. The DCT of Step 6 also involves potential loss of spectral envelope information depending on whether MFCC vectors are truncated. Performing interpolation in the mel-scale log-spectral domain indirectly through a high-resolution inverse DCT of highband MFCCs attempts to recover the information loss most detrimental to reconstructed highband speech quality—that resulting from the mel-scale binning. We note that no inversion is needed for the midband-equalized narrowband MFCCs; these are calculated from the available narrowband speech input only to be used for the MMSE estimation of highband parameters through the GMM tuple, $(\mathcal{G}_{\mathbf{x}\mathbf{C}}, \mathcal{G}_{\mathbf{x}G})$.

In performing MFCC parameterization of the highband content as described in Section 4.2.2, we used $K = 7$ mel-scale filters in the 4–8 kHz range.[99] Thus, given an untrun-

[99] See Step 6 and Figure 4.1 in Section 4.2.2.

cated set of MFCCs representing the highband content of a single frame and which also includes c_0, i.e., $\{c_n\}_{n\in\{0,\ldots,K-1\}}$, the highband mel-scale log-energies, $\{\log_e \varepsilon_k\}_{k\in\{0,\ldots,K-1\}}$, can be perfectly reconstructed by the conventional inverse of the Type-II DCT—the Type-III DCT—given by

$$\log_e \varepsilon_k = a \sum_{n=0}^{K-1} c_n \cos\left(n(k+\frac{1}{2})\frac{\pi}{K}\right), \text{ where } a = \begin{cases} \sqrt{\frac{1}{K}}, & \text{for } k = 0, \\ \sqrt{\frac{2}{K}}, & \text{for } k = 1,\ldots,K-1. \end{cases} \tag{5.1}$$

Since c_0 exclusively contains only information about the total energy of the signal, i.e., envelope gain, the shape of the spectral envelope—as represented by the values of mel-scale log-energies relative to each other—can still be perfectly reconstructed through Eq. (5.1) using only the coefficients $\{c_n\}_{n\in\{1,\ldots,K-1\}}$. In other words, discarding c_0 in Eq. (5.1) only results in shifting the reconstructed highband log-energies, $\{\log_e \varepsilon_k\}_{k\in\{0,\ldots,K-1\}}$, by a constant value, such that the overall highband spectral envelope shape is unaffected. This was partially the motivation for specifically using $K = 7$ mel-scale filters to represent the 4–8 kHz high band in Section 4.2.2, since this value allows us to use the highband $\{c_n\}_{n\in\{1,\ldots,6\}}$ MFCCs as described in Section 5.2.3 above, thereby ensuring consistency with the dimensionality choice we made earlier in Section 3.2.7 for our LSF representation of highband spectral envelope shapes—where 6 LSFs were used. We further note that, by discarding c_0 from our MFCC highband parameterization, we are also ensuring the best use of the dimensionalities available for highband envelope representation since redundancy with g—the highband excitation gain—is thus eliminated.

Given a highband MFCC feature vector, $\mathbf{c}_y$, obtained by MMSE estimation using $\mathcal{G}_{\mathbf{xc}}$, the IDCT of Eq. (5.1) thus provides an estimate of the corresponding highband envelope, consisting of 7 mel-scale log-energy values. Viewed as scaled samples of the log power spectrum at the centre frequencies of the mel-scale filters, it is clear that these few log-energy values are insufficient to recreate a smooth spectrum. Finer spectral detail can be obtained from these log-energies, however, by interpolating them indirectly through increasing the resolution of the IDCT of Eq. (5.1), per

$$\log_e \varepsilon_{k'} = a \sum_{n=0}^{K-1} c_n \cos\left(n(k'+\frac{1}{2})\frac{\pi}{KI}\right), \text{ where } a = \begin{cases} \sqrt{\frac{1}{K}}, & \text{for } k' = 0, \\ \sqrt{\frac{2}{K}}, & \text{for } k' = 1,\ldots,KI-1, \end{cases} \tag{5.2}$$

where an interpolation factor, I, was introduced in the denominator of the cosine frequencies. In essence, the high-resolution IDCT of Eq. (5.2) interpolates between the K mel-scale filterbank centres using the DCT basis functions themselves as the interpolating functions.

Corresponding to a mel-scale filterbank of KI overlapping filters rather than K, the interpolation factor, I, results in KI mel-scale log-spectral samples in the 4–8 kHz range, thus providing a fine and smooth representation of the highband power spectrum. Since the assumed interpolated KI filters partition the $f_{\mathrm{Hz}_l} = 4$ to $f_{\mathrm{Hz}_h} = 8$ kHz highband range into $KI+1$ intervals of equal length on the mel scale, then, using the linear-to-mel-scale frequency conversion of Eq. (4.1), the interpolation factor, I, can be calculated for a particular desired mel-scale resolution, δf_{mel}, through

$$I = \left\lceil \frac{1}{K} \left(\frac{f_{\mathrm{mel}_h} - f_{\mathrm{mel}_l}}{\delta f_{\mathrm{mel}}} - 1 \right) \right\rceil. \tag{5.3}$$

For a desired resolution of 1 mel, for example, Eq. (5.3) results in $I = 99$, with a total of $KI = 693$ mel-scale log-spectral points in the 4–8 kHz range. Based on BWE $\overline{d}_{\mathrm{LSD}}$ results, we found empirically that best reconstruction performance is achieved with a mel-scale resolution of $\delta f_{\mathrm{mel}} \approxeq 4$ mel, accompanied by an FFT length of 4096 for our 320-sample speech frames.[100,101] Per Eq. (5.3), this resolution translates into an interpolation factor of $I = 25$—i.e., $KI = 175$ equally-spaced mel-scale samples of the highband 4–8 kHz spectrum.

Finally, we note that, in practice, the high-resolution IDCT of Eq. (5.2) is applied through a pre-computed matrix with KI rows and K columns, and where the (i, j)th matrix element corresponds to the (k', n)th $a\cos(\cdot)$ term in Eq. (5.2).

5.2.5 Highband speech synthesis

By exponentiation of the interpolated mel-scale log-energies obtained by high-resolution IDCT, i.e., $\{\log_e \varepsilon_{k'}\}_{k' \in \{0,\ldots,KI-1\}}$, we obtain single-sided highband power spectra consisting of KI samples that are equally spaced on the mel scale as well as being scaled by the areas under the mel-scale triangular filters. Thus, to obtain linear-frequency spectra, $P(\omega)$, we

[100] As described in Sections 3.2.8 and 3.2.1, we employ 20 ms windowing and a sampling rate conversion from 8 to 16 kHz applied during preprocessing.

[101] As noted in Section 5.2.3 above, in addition to performing highband speech reconstruction in the extension stage by inverting MFCCs through high-resolution IDCT, we apply a similar MFCC-based reconstruction during the training stage in order to generate the excitation gain, g, values to be used for the maximum-likelihood training of the $\mathcal{G}_{\mathbf{x}_G}$ GMM.

first apply mel-to-linear frequency scale conversion using the inverse of Eq. (4.1),[102] followed by scaling by the inverse of the mel-filterbank areas to equalize the mel-scale spectral tilt.

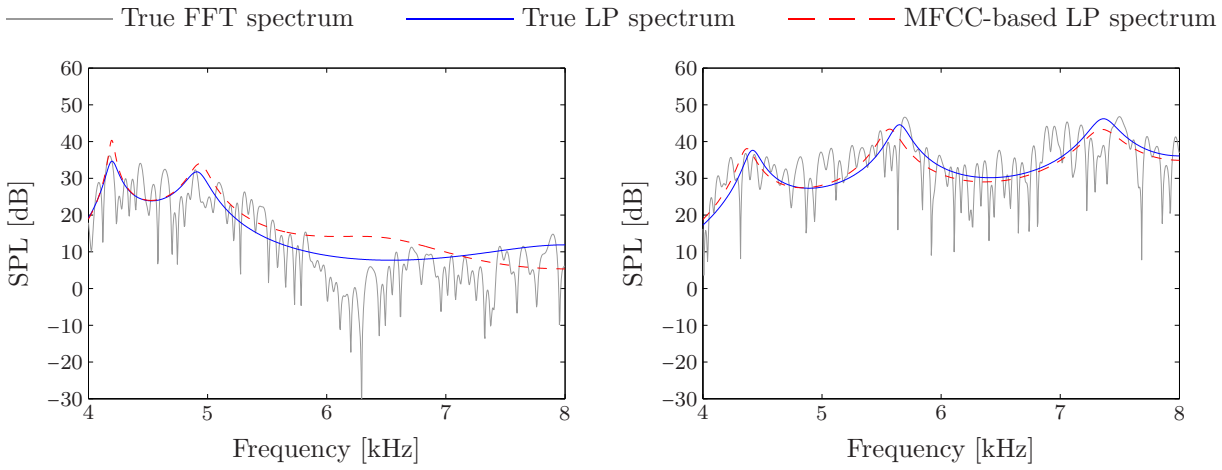
Per the Wiener-Khintchine theorem, computing the inverse Fourier transform of the two-sided power spectra—obtained by reflecting the single-sided spectra—results in the autocorrelation coefficients $\{r_{yy}(l)\}_{l\in\{0,\ldots,N_{\text{IFFT}}-1\}}$, where N_{IFFT} is the inverse fast Fourier transform (IFFT) length [47, Section 4.3.2]. As described in Section 2.3.1 for the source-filter speech production model, the $p+1$ highband autocorrelation coefficients $\{r_{yy}(l)\}_{l\in\{0,\ldots,p\}}$ can then be used to solve the corresponding $p+1$ Yule-Walker equations by means of the Levinson-Durbin recursion, resulting in p highband LPCs, $\{a_y(k)\}_{k\in\{1,\ldots,p\}}$, and an estimate for the minimum mean-square forward prediction error. The LPCs minimizing the forward predictor MSE represent the coefficients of the all-pole vocal tract filter corresponding to the shape of the KI-sample MFCC-based highband power spectrum, while the average power of the spectral envelope is determined either directly using $r_{yy}(0)$ or indirectly via the prediction error variance in conjunction with the LPCs. Consistent with the prediction order used in Section 3.2.7 for our LSF-based dual-mode BWE system, we use p = 6th-order linear prediction for our MFCC-based spectra. Adapted from the work in [151] and [152], both of which were concerned rather with DSR-backend speech reconstruction, our technique for the conversion of highband MFCCs to LPCs for the purpose of BWE is summarized in Figure 5.1(d).

Figure 5.2 illustrates the high quality of our MFCC-based highband power spectral LP approximations by comparing two such approximations to those of the conventional LP spectra of the same order—where autocorrelation coefficients are calculated directly from the input speech samples. Superimposed on the original non-smoothed FFT power spectra, the MFCC-based and conventional LP spectral approximations are shown for a vowel, /e/, and a fricative, /s/, in Figures 5.2(a) and 5.2(b), respectively.[103] It can be seen that our MFCC-based spectra closely match the true LP approximations, particularly so for the more important fricative highband spectra. Figure 5.2 also shows, however, that, despite the success of our interpolation-based approach in generating generally accurate spectral envelope reconstructions, the reconstructed envelopes nevertheless still exhibit some errors due to the non-invertibility of mel-scale filterbank binning. The most notable of these errors

[102] Mel-to-linear frequency scale conversion is given by $f_{\text{Hz}} = 700\big[10^{\frac{f_{\text{mel}}}{2595}} - 1\big]$.

[103] In obtaining the MFCC-based LP approximations of Figure 5.2, the gains of the pre-LP interpolated spectra where determined using the 0th coefficient, c_0, rather than the excitation gain, g.

are those in the spectral valley near 6.5 kHz of the vowel spectrum in Figure 5.2(a), and in the formant near 5.7 kHz for the fricative spectrum in 5.2(b). The effects of such errors on the overall objective BWE performance are discussed in Section 5.2.6 below.



(a) /e/ (b) /s/

Fig. 5.2: Comparing MFCC-based LP approximations of highband power spectra—obtained through MFCC inversion with interpolation via high-resolution IDCT—to those of conventional LP spectra for two non-windowed 20 ms highband speech frames corresponding to the mid-regions of a vowel, /e/, and a fricative, /s/. The non-smoothed FFT-based power spectra are shown as the reference for the approximations. Power spectra are mapped to sound pressure level (SPL) on the ordinate using an SPL value of 90.3 dB for the maximum attainable value of the 16-bit linear PCM-coded speech frames.

With the MFCC-based LP spectral estimates obtained as described above, highband speech can be reconstructed using an appropriate excitation signal. In the DSR approaches of [151–153], the excitation signal is generated using voicing-based models which require an estimate of the pitch. A pitch parameter is thus added to MFCC feature vectors as side-information in these techniques. In the context of BWE, however, a superior highband excitation signal can be generated using the narrowband signal readily available as BWE input. As previously described for the LSF-based dual-model BWE system in Section 3.2.4, modulating white Gaussian noise with the 3–4 kHz midband-equalized narrowband signal, in particular, provides such a superior excitation signal. This EBP-MGN excitation mirrors the narrowband harmonic structure into the high band, resulting in pitch harmonics for vowel-like voiced sounds, noise for unvoiced sounds, and a mixture of both for mixed sounds.

Furthermore, due to its phase-coherence with the narrowband signal in the 3–4 kHz range, the EBP-MGN excitation partially mitigates the loss of phase information in Step 3 of MFCC parameterization, noting that a more accurate—and consequently more complex—estimation of phase is unwarranted due to the relative unimportance of phase for speech intelligibility [162].

5.2.6 Memoryless baseline performance

Through the spectral interpolation performed via high-resolution IDCT and the coherence of the EBP-MGN excitation signal phase with that of the narrow band, we have addressed the loss of spectral envelope and phase information associated with Steps 4 and 3 of MFCC parameterization, respectively. By further aligning the number of mel-scale filters in the 4–8 kHz range with our baseline highband MFCC feature vector dimensionality, we have also precluded the loss of spectral information as a result of MFCC truncation. As such, we were able to reconstruct high-quality highband speech from MFCCs, thereby enabling us to adapt our LSF-based dual-mode BWE system to MFCCs, as summarized in Figure 5.1. This, in turn, allows us to potentially exploit the superior highband certainty properties of MFCCs—shown in Sections 4.3.4 and 4.4.3.2—to improve BWE performance. Table 5.1 below lists our MFCC-based memoryless BWE baseline performance obtained for the TIMIT core test set with $N_\mathrm{f} \cong 58 \times 10^3$ frames.[104,105]

Table 5.1: Speaker-independent memoryless BWE baseline performance using full-covariance GMMs with $M = 128$, and MFCC parameterization with $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{c}_y\end{bmatrix}\right) = 16$ and $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{x}\\ G\end{bmatrix}\right) = 11$.

$\overline{d}_{\mathrm{LSD}}$ [dB]	$\overline{d}_{\mathrm{LSD(RMS)}}$ [dB]	$\overline{Q}_{\mathrm{PESQ}}$	$\overline{d}^{*}_{\mathrm{IS}}$ [dB]	$\overline{d}^{*}_{\mathrm{I}}$ [dB]
5.17	5.89	3.01	12.32	0.5820

By comparing the MFCC-based performance figures of Table 5.1 to those of the LSF-based baseline performance in Table 3.1, we can conclude that our attempts at mitigating the spectral envelope information losses associated with MFCC parameterization were largely successful, resulting in an overall highband speech reconstruction quality that is comparable to that obtained using LSFs. In particular, the $\overline{d}_{\mathrm{LSD}}$, $\overline{d}_{\mathrm{LSD(RMS)}}$, and $\overline{Q}_{\mathrm{PESQ}}$

[104] See Footnote 77 regarding GMM-derived results.
[105] See Section 3.2.10 for description of the training and test data.

measures—measuring distortions in both the shape and gain of reconstructed spectral envelopes—show a relative decrease in performance of less than 2% using MFCCs, while the gain-independent $\overline{d}_{\text{I}}^{*}$ measure shows nearly identical performance for the reconstruction of envelope shapes using both LSFs and MFCCs. This indicates that, in the context of our dual-model BWE implementation, MFCC-based BWE marginally lags that based on LSFs only in terms of spectral envelope gain estimation.

We note, however, that the superior certainty properties of MFCCs in the memoryless case—shown in Table 4.1 for the reference $\text{Dim}(\mathbf{X}, \mathbf{Y}) = (10, 7)$ LSF- and MFCC-based memoryless spaces—did not translate into corresponding BWE performance gains compared to our baseline LSF-based performance. Since our dual-model BWE implementation shares the same full-covariance GMM-based statistical modelling as well as the same parameterization type and dimensionality with our cross-band correlation modelling of Chapter 4, we conclude that the underlying MFCC-based certainty gains observed in Table 4.1 were offset by errors in reconstructing spectral envelopes through the MFCC-based spectral interpolation described above, rather than through LSFs. Using the performance lower bound of $\downarrow\overline{d}_{\text{LSD(RMS)}} = 4.62\,\text{dB}$ for the $\text{Dim}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_{\text{ref}}) = (10, 7, 7)$ baseline MFCC space in Table 4.2, we can in fact quantify the inoptimality of our MFCC-based dual-mode BWE system—including interpolation-based envelope reconstruction errors—as the equivalent to a distortion of $\overline{d}_{\text{LSD(RMS)}} = 1.27\,\text{dB}$.

Despite its inoptimality, our success in achieving a baseline MFCC-based BWE performance comparable to that based on LSFs motivates us to exploit the superior certainty advantages of memory inclusion based on MFCCs, rather than LSFs, for the purpose of improving BWE performance. In particular, we showed in Section 4.4.3.2 that including memory through delta features based on MFCCs results in considerably higher certainties about the high band than achieved by LSF-based memory inclusion. While reference high-band certainties for the memoryless LSF- and MFCC-based baselines differ by only $\approx 4.6\%$ (15.9% compared to 20.5% for LSFs and MFCCs, respectively, per Table 4.1), the difference between LSF- and MFCC-based certainties in the case of memory inclusion can potentially reach 19.5%–23.6% in favour of MFCCs, as shown in Table 4.4. More importantly, in the case of memory inclusion under fixed-dimensionality constraints (Case S-2 in Table 4.4), the focus of our work described below, MFCC-based cross-band correlation modelling was shown to be much less susceptible than its LSF-based counterpart to the adverse effects of the time-frequency information tradeoff; including memory as described for Case S-2

increases MFCC-based certainty by a relative 77.5%, compared to only 9.8% using LSFs.

Based on these observations, we will henceforth exclusively consider MFCC-based parameterization for the implementation of memory inclusion.

5.3 BWE with Frontend-Based Memory Inclusion

In this section, we present our first attempt to translate the highband certainty gains obtained in Section 4.4.3 as a result of memory inclusion—i.e., the inclusion of speech dynamics—into measurable MFCC-based dual-mode BWE performance improvements.

As discussed in the preamble of Section 4.4, transforming temporal sequences of conventional static feature vectors through a dimensionality-reducing transform represents the most compact and efficient—albeit lossy—means of memory inclusion, thereby providing the motivation for having employed delta features for the information-theoretic investigation of Section 4.4.3. For the purpose of improving BWE performance by exploiting the high cross-band correlations of speech dynamics, it follows that we similarly investigate memory inclusion through the use of delta features, although, as discussed in Section 4.4.2, such a frontend-based approach is by no means optimal by virtue of the non-invertibility of lossy dimensionality-reducing transforms in general. As such, we begin by reviewing the application of frontend-based memory inclusion in the literature.

5.3.1 Review of previous works on frontend-based memory inclusion

As described earlier in Section 1.4, previous attempts to exploit the information in speech dynamics for the purpose of improving BWE performance have primarily taken a modelling-based approach where the cross-band correlations of speech dynamics are modelled through HMMs. In contrast, exploiting memory through its inclusion into the parameterization frontend has been quite limited, not only in terms of use, but also in terms of scope where it has indeed been applied. In particular, except for the work of [132] discussed below, frontend-based memory inclusion has exclusively been applied merely as a secondary means for improved narrowband feature space parameterization, rather than as a means of capturing the important cross-band information about speech dynamics.

To the best of our knowledge, the use of frontend-based memory inclusion has only been applied in [87, 129, 132, 163]. In [129], where a neural network is used to model the cross-correlations between narrowband features and four mel-scale subband energies in the

4–8 kHz range, the ratio of signal energy in a speech frame to that in the previous frame—representing short-term narrowband speech dynamics—is included as a single parameter in narrowband feature vectors. Similarly, delta as well as delta-delta (second-order regression) features have been used in [87, 163] to incorporate dynamic information. To model the cross-band correlation of speech dynamics, however, both these approaches rely instead on the first-order HMM state transition probabilities as previously described in Section 2.3.3.4, with the latter technique to be further detailed in Section 5.4.1.3. In fact, the BWE technique of [87] incorporates dynamic features only for the narrow band, thereby including memory per Scenario 1 of our investigation in Section 4.4.3. As was shown therein, the inclusion of memory in such a scenario provides minimal to no benefits in terms of certainty about the high band.

Finally, we note the work of [132] where GMM-estimated short-term temporal envelopes of the 4–7 kHz band are used directly to reconstruct highband speech. First proposed in [164] as an alternative to the source-filter speech production model, Kim et alia subjectively show that the temporal envelope of the highband signal is an important perceptual cue of highband content while rapidly varying components—i.e., fine structure—in the temporal domain are not as important. Mimicking the temporal masking properties of speech, highband components in each 5 ms frame are represented by the shapes and gains of the temporal envelopes of four subband signals, with the assumption that these highband temporal envelopes are related to the temporal envelope in the *intermediate* 3–4 kHz band—obtained through a Hilbert transform—through a linear transformation. Using GMMs to estimate highband content (represented by the gains and transform filter coefficients of the four subband signals) using that of the narrow band (represented by LFCCs—linear-frequency cepstral coefficients), highband speech is then reconstructed per frame through a time-domain multiplication of the MMSE-estimated temporal envelopes with fine structure signals—obtained by full-wave rectification of the narrowband signal followed by a Hilbert transform—and, finally, summing the four time-domain products corresponding to the subband signals.

Although the speech production model of [132, 164] is based on a temporal representation of the signal, this BWE technique only considers temporal dynamics within frame-based intervals no longer than 5 ms. As such, it can not be considered as applying frontend-based memory inclusion, per se. Furthermore, while this temporal envelope-based technique is similar to the dual-mode BWE technique in that it relies on mapping the voic-

ing and noisiness characteristics of the signal in the intermediate 3–4 kHz range into the high band,[106] it assumes that speech content in that intermediate range is readily available in narrowband input. Since this assumption is not normally valid for conventional telephony, however, the conclusions and subjective results of [132, 164] must be correspondingly qualified; i.e., they don't account for highband temporal envelope distortions linearly mapped from an imperfectly reconstructed envelope in the 3.4–4 kHz subband. To conclude, we note that the BWE performance using the temporal envelope model in [132] was evaluated by comparing it to the performance of the conventional GMM- and source-filter model-based BWE system of [82]. Using the subjective MUSHRA [165] and ABX preference [166] tests, results in [132] show a slight preference for the wideband speech obtained using the proposed temporal-based technique.[107]

5.3.2 Fixed-dimensionality constraint

To render the comparisons of memory-inclusive and memoryless BWE performances—and any improvements achieved—as practical and fair as possible while also ensuring consistency with the information-theoretic investigation of Section 4.4.3, we restrict our work herein by imposing a fixed-dimensionality constraint; the inclusion of memory through delta features should not result in an increase of dimensionality for the dual-mode system's GMM with maximum dimensionality—$\mathcal{G}_{\mathbf{xc}}$[108]. This constraint guarantees that the same amount of data previously used to train the GMM tuple of the memoryless MFCC-based dual-mode BWE system can be used without increase with the memory-inclusive modifications. Furthermore, while only a slight increase in computational costs will be required during parameterization as a result of the additional processing needed for delta feature calculation, the fixed-dimensionality constraint ensures that all certainty and BWE per-

[106] The dual-mode BWE system of [55] generates voicing and noisiness characteristics for the high band indirectly by equalizing the 3.4–4 kHz range before using the 3–4 kHz subband to generate the EBP-MGN excitation signal; see Section 3.2.4, while the temporal envelope-based approach of [132] maps these characteristics from the temporal envelope in the 3–4 kHz range directly into the temporal envelope of the high band through a linear transform.

[107] In multiple-stimuli-with-hidden-reference-and-anchor (MUSHRA) tests, listeners assess the quality of multiple test stimuli—including a hidden reference and one or more hidden anchors—by assigning a score to each stimulus. In [132], the stimuli consist of two anchors, a hidden reference, and undistorted narrowband speech samples as well as the corresponding samples from the proposed temporal-envelope model-based and reference source-filter model-based BWE algorithms. In ABX tests, listeners determine which of the two test stimuli, A and B, is more identical to the reference stimulus, X.

[108] See Section 5.2.3.

formance improvements achieved are exclusively due to the substitution of static features by delta ones, rather than to any improvements in GMM training resulting from higher degrees of freedom for feature-space modelling or from the use of additional training data.

5.3.3 Exploiting the cross-correlation between narrowband and highband spectral envelope dynamics

5.3.3.1 Re-examining information-theoretic findings in the context of BWE for illustrative purposes

Using information-theoretic measures to quantify cross-band correlation, we showed in Section 4.4.3 that incorporating memory—in the form of delta features—into the parameterizations of both narrowband and highband spectral envelopes can increase such cross-band correlation considerably. For MFCCs with a fixed-dimensionality constraint, in particular, we showed that memory inclusion per Case S-2 increases certainty about the high band to 36.5% when represented by the dynamic vectors $\overset{\triangle}{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{\Delta_Y} \end{bmatrix}$, up from 20.5% with only the conventional static representation, $\mathbf{Y}$, corresponding to a potential 0.82 dB reduction in RMS-LSD BWE distortion.[109]

Translating these highband certainty gains obtained through the use of delta features into practical BWE performance improvements requires, however, that we re-examine the relevant conclusions of Section 4.4.3.2 in the context of BWE implementation, as follows:

(a) As shown by the results of Scenario 1, narrowband spectral dynamics represented by the delta features, $\mathbf{\Delta_X}$, provide minimal information about static highband spectra, $\mathbf{Y}$, and vice versa, i.e.; $I(\mathbf{\Delta_X};\mathbf{Y}) \ll H(\mathbf{Y})$ and $I(\mathbf{X};\mathbf{\Delta_Y}) \ll H(\mathbf{X})$. To simplify the analysis to follow as well as emphasize that these findings were made: (a) based on GMM-based estimates of MI,[110] and (b) using a joint-band GMM that only models the joint distribution of $\overset{\triangle}{\mathbf{X}}$ and $\mathbf{Y}$ (or $\mathbf{X}$ and $\overset{\triangle}{\mathbf{Y}}$),[111] we write

$$\hat{I}(\mathbf{\Delta_X};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}}) \approxeq 0 \quad \text{and} \quad \hat{I}(\mathbf{X};\mathbf{\Delta_Y}|\mathcal{G}_{\mathbf{X}\overset{\triangle}{\mathbf{Y}}}) \approxeq 0. \tag{5.4}$$

The assumption that these quantities equal zero implies that modifying the dual-mode BWE system—represented by the $\boldsymbol{\mathcal{G}} = (\mathcal{G}_{\mathbf{XC}}, \mathcal{G}_{\mathbf{X}G})$ GMM tuple—by using $\overset{\triangle}{\mathbf{X}}$,

[109] See Table 4.4.
[110] See Eq. (4.7).
[111] See Section 4.4.3.2.

rather than $\mathbf{X}$, as the representation of the narrow band while continuing to use only the static $\mathbf{Y}$ representation for the high band, will result in no improvement in performance.

(b) The results of Scenario 2 showed that appending delta features to the static feature vectors of both bands—i.e., Case A-2—increases cross-band correlation by up to 99% for MFCCs when all available delta features are used without truncation. Using delta features to replace some of the static features in both bands—i.e., Case S-2—also results in an overall increase in cross-band correlation, albeit lower than that of Case A-2 as a result of a time-frequency information tradeoff.

To illustrate the relations between the information content of the four feature vector spaces considered in Scenario 2—i.e., $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{\Delta_X}$, and $\mathbf{\Delta_Y}$—in a manner similar to that of Figure 4.4, we extend the assumption of Eq. (5.4) to the case of Scenario 2 as well where a joint-band GMM, $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}$, modelling the joint distribution of $\overset{\triangle}{\mathbf{X}}$ and $\overset{\triangle}{\mathbf{Y}}$, is used, rather than $\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}}$ as in Scenario 1. In other words,

$$\text{assume:} \quad \begin{aligned} \hat{I}(\mathbf{\Delta_X};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}) &= \hat{I}(\mathbf{\Delta_X};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}}) \approxeq 0, \\ \hat{I}(\mathbf{X};\mathbf{\Delta_Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}) &= \hat{I}(\mathbf{X};\mathbf{\Delta_Y}|\mathcal{G}_{\mathbf{X}\overset{\triangle}{\mathbf{Y}}}) \approxeq 0, \end{aligned} \tag{5.5}$$

then, the relations between the information content of the four feature vector spaces can be visualized through the Venn-like diagram[112] in Figure 5.3 below, which shows that

$$\begin{aligned} \hat{I}(\overset{\triangle}{\mathbf{X}};\overset{\triangle}{\mathbf{Y}}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}) \triangleq \hat{I}(\mathbf{X},\mathbf{\Delta_X};\mathbf{Y},\mathbf{\Delta_Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}) &\approxeq \mathcal{R}_1 \cup \mathcal{R}_2 \\ &= \hat{I}(\mathbf{X};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}) + \hat{I}(\mathbf{\Delta_X};\mathbf{\Delta_Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}). \end{aligned} \tag{5.6}$$

(c) Similar to most speech processing techniques, BWE operates on a frame-by-frame basis such that a time-domain highband signal can be reconstructed using quasi-stationary spectral envelope estimates obtained from the available narrowband input. Thus, without fundamental changes involving the source-filter speech production model and/or GMM-based statistical modelling, making use of information gained about the dynamics of highband spectral envelopes requires translating such information into corresponding information about static envelopes. Consequently, in the

[112] See Footnote 91.

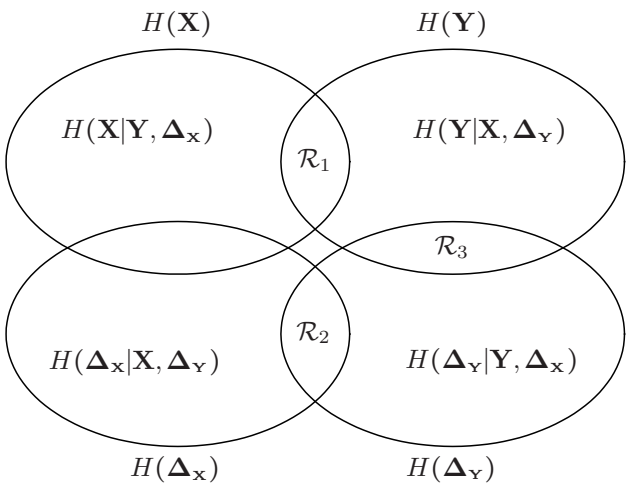


Fig. 5.3: Venn-like diagram representing the relations between the information content of the $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{\Delta_X}$ and $\mathbf{\Delta_Y}$ spaces, under the assumption that $\hat{I}(\mathbf{\Delta_X};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}) = \hat{I}(\mathbf{\Delta_X};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\mathbf{Y}}) \approxeq 0$ and $\hat{I}(\mathbf{X};\mathbf{\Delta_Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}) = \hat{I}(\mathbf{X};\mathbf{\Delta_Y}|\mathcal{G}_{\mathbf{X}\overset{\triangle}{\mathbf{Y}}}) \approxeq 0$.

context of BWE, the increase in highband certainty we achieved by memory inclusion using delta features can only be useful if the improved cross-band correlation between the $\overset{\triangle}{\mathbf{X}}$ and $\overset{\triangle}{\mathbf{Y}}$ representations is mapped into higher certainty about static $\mathbf{Y}$ feature vectors—more specifically, if the gained information about $\mathbf{\Delta_Y}$ feature vectors is mapped into improved $\mathbf{Y}$ vectors.

(d) As described in Sections 4.4.1 and 4.4.2, delta features are obtained by non-causal FIR filtering of static features with zeroes on the unit circle, and hence, are not practically invertible as the inverse filter is only marginally stable. Delta features can not thus be deterministically used for LP-based reconstruction of static envelopes. Accordingly, statistical mapping is the only means to convert the information attained about $\mathbf{\Delta_Y}$ features using the narrowband $\overset{\triangle}{\mathbf{X}}$ dynamic representation into additional information about the static $\mathbf{Y}$ spectral envelope representation.

(e) The information that can be used for obtaining better estimates for $\mathbf{Y}$ given $\mathbf{\Delta_Y}$—i.e., the information that is mutual to both $\mathbf{Y}$ and $\mathbf{\Delta_Y}$—is represented in Figure 5.3 by the region $\mathcal{R}_3$. This region, in addition to that denoted by $\mathcal{R}_1$, represents the information content that can be used to reconstruct $\mathbf{Y}$ in a practical frame-based BWE implementation as described in point (c) above. As a result of the assumptions made in Eq. (5.5), however, it is clear from Figure 5.3 that region $\mathcal{R}_3$ does not

overlap with either $H(\mathbf{X})$ or $H(\mathbf{\Delta_X})$—the information available via the narrowband input. In other words, neither $\mathbf{X}$ nor $\mathbf{\Delta_X}$ provides information about $\mathbf{\Delta_Y}$ that can, in turn, be used to improve estimates of $\mathbf{Y}$. Stated more formally, Eq. (5.6)—resulting from the assumptions of Eq. (5.5)—shows that the certainty gains measured in Scenario 2 are exclusively due to the additional $\hat{I}(\mathbf{\Delta_X};\mathbf{\Delta_Y}|\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{\mathbf{Y}}})$ term represented by the region $\mathcal{R}_2$. Since $\hat{I}(\mathbf{\Delta_X};\mathbf{Y}|\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{\mathbf{Y}}}) \approxeq 0$ per our assumptions, then, by the *data-processing inequality*[113], $\mathbf{Y}$ is also conditionally independent of any estimate, $\hat{\mathbf{\Delta}}_\mathbf{Y}$, that is probabilistically a function of only $\mathbf{\Delta_X}$—i.e., $\mathbf{Y}$, $\mathbf{\Delta_X}$, and $\hat{\mathbf{\Delta}}_\mathbf{Y}$ form the Markov chain $\mathbf{Y} \to \mathbf{\Delta_X} \to \hat{\mathbf{\Delta}}_\mathbf{Y}$—and hence,

$$\hat{I}(\hat{\mathbf{\Delta}}_\mathbf{Y};\mathbf{Y}|\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{\mathbf{Y}}}) \leq \hat{I}(\mathbf{\Delta_X};\mathbf{Y}|\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{\mathbf{Y}}}) = \hat{I}(\mathbf{\Delta_X};\mathbf{Y}|\mathcal{G}_{\overset{\vartriangle}{\mathbf{XY}}}) \approxeq 0, \tag{5.7}$$

showing that the certainty advantages measured in Scenario 2 as a result of memory inclusion can not—under the simplifying assumptions of Eq. (5.5)—be translated into practical BWE performance if such inclusion is applied using non-invertible delta features.

5.3.3.2 Exploiting highband dynamics to improve joint-band modelling

By facilitating the analysis above, the assumptions of Eq. (5.5) allowed us to gain a better understanding of the effect of memory inclusion using delta features on potential BWE performance. These assumptions, however, do not take into account an important advantage of $\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{\mathbf{Y}}}$ over $\mathcal{G}_{\overset{\vartriangle}{\mathbf{XY}}}$—the ability to exploit the $\mathbf{\Delta_Y}$ training data to obtain a better model of the underlying acoustic classes. This, in turn, should result in improved estimates for the true $I(\overset{\vartriangle}{\mathbf{X}};\mathbf{Y})$—the information actually made use of in a practical BWE system as discussed in point (c) above. As such, a BWE system based on $\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{\mathbf{Y}}}$, rather than $\mathcal{G}_{\overset{\vartriangle}{\mathbf{XY}}}$, will then generate better estimates for $\mathbf{Y}$ using the $\overset{\vartriangle}{\mathbf{X}} = \begin{bmatrix}\mathbf{X}\\ \mathbf{\Delta_X}\end{bmatrix}$ inputs—despite the fact that the $\mathbf{\Delta_Y}$ subspace model is, in fact, discarded during the extension stage—provided that the true $I(\mathbf{\Delta_X};\mathbf{Y})$ is, in fact, higher than our $\hat{I}(\mathbf{\Delta_X};\mathbf{Y}|\mathcal{G}_{\overset{\vartriangle}{\mathbf{XY}}})$ estimates. Indeed, although the results of Scenario 1 show only a modest correlation between the $\mathbf{\Delta_X}$ and $\mathbf{Y}$ feature

[113] The random variables X, Y, and Z, are said to form a Markov chain—denoted by $X \to Y \to Z$—if the conditional distribution of Z depends only on Y and is conditionally independent of X. The data-processing inequality states that, if $X \to Y \to Z$—which also implies $Z \to Y \to X$—then, $I(X;Y) \geq I(X;Z)$. See [64, Section 2.8] for proof and corollaries.

spaces—i.e., in contrast to our simplifying assumptions, $I(\boldsymbol{\Delta}_{\mathbf{X}};\mathbf{Y}) \not\approx 0$—the properties of speech discussed in Sections 1.1.3.1 and 1.2 suggest that the correlation between the two spaces should be higher. For tense and stressed vowels, for example, static features of the low-energy highband envelopes should exhibit a close relationship with the delta features of the narrow band as these vowels are characterized by relatively constant properties over longer durations—up to an average 130 ms for stressed vowels—compared to other manners of articulation.

As described in Section 2.3.3.4, the foremost motivation for using GMMs to model speech in general is their ability to model underlying sets of acoustic classes with an intuitive correspondence between such classes and the Gaussian component densities. As such, the components of the memoryless joint-band GMM $\mathcal{G}_{\mathbf{XY}}$ that is trained only on the static features of both bands—and with $M = 128$ as described in Section 3.5.3—will tend to model underlying classes corresponding to the fine spectral detail of quasi-stationary allophonic variations of phonemes.[114] Without an accompanying increase in the number of GMM components, the introduction of temporal features—e.g., delta features—in addition to their existing static counterparts during training will influence the iterative maximum-likelihood (ML) estimation of the mixture model towards salient properties along temporal axes, such that the underlying classes represented by the M components acquire temporal resolution at the cost of decreased spectral resolution. In a joint-band GMM, such as $\mathcal{G}_{\mathbf{XY}}$, $\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}}$, or $\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}}$, the two feature subspaces corresponding to the two frequency bands are modelled jointly and are assumed to share the same underlying acoustic classes—the basis of BWE. Thus, introducing temporal features into the representations of both bands ensures that the two corresponding ML- and jointly-trained subspace models are influenced uniformly in the same manner by temporal properties, thereby generating a better model of the underlying classes shared by the two feature subspaces. This, in turn, should result in more accurate estimates of the true correlation between temporal features in one band and static ones in the other than can be obtained by incorporating temporal features into the parameterization of only one band.

To summarize, we argue that the superior cross-band correlation between the dynamic $\overset{\triangle}{\mathbf{X}}$ and $\overset{\triangle}{\mathbf{Y}}$ vectors improves the overall ability of the dynamic GMM, $\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}}$, to model, and subsequently estimate, the cross-band correlation between $\overset{\triangle}{\mathbf{X}}$ and $\mathbf{Y}$—represented by

[114] See the discussion in Section 3.3.4 on the correspondence of the number of Gaussian components and type of acoustic features used in a GMM to the underlying classes modelled by the GMM.

$\hat{I}(\overset{\triangle}{\mathbf{X}};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}})$—since training for $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}$ is performed using the static highband features, $\mathbf{Y}$, jointly with their $\mathbf{X}$, $\mathbf{\Delta_X}$, and $\mathbf{\Delta_Y}$ counterparts, thereby making use of the correlations between all four quantities—particularly the strong correlations between $\mathbf{\Delta_X}$, and $\mathbf{\Delta_Y}$—rather than just those between $\overset{\triangle}{\mathbf{X}}$ and $\mathbf{Y}$. In other words,

$$\hat{I}(\overset{\triangle}{\mathbf{X}};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\mathbf{Y}}) \leq \hat{I}(\overset{\triangle}{\mathbf{X}};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}) \leq I(\overset{\triangle}{\mathbf{X}};\mathbf{Y}), \tag{5.8}$$

where $I(\overset{\triangle}{\mathbf{X}};\mathbf{Y})$ is the true mutual information. This indirect effect of using $\mathbf{\Delta_Y}$ data jointly with their $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{\Delta_X}$ counterparts to improve the overall Gaussian mixture model during training is similar in principle to the effect of training diagonal-covariance GMMs on any set of joint vectors; despite their lack of cross-covariances, diagonal-covariance GMMs still capture the underlying correlation between the modelled subspaces as a result of training on joint vectors.

To verify these arguments summed up by Eq. (5.8)—as well as assess the validity of our simplifying assumptions in Eq. (5.5)—we compare the certainty $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}})$, i.e., the certainty obtained for static $\mathbf{Y}$ highband features given the dynamic $\overset{\triangle}{\mathbf{X}} = \begin{bmatrix}\mathbf{X}\\ \mathbf{\Delta_X}\end{bmatrix}$ narrowband representation and a joint-band $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}$ GMM trained on joint $[\overset{\triangle}{\mathbf{X}}^T,\overset{\triangle}{\mathbf{Y}}^T]^T$ feature vectors, to the corresponding certainty $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\mathbf{Y}})$, obtained using $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\mathbf{Y}}$, for the same MFCC dimensionalities used in Section 4.4.3.1. As described in Section 4.3, certainties—rather than mutual information figures—are more relevant to BWE. Representing upper bounds on BWE performance, certainties take the self-information of highband features into account, as well as account for the effects of differences in highband dimensionality.

Reusing our earlier $\text{Dim}(\mathbf{X},\mathbf{\Delta_X},\mathbf{Y},\mathbf{\Delta_Y},\mathbf{Y}_{\text{ref}})$ representation of acoustic space dimensionalities introduced in Sections 4.3.4 and 4.4.3.1, we evaluate certainties relative to our memoryless MFCC-based $(10,0,7,0,7)$ baseline of Section 4.3.4 with the certainty and RMS-LSD lower bound performances shown in Table 4.2. To preserve dimensionality as described in Section 5.3.2, we only consider the inclusion of delta features under Context S[115] with the feature dimensionalities given by $(5,5,4,0,7)$ and $(5,5,4,4,7)$ for $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\mathbf{Y}})$ and $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}})$, respectively. Except for the difference in $\text{Dim}(\mathbf{Y}_{\text{ref}})$, we estimate $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\mathbf{Y}})$ for the $(5,5,4,0,7)$ MFCC dimensionalities in the same manner as

[115] See Section 4.4.3.1.

previously performed for the evaluation of memory inclusion per Case S-1;[116] i.e., for N feature vectors,

$$C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}}) = \frac{\hat{I}(\overset{\triangle}{\mathbf{X}};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}})}{H(\mathbf{Y})|_{\overline{d}_{\text{LSD}}=1\text{dB}}}, \text{ with } \hat{I}(\overset{\triangle}{\mathbf{X}};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}}) = \frac{1}{N}\sum_{n=1}^{N}\log_2\left(\frac{\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}}(\overset{\triangle}{\mathbf{x}}_n,\mathbf{y}_n)}{\mathcal{G}_{\overset{\triangle}{\mathbf{X}}}(\overset{\triangle}{\mathbf{x}}_n)\mathcal{G}_{\mathbf{Y}}(\mathbf{y}_n)}\right). \quad (5.9)$$

In contrast, we estimate $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}})$ by marginalizing $\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}}(\overset{\triangle}{\mathbf{x}},\overset{\triangle}{\mathbf{y}})$ over the $\boldsymbol{\Delta}_{\mathbf{Y}}$ subspace to obtain $\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}}(\overset{\triangle}{\mathbf{x}},\mathbf{y})$, such that

$$C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}}) = \frac{\hat{I}(\overset{\triangle}{\mathbf{X}};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}})}{H(\mathbf{Y})|_{\overline{d}_{\text{LSD}}=1\text{dB}}} \text{ and } \hat{I}(\overset{\triangle}{\mathbf{X}};\mathbf{Y}|\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}}) = \frac{1}{N}\sum_{n=1}^{N}\log_2\left(\frac{\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}}(\overset{\triangle}{\mathbf{x}}_n,\mathbf{y}_n)}{\mathcal{G}_{\overset{\triangle}{\mathbf{X}}}(\overset{\triangle}{\mathbf{x}}_n)\mathcal{G}_{\mathbf{Y}}(\mathbf{y}_n)}\right). \quad (5.10)$$

Illustrating the results obtained for both $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}})$ and $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}})$, Figure 5.4 shows the latter to be consistently higher. We also find that the difference increases as a function of the amount of memory incorporated into the delta features—represented by the number of neighbouring frames, L, used to calculate the delta features in Eq. (4.34)—reaching saturation at roughly $T = 200\,\text{ms}$, i.e., at the syllabic rate. Since Eqs. (5.9) and (5.10) differ only in terms of the joint-band GMM used to estimate $\hat{I}(\overset{\triangle}{\mathbf{X}};\mathbf{Y})$—i.e., the inclusion of $\boldsymbol{\Delta}_{\mathbf{Y}}$ is restricted to only that term—we are certain that the superior results for $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}})$ compared to $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}})$ in Figure 5.4 are exclusively due to the aforementioned influence of the $\boldsymbol{\Delta}_{\mathbf{Y}}$ training data in shaping the overall joint-band model during maximum-likelihood training.

Despite the superior results for $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}})$ compared to $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}})$, however, Figure 5.4 also shows that $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}},\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}\overset{\triangle}{}})$ is nevertheless still lower than the certainty $C(\mathbf{Y}|\mathbf{X},\mathcal{G}_{\mathbf{XY}})$ of the reference memoryless $(10,0,7,0,7)$ baseline with equivalent $\mathcal{G}_{\mathbf{XC}}/\mathcal{G}_{\mathbf{X\Omega}}$ dimensionality. In other words, the net time-frequency information tradeoff imposed by the chosen delta feature dimensionalities is, in fact, negative. Consequently, the performance of a practical GMM-based BWE system—e.g., our MFCC-based dual-mode BWE system of Section 5.2—that incorporates memory by replacing static spectral envelope features by delta ones per these dimensionalities will *not* improve. An optimization is thus required to find the optimal allocation of the available feature vector dimensionalities in each of the two frequency bands among static and delta features.

[116] See Table 4.3.

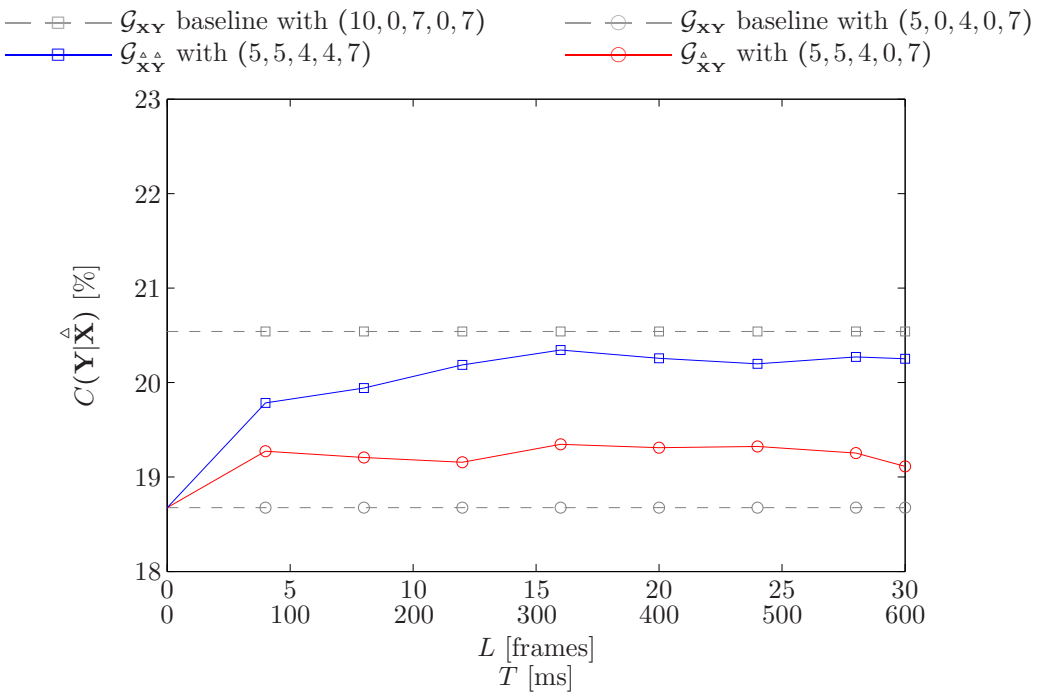


Fig. 5.4: Comparing the effects of memory inclusion using the $\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}}$ and $\mathcal{G}_{\overset{\triangle}{\mathbf{XY}}}$ joint-band GMMs on the MFCC-based static highband certainty, $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}})$, relative to the memoryless $\text{Dim}(\mathbf{X}, \mathbf{\Delta_X}, \mathbf{Y}, \mathbf{\Delta_Y}, \mathbf{Y}_{\text{ref}}) = (10,0,7,0,7)$ and $(5,0,4,0,7)$ baselines.

5.3.4 Optimization of the time-frequency information tradeoff

As discussed in Section 1.2, temporal information in speech has been shown to complement memoryless spectral information. In fact, the works of [167] and [168] on the role of temporal cues—briefly described in Sections A.1 and A.3, respectively—have shown temporal information to be even sufficient to maintain accurate word intelligibility and effective communication when spectral information is missing or severely degraded. However, the tradeoff between information in the time and frequency axes in the context of BWE where perceived quality—rather than intelligibility—is the measure of performance, is much more subtle, particularly at reduced dimensionalities. A case in point is the contrast between voiced and unvoiced fricatives in terms of the importance of frequency and temporal properties relative to each other. Since unvoiced fricatives, e.g., /s/ and /f/, are characterized by nearly flat highband spectra, including long-term memory through narrowband and highband delta features at the cost of reducing the corresponding static spectral features to only

a few parameters, allows the joint-band model to incorporate memory for improved fricative separation during model training. This, in turn, improves their identification during reconstruction while still retaining sufficient spectral information for the accurate reconstruction of their flat spectra. In contrast, a similar reduction in memoryless spectral information for phonemes with finer spectral detail, e.g., voiced fricatives with harmonics imposed on frication noise, may be higher than can be compensated by temporal information.

The objective, then, for a BWE system employing frontend-based memory inclusion through non-invertible features, is to operate at the optimal point of the time-frequency information tradeoff associated with the particular dimensionalities of that system. This optimal point corresponds to the maximum achievable certainty about static highband features, $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}}, \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}})$, resulting in the minimum achievable reconstructed highband spectral distortion as represented by $\downarrow\overline{d}_{\text{LSD(RMS)}}$—the RMS-LSD lower bound obtained by replacing $C(\mathbf{Y}|\mathbf{X}, \mathcal{G}_{\mathbf{XY}})$ in Eq. (4.32) by $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}}, \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}})$. The domain of this optimization problem is the three-dimensional (p, q, L) space of *static* narrowband and highband feature vector dimensionalities, p and q, respectively, and the length L of the window used to calculate delta features, with the optimized function being that of $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}}, \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}})$ or $\downarrow\overline{d}_{\text{LSD(RMS)}}$. Using $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}}, \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}})$, we can thus write the optimal point as the tuple

$$(\overset{*}{p}, \overset{*}{q}, \overset{*}{L}) = \arg\max_{p,q,L} C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}}, \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}), \text{ subject to: } \begin{cases} 1 \le p \le p_{\max}, \\ 1 \le q \le q_{\max}, \\ L \ge 0, \\ p, q, L \in \mathbb{Z}, \end{cases} \tag{5.11}$$

where the upper limits of the constraints imposed on p, q, and L, are determined as described below.

Since Figure 5.4 indicates that $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}}, \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}})$ is not convex at least as a function of L—with two separate maxima at $T \approxeq 320$ and 560 ms—we perform the optimization of Eq. (5.11) empirically. In particular, we estimate $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}}, \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}})$ using marginalization of $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}$ in the same manner as performed in Section 5.3.3.2 above, at (p, q, L) values spanning the constraint ranges of Eq. (5.11). The upper constraint limits are determined such that the fixed-dimensionality constraint of Section 5.3.2 is satisfied while ensuring consistency with our previous approach for the inclusion of delta features. Specifically:

- $p_{\max} = 9$.

As described in Section 4.3.4, the dimensionality of the memoryless baseline $\mathcal{G}_{\mathbf{XY}}$ GMM used for highband certainty estimation was determined as $\text{Dim}(\mathbf{X}, \mathbf{Y}) = (10, 7)$ in order to coincide with the baseline dual-mode BWE GMM tuple dimensionalities given by $\text{Dim}\left(\begin{bmatrix}\mathbf{X}\\ \mathbf{\Omega}_y\end{bmatrix}\right) = \begin{bmatrix}10\\6\end{bmatrix}$ for $\mathcal{G}_{\mathbf{X\Omega}_y}$ (or $\text{Dim}\left(\begin{bmatrix}\mathbf{X}\\ \mathbf{C}_y\end{bmatrix}\right) = \begin{bmatrix}10\\6\end{bmatrix}$ for $\mathcal{G}_{\mathbf{XC}_y}$ in the MFCC case) and $\text{Dim}\left(\begin{bmatrix}\mathbf{X}\\ G\end{bmatrix}\right) = \begin{bmatrix}10\\1\end{bmatrix}$ for $\mathcal{G}_{\mathbf{X}G}$—i.e., 10 narrowband features shared by both $\mathcal{G}_{\mathbf{X\Omega}_y}$ (or $\mathcal{G}_{\mathbf{XC}_y}$) and $\mathcal{G}_{\mathbf{X}G}$, and 7 highband features divided into 6 envelope shape parameters in $\mathcal{G}_{\mathbf{X\Omega}_y}$ (or $\mathcal{G}_{\mathbf{XC}_y}$) and 1 envelope gain parameter in $\mathcal{G}_{\mathbf{X}G}$.

With the inclusion of delta features and focusing only on the MFCC-based parameterization, we represent the dual-mode system's GMM tuple by $\overset{\triangle}{\boldsymbol{\mathcal{G}}} \triangleq (\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}}, \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}})$, with both GMMs sharing the same $\overset{\triangle}{\mathbf{X}}$ narrowband representation and $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}}$ having the maximum dimensionality of 16 per our fixed-dimensionality constraint. Thus, in order for $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}$—the single GMM used in our highband certainty investigation—to coincide with the $(\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}}, \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}})$ tuple in the same manner as described above for the memoryless baseline, a dynamic $\overset{\triangle}{\mathbf{X}}$ narrowband dimensionality of 10 is used. To further conform with our earlier approach for incorporating delta features where priority was given to static envelope gain parameters and their delta features,[117] the minimal inclusion of delta features in each band should consist of a single log-energy delta feature—i.e., $\delta_{c_{x_0}}$ and $\delta_{c_{y_0}}$ for the narrow and high bands, respectively. As such, the maximum dimensionality of *static* narrowband features with the inclusion of delta features is given by $p_{\max} = 9$, in which case the overall dynamic narrowband feature vectors consist of $[c_{x_1}, \ldots, c_{x_8}, c_{x_0}, \delta_{c_{x_0}}]^T$. For $p < p_{\max}$, higher-order static envelope shape parameters—i.e., c_{x_i} where $i > 0$—are replaced by the delta features of shape parameters with increasing order, i.e., for $p = 7$, for example, c_{x_7} and c_{x_8} are replaced by δ_{c_1} and δ_{c_2}, respectively.

- $q_{\max} = 7$.
  Conforming with the memoryless case, highband modelling in the dynamic BWE case is divided among the two $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}}$ and $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}$ GMMs; envelope shape parameters in $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}}$ while those of the gain in $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}$. Given the priority of both static and delta gain parameters as noted above, we include both g and δ_g in the $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}$ model, such that the overall dimensionality of the joint-band space modelled by $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}$ is given by

[117] See Section 4.4.3.1.

$\text{Dim}\left(\begin{bmatrix}\overset{\triangle}{\mathbf{X}}\\ \overset{\triangle}{G}\end{bmatrix}\right) = \begin{bmatrix}10\\2\end{bmatrix}$. Since this dimensionality is still lower than that of $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}}$, the increase relative to the dimensionality of the memoryless $\mathcal{G}_{\mathbf{X}G}$ involves no additional training data requirements.

With the additional highband gain delta feature, the overall dimensionality of the dynamic highband space to be modelled by $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{Y}}}$ for certainty estimation increases by 1, i.e., $\text{Dim}(\overset{\triangle}{\mathbf{Y}}) = 8$. This results in a maximum *static* highband feature dimensionality of $q_{\max} = 7$, in which case highband feature vectors consist of $[c_{y_1}, \ldots, c_{y_6}, c_{y_0}, \delta_{c_{y_0}}]^T$. As for p, higher-order static envelope shape parameters of the high band are replaced by lower-order shape delta features when $q < q_{\max}$.

Figure 5.5 illustrates the results obtained by empirically optimizing Eq. (5.11) over the $5 \le p \le 9$, $4 \le q \le 7$, and $0 \le L \le 30$ ranges, relative to our memoryless baseline with $\text{Dim}(\mathbf{X}, \mathbf{\Delta}_\mathbf{X}, \mathbf{Y}, \mathbf{\Delta}_\mathbf{Y}, \mathbf{Y}_{\text{ref}}) = (10, 0, 7, 0, 7)$.[118] Inspecting the $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}})$ certainty results in Figures 5.5(a)–5.5(e) as a function of L confirms our earlier finding in Section 4.4.3.2 that the effects of memory inclusion on cross-band correlation saturate roughly around $T \approxeq 200$ ms—corresponding to the syllabic rate—regardless of feature vector dimensionalities. Conversely, the effects of the static p and q dimensionalities on highband certainty—i.e., the time-frequency information tradeoff—are evident by comparing the results in Figures 5.5(a)–5.5(e) independently of L. In particular, we conclude that certainty is generally maximized at $\overset{*}{p} = 8$ and $\overset{*}{q} = 6$, i.e., when only one spectral shape delta feature, δ_{c_1}, is included in each band's features in addition to the minimal spectral gain delta feature, δ_{c_0}, with saturation reached at $L \approxeq 8$ corresponding to 160 ms of two-sided memory. The corresponding effect on $\downarrow\overline{d}_{\text{LSD(RMS)}}$, the RMS-LSD lower distortion bound on achievable BWE performance, is shown in Figure 5.5(f). Table 5.2 further summarizes the results obtained using frontend-based memory inclusion at the optimal $(\overset{*}{p}, \overset{*}{q}, \overset{*}{L})$ tuple.

Although the results of Figure 5.5 and Table 5.2 confirm that the substitution of static features by delta ones can indeed improve the highband certainty $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}})$ even with a fixed-dimensionality constraint, the improvements attained relative to the memoryless baseline represent only a small fraction of the significant $C(\overset{\triangle}{\mathbf{Y}}|\overset{\triangle}{\mathbf{X}})$ certainty gains observed in Section 4.4.3.2. While the MFCC-based $(5, 5, 4, 4, 7)$ model of Case S-2 achieves a certainty increase of 77.5% relative to the $(10, 0, 7, 0, 7)$ baseline when all information about the high

[118] See Footnote 77 regarding GMM-derived results.

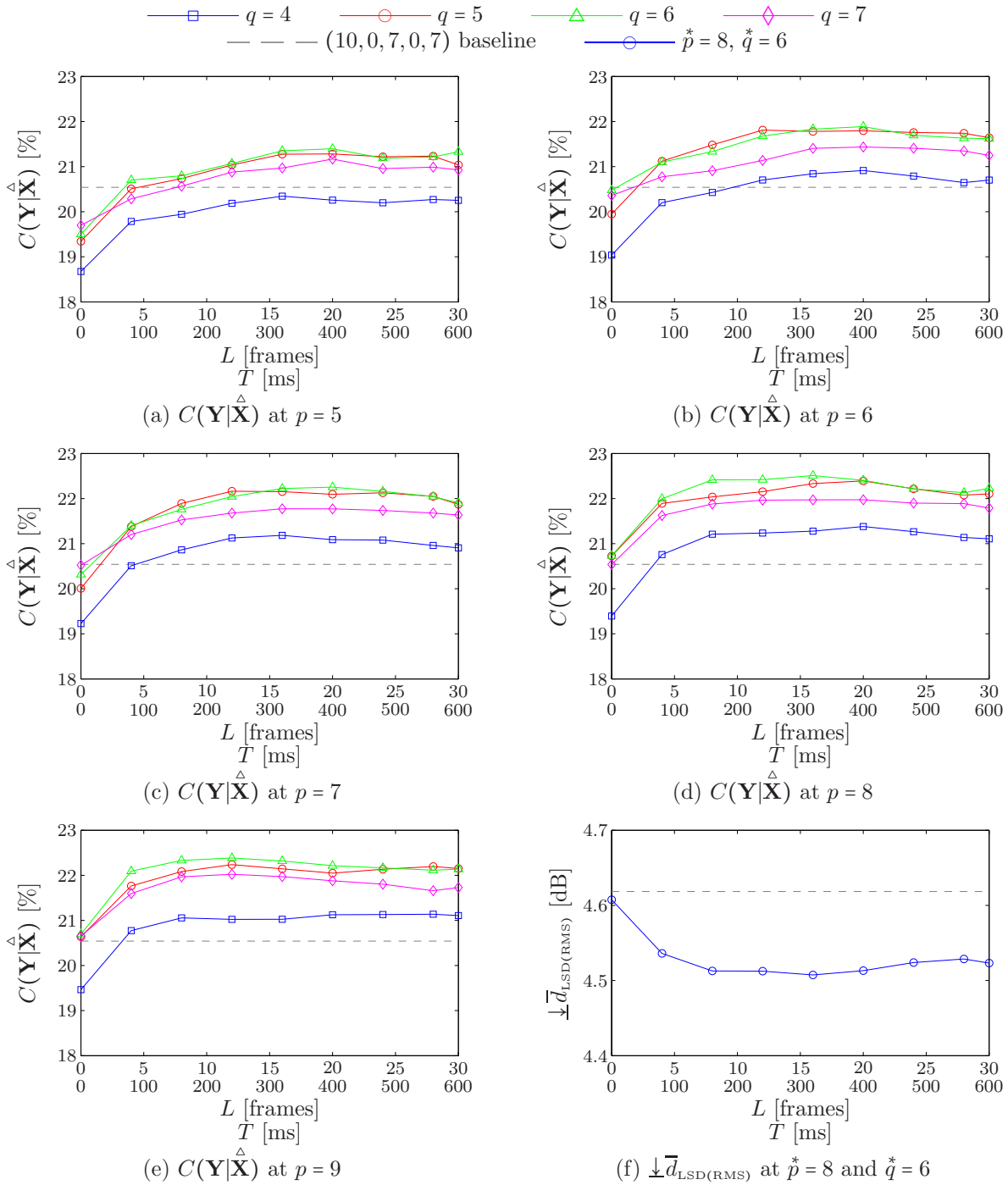


Fig. 5.5: Empirical optimization over the frontend-based memory inclusion's (p, q, L) variable space, relative to the memoryless $\text{Dim}(\mathbf{X}, \mathbf{\Delta}_\mathbf{X}, \mathbf{Y}, \mathbf{\Delta}_\mathbf{Y}, \mathbf{Y}_{\text{ref}}) = (10, 0, 7, 0, 7)$ baseline. Subfigures (a)–(e) show $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}})$ performance for $5 \leq p \leq 9$, $4 \leq q \leq 7$, and $0 \leq L \leq 30$, with the result that $p^* = 8$ and $q^* = 6$. Subfigure (f) shows $\downarrow\overline{d}_{\text{LSD(RMS)}}$ performance against L at p^* and q^*.

Table 5.2: Effect of frontend-based memory inclusion at the optimal $(\overset{*}{p}, \overset{*}{q}, \overset{*}{L})$ value on highband certainty and RMS-LSD lower bound. The Δ_C and $\Delta_{\downarrow \overline{d}_{\text{LSD(RMS)}}}$ differences are estimated relative to the results of the memoryless $(10, 0, 7, 0, 7)$ baseline shown in Table 4.2.

$\overset{*}{p}$	$\overset{*}{q}$	$\overset{*}{L}$	$\max\left[C(\mathbf{Y}\|\overset{\triangle}{\mathbf{X}})\right]$	$\max\left[\frac{\Delta_C}{C(\mathbf{Y}\|\mathbf{X})}\right]$	$\min\left[\downarrow \overline{d}_{\text{LSD(RMS)}}\right]$	$\max\left[\Delta_{\downarrow \overline{d}_{\text{LSD(RMS)}}}\right]$
8	6	16	22.5%	9.6%	4.51 dB	0.11 dB

band—delta as well as static—is taken into account,[119] the optimized $(8, 2, 6, 2, 7)$ model achieves a maximum relative increase of only 9.6% when certainty is estimated based on only static highband features. In terms of the RMS-LSD lower bound on BWE performance, the maximum absolute improvement for the optimized model is only 0.11 dB, compared to 0.82 dB for the model of Case S-2.

Thus, despite their advantages, the non-invertibility of delta features—restricting us to the use of statistical mapping for the implementation of frontend-based memory inclusion—has considerably hampered our ability to convert the information attained about the temporal properties of the high band—represented by $\mathbf{\Delta}_{\mathbf{Y}}$—given the dynamic narrowband representation $\overset{\triangle}{\mathbf{X}}$, into static envelope information that can, in practice, be used for the LP-based reconstruction of highband content. This, in turn, suggests that the BWE performance gains to be obtained as a result of frontend-based memory inclusion—investigated in the following section—are expected to be modest.

In conclusion, we note that as the overall joint-band model dimensionalities change, the optimal $(\overset{*}{p}, \overset{*}{q}, \overset{*}{L})$ tuple changes as well. However, the time-frequency information tradeoff becomes much less of a concern at higher dimensionalities—corresponding to increasingly finer spectral detail—since the advantages gained by the inclusion of temporal information increasingly outweigh the accompanying reductions in static spectral envelope information.

5.3.5 BWE performance with optimized frontend-based memory inclusion

5.3.5.1 System description

With the delta feature inclusion scheme and the subsequent certainty results discussed above, we can now propose an optimized memory-inclusive BWE technique that requires

[119] See Table 4.4.

only minor modifications to our memoryless MFCC-based dual-mode baseline system of Section 5.2. Figure 5.6 illustrates these modifications, namely:

- the integration of delta feature calculation into the parameterization frontend,[120] and,
- the substitution of the memoryless $\mathcal{G} = (\mathcal{G}_{\mathbf{X}\mathbf{C}_y}, \mathcal{G}_{\mathbf{X}G})$ GMM tuple by the dynamic $\overset{\triangle}{\mathcal{G}} = \left(\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}_y}(\overset{\triangle}{\mathbf{x}}, \mathbf{c}_y), \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}(\overset{\triangle}{\mathbf{x}}, g)\right)$ tuple, where $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}_y}(\overset{\triangle}{\mathbf{x}}, \mathbf{c}_y)$ and $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}(\overset{\triangle}{\mathbf{x}}, g)$ represent the GMMs obtained by marginalizing $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}_y}(\overset{\triangle}{\mathbf{x}}, \overset{\triangle}{\mathbf{c}}_y)$ and $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}(\overset{\triangle}{\mathbf{x}}, \overset{\triangle}{g})$ over the $\mathbf{\Delta}_{\mathbf{C}_y}$ and $\mathbf{\Delta}_G$ subspaces, respectively.

With these minor modifications, the MMSE-based reconstruction of highband speech can then be performed using the same formulae previously detailed in Section 3.3.1—namely, Eqs (3.12), (3.16), and (3.17).

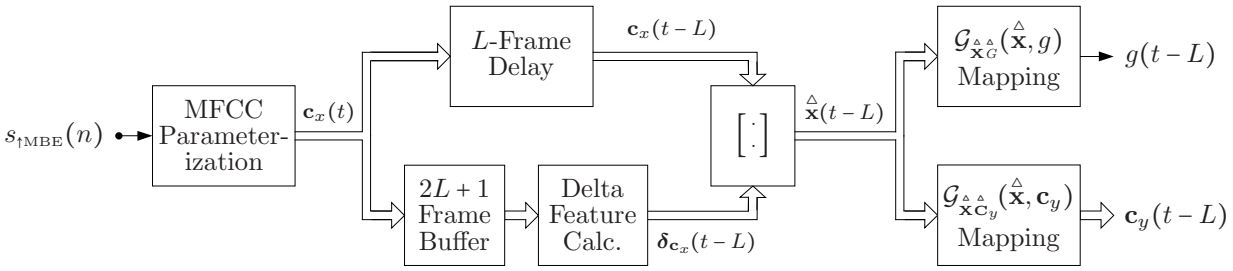


Fig. 5.6: Frontend-based memory inclusion modifications to the baseline MFCC-based dual-model BWE system of Figure 5.1. The modifications are applied to the upper-most path of the main processing block in Figure 5.1(b) and to the GMM-based MMSE estimation block in Figure 5.1(c). With n and t representing the sample and frame time indices, respectively, the input signal, $s_{\uparrow\text{MBE}}(n)$, is that of the midband-equalized and interpolated narrowband speech, while L is the number of neighbouring frames—on each side of the tth static frame being processed—used to calculate delta features.

For the optimized $\text{Dim}(\mathbf{X}, \mathbf{\Delta}_{\mathbf{X}}, \mathbf{Y}, \mathbf{\Delta}_{\mathbf{Y}}, \mathbf{Y}_{\text{ref}}) = (8, 2, 6, 2, 7)$ dimensionalities, the dynamic $\overset{\triangle}{\mathcal{G}}$ GMM tuple has the joint-band dimensionalities of $\text{Dim}(\mathbf{X}, \mathbf{\Delta}_{\mathbf{X}}, \mathbf{C}_y) = (8, 2, 5)$ and $\text{Dim}(\mathbf{X}, \mathbf{\Delta}_{\mathbf{X}}, G) = (8, 2, 1)$ for the marginal $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}_y}(\overset{\triangle}{\mathbf{x}}, \mathbf{c}_y)$ and $\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}(\overset{\triangle}{\mathbf{x}}, g)$ GMMs, respectively. Using Eqs. (4.34), (4.2), and (3.34), the per-frame computational cost of integrating frontend-based memory inclusion during the extension stage as shown Figure 5.6 can, thus, be calculated as—relative to the $(10, 0, 7, 0, 7)$ baseline:

- an additional $L \cdot \text{Dim}(\mathbf{\Delta}_{\mathbf{X}})$ multiplication and $L \cdot \text{Dim}(\mathbf{\Delta}_{\mathbf{X}})$ subtraction operations for the calculation of delta features per Eq. (4.34), for a total of $4L$ additional FLOPs;

[120] See Eq. (4.34).

- a decrease of 58 FLOPs for the calculation of 8 narrowband MFCCs per Eq. (4.2), rather than 10,[121]; and,
- a decrease of $M^{\text{full}}[21]+1$ FLOPs for the MMSE estimation of 5 highband MFCCs per Eq. (3.34), rather than 6—a total decrease of 2689 FLOPs for $M^{\text{full}} = 128$ component densities per GMM as selected in Section 3.5.3.

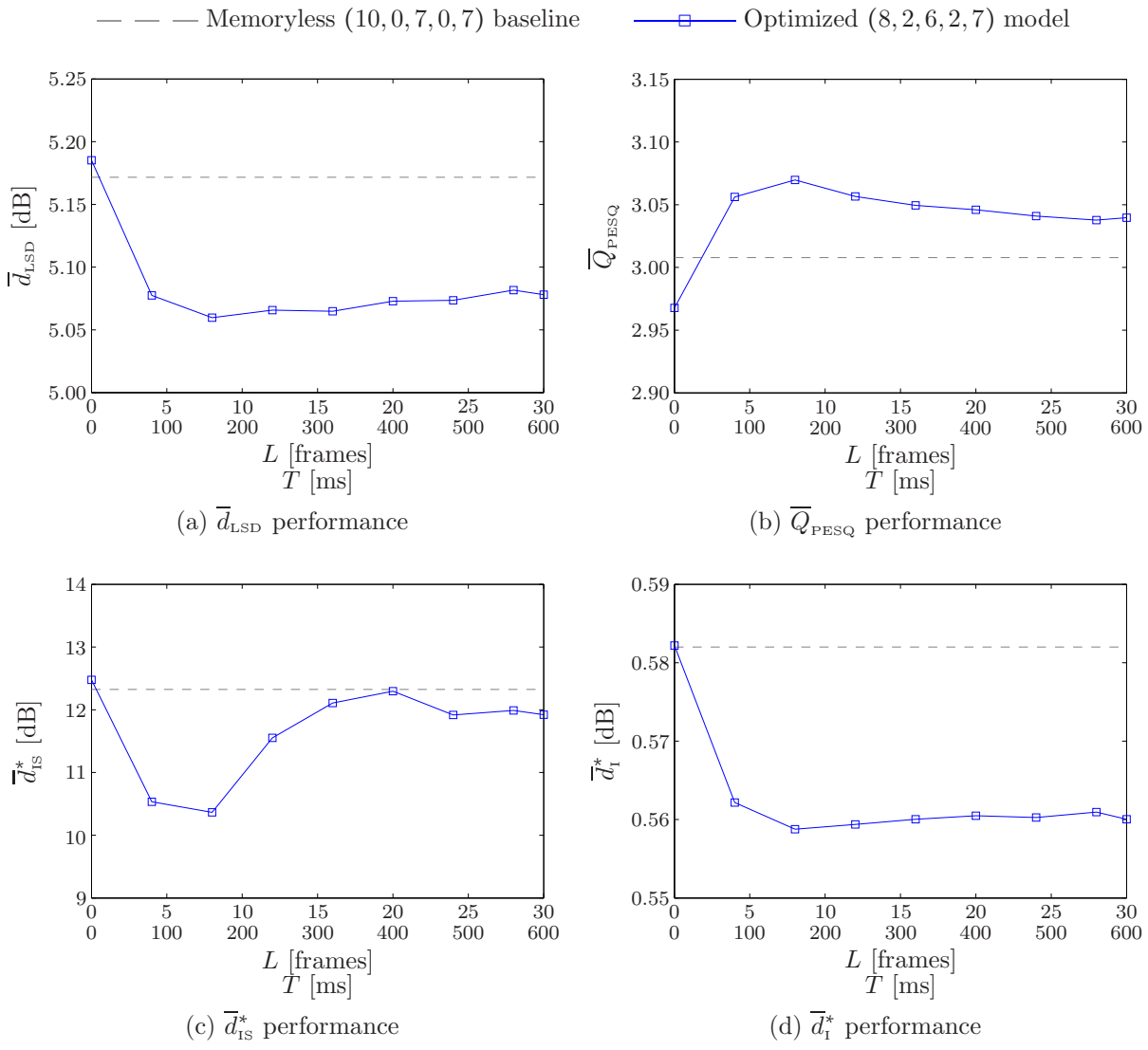
Thus, for all practical and reasonable values of L—the radius of the delta calculation window—including the full $L \in [0, 30]$ range considered in our previous investigations, the inclusion of memory into BWE using delta features with our fixed-dimensionality constraint of Section 5.3.2 results, in fact, in slightly lower run-time computational cost, compared to the memoryless dual-mode baseline system.

More importantly, however, the inclusion of memory via the non-causal delta features as shown in Figure 5.6 imposes an overall algorithmic delay of L frames—corresponding to $10L$ ms given our 10 ms parameterization step discussed in Section 3.2.8. Since real-time two-way speech communication typically requires a maximum 150 ms end-to-end transmission delay, the algorithmic delay due to speech processing should not exceed 20–30 ms in order to guarantee acceptable interactive speech communication when all other sources of latency—namely computational and network delays—are taken into account [169, Section 18.4]. For our modified MFCC-based dual-mode BWE system in Figures 5.1 and 5.6, this corresponds to $L \leq 3$, considerably lower than $L \cong 8$—the point at which the certainty saturation plateau is reached for the optimal $(8, 2, 6, 2, 7)$ model, as shown in Figure 5.5(d). Thus, provided that BWE performance—to be measured in the section below—does, indeed, coincide with our highband certainty results in terms of the effect of L, the ability to realize the full performance improvement potential of our optimal frontend-based memory inclusion scheme would, nevertheless, be limited by network channel and hardware factors. In other words, only under favourable channel and computational hardware latency conditions—allowing a higher algorithmic delay of $L \cong 8$, i.e., 80 ms—can the maximum BWE performance improvements be attained. Finally, it is worth noting that this delay associated with delta feature calculation is, in fact, the only source of algorithmic delay introduced by our memory inclusion modifications discussed above.

[121] In practice, the $a\cos\left(n(k+\frac{1}{2})\frac{\pi}{K}\right)$ terms in Eq. (4.2) are pre-calculated and applied directly during extension-stage run time, while the $\log_e \varepsilon_k$ terms are calculated once per frame during run time. Thus, for $K = 15$ mel-scale filters in the midband-equalized 0–4 kHz narrowband range (see Step 4 of MFCC parameterization in Section 4.2.2), the calculation of each cepstral parameter in Eq. (4.2) requires 15 multiplication and 14 addition operations, for a total of 29 FLOPs.

5.3.5.2 Performance and analysis

Figure 5.7 illustrates the BWE performance obtained for our MFCC-based dual-mode BWE system with frontend-based memory inclusion at the empirically-optimized $(8, 2, 6, 2, 7)$ dimensionalities, as a function of the delta feature calculation window radius, L.[122]



(a) $\overline{d}_{\text{LSD}}$ performance

(b) $\overline{Q}_{\text{PESQ}}$ performance

(c) $\overline{d}^*_{\text{IS}}$ performance

(d) $\overline{d}^*_{\text{I}}$ performance

Fig. 5.7: MFCC-based dual-mode BWE performance with optimized frontend-based memory inclusion, i.e., with $\text{Dim}(\mathbf{X}, \mathbf{\Delta}_{\mathbf{X}}, \mathbf{Y}, \mathbf{\Delta}_{\mathbf{Y}}, \mathbf{Y}_{\text{ref}}) = (8, 2, 6, 2, 7)$, relative to the memoryless $(10, 0, 7, 0, 7)$ baseline.

[122] See Footnote 77 regarding GMM-derived results.

Based on the results of Figure 5.7, we can itemize our findings and conclusions as follows:

- Conforming with our earlier information-theoretic findings in Figure 5.5(d), the inclusion of memory using delta features at the optimal dimensionalities does, indeed, result in an overall BWE performance improvement relative to the memoryless baseline, across all performance evaluation measures, and regardless of the extent of memory used, i.e., L. Since the reconstruction of highband content is based on a lower static highband feature dimensionality as imposed by the fixed-dimensionality constraint, the ability of memory inclusion to provide an overall-beneficial time-frequency information tradeoff in terms of measurable BWE performance is, thus, confirmed.

- As a function of L, the $\overline{d}_{\text{LSD}}$, $\overline{Q}_{\text{PESQ}}$, and $\overline{d}_{\text{I}}^{\,*}$ BWE performances generally mirror the $C(\mathbf{Y}|\overset{\triangle}{\mathbf{X}})$ certainty and $\downarrow\overline{d}_{\text{LSD(RMS)}}$ lower bound performances at the optimal $\overset{*}{p}$ and $\overset{*}{q}$ static feature dimensionalities in Figures 5.5(d) and 5.5(f), respectively, with the $\overline{d}_{\text{LSD}}$ performance, in particular, being a near-perfect match.

- As suggested by the certainty results in Table 5.2 for our empirically-optimized model, the BWE performance improvements achieved by frontend-based memory inclusion are generally modest, reaching their best at $\overset{*}{L} = 8$—the point at which highband certainty reaches saturation in Figure 5.5(d) for $\overset{*}{p} = 8$ and $\overset{*}{q} = 6$. Table 5.3 lists these improvements.

  Table 5.3: Highest BWE performance improvements achieved using frontend-based memory inclusion with $\overset{*}{L} = 8$—corresponding to 160 ms of two-sided memory and 80 ms algorithmic delay—and the optimal MFCC-based $\text{Dim}(\mathbf{X}, \mathbf{\Delta}_{\mathbf{X}}, \mathbf{Y}, \mathbf{\Delta}_{\mathbf{Y}}, \mathbf{Y}_{\text{ref}}) = (8, 2, 6, 2, 7)$ dimensionalities, relative to the memoryless $(10, 0, 7, 0, 7)$ baseline of Table 5.1.

	$\overline{d}_{\text{LSD}}$ [dB]	$\overline{Q}_{\text{PESQ}}$	$\overline{d}_{\text{IS}}^{\,*}$ [dB]	$\overline{d}_{\text{I}}^{\,*}$ [dB]
$(10, 0, 7, 0, 7)$ baseline	5.17	3.01	12.32	0.5820
$(8, 2, 6, 2, 7)$ model	5.06	3.07	10.37	0.5588
Improvement	0.11 (2.2%)	0.06 (2.1%)	1.95 (15.9%)	0.0232 (4.0%)

- Using the knowledge described in Section 3.4 about the perceptual principles underlying the formulation and calculation of all four performance evaluation measures, we can further interpret the results of Figure 5.7 to obtain a more detailed under-

standing of the effect of memory inclusion on the reconstruction accuracy of highband envelopes, as follows:

- As described in Section 3.4.1, the $\overline{d}_{\text{LSD}}$ measures weight all deviations in log spectra equally. The $\overline{Q}_{\text{PESQ}}$ measure, on the other hand, is asymmetric in the sense that it focuses on over-estimation disturbances rather than under-estimations, explicitly employing an *asymmetry factor* in its calculation of perceptual disturbances as described in Section B.1. From the observation that the $\overline{d}_{\text{LSD}}$ and $\overline{Q}_{\text{PESQ}}$ performances in Figures 5.7(a) and 5.7(b), respectively, generally coincide as a function of L, we can then conclude that the extent to which the duration of included memory mitigates over- and under-estimations in highband envelopes is consistent for both types of disturbances across L. In other words, at each particular value for L, memory inclusion mitigates over- and under-estimations by the same relative extent, with the duration of included memory having no effect in terms of favouring the alleviation of one type over the other. Furthermore, the nearly-identical relative $\overline{d}_{\text{LSD}}$ and $\overline{Q}_{\text{PESQ}}$ improvements at $\overset{*}{L} = 8$, as shown in Table 5.3, indicate that, in fact, frontend-based memory inclusion improves envelope over- and under-estimations equally.
- Secondly, as described in Section 3.4.2, the symmetrized $\overline{d}^{*}_{\text{IS}}$ and $\overline{d}^{*}_{\text{I}}$ measures weight larger deviations in log spectra more heavily than does the $\overline{d}_{\text{LSD}}$ measure. As such, the observation that the gain-independent $\overline{d}^{*}_{\text{I}}$ performance in Figure 5.7(d) matches that of $\overline{d}_{\text{LSD}}$ in Figure 5.7(a) as a function of L, indicates that frontend-based memory inclusion mitigates all degrees of deviations in envelope shapes in a consistent manner across L. In other words, at each particular value for L, memory inclusion mitigates all deviations by the same relative extent, with the duration of included memory again having no effect in terms of favouring the alleviation of one type over the other. The larger relative $\overline{d}^{*}_{\text{I}}$ improvement at $\overset{*}{L} = 8$, relative to that of $\overline{d}_{\text{LSD}}$, further indicates that frontend-based memory inclusion is, in fact, more successful in mitigating the more perceptually-relevant larger envelope shape deviations.
- In contrast, the $\overline{d}^{*}_{\text{IS}}$ performance in Figure 5.7(c)—taking into account envelope gain deviations as well as those of the shape—exhibits rapidly-falling performance improvements for $L > 8$. Since, as discussed immediately above, the

similarly-derived but gain-independent $\overline{d}_{\text{I}}^{\star}$ measure shows envelope shape reconstruction to be rather consistent as a function of L, we can conclude that the decline in $\overline{d}_{\text{IS}}^{\star}$ performance for $L > 8$ is attributed solely to the decreased ability of the joint-band MMSE estimation using $\{\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{G}}(\overset{\vartriangle}{\mathbf{x}}, g)\}_{\forall L>8}$ to mitigate large deviations in the reconstruction of the highband envelope gain. We note that this conclusion is independent of those made above regarding the consistency of the $\overline{d}_{\text{LSD}}$ and $\overline{Q}_{\text{PESQ}}$ performances as a function of L since, as mentioned above, both $\overline{d}_{\text{IS}}^{\star}$ and $\overline{d}_{\text{I}}^{\star}$ measures weight envelope deviations rather differently from both $\overline{d}_{\text{LSD}}$ and $\overline{Q}_{\text{PESQ}}$. We should also note that this unexpected inconsistency in addressing large deviations in envelope gain estimation could not be observed through our highband certainty investigation since:

1. As described in Section 4.3.1, the estimation of the mutual information, e.g., $I(\overset{\vartriangle}{\mathbf{X}}; \mathbf{Y})$, is performed using GMM-based likelihoods where feature vector deviations are weighted equally by the relevant GMM inverse covariance, regardless of the extent or direction of the deviation.
2. As described in Section 4.3.2, the estimation of the discrete highband entropy $H(\mathbf{Y})\big|_{\overline{d}_{\text{LSD}}=1\text{dB}}$ using vector quantization treats all deviations of data points from their respective Voronoi centroids equally.
3. While highband envelope gains are modelled in both $\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{\mathbf{Y}}}$ and $\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{G}}$ joint-band GMMs used for certainty estimation and dual-mode BWE, respectively, the excitation gain g—used in $\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{G}}$—represents highband energy rather indirectly through a ratio that depends on the gain in the equalized 3–4 kHz midband range as well as that of the 4–8 kHz high band, whereas c_{y_0}—used in $\mathcal{G}_{\overset{\vartriangle}{\mathbf{X}}\overset{\vartriangle}{\mathbf{Y}}}$—only models the latter.[123] As such, the $\overline{d}_{\text{IS}}^{\star}$ performance of Figure 5.7(c) is particularly sensitive to errors in midband equalization while the certainty evaluations of Figure 5.5 are not.

- As shown in Table 5.4 below, the BWE performance improvements achieved at $L = 4$—corresponding to an algorithmic delay of 40 ms—represent 78–91% of the highest improvements achieved at $\overset{\star}{L} = 8$. As such, despite their modest values, most of

[123] As described in Section 5.2.3, the excitation gain, g, is obtained by artificially synthesizing the highband signal using the EBP-MGN excitation derived from the 3–4 kHz band and the LPCs obtained by high-resolution IDCT of the true 4–8 kHz highband MFCCs.

the improvements obtained using our frontend-based memory inclusion scheme can still be attained in strict or unfavourable conditions of network and computational latencies.

Table 5.4: BWE performance improvements achieved using frontend-based memory inclusion with $L = 4$—corresponding to an algorithmic delay of 40 ms—as a percentage of the maximum improvements of Table 5.3 achieved at $\overset{*}{L} = 8$.

$\overline{d}_{\text{LSD}}$ [dB]	$\overline{Q}_{\text{PESQ}}$	$\overline{d}^{*}_{\text{IS}}$ [dB]	$\overline{d}^{*}_{\text{I}}$ [dB]
84.1%	78.1%	91.4%	85.4%

We have thus proposed a BWE system implementing frontend-based memory inclusion for the purpose of improving BWE performance. Although the presented scheme attains only a fraction of the potential improvements achievable by fully converting the information about highband dynamics—shown in Chapter 4 to be highly-correlated with those of the narrow band—into static envelope information, the modest improvements achieved are obtained with minimal changes to the baseline memoryless BWE system, with no additional run-time computational cost, and with no increase in training data requirements, thereby providing an easy and convenient means for exploiting speech dynamics to improve BWE performance.

5.3.5.3 Comparisons to relevant approaches

As discussed in Section 5.3.1, the inclusion of memory exclusively into the frontend—as implemented in our scheme above—for the purpose of improving BWE performance has been quite limited in both scope and application in the literature. Nevertheless, we attempt to review and interpret the results of the relevant works previously discussed in Section 5.3.1 within the context at hand. To simplify the comparison of performances against our frontend-based memory inclusion approach, we will assume that the test sets used by the cited techniques are sufficiently diverse—phonetically as well as in terms of speaker gender and dialects—such that the results reported therein can be considered general enough for direct comparison against our results in Tables 5.3 and 5.4. In other words, we preclude any effects that the differences in testing data—relative to the TIMIT core test described in Section 3.2.10—may have on the generality, and hence the comparability, of reported performances.

In [129] where a single parameter was used to model the ratio of narrowband signal energy in immediately successive frames, a subjective performance improvement was shown relative to the reference BWE system of [43] employing a spectral folding technique for highband spectral envelope reconstruction. No absolute subjective or objective evaluations were performed in [129]. However, on a customized 7-point absolute superiority scale derived from CMOS results, a relative improvement of approximately 0.48 points was shown for the BWE technique of [129] over that of [43].[124] With the latter system showing an improvement of 0.6 points on the same scale over narrowband speech, combined with a corresponding $\overline{Q}_{\text{PESQ}}$ score increase of 0.2 over narrowband speech as reported in [43], we estimate the 0.48-point improvement of the BWE technique of [129] to correspond to 0.16 $\overline{Q}_{\text{PESQ}}$ points. While this estimated improvement is higher than the 0.06 $\overline{Q}_{\text{PESQ}}$ points shown in Table 5.3 for our technique, it is not exclusively attributed to the aforementioned temporal energy ratio. The estimated improvement of the BWE system of [129] is rather attributed primarily to several structural modifications implemented in order to improve upon the system of [43], most notably the use of neural networks to model cross-band spectral envelope correlations rather than spectral folding, with four mel-scale subband energies representing the 4–8 kHz band. In contrast, our $\overline{Q}_{\text{PESQ}}$ improvement reported in Table 5.3 is, in fact, exclusively attributed to the use of delta features.

Similarly, while both approaches proposed in [87] and [163] employ delta—as well as delta-delta—narrowband features, they rely primarily on first-order HMMs to exploit speech dynamics for improved cross-band correlation modelling. In fact, the delta and delta-delta window radii—i.e., L—used in both works are not even reported, presumably due to the minor role of these feature vector derivatives in the BWE approaches presented therein. Nevertheless, an average 0.28-point $\overline{Q}_{\text{PESQ}}$ improvement is reported in [87] for the proposed HMM-based BWE system—described in detail in Section 2.3.3.4—relative to the baseline system of [60]—also discussed in Section 2.3.3.1—which uses a much-less sophisticated piecewise-linear mapping technique for the estimation of highband envelopes. For the relatively more advanced multi-stream HMM-based system of [163], described in more

[124] An implementation of the comparison category rating (CCR) standard in [30], the comparison mean opinion score (CMOS) involves listener ratings of a processed test sample relative to an unprocessed sample on a range from −3 (much worse) to 3 (much better), with 0 representing similar quality (about the same). The testing procedure differs from that of DCR—see Footnote 17—in that the order of presentation of the two test samples being compared is randomized in CCR, whereas the reference undegraded signal is always presented first to the listeners in DCR.

detail in Section 5.4.1.3 below, larger $\overline{Q}_{\text{PESQ}}$ improvements ranging $\approx$0.6–0.8 points—as well as $\overline{d}_{\text{LSD}}$ improvements ranging $\approx$1.2–1.8 dB—are reported relative to a rather simple BWE system that is based on the single-codebook mapping technique described in Section 2.3.3.2.

In conclusion, we note that, although the performance improvement figures discussed above are superior to those obtained through our technique, these figures result from the joint evaluation of multiple significant system enhancements, rather than from the exclusive evaluation of frontend-based memory inclusion. Furthermore, it is notable that all works cited above use clearly inferior approaches to provide benchmark BWE performances. In contrast, the GMM-based system proposed in [132] is compared against a truly-comparable system, that of [82], with the comparison limited to a single proposed system enhancement—namely, the use of temporal-envelope modelling rather than the ubiquitous source-filter model. Subjective evaluations reported in [132] indicate only a *slight* preference for its proposed technique over that of [82], rather than *dramatic*, as put by the authors. In addition to limiting its modelling of temporal properties to frame-based intervals no longer than 5 ms, however, no objective results were reported in [132], thereby making a comparison to our technique for frontend-based memory inclusion rather difficult.

5.4 BWE with Model-Based Memory Inclusion

In this section, we investigate model-based alternatives to frontend-based memory inclusion. We showed in Section 5.3 above that employing delta features in a practical BWE context is suboptimal in the sense that it only succeeds in translating a modest proportion of the certainty gains achievable by memory inclusion into tangible BWE performance improvements. This followed as a result of the time-frequency information tradeoff imposed by the non-invertibility of delta features. Moreover, as delta features are, by conventional definition, non-causal, they result in an algorithmic delay that limits their usefulness in real-time BWE implementations.

These drawbacks provide the motivation to pursue memory inclusion through a different avenue. In particular, we seek a technique that preserves highband dimensionality, minimizes increases in training data requirements, and further considers only causal memory for the benefit of real-time implementation. Such a technique should also provide flexibility in regards to the extent of memory modelled—the primary advantage of delta features and simultaneously the deficiency of first-order HMM-based methods.

5.4.1 Review of previous works on model-based memory inclusion

5.4.1.1 GMM-based memory inclusion

Among the spectral envelope modelling techniques described in Section 2.3.3, GMMs have been the most successful in BWE due to their superior ability to represent the complex nonlinear cross-band correlations in speech. Aside from the secondary use of GMMs for state-conditional *pdf* modelling in HMM-based BWE implementations, however, the success of GMMs has been restricted to memoryless implementations of BWE. This follows from both computational and algorithmic complications associated with the Expectation-Maximization (EM) GMM training algorithm when used in high-dimensional settings where speech memory is incorporated directly into GMMs by modelling supervectors composed of temporal sequences of feature vectors—rather than just the conventional memoryless vectors corresponding to 10–30 ms frames. In Section 5.4.2.1 below, we discuss these GMM limitations in more detail to provide the insight behind our proposed temporal extension approach to the GMM framework.

5.4.1.2 Neural network-based memory inclusion

As noted in Section 2.3.3.3, neural networks, on the other hand, theoretically allow for such a straightforward means of memory inclusion where narrowband supervectors can be used directly as model inputs, although this particular application of neural networks has not been investigated in the literature. This ability of neural networks to model data with higher dimensionalities follows from their relatively lower computational requirements compared to GMMs.[125] As indicated in our review in Section 2.3.3.3, however, implementations of neural networks in the context of BWE—namely those of [41, 56, 70]—have only resulted in mixed and inconclusive performances relative to other techniques. Although the more recent work of [129] shows modest BWE performance improvements, these improvements are not exclusively attributed to the use of neural networks as discussed in Section 5.3.5.3 above, and secondly, they result from a comparison to the rather simple non-model-based spectral folding technique of Section 2.2.1. Finally, we note that the application of neural

[125] The back-propagation algorithm typically used for neural network training is computationally cheaper than the maximum likelihood-based EM algorithm used for GMMs. Similarly, the run-time feed-forward operation of neural networks during the extension stage is rather simple compared to the MMSE estimation used with GMMs.

networks in all cited techniques has been rather restricted to memoryless BWE.

5.4.1.3 HMM-based memory inclusion

In contrast to approaches based on GMMs and neural networks, approaches applying model-based memory inclusion where modelling is based on hidden Markov models (HMMs), have relatively been more successful. In addition to the detailed review in Section 2.3.3.4, these HMM-based approaches have hitherto been discussed with varying detail throughout the thesis. To our knowledge, all HMM-based approaches proposed in the literature—save the more recent work of [163], described below, as well as the earlier computationally-demanding approach of [84], detailed in Section 2.3.3.4—share the same idea underlying the work in [39] and [87]. To recapitulate the said idea, temporal sequences of narrowband feature vectors are used to train first-order HMMs where states comprise GMMs statistically modelling the narrowband envelopes. Cross-band correlation with highband—or wideband—envelopes is modelled indirectly within the HMM state transition probabilities by tying a VQ codebook of highband—or wideband—feature vectors to the narrowband-specific HMM states. BWE is then performed at run time via an iterative MMSE estimation of highband—or wideband—feature vectors as a function of the state posterior probabilities given the observed sequences of narrowband feature vectors in conjunction with the highband—or wideband—VQ codevectors associated with the HMM states.

Although the performance comparisons reported in [39] and [87] relative to other techniques are rather limited, the 0.28-point $\overline{Q}_{\text{PESQ}}$ objective performance improvement reported in [87]—relative to the piecewise-linear mapping technique of [60]—is nevertheless higher than the improvements reported for non-HMM-based approaches. The more recent HMM-based approach of [163] results in even higher performance improvements. This approach performs temporal clustering of narrowband feature vectors by training a multi-stream set of parallel single-state HMMs on joint narrowband-wideband feature vectors in an unsupervised manner. Using diagonal-covariance GMMs, the trained HMM states can then be split into separate narrowband and wideband models sharing the same state transition probabilities. At run time, sequences of input narrowband feature vectors are temporally segmented using Viterbi decoding [86] on the narrowband model to extract the most likely state sequence. Given the obtained narrowband state sequences, wideband features are then estimated by performing linear prediction on a dimensionality-reduced version of the

time-indexed narrowband features assigned by segmentation into each particular state, with the state-specific wideband feature means—derived from the most likely wideband state sequence corresponding to the narrowband sequence obtained by Viterbi decoding—used as additive bias terms in the linear prediction formulae.

While the approach of [163] improves on those of [39] and [87] by employing joint narrowband-wideband feature vectors for HMM training as well as by employing linear prediction rather than codebook mapping for the estimation of wideband features from the decoded state sequences, it still effectively incorporates memory using first-order HMMs. Thus, it is similar to the earlier HMM-based techniques in that it only accounts for short-term memory—ranging 20–40 ms of memory—through state-to-state and self transitions. Furthermore, using the Viterbi algorithm for state sequence decoding—rather than the real-time MMSE estimation of [39] and [87]—imposes algorithmic delays which limit its effectiveness for real-time BWE tasks. In particular, the Viterbi algorithm requires segmenting speech into blocks within each of which the whole observation trellis must first be accumulated before tracing back in order to determine the optimal state sequence for that particular speech segment.

Notwithstanding the algorithmic delay limitations, the aforementioned modelling improvements proposed in this approach make it more successful in translating the theoretical certainty gains corresponding to such short-term memory into measurable performance gains. In particular, objective $\overline{Q}_{\text{PESQ}}$ and $\overline{d}_{\text{LSD}}$ improvements ranging $\approx$0.6–0.8 points and $\approx$1.2–1.8 dB, respectively, are reported in [163] relative to a memoryless BWE system based on the single-codebook mapping technique described in Section 2.3.3.2.

5.4.1.4 Codebook-based memory inclusion

As described in Section 2.3.3.2, BWE techniques based on codebook mapping are generally quite simpler and much less computationally-demanding than HMM-based approaches in both the training and extension stages. Because of the limitations of codebook mapping in terms of temporal modelling, however, its application has been mostly restricted to memoryless BWE implementations. Two notable exceptions where the dynamics of speech are incorporated into the codebook-based mapping are the works of [130] and [131].

In the relatively early approach of [130], codebook-based classification is performed in three steps. Starting with an N-sized wideband feature vector codebook tied to a

similarly-sized shadow narrowband codebook as explained in Section 2.3.3.2, M wideband codevectors—where $1 < M < N$—corresponding to the M narrowband codevectors nearest to the narrowband input vector are selected. In the second step, the M potential wideband codevectors are further reduced to L—where $1 < L < M$—based on the cepstral distances of the M codevectors from the final wideband feature vector estimate obtained for the preceding frame. Finally, implementing the codevector interpolation technique described in Section 2.3.3.2, the L codevectors are linearly combined with weights based on the sums of the distances calculated for each of the L wideband codevectors in the two earlier classification steps. This approach thus improves upon conventional codebook-based techniques by incorporating memory—albeit only at the limited interframe level—into its estimation of wideband envelopes. Informal subjective evaluations reported in [130] show improved wideband signal quality due to the inclusion of memory in the second classification step. No formal subjective or objective results were presented, however.

Rather than incorporate interframe memory in the classification stage as described above, the approach of [131] incorporates such memory directly into codebook design and training using an extension of predictive VQ—a special case of memory VQ[126] [37]. In particular, a codebook is trained on a linear combination of two quantities calculated at each speech frame: (a) the difference between the current narrowband feature vector and a weighted version of the quantized or unquantized narrowband vector of the preceding frame—corresponding to closed-loop or open-loop prediction, respectively, and (b) the quantized or unquantized highband feature vector of the preceding frame. Despite the inclusion of memory only at the interframe level, it is reported in [131] that the use of predictive VQ results in an objective $\overline{d}_{\text{LSD}}$ performance improvement of 0.45 dB for the reconstructed highband signal, relative to conventional memoryless VQ with the same codebook size.

5.4.1.5 Non-HMM state space-based memory inclusion

To conclude this review, we note the insightful approach of [133] where a linear state space model treats narrowband feature vectors as the linear observations resulting from linearly-evolving hidden states representing the unknown wideband feature vectors. However, because of the assumption that narrowband and wideband feature vectors are linearly related, and since speech dynamics can not be all modelled by a single linear model, this

[126] See Footnote 23.

state space approach requires a large number of modes—where each mode is a different set of values for the linear model's parameters—with the model changing its mode every L frames. Parameters of the state space model are estimated at every L-frame mode using the forward recursion of the Kalman filter algorithm [170, Chapter 10]. With a frame step of 10 ms, values of $L \in [10, \ldots, 50]$—corresponding to 100–500 ms of memory—were investigated in [133]. This approach, thus, accounts for considerably longer-term speech memory than any of the other techniques discussed thus far. Moreover, as a result of the sequential nature of the Kalman forward recursion, it introduces no algorithmic delays.

With speech processed in blocks of $L = 30$ frames, i.e., modelling up to 300 ms of memory, this state space approach is reported in [133] to achieve objective $\overline{d}_{\text{LSD}}$ performance improvements of ≈ 0.06, 0.36 and 0.69 dB, relative to HMM-, GMM-, and codebook-based systems based on those of [171], [82], and [59], respectively. Thus, despite the considerably higher complexity of this approach relative to that of HMM-based systems as well as the longer-term memory it incorporates, it only succeeds in achieving modest performance improvements. Furthermore, in contrast to HMM-based approaches, it suffers from discontinuity effects resulting from the abrupt transitions between modes across the boundaries of the L-frame blocks.

5.4.2 Temporal-based extension of the GMM framework

5.4.2.1 On the limitations of GMMs in high-dimensional settings

For the purpose of incorporating memory into BWE speech modelling, a straightforward extension to the successful memoryless joint-band GMM-based approach is to directly expand the modelled feature vector space along temporal axes, whereby the conventional memoryless narrowband—and, optionally, highband or wideband—feature vectors used for model training and extension are replaced by supervectors consisting rather of temporal sequences of such memoryless feature vectors. As discussed in Section 5.4.1.1, however, the multiple-fold increase in dimensionality associated with using such supervectors—assuming that spectral resolution, i.e., memoryless feature vector dimensionality, is to be preserved—not only prohibitively increases the computational as well as data requirements associated with GMM training via the EM algorithm, but also results in severely degraded estimates for GMM parameters.

This *curse of dimensionality* follows as a direct result of the increase in parameters

required to model each mode of the temporally-extended multi-modal feature vector distribution (or *pdf*), as well as indirectly as more Gaussian kernels become required in order to adequately model the increase in the number of modes—the underlying acoustic classes.[127,128,129] Specifically, the exponential increase in the degrees of freedom of the GMM-based model, relative to the increase in dimensionality,[130] leads to the problems of *oversmoothing* and *overfitting*, which have been investigated in the fields of machine learning and speaker conversion in particular.

Oversmoothing refers to the effect where the spectral characteristics of the MMSE target data estimated via Eqs. (3.12), (3.16), and (3.17)—i.e., the MMSE estimates, $\{E[\mathbf{Y}|\mathbf{x}]\}$, of the highband feature vectors given those of the narrow band and the joint-band $\mathcal{G}_{\mathbf{XY}}$ model—are excessively smoothed due to the near-elimination of the source-data contribution given by the second term in Eq. (3.17), resulting, in turn, in low-quality highband speech signals. The near-elimination of the narrowband source-data contribution itself follows as a result of the tendency of the $\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}$ covariance ratios to decrease—in determinant or norm—with increasing dimensionality, with the result that, nearly regardless of the source data, $\mathbf{X}$, the variation in MMSE-estimated $\mathbf{Y}$ target vectors is minimal with the vectors scarcely differing from the $\boldsymbol{\mu}_i^{\mathbf{y}}$ means in Eq. (3.17).[131]

Also typically associated with increases in dimensionality, overfitting results from the disproportionate increase in the degrees of freedom allowed by a GMM-based model relative to the available amounts of training data. As dimensionality increases, the volume of the underlying space increases exponentially such that the available data becomes sparse. Such sparsity undermines the statistical reliability of the EM algorithm since it will often converge to a significantly suboptimal local maximum for the data's likelihood, which, in

[127] The term *curse of dimensionality* was coined by Bellman in [172].

[128] As discussed in Section 3.3.4, the increase in number of underlying acoustic classes itself follows from the additional degrees of freedom introduced along temporal axes.

[129] While the increase in feature vector dimensionalities also adversely affects runtime computational complexity, the effect is much less pronounced than that on training-stage complexity. In particular, we have shown in Section 3.5.1 that most of the computationally-demanding matrix operations associated with MMSE estimation can be performed offline, such that runtime complexity is reduced from $O(p^3)$ for full-covariance GMMs and narrowband dimensionality p, to $O(p^2)$, per Eq. (3.34).

[130] As shown in the right-hand-side denominator of Eq. (3.18), the number of parameters, N_{p}, of a full-covariance GMM, is related to the dimensionality, D, by $N_{\mathrm{p}} \propto D^2$.

[131] In the context of speaker conversion, Chen et alia showed in [159], for example, that—for a 40-dimensional $\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}$ square correlation matrix obtained for log spectrum features transformed via mel-scale DCT—more than 90% of the $\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}$ matrix elements are smaller than 0.1, and more than 40% are smaller than 0.01.

turn, voids the model of its generalization capability. The challenge then becomes finding the optimal balance between restricting a highly-dimensional GMM's degrees of freedom to avoid overfitting, while simultaneously ensuring the availability of sufficient degrees of freedom to adequately model the underlying modes, or classes, of the *pdf* being modelled.

As described in Section 4.4.2, an approach proposed and applied in [126] and [149] to circumvent the high-dimensionality limitation of GMMs is to employ dimensionality-reducing transforms in the frontend rather than to incorporate memory within GMMs themselves. Most notable of these transforms are those of linear discriminant analysis (LDA) and the Karhunen-Loève transform (KLT)—although LDA was only applied in [126] to reduce static feature vector dimensionalities, rather than to reduce those of temporal-based supervectors. Despite their well-known advantages in the context of classification, however, these transforms suffer the same time-frequency information tradeoff of delta features, thereby limiting their usefulness for practical memory-inclusive BWE.

Alternatively, several approaches have been proposed in the speaker conversion and machine learning literature to address the oversmoothing and overfitting problems. The common idea underlying these approaches is to impose some constraints on the parameters of a high-dimensional GMM in order to reduce the allowed degrees of freedom, to impose minimum thresholds on variances, or both. Approaches intended for the speaker conversion task address both problems by constraining the source-data contribution weights—i.e., $\{\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}\}_{i\in\{1,\ldots,M\}}$—themselves, as in [159] and [160],[132] for example, or by constraining the target-data covariances—i.e., $\{\mathbf{C}_i^{\mathbf{yy}}\}_{i\in\{1,\ldots,M\}}$—alone, as in [161].

In the context of machine learning where GMMs—referred to as *Gaussian graphical models* in the graphical model subcontext [173]—have been by far the most popular means of mixture model-based density estimation and clustering [174, Section 6.8], no source-target Gaussian-based transformations are involved. Thus, approaches concerned with GMM-based clustering in high-dimensional settings have only focused on addressing the problem of overfitting through constraining—or regularizing—GMM mean vectors [154], covariances [155]—$\{\mathbf{C}_i^{\mathbf{z}}\}_{i\in\{1,\ldots,M\}}$, where $\mathbf{Z} = \begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix}$, in our source-target context—or inverse covariance matrices [156]. Generally, the constraints imposed by regularization on an ill-posed problem are equivalent to incorporating or introducing prior information in order to achieve well-posedness, thereby allowing finding accurate approximate solutions to the

[132] In [159], the $\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}$ covariance ratios are assumed to be diagonal identity matrices, whereas in [160], they are tied to a global diagonal covariance.

problem.[133] In [156], for example, where sparsity is induced into the GMM through ℓ_1—or *lasso*—regularization [157], the introduced information is that the L_1-norm of the solution does not exceed a particular threshold. Thus, the regularization approaches cited above also modify the conventional implementation of the EM algorithm for GMMs in order to incorporate the added constraints.

Finally, we note that GMMs have also been used for the related task of subspace clustering where the objective is to localize the search for clusters in the high-dimensional space to lower-dimensionality subspaces along the most relevant dimensions, thereby circumventing many of the problems associated with the curse of dimensionality.[134] In this context, prior information is introduced through various means of regularization such that the parameters of the Gaussian kernels representing the subspace clusters are controlled by the dimensionalities of the potential latent factor spaces to be searched. In [158], for example, regularization is applied by tying subspace orientations—as defined by the Eigen space of the GMM covariances—or by tying the covariances themselves. Similar to the GMM regularization approaches described above for clustering in general, GMM-based subspace clustering techniques involve adapting the EM algorithm.

5.4.2.2 Integrating memory into GMMs through a state space approach

In the discussion above, we have identified the foremost flaw precluding the practicality and value of the aforementioned approach to incorporating memory into GMMs using simple extensions of the GMM modelling space along temporal axes—i.e., by simply modelling supervectors composed of temporal sequences of static feature vectors. In particular, it is practically impossible using such an approach to compile sufficiently large yet diverse amounts of training data in order to compensate for the continuous increases in the model's degrees of freedom associated with the attempt to model increasingly higher orders of feature vector memory—i.e., the attempt to model higher-dimensional supervectors corresponding to longer sequences of feature vectors.

[133] As defined by Hadamard in [175], the problem of solving the mapping $A: X \rightarrow Y$ for A is well-posed if: (a) a solution exists for every y, i.e., $\forall y \in Y$, $\exists x \in X$ such that $Ax = y$, (b) the solution is unique, i.e., if $Ax_1 = Ax_2$, then $x_1 = x_2$, and (c) the solution is stable, i.e., A^{-1} is continuous.

[134] As described in [176], subspace clustering is motivated by the fact that many of the dimensions for high-dimensional data are often irrelevant. These irrelevant dimensions confuse conventional clustering algorithms by hiding the underlying clusters in noisy data. In very high dimensions, it further becomes common for all the objects in a data set to be nearly equidistant from each other, thereby completely masking the clusters.

In comparison, however, we have also shown above how several approaches in the speaker conversion and machine learning domains have successfully addressed the dimensionality-related problems of Gaussian mixture modelling—namely through incorporating prior information into the modelling paradigm in the form of constraints or regularization. From this perspective, we can then characterize the flaw of the aforementioned GMM temporal extension approach more accurately as *the attempt to model high-dimensional feature vector distributions—in all dimensions simultaneously—without exploiting any prior knowledge about the properties of speech underlying these distributions*. Specifically, this GMM extension approach makes no use of the structure inherent in speech beyond the conventional quasi-stationary 10–30 ms frame durations. By quantifying the temporal information in speech in Chapter 4, we showed, however, that the structure of speech does, in fact, exhibit considerable predictability that extends to much longer durations. Consequently, if such considerable information about the structure of speech—in the form of temporal sequences of feature vectors of quasi-stationary segments—were to be properly exploited to constrain the degrees of freedom in the high-dimensional GMMs to be learned, the complications described in Section 5.4.2.1 above—namely those of oversmoothing and overfitting—could then be successfully mitigated.

Based on this analysis and inspired by the speaker conversion and machine learning techniques previously described, we have developed a novel temporal-based GMM extension approach that exploits the information and predictability in the structure of speech in a progressive manner in order to arrive at a model for the target high-order distributions at the desired temporal depth—i.e., the desired extent of memory inclusion. First proposed in [177], our approach essentially transforms the temporally-extended high-dimensional GMM-based modelling problem into a time-frequency state space modelling task with interpretations in the contexts of subspace and hierarchical clustering, [178] and [174, Section 14.3.12], respectively, as well as graphical model inference [179]. The crux of the approach is to effectively utilize and combine two previously-discussed and well-known properties of speech and GMMs:

The correspondence of GMM component densities to underlying acoustic classes

In Sections 2.3.3.4 and 3.3.4, we addressed the correspondence of the kernels—or component densities—of multi-modal Gaussian mixture models to the acoustic classes underlying the feature vector distributions being modelled. Indeed, as described in

Section 5.4.2.1 above, it is this very correspondence that provides the motivation for the use of GMMs as a generative approach to clustering as well as subspace clustering. In Section 5.3.3.2, we made use of this correspondence—in conjunction with the temporal information incorporated by delta features—to improve the ability of joint-band GMMs to model extensions of the original memoryless acoustic classes along temporal axes. In our model-based approach presented here, we exploit this correspondence as a means by which to *partition* or *cluster* training data into data subsets with varying degrees of overlap corresponding to the underlying complex and overlapping acoustic classes, with the data in each subset further assumed to be independent and identically distributed (i.i.d.). Stated alternatively, we use the aforementioned correspondence to *fuzzily* quantize the memoryless and temporally-extended feature vector spaces into overlapping frequency—in reference to the spectral characteristics specific to each acoustic class—and time-frequency regions, respectively. This, in combination with the strong correlation properties of neighbouring speech frames, allows us to break down the infeasible task of estimating increasingly higher-dimensional *pdf*s—where, for each particular order of temporal extension, a single multi-modal *pdf* modelled by a GMM spans the entire temporally-extended feature vector space—into a series of time-frequency-localized *pdf* estimation operations with considerably lower complexity and fewer degrees of freedom.

The strong correlation between neighbouring speech frames

As a result of the slow vocal tract movements relative to typical speech sampling rates, neighbouring speech frames exhibit a strong correlation. Indeed, as noted in Section 1.2, typical phonetic events last more than 50 ms, with rapid spectral changes being limited to stop onsets and releases or to phone boundaries involving a change in manner of articulation. This redundancy or predictability in speech has been exploited extensively for the purpose of coding speech at rates much lower than those of standard PCM.[135] We also indirectly made use of this property in our earlier frontend-based approach to memory inclusion; as shown in Eq. (4.34), delta features attempt to maximize their information content by increasingly emphasizing spectral differences at larger temporal separation. In our approach presented here, we employ the strong

[135] See [10, Table 7.2] for a comparison between a wide range of speech coders in terms of quality, bit rate, complexity, and frequency of use.

correlation between neighbouring frames in two ways. First, we exploit the correlation of the data with their past frames by carrying over time-frequency localization information obtained at a particular order of memory inclusion as described above into the process of *pdf* estimation at higher orders. As such, we progressively make use of and build upon the information obtained about the underlying time-frequency classes with increasing orders of memory inclusion, in order to better estimate the more difficult higher-dimensional *pdf*s at higher orders of memory inclusion. Secondly, as described below, we make use of the redundancy in feature vectors across time in order to limit the number of Gaussian kernels needed to model the *pdf* of each time-frequency state following the application of temporal extension. Conceptually, this is similar to the removal of speech redundancies in speech coding in order to maximize the information content of the available coding bits.

Depicting our application of these two properties, Figure 5.8 illustrates a state space representation of our proposed approach. Using the previous $\mathbf{X}$ and $\mathbf{Y}$ notations for the static—i.e., memoryless—narrowband and highband feature vectors, respectively, we temporally extend the spectral information in both bands by defining the feature vector sequences $\mathbb{X}_t^{(\tau,l)} = [\mathbf{X}_t^T, \mathbf{X}_{t-\tau}^T, \ldots, \mathbf{X}_{t-l\tau}^T]^T$ and $\mathbb{Y}_t^{(\tau,l)} = [\mathbf{Y}_t^T, \mathbf{Y}_{t-\tau}^T, \ldots, \mathbf{Y}_{t-l\tau}^T]^T$, with τ representing the *memory inclusion step*—the step, in number of frames, between the static frames included in a sequence—and l representing the *memory inclusion index*, or order—the number of past frames incorporated into a sequence in addition to the reference frame. With no temporal extension, i.e., $l = 0$, the feature vector sequences $\mathbb{X}^{(\tau,0)}$ and $\mathbb{Y}^{(\tau,0)}$ correspond to the conventional memoryless static vectors.[136]

Starting with the memoryless joint-band GMM, $\mathcal{G}_{\mathbf{XY}}$, which we now rewrite as $\mathcal{G}_{\mathbb{X}^{(\tau,0)}\mathbf{Y}}$, we progressively incorporate narrowband as well as highband memory—by extending the feature vector sequences $\mathbb{X}^{(\tau,l)}$ and $\mathbb{Y}^{(\tau,l)}$ using past frames at steps of τ—into the estimation of the Gaussian-based model of the now-temporally-extended feature vector *pdf* in steps, with each step corresponding to an increment of the index l. After each such step, the end result is a new GMM, $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}} := \mathcal{G}(\mathbb{x}^{(\tau,l)}, \mathbf{y}; M^{(l)}, \mathrm{A}^{(l)}, \Lambda^{(l)})$, modelling the temporally-extended $\mathbb{X}^{(\tau,l)}$ feature vector space jointly with the reference memoryless $\mathbf{Y}$

[136] In the sequel, we model time-frequency spaces of features vector sequences where each sequence is considered in isolation independently of its absolute temporal location within a speech signal. As such, we drop the time subscript t from all representations to follow unless otherwise needed for clarifying or disambiguating the temporal properties of one representation relative to another.

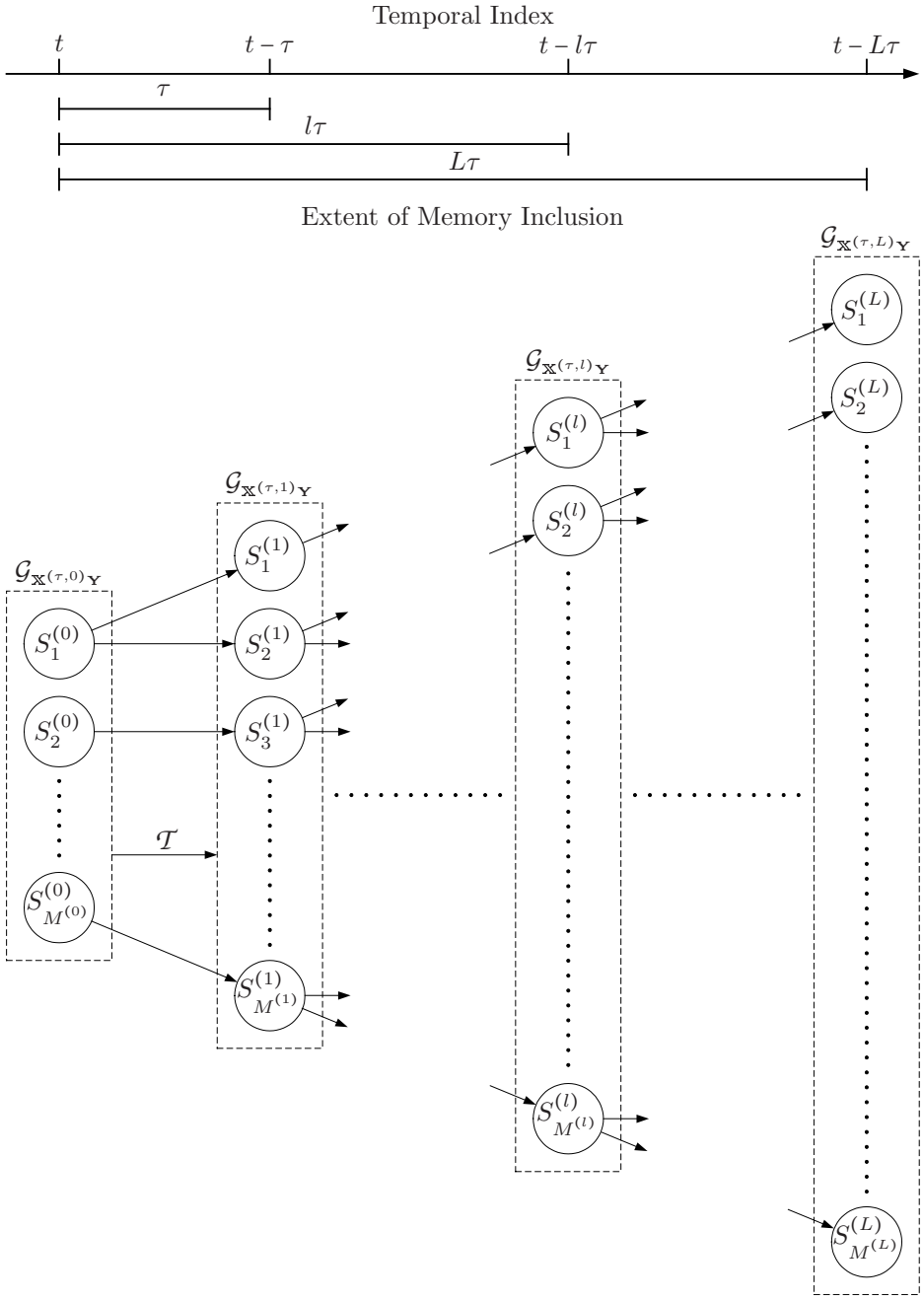


Fig. 5.8: A state space representation of our approach to the inclusion of memory into the GMM framework. Temporally-extended GMMs, given by $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}} \coloneqq \mathcal{G}(\mathbb{X}^{(\tau,l)}, \mathbf{y}; M^{(l)}, \mathrm{A}^{(l)}, \Lambda^{(l)})$, where $l = 0, \ldots, L$, model sequences of $l+1$ narrowband feature vectors—with step τ—jointly with their non-extended highband counterparts. The time-frequency states $\{S_i^{(l)}\}_{i\in\{1,\ldots,M^{(l)}\}}$—corresponding to the GMM kernels given by the tuples $\{(\alpha_i^{(l)}, \lambda_i^{(l)})\}_{i\in\{1,\ldots,M^{(l)}\}}$—are viewed as parent states at memory inclusion index l, with each of which extended into one or more child states at index $l+1$ through the transformation $\mathcal{T}$.

space. At the extension stage, the $\mathbb{X}^{(\tau,l)}$ features—to be used as the MMSE estimation input to $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}$—are readily available from the BWE system's causal narrowband speech input. Thus, unlike the non-causal $\mathbf{\Delta}_{\mathbf{X}}$ features, the computation of $\mathbb{X}^{(\tau,l)}$ features involves no algorithmic delay.

As previously described, incorporating memory into the GMM-based model merely through the temporal extension of feature vectors into sequences of time-indexed vectors followed by conventional stand-alone GMM training—i.e., independently of any previous information already incorporated into the GMMs trained for lower orders of memory inclusion—is computationally unsustainable, as well as practically flawed, at increasing orders of memory inclusion. Instead, we exploit the information previously incorporated into each GMM at a particular memory inclusion index to facilitate the temporal extension of the model into a new GMM at the immediately higher order of memory inclusion, while simultaneously ensuring the reliability, accuracy, and generalization capability of the extended GMMs. To that end, we employ the correspondence of GMM Gaussian components to underlying acoustics classes to identify time-frequency regions—or states—characterized by distinct static and/or dynamic acoustic properties. In particular, as illustrated in Figure 5.8, Gaussian kernels of a temporally-extended GMM $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}$—given by the tuples $\{(\alpha_i^{(l)}, \lambda_i^{(l)})\}_{i\in\{1,\ldots,M^{(l)}\}}$—are treated as distinct uni-modal time-frequency states $\{S_i^{(l)}\}_{i\in\{1,\ldots,M^{(l)}\}}$. We represent this correspondence by $\{S_i^{(l)} \triangleq (\alpha_i^{(l)}, \lambda_i^{(l)})\}_{i\in\{1,\ldots,M^{(l)}\}}$. Given the strong correlation we previously demonstrated between neighbouring speech frames, these distinct states, derived at a particular memory inclusion index l, can then be viewed as *parent* states from which the localized time-frequency information can be used to infer finer *children* states at the higher $(l+1)$th index. In this manner, the overall GMM-based *pdf* at index $l+1$, i.e., $\mathcal{G}_{\mathbb{X}^{(\tau,l+1)}\mathbf{Y}}$, can then be estimated by linearly combining all child state *pdf*s obtained at index $l+1$, rather than estimating it anew independently from the lower-order GMM, $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}$.

This time-frequency state-specific extension or *growth* approach illustrated in Figure 5.8, becomes intuitive when the underlying classes are viewed from the multi-dimensional spatial perspective of the temporally-extended feature vector space. Since the underlying classes represented by the states $\{S_i^{(l+1)}\}_{i\in\{1,\ldots,M^{(l+1)}\}}$ at memory inclusion index $l+1$ can be viewed as finer realizations of the lth-order temporally-extended acoustic representation of speech along a new additional temporal axis, these classes at index $l+1$ are, in fact, subclasses of those at l. Conversely, the lth-order classes represented by $\{S_i^{(l)}\}_{i\in\{1,\ldots,M^{(l)}\}}$ can be viewed as

the lower-resolution subspace projections of the $(l+1)$th-order classes onto the temporally-extended subspace at memory inclusion index l. This incremental approach for partitioning increasingly high-dimensional feature vector spaces by building upon partitions in their lower-dimensional subspaces is further motivated by the observation that real-world high-dimensional data tend to concentrate in subspace manifolds with dimensionalities lower than that of the original space [158]. We also note that, conceptually, this hierarchy of temporally-extended classes/states across time is similar to that described in Section 3.3.4 for memoryless acoustic classes, except that the hierarchy for the latter is rather a function of the number of memoryless GMM components; classes corresponding to phonemes can be viewed as subclasses of those representing place of articulation, which, in turn, are subclasses of those representing manner of articulation.[137]

Another intuitive interpretation of our approach is that obtained from the perspective of top-down—or divisive—hierarchical clustering [174, Section 14.3.12]. In particular, Figure 5.8 can be viewed as a top-down *dendrogram* where the *root* nodes at memory inclusion index $l = 0$ represent rough clusters of the fully-extended $\begin{bmatrix}\mathbb{X}^{(\tau,L)}\\ \mathbf{Y}\end{bmatrix}$ joint-band data with the clustering performed by applying a *distance metric* only to the $\begin{bmatrix}\mathbb{X}^{(\tau,0)}\\ \mathbf{Y}\end{bmatrix}$ data. By further considering the new $\mathbb{X}^{(\tau,l)}$ data available with each increment of l, the $\begin{bmatrix}\mathbb{X}^{(\tau,l-1)}\\ \mathbf{Y}\end{bmatrix}$ clusters are split into finer and more accurate *daughter* clusters.[138] Depending on the *linkage criterion* used to measure the similarity—or lack thereof—of the new incremental features within each parent cluster, the variability of the incremental data may not warrant splitting, in which case the dendrogram branch is extended by simply augmenting the data samples assigned to the parent cluster with their respective incremental features. As described in the following section, we use GMM-based measures for the distance metric as well as for the linkage criterion.

To summarize, we *grow* our model in steps across time in a tree-like fashion—starting from the memoryless $\mathcal{G}_{\mathbb{X}^{(\tau,0)}\mathbf{Y}}$—until the desired level of memory inclusion—denoted by L, the maximum value for l—is achieved. The exact means by which parent states are extended into child states—represented by the transformation $\mathcal{T}$ in Figure 5.8—is detailed

[137] See Table 1.1.

[138] As detailed in Section 5.4.2.3, we, in fact, use both $\mathbb{X}^{(\tau,l)}$ and $\mathbb{Y}^{(\tau,l)}$ features to split a parent $\begin{bmatrix}\mathbb{X}^{(\tau,l-1)}\\ \mathbf{Y}\end{bmatrix}$ cluster into daughter clusters at order l. After estimating the localized $\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ \mathbb{Y}^{(\tau,l)}\end{bmatrix}$ *pdf*s corresponding to the daughter clusters in an intermediate step, the marginal $\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ \mathbf{Y}\end{bmatrix}$ *pdf*s are then extracted.

in the following section. We note here, however, that the validity and the success of such a transformation relies on the aforementioned correlation between neighbouring frames. Our second use of the redundancy in speech frames is also depicted in Figure 5.8 by the variability in number of child states per parent state. Detailed in the following sections, incorporating such variability in our tree-like modelling approach is intended to model the variations in the range of spectral changes across time for different classes, while simultaneously taking advantage of redundancies across time to simplify our temporal Gaussian-based model and maximize its information content.[139] Thus, in a manner akin to the GMM regularization approaches described in Section 5.4.2.1, we use the information already incorporated into lower-order GMMs—namely, the information between neighbouring speech frames as well as that represented by the correspondence of Gaussian kernels to underlying classes—to constrain the complexity and parameter space of the higher-order GMMs.

As noted in Section 5.4.2.1 above, GMMs have long represented the most popular means for mixture model-based clustering [174, Section 6.8], with the vast majority of techniques employing the correspondence of Gaussian components to underlying classes in order to perform a *hard-decision* Bayesian classification of data. This hard-decision discretization approach of the feature vector space discards the degree of overlap between the classes modelled by the *mixture* model, and hence, its classification performance depends heavily on the actual amount of overlap between the underlying classes. As described above and further detailed in Section 5.4.2.3 below, we exploit the same idea underlying GMM-based clustering to group training data at each memory inclusion index l into $M^{(l)}$ time-frequency data subsets corresponding to the states $\{S_i^{(l)}\}_{i\in\{1,\ldots,M^{(l)}\}}$ shown in Figure 5.8. Viewing these subsets as realizations of the distinct time-frequency classes in the space of the temporally-extended joint-band random feature vector at index l, we can then use such subsets—after extending them temporally—to estimate a transformation $\mathcal{T}$ in order to temporally extend the parent state uni-modal *pdf*s—representing time-frequency classes at index l—into multi-modal children *pdf*s that represent new finer states and subclasses at index $l+1$, with the transformation performed for each parent state independently of all other states at the same memory inclusion index l. Since the speech time-frequency classes underlying these parent states do, in fact, overlap considerably, with the extent of overlap further increasing

[139] Temporally-extended classes corresponding to vowels, for example, exhibit much less spectral variability throughout the durations of the vowels, while plosives, on the other hand, are characterized by short intervals of rapid spectral change preceded and followed by longer intervals of considerably lower spectral variation across time.

with dimensionality per the *empty space phenomenon*,[140] using the aforementioned conventional hard-decision approach would result in data subsets that are increasingly limited in terms of their representation of the underlying overlapping classes, and hence, increasingly insufficient for the reliable estimation of child subclasses—i.e., leading to a higher risk of overfitting. This follows from the increasing importance of Gaussian tails in higher dimensional spaces as densities become more spread out, combined with the fact that the *zero-one loss function* underpinning Bayes' decision rule discards information in such tails regarding the extent of class overlap.[141] We illustrate this effect through a simple example in Section 5.4.2.3 below.

Instead of the conventional hard-decision Bayesian classification, we thus propose and employ a novel *fuzzy* approach to GMM-based clustering. While the idea of fuzzy, or soft, mixture-based clustering is itself not new,[142] our proposed algorithm is novel in that it introduces a *fuzziness factor* to selectively control GMM-based classification fuzziness, with the soft membership weights associated to input data by clustering normalized in a manner that ensures the probabilistic consistency of the resulting partitioned subsets regardless of the value used for the fuzziness factor. In effect, our proposed algorithm thus improves upon the blanket fuzziness employed by the Expectation-Maximization (EM) GMM training algorithm—where *all* classes in the mixture partly share the membership of all data points—by incorporating the selectiveness of the well-known non-GMM-based fuzzy K-means approach of [183]. More specifically, we relax the conventional conditions defining the class membership of data points to include data from K neighbouring clusters—rather than from all clusters—in a qualitative manner. Careful choices for the fuzziness factor, K, allow us to partially alleviate the adverse effects of class overlap in higher-dimensional spaces while still allowing us to break down the estimation of the temporal extension transformation into localized time-frequency regions centred near the high-density means of the subspace parent classes. The *selective fuzziness* of our classification approach—partly inspired by the relative success of the notion of fuzzy pattern classification in general—thus represents a compromise between: (a) minimizing the risk of overfitting, and (b) maximizing

[140] Aptly illustrated in [180], the empty space phenomenon refers to the fact that high-dimensional spaces are inherently sparse. As dimensionality increases, distances between points in the space tend to be more uniform, with the result that densities become more spread out, and hence, increasingly overlapping.

[141] See [71, Sections 2.2–2.4] for a detailed description of Bayesian decision theory.

[142] See [181] for a wide and detailed literature review of fuzzy pattern recognition techniques in general, as well as [182] for a review including fuzzy mixture-based clustering in particular.

the ability to compartmentalize, and hence simplify, the task of modelling high-dimensional distributions by reducing the size of the data subsets to be used for estimating child state *pdf*s. Introducing greater overlap in data subsets increases the training computational cost as well as the size of the resulting temporally-extended GMMs, while discarding the underlying overlap altogether will likely result in overfitting. Despite the data subset overlap introduced by our fuzzy clustering approach, we show in Section 5.4.2.3 that the qualitative technique by which we expand the time-frequency data subsets does not, in fact, increase the risk of oversmoothing.

To incorporate the *soft* data classification resulting from our proposed fuzzy clustering approach into the aforementioned estimation of child state *pdf*s, we also propose and derive in Section 5.4.2.3 a *weighted* implementation of the conventional EM algorithm used for GMM training. In particular, we derive iterative update formulae taking account of the soft membership weights such that a weighted log-likelihood function is maximized, and further prove the convergence of our iterative weighted algorithm. Similar to the idea underlying our fuzzy GMM-based clustering algorithm proposed above, however, the idea underlying our weighted EM implementation—namely, incorporating weights that quantify the importance of training data points relative to each other—is itself not novel. Indeed, several weighted implementations of the EM algorithm have previously been proposed in the literature to address training data limitations in terms of number or unevenness, e.g., [184], or, among others, to improve the speed of EM convergence, e.g., [185]. As motivated and detailed in Operation (c) of Section 5.4.2.3 below, however, our proposed weighted EM implementation differs from previous EM approaches in introducing a two-stage training approach that allows us to target, or localize, the density estimation power of the EM algorithm towards any particular subspace of interest—e.g., those subspaces underlying highband feature vectors that, relative to an arbitrary time-indexed reference point, occur at varying instances, or indices, in the past. In contrast, conventional and weighted EM implementations encountered in the literature treat all dimensions of the spaces underlying the input training data equally in terms of density estimation.

By implementing our weighted EM-based density estimation independently for each of the fuzzily clustered parent data subsets, its computational complexity is significantly reduced compared to the infeasible approach of performing stand-alone conventional EM as previously described; first, the EM training procedure inherits the time-frequency localization inherent in the corresponding parent data subsets, thereby considerably restricting the

number of Gaussian components—representing child states—needed to model the localized variability of training data, and secondly, the update formulae themselves can be applied to potentially much smaller amounts of training data. In Section 5.4.2.4, we examine the performance of our fuzzy GMM-based clustering approach combined with weighted EM-based density estimation by assessing the reliability of the final obtained temporally-extended GMMs, $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}\}$, in terms of both oversmoothing and overfitting.

To conclude, we note that, in addition to the correspondence of the idea underlying our tree-like growth approach to those of subspace clustering techniques, the state space representation of Figure 5.8 closely resembles that of a directed graphical model [179]. In particular, we demonstrate in Section 5.4.2.3 below that the states $\{S_i^{(l)}\}_{\forall i,l}$ can be viewed as graphical model nodes, each of which representing a variable in a linear vector subspace of $\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ \mathbf{Y}\end{bmatrix}$, the global temporally-extended joint-band feature vector space, with the subspace variable's *pdf* given by $(\alpha_i^{(l)}, \lambda_i^{(l)})$. Moreover, we show that the conditional independence properties of these variables follow the definition of *Markov blankets*.[143]

5.4.2.3 Implementation

Having presented above a conceptual description of our state space tree-like GMM extension approach, we now describe the details of its implementation.

As described above, we incorporate memory into the joint-band Gaussian mixture model incrementally starting with the memoryless GMM, $\mathcal{G}_{\mathbb{X}^{(\tau,0)}\mathbf{Y}}$, resulting in the set $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}\}_{l\epsilon\{0,\ldots,L\}}$ where $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}$ represents the temporally-extended GMM obtained at the lth step and τ represents the frame step used in the construction of data in $\mathbb{X}^{(\tau,l)}$ and $\mathbb{Y}^{(\tau,l)}$—the temporally-extended narrowband and highband feature vector spaces, respectively. Through quantitatively measuring the effect of memory inclusion on highband certainty in Section 4.4.3, we showed, however, that incorporating the spectral dynamics of both bands into joint-band modelling clearly outperforms incorporating the dynamics of only the narrow band in terms of the certainty gains achievable about the target static highband spectra. Concisely stated in Eq. (5.8), it is, indeed, such joint-band inclusion of memory that represented the basis of our frontend-based approach to improving BWE. Reiterating our conclusion from Section 5.3.3.2, the objective then in the context herein is to achieve the best possible estimates of the underlying temporally-extended *joint-*

[143] See Footnote 161 for the definition of Markov blankets.

band distributions where the temporal extension is applied to the representations of *both* bands. Accordingly, we implement our state space approach in the joint-band spaces of $\left\{\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ \mathbb{Y}^{(\tau,l)}\end{bmatrix}\right\}_{l\in\{0,\ldots,L\}}$, rather than those of $\left\{\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ \mathbf{Y}\end{bmatrix}\right\}_{l\in\{0,\ldots,L\}}$, with the subspace models to be used for BWE—i.e., $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}\}_{l\in\{0,\ldots,L\}}$—extracted by marginalization in a post-processing step. As such, we define $\mathbb{Z}^{(\tau,l)}$, representing the lth-order temporally-extended joint-band feature vector space with step τ, i.e., $\mathbb{Z}^{(\tau,l)} = [\mathbf{Z}_t^T, \mathbf{Z}_{t-\tau}^T, \ldots, \mathbf{Z}_{t-l\tau}^T]^T$ where $\mathbf{Z}_t = \begin{bmatrix}\mathbf{X}_t\\ \mathbf{Y}_t\end{bmatrix}$.

At each extension step, we perform five main operations. In order to simplify the presentation, we first detail these operations individually before describing how we apply and integrate them together:

(a) Fuzzy GMM-based clustering of training data

As described in Section 5.4.2.2 above, we break down the difficult GMM temporal extension task at each memory inclusion step into simpler time-frequency-localized extension operations by exploiting the correspondence of Gaussian kernels to underlying time-frequency classes. This is achieved by progressively clustering temporally-extended joint-band training data—representing realizations in the $\mathbb{Z}^{(\tau,l)}$ vector spaces for $l \in \{0,\ldots,L\}$—into overlapping subsets with the clustering performed as a function of joint-band GMM components, thereby taking advantage of the considerable cross-band correlation of temporal information shown earlier in Section 4.4.3.

Let $\mathcal{G}_{\mathbb{Z}^{(\tau,l)}{}_i}$ represent a *localized* GMM modelling the *pdf* underlying a subset

$$\mathcal{V}_i^{\mathbb{Z}^{(\tau,l)}} \subseteq \mathcal{V}^{\mathbb{Z}^{(\tau,l)}}, \tag{5.12}$$

where $\mathcal{V}^{\mathbb{Z}^{(\tau,l)}}$ represents the set of all training data in the $\mathbb{Z}^{(\tau,l)}$ space, and $i \in \mathcal{I}^{(l)}$—an integer index set given by $\mathcal{I}^{(l)} = \{1,\ldots,|\mathcal{I}^{(l)}|\}$. Per our GMM notation introduced in Eq. (2.13), $\mathcal{G}_{\mathbb{Z}^{(\tau,l)}{}_i}$ is given by $\mathcal{G}_{\mathbb{Z}^{(\tau,l)}{}_i} = \mathcal{G}(\mathbb{z}^{(\tau,l)}; M_i^{\mathbb{Z}^{(\tau,l)}}, \mathrm{A}_i^{\mathbb{Z}^{(\tau,l)}}, \Lambda_i^{\mathbb{Z}^{(\tau,l)}})$. To simplify notation in the sequel, however, we drop the memory inclusion step τ from notation—unless required for clarity—since τ is assumed to be fixed in the presentation below, thus rewriting $\mathcal{V}_i^{\mathbb{Z}^{(\tau,l)}}$ as $\mathcal{V}_i^{\mathbb{Z}^{(l)}}$, for example. In addition, we rewrite the GMM $\mathcal{G}_{\mathbb{Z}^{(\tau,l)}{}_i}$ as $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$ to make the notation below consistent in the sense that l can be viewed as denoting an incremental index of temporal extension applicable to the underlying feature vector space as well as to the quantities being estimated. As such, we write $\mathcal{G}_{\mathbb{Z}_i}^{(l)} := \mathcal{G}_{\mathbb{Z}^{(\tau,l)}{}_i} = \mathcal{G}(\mathbb{z}^{(l)}; M_i^{\mathbb{Z}^{(l)}}, \mathrm{A}_i^{\mathbb{Z}^{(l)}}, \Lambda_i^{\mathbb{Z}^{(l)}})$, where $\Lambda_i^{\mathbb{Z}^{(l)}} = \{\lambda_{ij}^{\mathbb{Z}^{(l)}} := (\boldsymbol{\mu}_{ij}^{\mathbb{Z}^{(l)}}, \mathbf{C}_{ij}^{\mathbb{Z}\mathbb{Z}^{(l)}})\}_{j\in\mathcal{J}_i^{(l)}}$,

$\mathrm{A}_i^{\mathbb{z}^{(l)}} = \{\alpha_{ij}^{\mathbb{z}^{(l)}} := P(\lambda_{ij}^{\mathbb{z}^{(l)}})\}_{j\in\mathcal{J}_i^{(l)}}$, and $\mathcal{J}_i^{(l)} = \{1,\dots,M_i^{\mathbb{z}^{(l)}}\}$.[144]

Given the correspondence of the $\Lambda_i^{\mathbb{z}^{(l)}}$ kernels of $\mathcal{G}_{\mathbb{z}_i}^{(l)}$ to localized classes in the time-frequency $\mathbb{Z}^{(l)}$ space, we further localize the temporal extension task by partitioning the data in the *parent* subset $\mathcal{V}_i^{\mathbb{z}^{(l)}}$ into $|\mathcal{J}_i^{(l)}| = M_i^{\mathbb{z}^{(l)}}$ *child* subsets, $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$, corresponding to the kernels $\{\lambda_{ij}^{\mathbb{z}^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$. As described in Section 5.4.2.2 above, GMM-based clustering approaches—e.g, [154–156]—typically follow Bayesian decision theory to determine the class membership of data, where classification is performed in a hard-decision manner using the maximum *a posteriori* probabilities of the underlying classes—represented by the $\{\lambda_{ij}^{\mathbb{z}^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$ component Gaussians—given the data; i.e.,

$$\forall m \in \mathcal{J}_i^{(l)}\colon \quad \mathcal{V}_{im}^{\mathbb{z}^{(l)}} = \left\{ \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}} \colon \arg\max_{\lambda_{ij}^{\mathbb{z}^{(l)}}\in\Lambda_i^{\mathbb{z}^{(l)}}} P(\lambda_{ij}^{\mathbb{z}^{(l)}}|\mathbb{z}_n^{(l)}) = \lambda_{im}^{\mathbb{z}^{(l)}} \right\}, \tag{5.13}$$

where $n \in \{1,\dots,|\mathcal{V}^{\mathbb{z}^{(l)}}|\}$ indexes all training data points available in the $\mathbb{Z}^{(l)}$ space. As shown in Section 3.3.1, applying Bayes' rule per Eq. (3.13) for GMMs results in the posterior probabilities given by Eq. (3.16)—rewritten for the variables herein as

$$P(\lambda_{ij}^{\mathbb{z}^{(l)}}|\mathbb{z}_n^{(l)}) = \frac{\alpha_{ij}^{\mathbb{z}^{(l)}}\,\mathcal{N}(\mathbb{z}_n^{(l)};\boldsymbol{\mu}_{ij}^{\mathbb{z}^{(l)}},\mathbf{C}_{ij}^{\mathbb{z}\mathbb{z}^{(l)}})}{\sum_{k=1}^{M_i^{\mathbb{z}^{(l)}}}\alpha_{ik}^{\mathbb{z}^{(l)}}\,\mathcal{N}(\mathbb{z}_n^{(l)};\boldsymbol{\mu}_{ik}^{\mathbb{z}^{(l)}},\mathbf{C}_{ik}^{\mathbb{z}\mathbb{z}^{(l)}})}. \tag{5.14}$$

Classifying data as such results in pairwise-disjoint $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$ subsets, where

$$\bigcup_{j\in\mathcal{J}_i^{(l)}}\mathcal{V}_{ij}^{\mathbb{z}^{(l)}} = \mathcal{V}_i^{\mathbb{z}^{(l)}}, \tag{5.15a}$$

$$\forall j,k \in \mathcal{J}_i^{(l)} \text{ and } j \neq k\colon\ \mathcal{V}_{ij}^{\mathbb{z}^{(l)}} \cap \mathcal{V}_{ik}^{\mathbb{z}^{(l)}} = \phi. \tag{5.15b}$$

As previously discussed, however, the classification error—or Bayes risk—associated with Bayes' decision rule of Eq. (5.13) increases with greater overlap in the underlying

[144] As detailed in Operations (c) and (d) below, the subscript i in $\mathcal{J}_i^{(l)}$ is intended to denote the dependence of the number of Gaussian kernels—$|\mathcal{J}_i^{(l)}|$—in the GMM $\mathcal{G}_{\mathbb{z}_i}^{(l)}$ on the particular index of the GMM; i.e., $|\mathcal{J}_i^{(l)}|$ is not a fixed value for all i.

classes, and is particularly exacerbated with increasing dimensionality as a result of the accompanying increase in data sparsity. More importantly for our task, the hard-decision classification increases the risk of overfitting in higher-dimensional spaces since it results in subsets that are increasingly insufficient to reliably estimate the child subclasses of the parent underlying classes corresponding to $\Lambda_i^{\mathbb{z}^{(l)}}$. Thus, to mitigate this dimensionality effect, we relax the hard-decision classification rule of Eq. (5.13) by qualitatively including all data points for which the likelihood of the class in question—i.e., $P(\lambda_{im}^{\mathbb{z}^{(l)}}|\mathbb{z}_n^{(l)})$—is, not only the highest as in Eq. (5.13), but also among the top K, where $1 \leq K \leq M_i^{\mathbb{z}^{(l)}}$.

Let $\overset{*}{\lambda}{}_{ij_k,n}^{\mathbb{z}^{(l)}}$, where $j_k \in \mathcal{J}_i^{(l)}$ and $k \in \{1,\dots,K\}$, denote the kth most-likely class for the nth data point, $\mathbb{z}_n^{(l)}$; i.e.,

$$\overset{*}{\lambda}{}_{ij_k,n}^{\mathbb{z}^{(l)}} := \begin{cases} \arg \max\limits_{\lambda_{ij}^{\mathbb{z}^{(l)}} \in \Lambda_i^{\mathbb{z}^{(l)}}} P(\lambda_{ij}^{\mathbb{z}^{(l)}}|\mathbb{z}_n^{(l)}), & \text{for } k = 1, \\ \arg \max\limits_{\lambda_{ij}^{\mathbb{z}^{(l)}} \in \Lambda_i^{\mathbb{z}^{(l)}} - \{\overset{*}{\lambda}{}_{ij_1,n}^{\mathbb{z}^{(l)}},\dots,\overset{*}{\lambda}{}_{ij_{k-1},n}^{\mathbb{z}^{(l)}}\}} P(\lambda_{ij}^{\mathbb{z}^{(l)}}|\mathbb{z}_n^{(l)}), & \text{for } 1 < k \leq K \leq M_i^{\mathbb{z}^{(l)}}. \end{cases} \tag{5.16}$$

Then, rather than partition data based on only $\overset{*}{\lambda}{}_{ij_1,n}^{\mathbb{z}^{(l)}}$, for each $\mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}$, as in Eq. (5.13), which we now simplify as

$$\forall m \in \mathcal{J}_i^{(l)}\colon \quad \mathcal{V}_{im}^{\mathbb{z}^{(l)}} = \left\{\mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\colon \overset{*}{\lambda}{}_{ij_1,n}^{\mathbb{z}^{(l)}} = \lambda_{im}^{\mathbb{z}^{(l)}}\right\}, \tag{5.17}$$

we relax the conditions for class membership by considering the top K most-likely classes; i.e.,

$$\forall m \in \mathcal{J}_i^{(l)}\colon \quad \mathcal{V}_{im}^{\mathbb{z}^{(l)}} = \bigcup_{k=1}^{K} \left\{\mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\colon \overset{*}{\lambda}{}_{ij_k,n}^{\mathbb{z}^{(l)}} = \lambda_{im}^{\mathbb{z}^{(l)}}\right\}. \tag{5.18}$$

This expands the resulting $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)}}\}_{j \in \mathcal{J}_i^{(l)}}$ subsets *quantitatively*, or spatially, and introduces overlap as each training data point is now assigned to K different subsets—i.e., Eq. (5.15b) no longer holds while Eq. (5.15a) still does. Hence, we will refer to K as the *fuzziness factor*.

Since the class memberships of data are now non-unique, a set of K soft continuous *membership weights* must also be attached to each data point as measures of the degree by which the data point belongs to each of the K underlying classes; points

near the boundaries of a class belong to that class to a lesser degree than those near its centre. This notion of soft membership contrasts with the hard binary memberships underlying the conventional Bayesian decision rule of Eqs. (5.13) and (5.17). Given the intuitive probabilistic nature of such membership weights [186, Section 6.2.2], we use the posterior probabilities of Eq. (5.14) as membership weights after adequate normalization. In particular, let $w_{ij_k,n}^{(l)}$ represent the membership weight associated with $\overset{*}{\lambda}{}_{ij_k,n}^{\mathbb{z}^{(l)}}$, the kth most-likely class in the ith time-frequency region represented by the GMM $\mathcal{G}_{\mathbb{z}_i}^{(l)}$, given the nth data point, $\mathbb{z}_n^{(l)}$.[145] Then, we define $w_{ij_k,n}^{(l)}$ as

$$w_{ij_k,n}^{(l)} = \frac{P(\overset{*}{\lambda}{}_{ij_k,n}^{\mathbb{z}^{(l)}}|\mathbb{z}_n^{(l)})}{\sum_{m=1}^{K} P(\overset{*}{\lambda}{}_{ij_m,n}^{\mathbb{z}^{(l)}}|\mathbb{z}_n^{(l)})}, \tag{5.19}$$

where $k \in \{1,\ldots,K\}$. In addition to ensuring that membership weights for any particular data point always sum to 1, we note that, for $K = 1$ where our fuzzy clustering approach reduces to that based on Bayes' decision rule, Eq. (5.19) results in the desired binary membership weights. We also note that, as shown by the illustrative example of Figure 5.9 below, incorporating the weights of Eq. (5.19) into child density estimation enables us to balance mitigating the risk of overfitting against increased computational cost through the fuzziness factor, K.

Weighting class memberships per Eq. (5.19) renders the subset quantitative expansion of Eq. (5.18) a *qualitative* one as well. This is necessary in order to preserve distinctions between the expanded subsets—i.e., prevent them from being similar—as well as to reduce overall classification error rate rather than increase it if quantitative expansion through Eq. (5.18) alone is applied. Indeed, the illustrative example described below shows that introducing subset overlap—and hence multiple class memberships for data—without accounting for a *degree* of membership *lobotomizes*—or oversmoothes—the resulting subsets.

We have thus partitioned the base lth-order parent subset, $\mathcal{V}_i^{\mathbb{z}^{(l)}}$, into $M_i^{\mathbb{z}^{(l)}}$ over-

[145] We note that the $\mathbb{z}$ superscript is dropped from the notation for membership weights since, as described in Operation (c) below, $\mathbb{z}_n^{(l)}$, $\mathbb{x}_n^{(l)}$, $\mathbb{y}_n^{(l)}$, $\mathbf{x}_{n,t}$, $\mathbf{y}_{n,t-l\tau}$, et cetera, are all time-frequency representations referenced to the same nth wideband speech data point with reference time t, and hence, should share the same weight for membership in any particular underlying time-frequency class.

lapping child subsets, $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$, with corresponding sets of membership weights $\{\mathcal{V}_{ij}^{w^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$ given by

$$\forall m \in \mathcal{J}_i^{(l)}\colon \quad \mathcal{V}_{im}^{w^{(l)}} = \bigcup_{k=1}^{K}\left\{w_{ij_k,n}^{(l)}\colon \overset{*}{\lambda}{}_{ij_k,n}^{\mathbb{z}^{(l)}} = \lambda_{im}^{\mathbb{z}^{(l)}},\ \forall \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\right\}. \tag{5.20}$$

For easier reference in the sequel, we will often combine the pairs of corresponding $\mathcal{V}_{ij}^{\mathbb{z}^{(l)}}$ and $\mathcal{V}_{ij}^{w^{(l)}}$ sets—given by Eqs. (5.18) and (5.20), respectively—through the pairwise-disjoint sets of unique $(\mathbb{z}_n^{(l)}, w_{ij_k,n}^{(l)})$ tuples, given by

$$\forall m \in \mathcal{J}_i^{(l)}\colon \quad \mathcal{V}_{im}^{\mathbb{z}^{(l)},w^{(l)}} = \bigcup_{k=1}^{K}\left\{(\mathbb{z}_n^{(l)}, w_{ij_k,n}^{(l)}) \in \left(\mathcal{V}_i^{\mathbb{z}^{(l)}}, (0,1]\right)\colon \overset{*}{\lambda}{}_{ij_k,n}^{\mathbb{z}^{(l)}} = \lambda_{im}^{\mathbb{z}^{(l)}}\right\}. \tag{5.21}$$

The net result of this fuzzy classification approach is that the child time-frequency state *pdf*s to be estimated at index $l+1$—i.e., in the $\mathbb{Z}^{(l+1)}$ space—based on the $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)},w^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$ subsets, will better account for the overlap between the underlying time-frequency classes in the $\mathbb{Z}^{(l)}$ subspace when $K > 1$, and hence, ultimately result in a better model for the ith time-frequency-localized region of the $\mathbb{Z}^{(l+1)}$ space represented by the data in $\mathcal{V}_i^{\mathbb{z}^{(l+1)}}$, at the cost of increased computations.

To conclude, we also note that, in the development presented above, we did not explicitly incorporate the previously-obtained localization information—i.e., the information represented by the $(l-1)$th-order membership weights in the $\{\mathcal{V}_i^{w^{(l-1)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets, associated with the $\{\mathcal{V}_i^{\mathbb{z}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ parent subsets—into the construction of the new lth-order $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)},w^{(l)}}\}_{\forall i,j}$ child subsets. In particular, the $(l-1)$th-order membership weight information does not explicitly appear in Eqs. (5.19) or (5.21). As will be discussed below in Operations (c) and (d), however, this $(l-1)$th-order information is incorporated, rather implicitly, through the maximum *weighted* log-likelihood estimation of the $|\mathcal{J}_i^{(l)}| = M_i^{\mathbb{z}^{(l)}}$ component Gaussians—i.e., $\{\lambda_{ij}^{\mathbb{z}^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$, themselves—of each $\mathcal{G}_{\mathbb{z}_i}^{(l)}$ model used to obtain $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)},w^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$ as shown above.

The advantage of fuzzy clustering: An illustrative example

To demonstrate the soft-decision advantage of our fuzzy clustering technique in terms of improving child state *pdf* estimation, we consider a simple single-child density estimation problem. Let X represent a scalar random variable with a true underlying distribution given by the highly-overlapping 7-component GMM, $\mathcal{G}_X = \sum_{i=1}^{7} \alpha_i p(x|\lambda_i)$, shown in Figure 5.9, with $\mathcal{V}^x$ representing a training data set spanning the space of X, generated randomly per $\mathcal{G}_X$. Viewing the Gaussian components of $\mathcal{G}_X$ as parent states or classes defined by the tuples $\{(\alpha_i, \lambda_i)\}_{i\in\{1,\ldots,7\}}$, we assume, for the purpose of this example, that the child *pdf*s to be estimated—$\{(\alpha_{ij}, \lambda_{ij})\}$ where $i \in \{1, \ldots, 7\}$ and, $\forall i$, $j \in \mathcal{J}_i$—are related to their respective parent states via an identity transformation, i.e., $\mathcal{T}\colon X \to X$, rather than the transformation intended to incorporate incremental memory described in Section 5.4.2.2 and detailed in this section. Since the identity transformation translates to single child states—i.e., $\forall i \in \{1, \ldots, 7\}$, $\mathcal{J}_i = \{1\}$, thereby making the index j redundant—with true *pdf*s identical to those of their respective parent states, we denote estimated child state *pdf*s by the simpler tuples $\left\{(\hat{\alpha}_i, \hat{\lambda}_i)\right\}_{i\in\{1,\ldots,7\}}$.

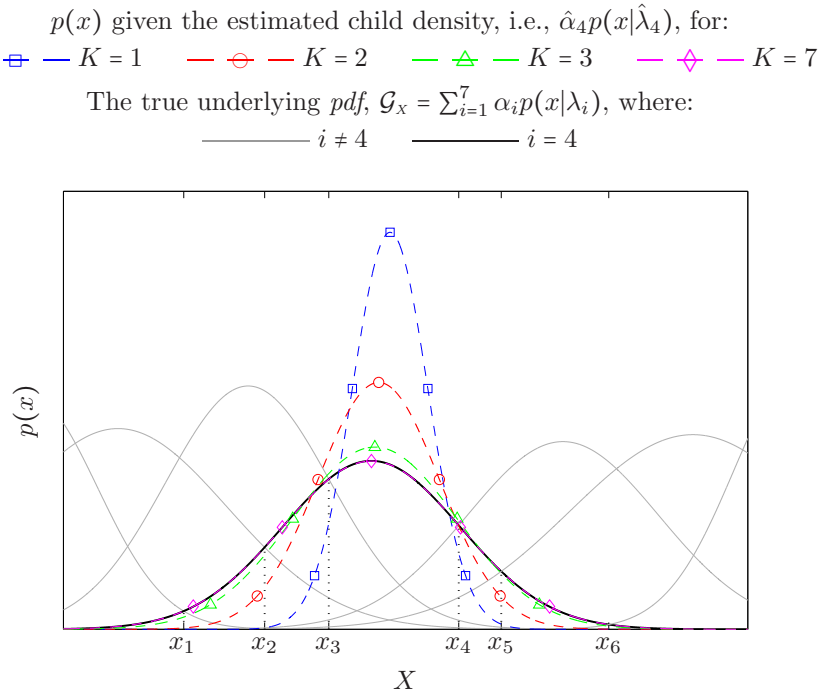
Focusing only on the $i = 4$th parent class represented in Figure 5.9 by the Gaussian component tuple (α_4, λ_4), we illustrate the effect of fuzzily determining the subset $\mathcal{V}_4^{x,w}$ on the estimation of the child state *pdf* given by $(\hat{\alpha}_4, \hat{\lambda}_4)$. To estimate the parameters of this child density—namely, $\hat{\alpha}_4$ and $\hat{\lambda}_4 := (\hat{\mu}_4, \hat{\sigma}_4)$—at a particular fuzziness factor, K, we first determine $\mathcal{V}_4^{x,w}$ per:

$$\mathcal{V}_4^x = \bigcup_{k=1}^{K} \left\{ x_n \in \mathcal{V}^x \colon \overset{*}{\lambda}_{i_k,n} = \lambda_4 \right\}, \tag{5.22a}$$

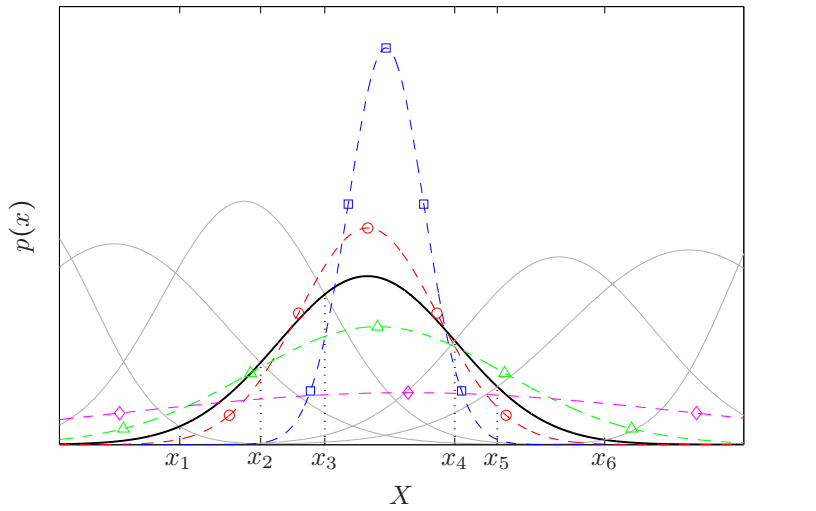
$$\forall x_n \in \mathcal{V}_4^x \colon \quad w_{4,n} = \frac{P(\lambda_4|x_n)}{\sum_{k=1}^{K} P(\overset{*}{\lambda}_{i_k}|x_n)}. \tag{5.22b}$$

Then, based on Eqs. (5.66) and (5.67) derived and detailed in Operations (d) and (e) below, respectively, we estimate $\hat{\alpha}_4$, $\hat{\mu}_4$, and $\hat{\sigma}_4$ as

$$\hat{\alpha}_4 = \alpha_4 \cdot 1 = \alpha_4, \tag{5.23a}$$



(a) Fuzzy clustering with membership weights



(b) Fuzzy clustering without membership weights

Fig. 5.9: Illustrating the advantage of fuzzy clustering in terms of improving *pdf* estimation, as well as the effect of membership weights, using a scalar random variable, X, with a randomly-determined highly-overlapping underlying *pdf*, $\mathcal{G}_X = \sum_{i=1}^{7} \alpha_i p(x|\lambda_i)$. With child densities assumed to be related to parent densities through an identity transformation, i.e., $\mathcal{T}{:}\, X \rightarrow X$, we estimate the $(\hat{\alpha}_4, \hat{\lambda}_4)$th child *pdf* from $\mathcal{G}_X$ using Eqs. (5.22) and (5.23), based on $|\mathcal{V}^x| = 10^6$ training samples spanning the range of X and generated randomly per $\mathcal{G}_X$.

$$\hat{\mu}_4 = \frac{\sum\limits_{\{n:\, x_n \in \mathcal{V}_4^x\}} w_{4,n} x_n}{\sum\limits_{\{n:\, x_n \in \mathcal{V}_4^x\}} w_{4,n}}, \tag{5.23b}$$

$$\hat{\sigma}_4 = \frac{\sum\limits_{\{n:\, x_n \in \mathcal{V}_4^x\}} w_{4,n} \left(x_n - \hat{\mu}_4\right)^2}{\sum\limits_{\{n:\, x_n \in \mathcal{V}_4^x\}} w_{4,n}}. \tag{5.23c}$$

Figures 5.9(a) and 5.9(b) illustrate the effect of performing fuzzing clustering per Eqs. (5.22) and (5.23) at varying values of K on the $(\hat{\alpha}_4, \hat{\lambda}_4)$ estimates, with and without the use of membership weights, respectively. At $K = 1$ where only the training data in the $[x_3, x_4]$ range are included in $\mathcal{V}_4^x$, i.e., where Eq. (5.22a) reduces to $\mathcal{V}_4^x = \{x \in \mathcal{V}^x\colon x \in [x_3, x_4]\}$, our fuzzy clustering technique reduces to the conventional hard-decision approach based on Bayes' rule with binary membership weights, thereby resulting in identical $(\hat{\alpha}_4, \hat{\lambda}_4)$ child *pdf* estimates regardless of the use of membership weights, as shown in Figures 5.9(a) and 5.9(b).

More importantly, Figure 5.9(a) clearly illustrates the adverse effects of the high overlap between the parent classes on the quality of the estimated child *pdf*; at $K = 1$, $(\hat{\alpha}_4, \hat{\lambda}_4)$ is significantly overfitted. By increasing the value of the fuzziness factor, K, Figure 5.9(a) shows our soft-decision technique to be quite successful in alleviating the problem of overfitting, albeit at increased computational costs due to the expansion of $\mathcal{V}_4^x$. In fact, we observe that, at the low value of $K = 3$ where $1 \le K \le 7$, a highly accurate $(\hat{\alpha}_4, \hat{\lambda}_4)$ estimate is achieved, demonstrating the power of fuzzy clustering in mitigating overfitting. This follows as a result of the quantitative expansion of $\mathcal{V}_4^x$ in conjunction with qualitative measures of membership, i.e., membership weights, as the range of training data considered for inclusion into $\mathcal{V}_4^x$ is increasingly extended at $K = 2$ and 3 to $x \in [x_2, x_5]$ and $x \in [x_1, x_6]$, respectively.

At $K = 7$, all available training data is included in $\mathcal{V}_4^x$, resulting in a nearly-perfect estimate for the child density $(\hat{\alpha}_4, \hat{\lambda}_4)$ as shown in Figure 5.9(a). This, however, is achieved at the cost of eliminating data localization for child *pdf* estimation altogether, translating into increased computational costs. Although data localization itself does not affect the time-frequency state localization described in Section 5.4.2.2

as a cornerstone of our tree-like memory inclusion technique, the importance of data localization in terms of limiting computational cost increases will become quite apparent in Section 5.4.3; with each incremental increase of the memory inclusion index, l, the higher cardinalities of $\mathcal{J}_i$ result in an exponential increase in the number of $\mathcal{G}_{\mathbb{Z}}^{(l)}$ Gaussian components. Given the highly accurate child *pdf* estimate obtained at $K = 3$ as observed above, this illustrative example thus demonstrates that excellent estimates for child state *pdf*s can, indeed, be achieved at low values for K, i.e., for $1 < K \ll M$ where M denotes the maximum number of parent states that can be considered for fuzzy clustering, thereby also largely preserving data localization ability, and accordingly, limiting increases in computational cost.

Finally, Figure 5.9(b) emphasizes the importance of the qualitative contribution of membership weights. In the absence of such weights, the inclusion of training data outside the $[x_3, x_4]$ range leads to oversmoothed $(\hat{\alpha}_4, \hat{\lambda}_4)$ *pdf* estimates, with the oversmoothing increasing as more data is considered with higher values for K. In particular, we point to the lower quality of the child state *pdf* estimate at $K = 3$ in Figure 5.9(b), compared to the corresponding estimate shown in Figure 5.9(a). At $K = 7$, the lack of the qualitative membership weights leads to a nearly flat, or *lobotomized*, estimate for $(\hat{\alpha}_4, \hat{\lambda}_4)$.

(b) Incremental temporal extension of training data

Partitioning the $\mathcal{V}^{\mathbb{Z}^{(l)}}$ training data spanning the entire $\mathbb{Z}^{(l)}$ space into the child subsets $\{\mathcal{V}_{ij}^{\mathbb{Z}^{(l)},w^{(l)}}\}$—where $i \in \mathcal{I}^{(l)}$ and, $\forall i,\ j \in \mathcal{J}_i^{(l)}$—per the fuzzy clustering technique described above represents the first of two steps in preparation for modelling the distribution in the $\mathbb{Z}^{(l+1)}$ space. In that first step, all information about the distribution of the data in the $\mathbb{Z}^{(l)}$ space has been incorporated into $\{\mathcal{V}_{ij}^{\mathbb{Z}^{(l)},w^{(l)}}\}_{\forall i,j}$; this information implicitly includes all previously-obtained information about distributions in the $\{\mathbb{Z}^{(m)}\}_{m\in\{0,\ldots,l-1\}}$ subspaces as well. Viewing $\{\mathcal{V}_{ij}^{\mathbb{Z}^{(l)},w^{(l)}}\}_{\forall i,j}$ as the subspace projections of the $(l+1)$th-order parent subsets in the $\mathbb{Z}^{(l+1)}$ space onto the $\mathbb{Z}^{(l)}$ space, the second step consists of temporally extending these lth-order subsets into their corresponding $(l+1)$th-order versions.

Prior to extending the $\{\mathcal{V}_{ij}^{\mathbb{Z}^{(l)},w^{(l)}}\}_{\forall i,j}$ subsets, however, we note that the ancestry information represented by the $\mathcal{I}^{(l)}$ and $\{\mathcal{J}_i^{(l)}\}_{i\in\mathcal{I}^{(l)}}$ parent and child integer index sets

is no longer needed. Thus, in order to make the notation tractable as we progressively incorporate more memory, we replace $\mathcal{I}^{(l)}$ and $\{\mathcal{J}_i^{(l)}\}_{i\in\mathcal{I}^{(l)}}$ with a single integer index set, $\mathcal{K}^{(l)}$ where $k \in \mathcal{K}^{(l)} = \{1,\dots,|\mathcal{K}^{(l)}|\}$, with the mapping given by

$$\forall i \in \mathcal{I}^{(l)}, j \in \mathcal{J}_i^{(l)}: \quad k = j + \sum_{m<i} |\mathcal{J}_m^{(l)}|, \tag{5.24a}$$

$$\mathcal{V}_k^{\mathbb{z}^{(l)},w^{(l)}} \xleftarrow{=} \mathcal{V}_{ij}^{\mathbb{z}^{(l)},w^{(l)}}. \tag{5.24b}$$

Noting that the child states and subsets obtained at index l also become the parents at index $l+1$, we have

$$\mathcal{I}^{(l+1)} \xleftarrow{=} \mathcal{K}^{(l)}. \tag{5.25}$$

Given the subsets $\{\mathcal{V}_k^{\mathbb{z}^{(l)},w^{(l)}}\}_{k\in\mathcal{K}^{(l)}}$ defined by Eqs. (5.21), (5.19), and (5.24), we now temporally extend the training data by simply augmenting the lth-order joint-band feature vector sequences in each subset with their corresponding static joint-band feature vectors at a relative temporal delay of $(l+1)\tau$ frames. In particular, for each lth-order sequence, $\mathbb{z}_{n,t}^{(\tau,l)} = [\mathbf{z}_{n,t}^T, \mathbf{z}_{n,t-\tau}^T, \dots, \mathbf{z}_{n,t-l\tau}^T]^T$, where we have reintroduced the memory inclusion step τ in $\mathbb{z}_{n,t}^{(\tau,l)}$ for clarity as well as introduced t to provide a local temporal frame of reference between the concatenated frames, we construct the $(l+1)$th-order sequence, $\mathbb{z}_{n,t}^{(\tau,l+1)} = \begin{bmatrix} \mathbb{z}_{n,t}^{(\tau,l)} \\ \mathbf{z}_{n,t-(l+1)\tau} \end{bmatrix} = [\mathbf{z}_{n,t}^T, \mathbf{z}_{n,t-\tau}^T, \dots, \mathbf{z}_{n,t-(l+1)\tau}^T]^T$. Dropping the τ and t from $\mathbb{z}_{n,t}^{(\tau,l)}$ and noting the equivalence of $\mathcal{I}^{(l+1)}$ and $\mathcal{K}^{(l)}$, we can thus express the incrementally extended data subsets as

$$\forall k \in \mathcal{K}^{(l)} \Leftrightarrow i \in \mathcal{I}^{(l+1)}:$$

$$\mathcal{V}_i^{\mathbb{z}^{(l+1)}} \xleftarrow{\text{Eq.(5.25)}} \mathcal{V}_k^{\mathbb{z}^{(l+1)}} = \left\{ \mathbb{z}_n^{(l+1)}: \mathbb{z}_n^{(l)} \in \mathcal{V}_k^{\mathbb{z}^{(l)}} \wedge \exists \mathbf{z}_{n,t-(l+1)\tau} \right\}, \tag{5.26}$$

where the last condition accounts for edge cases at the boundaries of training audio samples.

To conclude, we note that, at this step, the $\mathbb{Z}^{(l)} \xrightarrow{\text{Eq.(5.26)}} \mathbb{Z}^{(l+1)}$ extension is applied only to the training data. The association of the now-$(l+1)$-order data points in the subsets $\{\mathcal{V}_i^{\mathbb{z}^{(l+1)}}\}_{i\in\mathcal{I}^{(l+1)}} \xleftarrow{\text{Eq.(5.26)}} \{\mathcal{V}_k^{\mathbb{z}^{(l)}}\}_{k\in\mathcal{K}^{(l)}}$ with the lth-order membership weights in the sets $\{\mathcal{V}_i^{w^{(l)}}\}_{i\in\mathcal{I}^{(l+1)}} \xleftarrow{\text{Eq.(5.25)}} \{\mathcal{V}_k^{w^{(l)}}\}_{k\in\mathcal{K}^{(l)}}$ is unchanged since: (a) the degree of membership of the $(l+1)$th-order representations of data points to the lth-order

subspace projections of the underlying classes is the same as that of the corresponding lth-order representations of the same data points, and (b) the child state *pdf*s in the $\mathbb{Z}^{(l+1)}$ space, required to update the membership weights per Eq. (5.19), are yet to be estimated. The operation of temporally extending the data can thus be summarized as

$\forall k \in \mathcal{K}^{(l)} \Leftrightarrow i \in \mathcal{I}^{(l+1)}$:

$$\mathcal{V}_i^{\mathbb{z}^{(l+1)},w^{(l)}} \xleftarrow{=} \mathcal{V}_k^{\mathbb{z}^{(l+1)},w^{(l)}} = \left\{(\mathbb{z}_n^{(l+1)}, w_{i,n}^{(l)})\colon\ (\mathbb{z}_n^{(l)}, w_{k,n}^{(l)}) \in \mathcal{V}_k^{\mathbb{z}^{(l)},w^{(l)}} \wedge \exists \mathbf{z}_{n,t-(l+1)\tau}\right\}. \quad (5.27)$$

As discussed in Section 5.4.2.2, Eq. (5.27) effectively carries over the localization of the time-frequency information obtained at memory inclusion index l into the higher $l+1$ step as well as future ones. As such, we have implicitly made use of the strong correlation of speech characteristics across time by inferring the localization represented by the child Gaussian components to be estimated in the $\mathbb{Z}^{(l+1)}$ space based on the localization already obtained in the $\mathbb{Z}^{(l)}$ subspace.

(c) Extending parent states using weighted Expectation-Maximization

With l representing the memory inclusion index of the temporal extension step at hand, i.e., replacing $l+1$ in the discussion above by l for notational convenience and equation compactness below, we now describe our technique for estimating uni-modal child state *pdf*s at index l, based on the information obtained at the previous memory inclusion index, $l-1$.

In addition to comprising all previously-obtained information about the distribution of data in the $\{\mathbb{Z}^{(m)}\}_{m\in\{0,\ldots,l-1\}}$ subspaces as noted in Operation (b) above, the pair-wise-disjoint time-frequency-localized $\{\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}_{i\in\mathcal{I}^{(l)}} \xleftarrow{\text{Eq.(5.27)}} \{\mathcal{V}_k^{\mathbb{z}^{(l-1)},w^{(l-1)}}\}_{k\in\mathcal{K}^{(l-1)}}$ subsets also incorporate partial information about the distribution of the new static data in the time-dependent $\mathbf{Z}_{t-l\tau}$ static subspace by virtue of the incremental temporal

extension performed via Eq. (5.27).[146] Specifically, this partial information consists of the frequency-only localization information carried over from the parent states, $\{S_k^{(l-1)} \triangleq (\alpha_k^{\mathbb{Z}^{(l-1)}}, \lambda_k^{\mathbb{Z}^{(l-1)}})\}_{k\in\mathcal{K}^{(l-1)}}$, through the subsets $\{\mathcal{V}_k^{\mathbb{Z}^{(l-1)},w^{(l-1)}}\}_{k\in\mathcal{K}^{(l-1)}}$. Exploiting this partial localization information, we estimate finer child densities spanning the entire $\mathbb{Z}^{(l)}$ space in two steps:

(a) We first model the distribution of the new incremental data in the static $\mathbf{Y}_{t-l\tau}$ highband subspace, rather than in the entire joint-band $\mathbf{Z}_{t-l\tau}$ subspace, as motivated below. By applying a *pruning* condition to reduce potential modelling redundancies prior to estimating the new child state densities in the $\mathbf{Y}_{t-l\tau}$ subspace, *pdf* estimation is performed only for a subset of the localized frequency regions defined by the subsets $\{\mathcal{V}_i^{\mathbb{Z}^{(l)},w^{(l-1)}}\}_{i\in\mathcal{I}^{(l)}}$, with the estimation further performed *individually* for each region, i.e., *independently* of the others.

(b) In a second step, we extrapolate the new child densities thus obtained—representing projections of the underlying lth-order time-frequency classes onto the time-dependent $\mathbf{Y}_{t-l\tau}$ subspace—into the $\mathbb{Z}^{(l)}$ space by integrating them with the corresponding parent localized densities spanning the $\mathbb{Z}^{(l-1)}$ subspace. An equivalent extension into the $\mathbb{Z}^{(l)}$ space is also applied as a separate step to those $(l-1)$th-order parent states skipped by pruning in the first stage.

This latter integration—which also captures the cross-band correlation between spectral envelope representations in the $\mathbf{X}_{t-l\tau}$ and $\mathbf{Y}_{t-l\tau}$ subspaces, as shown below—is achieved by simply taking account of the information that has already been incorporated into the parent $\{\mathcal{V}_i^{\mathbb{Z}^{(l)},w^{(l-1)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets in prior extension steps.

The partial localization information carried over from parent states, thus, represents regularization information that allows us to break down the intractable task of esti-

[146] In the context described here, the time-dependency of a static subspace follows from directly or indirectly using the temporal information represented by neighbouring time-shifted static feature vectors in modelling the variability of training data along the frequency-only axis corresponding to that particular static subspace. Extracting models in the $\mathbf{Z}_{t-l\tau}$ subspace by marginalizing models of the distribution of lth-order temporally-extended $\mathbb{Z}^{(l)}$ feature vectors, for example, involves the introduction of time-dependency through the *direct* use of temporal information. In contrast, estimating independent models for localized regions in the static $\mathbf{Z}_{t-l\tau}$ subspace by modelling variability within disjoint subsets obtained by time-frequency localization along lower-order temporal axes, as implemented in our algorithm discussed here, represents an example of introducing time-dependency through the *indirect* use of temporal information. Without the use of temporal information as such, we consider the static spaces underlying our models to be time-independent, as is the case for the weighted EM initialization model described later in this section.

mating $\mathcal{G}_{\mathbb{Z}}^{(l)}$, a single global high-dimensional *pdf* for training data in the $\mathbb{Z}^{(l)}$ space, into $|\mathcal{I}^{(l)}|$ independent and localized $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$ *pdf* estimation tasks. For each of these *pdf*s to be estimated, i.e., $\forall i \in \mathcal{I}^{(l)}$, let $j \in \mathcal{J}_i^{(l)} = \{1, \ldots, |\mathcal{J}_i^{(l)}|\}$ represent the indices of the component Gaussian densities. Then, with the $|\mathcal{I}^{(l)}|$ *pdf*s representing finer $|\mathcal{J}_i^{(l)}|$-modal models of the distribution of the lth-order data in the corresponding $\{\mathcal{V}_i^{\mathbb{Z}^{(l)},w^{(l-1)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets, the $|\mathcal{J}_i^{(l)}|$ component densities of each ith *pdf*, $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$, constitute the lth-order uni-modal child states, $\{S_{ij}^{(l)} \triangleq (\alpha_i^{\mathbb{Z}^{(l)}} \cdot \alpha_{ij}^{\mathbb{Z}^{(l)}}, \lambda_{ij}^{\mathbb{Z}^{(l)}})\}_{j\in\mathcal{J}_i^{(l)}}$, descended from the corresponding kth $(l-1)$th-order uni-modal child state, $S_k^{(l-1)} \triangleq (\alpha_k^{\mathbb{Z}^{(l-1)}}, \lambda_k^{\mathbb{Z}^{(l-1)}})$, where $\alpha_i^{\mathbb{Z}^{(l)}} \xleftarrow{=} \alpha_k^{\mathbb{Z}^{(l-1)}}$ per Eq. (5.25), for all $i \in \mathcal{I}^{(l)} \Leftrightarrow k \in \mathcal{K}^{(l-1)}$, as illustrated in Figure 5.8.[147]

i. Two-stage child density estimation

To estimate the new lth-order $|\mathcal{I}^{(l)}|$ *pdf*s, $\{\mathcal{G}_{\mathbb{Z}_i}^{(l)}\}_{i\in\mathcal{I}^{(l)}}$, we employ the Expectation-Maximization (EM) algorithm. Our implementation of EM, however, differs from the conventional algorithm ubiquitously used for GMM training in the following (for notational consistency and brevity, we denote the time-dependent $\mathbf{X}_{t-l\tau}$, $\mathbf{Y}_{t-l\tau}$, and $\mathbf{Z}_{t-l\tau}$ static subspaces by $\mathbf{X}^{(l)}$, $\mathbf{Y}^{(l)}$, and $\mathbf{Z}^{(l)}$, respectively):

1. We perform EM in two distinct stages corresponding to the two modelling steps listed above. In the first and primary stage, we model the distribution of the localized incremental data in the static $\mathbf{Y}^{(l)}$ highband subspace. Combined with the child states obtained via the aforementioned pruning, the first stage effectively generates the $|\mathcal{J}_i^{(l)}|$-modal *pdf*s $\{\mathcal{G}_{\mathbf{Y}_i}^{(l)} := \mathcal{G}(\mathbf{y}^{(l)}; |\mathcal{J}_i^{(l)}|, \mathrm{A}_i^{\mathbf{y}^{(l)}}, \Lambda_i^{\mathbf{y}^{(l)}})\}_{i\in\mathcal{I}^{(l)}}$, where $\forall i \in \mathcal{I}^{(l)}$, $\Lambda_i^{\mathbf{y}^{(l)}} = \{\lambda_{ij}^{\mathbf{y}^{(l)}} := (\boldsymbol{\mu}_{ij}^{\mathbf{y}^{(l)}}, \mathbf{C}_{ij}^{\mathbf{y}^{(l)}})\}_{j\in\mathcal{J}_i^{(l)}}$, $\mathrm{A}_i^{\mathbf{y}^{(l)}} = \{\alpha_{ij}^{\mathbf{y}^{(l)}} := P(\lambda_{ij}^{\mathbf{y}^{(l)}})\}_{j\in\mathcal{J}_i^{(l)}}$, and $\mathcal{J}_i^{(l)} = \{1, \ldots, |\mathcal{J}_i^{(l)}|$. The motivation for constraining our focus to highband data is to increase the influence of the variability of localized static data in the

[147] As discussed in Operation (e), we weight the $\{\alpha_{ij}^{\mathbb{Z}^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$ priors of the $|\mathcal{J}_i^{(l)}|$ Gaussian components of each localized $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$ model before consolidating all $\{\mathcal{G}_{\mathbb{Z}_i}^{(l)}\}_{i\in\mathcal{I}^{(l)}}$ models into a single global $\mathcal{G}_{\mathbb{Z}}^{(l)}$ *pdf*. Since it is these final weighted components of the global $\mathcal{G}_{\mathbb{Z}}^{(l)}$ which, in fact, correspond to the lth-order states in our state space-based interpretation as illustrated in Figure 5.8, we can simplify our $S_{ij}^{(l)}$ state notation here and elsewhere by representing the correspondence with Gaussian component densities through the simpler $\{S_{ij}^{(l)} \triangleq (\alpha_{ij}^{\mathbb{Z}^{(l)}}, \lambda_{ij}^{\mathbb{Z}^{(l)}})\}_{j\in\mathcal{J}_i^{(l)}}$, rather than the more accurate but elaborate $\{S_{ij}^{(l)} \triangleq (\alpha_i^{\mathbb{Z}^{(l)}} \cdot \alpha_{ij}^{\mathbb{Z}^{(l)}}, \lambda_{ij}^{\mathbb{Z}^{(l)}})\}_{j\in\mathcal{J}_i^{(l)}}$ indicated by Eq. (5.67b).

high band on the number as well as the shape of the child state densities ultimately achieved. In other words, our objective in this first EM stage is to rather model the variability in the target frequency band of BWE—the 4–8 kHz high band—as accurately and finely as possible. The influence of variability in the static $\mathbf{X}^{(l)}$ narrowband subspace and its cross-correlation with that of the high band are modelled in the second extrapolation stage discussed in Item 3 below. Moreover, as shown in the EM formulae derived below, the influence of variability in the temporally-extended $\mathbb{Z}^{(l-1)}$ joint-band subspace is accounted for directly in this first EM stage through incorporating the $(l-1)$th-order parent membership weights, $\{\mathcal{V}_i^{w^{(l-1)}}\}_{i\in\mathcal{I}^{(l)}}$, in the iterative update equations for estimating $\{(\mathrm{A}_i^{\mathbf{y}^{(l)}}, \Lambda_i^{\mathbf{y}^{(l)}})\}_{i\in\mathcal{I}^{(l)}}$.

The focus on modelling variability in the $\mathbf{Y}^{(l)}$ highband subspace, rather than in the $\mathbf{Z}^{(l)}$ joint-band subspace, follows from the lower intra- and inter-frame variability of highband spectral envelopes, relative to those of the narrow band.[148] Given that narrowband variability and cross-band correlation are accounted for in the second extrapolation modelling stage described below, such lower highband variability motivates us to reduce the influence of narrowband variability on EM-based density estimation in this first modelling stage for the benefit of ultimately obtaining lth-order joint-band child states, $\{S_{ij}^{(l)}\}_{\forall i,j}$, that are more *attuned* to the distributions of the underlying classes in the target high frequency band, albeit at the cost of lower modelling accuracy for variability in the $\mathbf{X}^{(l)}$ narrowband subspace. Estimating such band-attuned joint-band *pdf*s by constraining the modelled feature space in an intermediate EM step, is the recip-

[148] As discussed in Sections 1.1.3 and 3.2.7, the 4–8 kHz range is dominated by unvoiced sounds with flat spectra, with the high-frequency formants of voiced sounds further characterized by wide bandwidths. In contrast, spectral envelopes in the 0.3–3.4 kHz narrow band typically exhibit a much larger intra-frame variability since the first three formants, for example, generally occur in the 250–3300 Hz range with larger variations in frequency, energy, and bandwidth, across the different sound classes, compared to highband formants [10, Section 3.4]. Indeed, it is this low intra-frame variability for highband spectral envelopes, compared to those in the narrow band, that allows the parameterization of these highband envelopes using fewer parameters.
Similarly, the low inter-frame variability of highband envelopes, compared to the narrow band, follows from the fact that distinctions between different sound classes in the high band tend to be more restricted to variability in overall energy level across the entire 4–8 kHz band rather than to variability of energy as a function of frequency, as illustrated, for example, by the difference between the alveolar /s/ and labial /f/ fricatives in Figure 1.2. Furthermore, as noted above, such energy variations between and within different sounds classes are generally lower in the high band than in the narrow band.

rocal to the idea exploited in Section 5.3.3.2 to improve frontend-based memory inclusion—namely, expanding the modelled feature space by the inclusion of the highband delta feature space, $\mathbf{\Delta}_{\mathbf{Y}}$, in order to capture the influence of highband dynamics to ultimately obtain an improved model of the underlying classes in the $\begin{bmatrix} \overset{\triangle}{\mathbf{X}} \\ \mathbf{Y} \end{bmatrix}$ joint-band subspace, as summarized in Eq. (5.8).

2. In the conventional EM algorithm derived for GMM-based density estimation, and for mixture models in general, the objective is to find the set of model parameters with maximum likelihood—or typically, log-likelihood—given the training data.[149] In our context, this corresponds to maximizing the log-likelihood of localized model parameters given the parent data subsets constrained to the $\mathbf{Y}^{(l)}$ subspace; i.e.,

$$\forall i \in \mathcal{I}^{(l)}: \quad \overset{*}{\Theta}{}_i^{\mathbf{y}^{(l)}} := (\overset{*}{\mathrm{A}}{}_i^{\mathbf{y}^{(l)}}, \overset{*}{\Lambda}{}_i^{\mathbf{y}^{(l)}}) = \arg\max_{\Theta^{\mathbf{y}^{(l)}}} \log\Big[L(\Theta^{\mathbf{y}^{(l)}}|\mathcal{V}_i^{\mathbf{y}^{(l)}})\Big], \tag{5.28}$$

where $\forall i \in \mathcal{I}^{(l)}$, $\Theta_i^{\mathbf{y}^{(l)}} \triangleq \mathcal{G}^{(l)}_{\mathbf{Y}_i}$ and,

$$\forall i \in \mathcal{I}^{(l)}: \quad \mathcal{V}_i^{\mathbf{y}^{(l)}} = \Big\{\mathbf{y}_{n,t-l\tau}: \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\Big\}, \tag{5.29}$$

with the model parameters $\{\mathrm{A}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ and $\{\Lambda_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ defined as in Item 1 above. Estimating localized model parameters as such does not, however, account for the fuzzy qualitative expansion of training data subsets described in Operation (a), and hence, will ultimately result in oversmoothed localized *pdf*s as shown in the illustrative example of Figure 5.9(b). Consequently, the maximization of model parameter log-likelihoods through EM should incorporate the qualitative membership of data in the localized $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ parent subsets. Since the static incremental data in the $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets are merely the projections of the corresponding temporally-extended lth-order data in the $\{\mathcal{V}_i^{\mathbb{z}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ parent subsets onto the $\mathbf{Y}^{(l)}$ subspace, i.e., static highband data points are referenced in time to the same wideband training frames used to construct the corresponding lth-order points, the fuzzy membership of these static data points to the $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets is defined by the same weights associated with $\{\mathcal{V}_i^{\mathbb{z}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$

[149] Since $\log(\cdot)$ is a strictly increasing function, the value $X = \overset{*}{x}$ that maximizes $\log[f(X)]$ also maximizes $f(X)$. Most implementations of EM use log-likelihood—rather than likelihood—since it typically makes the maximum-likelihood estimation of density parameters more tractable.

per Eqs. (5.27) and (5.21). Thus, to account for membership weights, we modify Eq. (5.28) such that

$$\forall i \in \mathcal{I}^{(l)}: \quad \overset{*}{\Theta}{}_i^{\mathbf{y}^{(l)}} = \arg\max_{\Theta^{\mathbf{y}^{(l)}}} f\left(\mathcal{V}_i^{w^{(l-1)}}, \log\left[L(\Theta^{\mathbf{y}^{(l)}}|\mathcal{V}_i^{\mathbf{y}^{(l)}})\right]\right), \tag{5.30}$$

where the cost function, $f\left(\mathcal{V}_i^{w^{(l-1)}}, \log\left[L(\Theta^{\mathbf{y}^{(l)}}|\mathcal{V}_i^{\mathbf{y}^{(l)}})\right]\right)$, is a weighted version of the log-likelihood function that guarantees the convergence of the derived iterative EM algorithm, similar to the convergence obtained using the conventional non-weighted log-likelihoods. By solving Eq. (5.30) through minor modifications to the conventional derivation of the EM algorithm for GMMs, we introduce below a *weighted* implementation of EM where the derived iterative update equations explicitly incorporate membership weights.

3. In the second EM stage, we exploit the information previously incorporated into the $\{\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets about the distribution of temporally-extended data in the parent $\mathbb{Z}^{(l-1)}$ subspace in order to extrapolate the $|\mathcal{I}^{(l)}|$ localized $\mathbf{Y}^{(l)}$ subspace densities, $\{(\overset{*}{\Lambda}{}_i^{\mathbf{y}^{(l)}}, \overset{*}{\Lambda}{}_i^{\mathbf{y}^{(l)}})\}_{i\in\mathcal{I}^{(l)}}$, obtained through Eq. (5.30), into the $|\mathcal{J}_i^{(l)}|$-modal densities, $\{(\overset{*}{\Lambda}{}_i^{\mathbb{z}^{(l)}}, \overset{*}{\Lambda}{}_i^{\mathbb{z}^{(l)}})\}_{i\in\mathcal{I}^{(l)}}$, spanning the entire $\mathbb{Z}^{(l)}$ space. In particular, we perform a single EM iteration to estimate the maximum-likelihood $\{(\overset{*}{\Lambda}{}_i^{\mathbb{z}^{(l)}}, \overset{*}{\Lambda}{}_i^{\mathbb{z}^{(l)}})\}_{i\in\mathcal{I}^{(l)}}$ parameters given the lth-order joint-band data and their membership weights, $\{\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}_{i\in\mathcal{I}^{(l)}}$, but with the posterior probabilities of the $\mathbb{Z}^{(l)}$ child densities determined in the Expectation step based entirely on the $\mathbf{Y}^{(l)}$ subspace densities previously estimated in the first EM stage of Item 1. With the new information about the distribution of data in the $\mathbf{Y}^{(l)}$ subspace now incorporated into the $\mathbb{Z}^{(l)}$ child density posterior probabilities, we can then—through a final Maximization step—easily extend the information readily available in the $\{\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets about the distribution of data in the $\mathbb{Z}^{(l-1)}$ subspace in order to obtain the maximum-likelihood estimates for the sought-after lth-order joint-band child states/densities, $\{S_{ij}^{(l)} \triangleq (\alpha_{ij}^{\mathbb{z}^{(l)}}, \lambda_{ij}^{\mathbb{z}^{(l)}})\}_{\forall i,j}$. Worthy of note is that, since the latter final Maximization step is applied using lth-order joint-band data in the $\mathbb{Z}^{(l)}$ space, the extension of the $\mathbb{Z}^{(l-1)}$ subspace densities—or, alternatively, the extrapolation of the $\mathbf{Y}^{(l)}$ subspace densities—implicitly incorporates the cross-correlation between data distributions in the static $\mathbf{X}^{(l)}$ and $\mathbf{Y}^{(l)}$ subspaces—as well the cross-correlation between

all $\{\mathbf{X}^{(m)}\}_{m\in\{0,\ldots,l-1\}}$ and $\{\mathbf{Y}^{(m)}\}_{m\in\{0,\ldots,l-1\}}$ subspaces—into the final model at memory inclusion index l represented by the $\{S_{ij}^{(l)}\}_{\forall i,j}$ states.

The focus on modelling the incremental variability in the static $\mathbf{Y}^{(l)}$ subspace when estimating child state densities, as described for our two-stage EM approach above, is similar in concept to the *shadowing* of the variability in one band into the other as employed in both codebook- and HMM-based BWE techniques. In the more-advanced class of codebook-based mapping techniques discussed in Section 2.3.3.2, variability of training data is first quantized in the narrow band before constructing a shadow highband—or wideband—codebook. Similarly, in the second class of HMM-based approaches discussed in Section 2.3.3.4, a VQ codebook of highband spectral envelopes is associated to HMM states modelling the corresponding envelopes of narrowband spectra. We note, in particular, the parallelism of our two-stage EM technique to the HMM-based approach of [39], where highband variability is first modelled through a highband VQ codebook before estimating narrowband mixture models in each HMM state based on the correspondence of the training narrowband envelopes to those of the quantized highband spectra.

ii. Deriving the weighted Expectation-Maximization formulae

To derive our weighted EM procedure and prove its convergence, we use the EM tutorials of Bilmes and Borman, in [187] and [188], respectively, as references. Rather than repeat the complete EM derivation detailed in [187], however, we focus only on detailing those steps and formulae impacted by the inclusion of membership weights per Eq. (5.30).

For generality, let $\mathcal{X} = \{\mathbf{x}_n\}_{n=\{1,\ldots,N\}}$ represent data observations of the random vector, $\mathbf{X}$, whose underlying multi-variate *pdf* we wish to model using an M-modal mixture model given by $\Theta = \{(\alpha_m, \lambda_m)\}_{m\in\{1,\ldots,M\}}$, such that

$$p(\mathbf{x}|\Theta) = \sum_{m=1}^{M} \alpha_m p(\mathbf{x}|\lambda_m), \tag{5.31}$$

where λ_m and $\alpha_m := P(\lambda_m)$ denote, respectively, the parameters and the mixing weight of the Mth component density. The EM algorithm attempts to find the set of parameters, $\overset{*}{\Theta}$, which maximize the log-likelihood function $\log[L(\Theta|\mathcal{X})] :=$

$\log[p(\mathcal{X}|\Theta)]$. Assuming the observations $\mathcal{X}$ to be drawn from $p(\mathbf{x}|\Theta)$ are i.i.d., the log-likelihood function can then be written as

$$\log[L(\Theta|\mathcal{X})] \coloneqq \log[p(\mathcal{X}|\Theta)] = \log \prod_{n=1}^{N} p(\mathbf{x}_n|\Theta) = \sum_{n=1}^{N} \log \left(\sum_{m=1}^{M} \alpha_m p(\mathbf{x}_n|\lambda_m) \right). \quad (5.32)$$

The log-likelihood given as thus, however, is difficult to optimize due to the right-hand-side logarithm-of-sums term. To make the maximum-likelihood estimation of Θ tractable, a *hidden* variable, Y where $y \in \{1, \ldots, M\}$, is introduced, with each of the *unobserved* realizations $\mathcal{Y} = \{y_n\}_{n \in \{1,\ldots,N\}}$ of Y representing the index of the generative mixture-model's component density underlying a corresponding observation among $\mathcal{X}$. By introducing Y, the *incomplete-data* log-likelihood of Eq. (5.32), to be optimized through EM, can be replaced by the *complete-data* log-likelihood,

$$\begin{aligned} \log[L(\Theta|\mathcal{X},\mathcal{Y})] &\coloneqq \log[p(\mathcal{X},\mathcal{Y}|\Theta)] = \log \prod_{n=1}^{N} p(\mathbf{x}_n, y_n|\Theta) \\ &= \sum_{n=1}^{N} \log[p(\mathbf{x}_n, y_n|\Theta)] = \sum_{n=1}^{N} \log[p(\mathbf{x}_n|y_n,\Theta)p(y_n|\Theta)] = \sum_{n=1}^{N} \log[\alpha_{y_n} p(\mathbf{x}_n|\lambda_{y_n})]. \end{aligned} \quad (5.33)$$

Now, let $\mathbf{y} = [y_1, \ldots, y_N]$ represent a realization of the random vector $\mathbf{Y}$ whose space $\Omega_{\mathbf{y}}$ comprises all the possible values that the N unobserved i.i.d. data in the subset $\mathcal{Y}$ can jointly take. The conventional EM algorithm solves the problem of finding $\overset{*}{\Theta} = \arg\max_{\Theta} \log[L(\Theta|\mathcal{X})]$ by iteratively maximizing an equivalent function, $Q(\Theta, \Theta^{(k)})$, where $\Theta^{(k)}$ represents the model estimates obtained at the kth EM iteration. In particular, using Eq. (5.33), the EM algorithm can be summarized as

$$\begin{aligned} \Theta^{(k+1)} &= \arg\max_{\Theta} Q(\Theta, \Theta^{(k)}) \\ &= \arg\max_{\Theta} E\left[\log[L(\Theta|\mathcal{X},\mathbf{Y})] \middle| \mathcal{X}, \Theta^{(k)} \right] \\ &= \arg\max_{\Theta} \sum_{\mathbf{y} \in \Omega_{\mathbf{y}}} \log[p(\mathcal{X},\mathbf{y}|\Theta)]\, P(\mathbf{y}|\mathcal{X},\Theta^{(k)}) \\ &= \arg\max_{\Theta} \sum_{\mathbf{y} \in \Omega_{\mathbf{y}}} \sum_{n=1}^{N} \log[\alpha_{y_n} p(\mathbf{x}_n|\lambda_{y_n})] \prod_{l=1}^{N} P(y_l|\mathbf{x}_l,\Theta^{(k)}) \\ &= \arg\max_{\Theta} \sum_{y_1=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{n=1}^{N} \log[\alpha_{y_n} p(\mathbf{x}_n|\lambda_{y_n})] \prod_{l=1}^{N} P(y_l|\mathbf{x}_l,\Theta^{(k)}), \end{aligned} \quad (5.34)$$

where we have made use of the fact that maximizing the incomplete-data log-likelihood, $\log[L(\Theta|\mathcal{X})]$, is equivalent to maximizing the expectation of the complete-data log-likelihood, $\log[L(\Theta|\mathcal{X},\mathbf{Y})]$, given the observed data and the previous model estimates, as shown in the development leading up to and including Eq. (15) in [188], and further proven for our weighted EM algorithm in Eq. (5.51) below.

Let w_n represent a prior membership weight associated with $\mathbf{x}_n$, the nth observation in $\mathcal{X}$, independent of $p(\mathbf{x}|\Theta)$. To incorporate the effects of all such weights, i.e., $\{w_n\}_{n\in\{1,\ldots,N\}}$, into the EM algorithm, we maximize the expectation of a *weighted* log-likelihood function, rather than the expectation of the conventional non-weighted log-likelihood as in Eq. (5.34). In particular, we replace $Q(\Theta,\Theta^{(k)})$ in Eq. (5.34) by $Q^w(\Theta,\Theta^{(k)})$, where

$$Q^w(\Theta,\Theta^{(k)}) = \sum_{y_1=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{n=1}^{N} w_n \log[\alpha_{y_n} p(\mathbf{x}_n|\lambda_{y_n})] \prod_{l=1}^{N} P(y_l|\mathbf{x}_l,\Theta^{(k)}). \tag{5.35}$$

By manipulating the right-hand-side of Eq. (5.35) in the same manner shown in Eqs. (3) and (4) of [187], $Q^w(\Theta,\Theta^{(k)})$ can be rewritten as

$$\begin{aligned} Q^w(\Theta,\Theta^{(k)}) &= \sum_{m=1}^{M}\sum_{n=1}^{N} w_n \log[\alpha_m p(\mathbf{x}_n|\lambda_m)]\, P(m|\mathbf{x}_n,\Theta^{(k)}) \\ &= \sum_{m=1}^{M}\sum_{n=1}^{N} w_n \log(\alpha_m) P(m|\mathbf{x}_n,\Theta^{(k)}) \\ &\quad + \sum_{m=1}^{M}\sum_{n=1}^{N} w_n \log[p(\mathbf{x}_n|\lambda_m)] P(m|\mathbf{x}_n,\Theta^{(k)}), \end{aligned} \tag{5.36}$$

from which it is clear that the Expectation step—in both the conventional EM and our weighted EM algorithms—reduces to evaluating $P(m|\mathbf{x}_n,\Theta^{(k)})$, for all combinations of $n \in \{1,\ldots,N\}$ and $m \in \{1,\ldots,M\}$.

Through independently maximizing the first and second terms of Eq. (5.36) relative to each of the $\{\alpha_m\}$ and $\{\lambda_m\}$ parameters, respectively, the expressions for the optimal $(k+1)$th-iteration model parameters, i.e., $\Theta^{(k+1)} = \{(\alpha_m^{(k+1)},\lambda_m^{(k+1)})\}_{m\in\{1,\ldots,M\}}$, can then be obtained. In particular, the component density priors, $\{\alpha_m^{(k+1)}\}_{m\in\{1,\ldots,M\}}$, are obtained using Lagrange optimization[150] of the first term, as shown in [187].

[150] See [71, Section A.3] for details regarding Lagrange optimization.

Introducing the scalar Lagrange multiplier γ[151] with the constraint that $\sum_m \alpha_m = 1$ and taking the derivative with respect to α_m, we obtain the following Lagrangian function for each of the M priors:

$$\frac{\partial}{\partial \alpha_m}\left[\sum_{m=1}^{M}\sum_{n=1}^{N} w_n \log(\alpha_m) P(m|\mathbf{x}_n, \Theta^{(k)}) + \gamma\left(\sum_{m=1}^{M}\alpha_m - 1\right)\right] = 0, \tag{5.37}$$

which reduces to

$$\sum_{n=1}^{N}\frac{1}{\alpha_m} w_n P(m|\mathbf{x}_n, \Theta^{(k)}) + \gamma = 0. \tag{5.38}$$

Given that $\sum_{m=1}^{M} P(m|\mathbf{x}_n, \Theta^{(k)}) = 1$, summing Eq. (5.38) over m results in the solution that $\gamma = -\sum_{n=1}^{N} w_n$, and hence,

$$\alpha_m^{(k+1)} \xleftarrow{=} \alpha_m = \frac{\sum_{n=1}^{N} w_n P(m|\mathbf{x}_n, \Theta^{(k)})}{\sum_{n=1}^{N} w_n}. \tag{5.39}$$

Up to this point, we made no assumptions in the development above about the shape of the kernel density, $p(\mathbf{x}|\lambda_m)$, representing the modes of the mixture model in Eq. (5.31). To obtain the optimal $\{\lambda_m^{(k+1)}\}_{m\in\{1,\ldots,M\}}$ density parameters, however, we now substitute the generic $p(\mathbf{x}|\lambda_m)$ in the second term of Eq. (5.36) by the Gaussian *pdf* denoted by $\mathcal{N}\big(\mathbf{x}; \lambda_m := (\boldsymbol{\mu}_m, \mathbf{C}_m)\big)$ and given as shown in Eq. (2.13). In particular, we rewrite the second term of Eq. (5.36) as

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{n=1}^{N} w_n \log[p(\mathbf{x}_n|\lambda_m)] P(m|\mathbf{x}_n, \Theta^{(k)}) \\ &\quad = \sum_{m=1}^{M}\sum_{n=1}^{N} w_n \left(\frac{1}{2}\log(|\mathbf{C}_m^{-1}|) - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_m)^T \mathbf{C}_m^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_m)\right) P(m|\mathbf{x}_n, \Theta^{(k)}),\end{aligned} \tag{5.40}$$

where we have dropped the constant $-\frac{\text{Dim}(\mathbf{X})}{2}\log(2\pi)$ term since it disappears after taking derivatives, and made use of the determinant property that $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$. By now taking the derivative with respect to $\boldsymbol{\mu}_m$, for all $m \in \{1, \ldots, M\}$, and setting it

[151] The Lagrange multiplier is typically denoted by λ. To avoid confusion with our λ notation for component density parameters, however, we denote the multiplier here by γ.

to zero, we obtain

$$\sum_{n=1}^{N} w_n \mathbf{C}_m^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_m) P(m|\mathbf{x}_n, \Theta^{(k)}) = 0, \tag{5.41}$$

which is easily solved for $\boldsymbol{\mu}_m$ to obtain

$$\boldsymbol{\mu}_m^{(k+1)} \xleftarrow{=} \boldsymbol{\mu}_m = \frac{\sum_{n=1}^{N} w_n \, P(m|\mathbf{x}_n, \Theta^{(k)}) \, \mathbf{x}_n}{\sum_{n=1}^{N} w_n \, P(m|\mathbf{x}_n, \Theta^{(k)})}. \tag{5.42}$$

Finally, as detailed in [187], making use of the matrix properties of the square and symmetric $\{\mathbf{C}_m\}_{m\in\{1,\ldots,M\}}$ covariance matrices allows us to reduce the derivative of Eq. (5.40) with respect to $\mathbf{C}_m^{-1}$, for all $m \in \{1, \ldots, M\}$, to

$$\sum_{n=1}^{N} w_n \, P(m|\mathbf{x}_n, \Theta^{(k)}) \left(\mathbf{C}_m - [\mathbf{x}_n - \boldsymbol{\mu}_m][\mathbf{x}_n - \boldsymbol{\mu}_m]^T\right) = 0, \tag{5.43}$$

from which we obtain

$$\mathbf{C}_m^{(k+1)} \xleftarrow{=} \mathbf{C}_m = \frac{\sum_{n=1}^{N} w_n \, P(m|\mathbf{x}_n, \Theta^{(k)}) \, [\mathbf{x}_n - \boldsymbol{\mu}_m][\mathbf{x}_n - \boldsymbol{\mu}_m]^T}{\sum_{n=1}^{N} w_n \, P(m|\mathbf{x}_n, \Theta^{(k)})}. \tag{5.44}$$

Eqs. (5.39), (5.42), and (5.44) represent the Maximization step.

iii. Convergence of the weighted Expectation-Maximization algorithm

Following [188], we prove the convergence of our weighted EM algorithm by showing that the weighted log-likelihood function to be maximized is a non-decreasing function of the iteration index k. As described above, the objective of the conventional EM algorithm is to find the model parameters, $\overset{\star}{\Theta}$, that maximize the log-likelihood of the observations, $\mathcal{X}$. For mixture models and i.i.d. realizations, this log-likelihood function—which we now denote by $\mathcal{L}(\Theta|\mathcal{X})$ for notational convenience—was shown in Eq. (5.32) to be

$$\mathcal{L}(\Theta|\mathcal{X}) \triangleq \log[L(\Theta|\mathcal{X})] = \sum_{n=1}^{N} \log[p(\mathbf{x}_n|\Theta)] = \sum_{n=1}^{N} \log\left(\sum_{m=1}^{M} p(\lambda_m|\Theta) p(\mathbf{x}_n|\lambda_m, \Theta)\right), \tag{5.45}$$

where we have rewritten α_m and $p(\mathbf{x}_n|\lambda_m)$ in Eq. (5.32) as $p(\lambda_m|\Theta)$ and $p(\mathbf{x}_n|\lambda_m,\Theta)$, respectively.

In comparison, by introducing the weights $\{w_n\}$ into the EM cost function as shown in Eqs. (5.35) and (5.36), our modified EM algorithm maximizes, rather, a weighted version of the observation log-likelihoods. Compared to the log-likelihood function of Eq. (5.45) above, our weighted modification of the log-likelihood, shown in Eq. (5.35), can be written as

$$\mathcal{L}^w(\Theta|\mathcal{X}) \triangleq \sum_{n=1}^{N} w_n \log[p(\mathbf{x}_n|\Theta)] = \sum_{n=1}^{N} w_n \log\left(\sum_{m=1}^{M} p(\lambda_m|\Theta)p(\mathbf{x}_n|\lambda_m,\Theta)\right). \tag{5.46}$$

As an iterative procedure, the conventional EM algorithm translates the problem of finding $\overset{*}{\Theta} = \arg\max_\Theta \mathcal{L}(\Theta|\mathcal{X})$ into the equivalent problem of finding $\overset{*}{\Theta}$ in steps—indexed on k to generate $\Theta^{(k)}$ estimates, with the initial $\Theta^{(0)}$ estimate given *a priori*—such that, $\forall k \geq 0$, $\mathcal{L}(\Theta^{(k+1)}|\mathcal{X}) \geq \mathcal{L}(\Theta^{(k)}|\mathcal{X})$, or, alternatively, such that the difference in log-likelihoods is maximized, i.e., $\Theta^{(k+1)} = \arg\max_\Theta \mathcal{L}(\Theta|\mathcal{X}) - \mathcal{L}(\Theta^{(k)}|\mathcal{X})$. For our weighted log-likelihood function, $\mathcal{L}^w(\Theta|\mathcal{X})$, this corresponds to

$$\Theta^{(k+1)} = \arg\max_{\Theta} \mathcal{L}^w(\Theta|\mathcal{X}) - \mathcal{L}^w(\Theta^{(k)}|\mathcal{X}). \tag{5.47}$$

Thus, to prove the convergence of our weighted algorithm, we need only show that, $\forall k \geq 0$, $\mathcal{L}^w(\Theta^{(k+1)}|\mathcal{X}) - \mathcal{L}^w(\Theta^{(k)}|\mathcal{X}) \geq 0$. Making use of the weighted log-likelihood definition in Eq. (5.46), as well as Jensen's inequality combined with the facts that, $\forall m,n$, $P(\lambda_m|\mathbf{x}_n,\Theta^{(k)}) \geq 0$ and $\sum_m P(\lambda_m|\mathbf{x}_n,\Theta^{(k)}) = 1$,[152] we have, $\forall k \geq 0$,

$$\begin{aligned}&\mathcal{L}^w(\Theta|\mathcal{X}) - \mathcal{L}^w(\Theta^{(k)}|\mathcal{X})\\ &\quad= \sum_{n=1}^{N} w_n \log \sum_{m=1}^{M} p(\lambda_m|\Theta)p(\mathbf{x}_n|\lambda_m,\Theta) - \sum_{n=1}^{N} w_n P(\mathbf{x}_n|\Theta^{(k)})\end{aligned}$$

[152] Jensen's inequality states that, for the constants $\{c_i\}_{i\in\{1,\ldots,I\}}$ satisfying $c_i \geq 0\ \forall i$, and $\sum_i c_i = 1$,

$$\log\left(\sum_{i=1}^{I} c_i x_i\right) \geq \sum_{i=1}^{I} c_i \log(x_i).$$

See [188, Section 2] for a detailed proof.

$$
\begin{aligned}
&= \sum_{n=1}^{N} w_n \log \sum_{m=1}^{M} P(\lambda_m|\mathbf{x}_n, \Theta^{(k)}) \left(\frac{p(\lambda_m|\Theta)p(\mathbf{x}_n|\lambda_m, \Theta)}{P(\lambda_m|\mathbf{x}_n, \Theta^{(k)})} \right) - \sum_{n=1}^{N} w_n P(\mathbf{x}_n|\Theta^{(k)}) \\
&\geq \sum_{n=1}^{N} w_n \sum_{m=1}^{M} P(\lambda_m|\mathbf{x}_n, \Theta^{(k)}) \log \left(\frac{p(\lambda_m|\Theta)p(\mathbf{x}_n|\lambda_m, \Theta)}{P(\lambda_m|\mathbf{x}_n, \Theta^{(k)})} \right) - \sum_{n=1}^{N} w_n P(\mathbf{x}_n|\Theta^{(k)}) \\
&= \sum_{n=1}^{N} w_n \sum_{m=1}^{M} P(\lambda_m|\mathbf{x}_n, \Theta^{(k)}) \log \left(\frac{p(\lambda_m|\Theta)p(\mathbf{x}_n|\lambda_m, \Theta)}{P(\lambda_m|\mathbf{x}_n, \Theta^{(k)})P(\mathbf{x}_n|\Theta^{(k)})} \right) \\
&= \sum_{n=1}^{N} w_n \sum_{m=1}^{M} P(\lambda_m|\mathbf{x}_n, \Theta^{(k)}) \log \left(\frac{p(\mathbf{x}_n, \lambda_m|\Theta)}{P(\mathbf{x}_n, \lambda_m|\Theta^{(k)})} \right) \\
&\triangleq \Delta(\Theta|\Theta^{(k)}).
\end{aligned} \tag{5.48}
$$

Equivalently, by defining

$$
l(\Theta|\Theta^{(k)}) \triangleq \mathcal{L}^w(\Theta^{(k)}|\mathcal{X}) + \Delta(\Theta|\Theta^{(k)}), \tag{5.49}
$$

Eq. (5.48) can be stated as

$$
\mathcal{L}^w(\Theta|\mathcal{X}) \geq l(\Theta|\Theta^{(k)}), \tag{5.50}
$$

i.e., $\forall k \geq 0$, $l(\Theta|\Theta^{(k)})$ is bounded from above by $\mathcal{L}^w(\Theta|\mathcal{X})$. Secondly, we note that the $\log\left(\frac{p(\mathbf{x}_n,\lambda_m|\Theta)}{P(\mathbf{x}_n,\lambda_m|\Theta^{(k)})}\right)$ term in the expression for $\Delta(\Theta|\Theta^{(k)})$ in Eq. (5.48) reduces to zero for $\Theta = \Theta^{(k)}$; i.e., the two functions, $l(\Theta|\Theta^{(k)})$ and $\mathcal{L}^w(\Theta|\mathcal{X})$, are equal at $\Theta = \Theta^{(k)}$. Based on both these properties of the relationship between $l(\Theta|\Theta^{(k)})$ and $\mathcal{L}^w(\Theta|\mathcal{X})$,[153] we can then conclude that any value for Θ that increases $l(\Theta|\Theta^{(k)})$, also increases $\mathcal{L}^w(\Theta|\mathcal{X})$, and hence, maximizing $\mathcal{L}^w(\Theta|\mathcal{X})$—the objective of our weighted EM algorithm—is equivalent to maximizing $l(\Theta|\Theta^{(k)})$. In turn, given that the weighted log-likelihood maximized in the previous EM iteration, i.e., $\mathcal{L}^w(\Theta^{(k)}|\mathcal{X})$, is constant with respect to Θ, then, as indicated by Eq. (5.49), maximizing $l(\Theta|\Theta^{(k)})$ itself reduces to maximizing $\Delta(\Theta|\Theta^{(k)})$, thereby proving our earlier statement regarding the equivalence of maximizing the weighted log-likelihood difference—as shown in Eq. (5.47)—to the original objective of maximizing the weighted log-likelihood

[153] See [188, Figure 2] for an illustration of the relationship between $l(\Theta|\Theta^{(k)})$ and $\mathcal{L}(\Theta|\mathcal{X})$.

function per se. Thus, the weighted EM algorithm can be formally expressed as

$$\begin{aligned}\Theta^{(k+1)} &= \arg\max_{\Theta} \sum_{n=1}^{N} w_n \sum_{m=1}^{M} P(\lambda_m|\mathbf{x}_n, \Theta^{(k)}) \log\left(\frac{p(\mathbf{x}_n, \lambda_m|\Theta)}{P(\mathbf{x}_n, \lambda_m|\Theta^{(k)})}\right) \\ &= \arg\max_{\Theta} \sum_{m=1}^{M} \sum_{n=1}^{N} w_n \log[p(\mathbf{x}_n, \lambda_m|\Theta)] P(\lambda_m|\mathbf{x}_n, \Theta^{(k)}) \\ &\equiv \arg\max_{\Theta} E\left[\sum_{n=1}^{N} w_n \log p(\mathbf{x}_n, Y|\Theta) \Big| \mathcal{X}, \Theta^{(k)}\right], \end{aligned} \tag{5.51}$$

where the second step is obtained by dropping all the additive terms that are constant with respect to Θ, and where we have rewritten the random variable λ_m in the final step as Y to obtain an expression similar to that used in Eqs. (5.34) and (5.35) to derive $Q^w(\Theta, \Theta^{(k)})$.

Since $\Theta^{(k+1)}$ is chosen to maximize the weighted log-likelihood difference $\Delta(\Theta|\Theta^{(k)})$, then, given that $\Delta(\Theta^{(k)}|\Theta^{(k)}) = 0$ as noted above, we have, $\forall k \geq 0$,

$$\Delta(\Theta^{(k+1)}|\Theta^{(k)}) \geq \Delta(\Theta^{(k)}|\Theta^{(k)}) = 0; \tag{5.52}$$

i.e., the weighted log-likelihood function, $\mathcal{L}^w(\Theta|\mathcal{X})$, is consistently non-decreasing, thereby proving the convergence our weighted EM algorithm.

iv. Estimating child state densities through two-stage localized weighted EM

Using the weighted EM iterative update formulae derived above, we can now directly exploit the fuzzy membership and localization information captured in the $\{\mathcal{V}_i^{\mathbb{Z}^{(l)}, w^{(l-1)}}\}_{i \in \mathcal{I}^{(l)}}$ subsets to estimate the maximum weighted log-likelihood estimates of $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$—or, more specifically, of the $|\mathcal{J}_i^{(l)}|$ child state densities, $\{S_{ij}^{(l)} \triangleq (\alpha_{ij}^{\mathbb{Z}^{(l)}}, \lambda_{ij}^{\mathbb{Z}^{(l)}})\}_{j \in \mathcal{J}_i^{(l)}}$—modelling each ith region of the $\mathbb{Z}^{(l)}$ space. With the EM density estimation performed independently for each $i \in \mathcal{I}^{(l)}$, we proceed as follows:

(1) Initialization

As described above, we perform weighted EM in two stages, first modelling the variability of the incremental highband data in the static $\mathbf{Y}^{(l)}$ subspace, followed by extrapolating the obtained finer child subclasses into the entire $\mathbb{Z}^{(l)}$ space. Using k to denote the weighted EM iteration index spanning the two stages, we extend the

notation for the child state densities to be iteratively estimated at the lth memory inclusion index in the first and second weighted EM stages to $\{(\alpha_{ij}^{\mathbf{y}^{(l,k)}}, \lambda_{ij}^{\mathbf{y}^{(l,k)}})\}_{\forall i,j}$ and $\{(\alpha_{ij}^{\mathbb{z}^{(l,k)}}, \lambda_{ij}^{\mathbb{z}^{(l,k)}})\}_{\forall i,j}$, respectively. To initialize the first EM stage, we independently train a single J-modal GMM covering the entire time-independent static highband space, $\mathbf{Y} \equiv \mathbf{Y}^{(0)}$.[154] Given the extended notation above where the 2-tuple superscript $(\cdot,\cdot)$ denotes order of memory inclusion and iteration index, respectively, while the non-extended 1-tuple $(\cdot)$ denotes memory inclusion index only,[155] we denote the initialization GMM by $\mathcal{G}_{\mathbf{Y}}^{(0)} \coloneqq \mathcal{G}(\mathbf{y}; J, \mathrm{A}^{\mathbf{y}^{(0)}}, \Lambda^{\mathbf{y}^{(0)}})$, where $\mathrm{A}^{\mathbf{y}^{(0)}} = \{\alpha_j^{\mathbf{y}^{(0)}}\}_{j\in\{1,\ldots,J\}}$ and $\Lambda^{\mathbf{y}^{(0)}} = \{\lambda_j^{\mathbf{y}^{(0)}}\}_{j\in\{1,\ldots,J\}}$. This GMM represents the single 0th-iteration model to be used to initialize our localized weighted EM in all $|\mathcal{I}^{(l)}|$ regions of the $\mathbf{Y}^{(l)}$ subspace, at all values for the memory inclusion index l, rather than perform K-means clustering independently on each of the $\{\mathcal{V}_i^{\mathbf{y}^{(l)}} \xleftarrow{\text{Eq.(5.29)}} \mathcal{V}_i^{\mathbb{z}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets, as would typically be done to initialize EM training.[156] Since J, the number of Gaussian components in $\mathcal{G}_{\mathbf{Y}}^{(0)}$, thus also determines the number of uni-modal child states to be derived from each uni-modal parent state, we will often refer to it as the *splitting factor*.

The motivation for initializing the weighted EM training using $\mathcal{G}_{\mathbf{Y}}^{(0)}$ covering the entire $\mathbf{Y}$ space, rather than frequency-localized regions corresponding to each of the $\{\mathcal{V}_i^{\mathbb{z}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets, is detailed in Operation (d) below. As described in Section 5.4.2.2, however, we note here that initializing our localized EM training through $\mathcal{G}_{\mathbf{Y}}^{(0)}$ is intended to simultaneously capture the degree of variability in spectral characteristics across time for different sounds while also exploiting this variability to reduce redundancy in our overall tree-like model prior to performing weighted EM, and thereby, maximizing the model's information content. As described in Operation (d), reducing redundancies as such is equivalent to *pruning* $|\mathcal{J}_i^{(l)}|$—the number of the Gaussian components of $\mathcal{G}_{\mathbf{Y}_i^{(l)}}$ needed to model the variability of localized data in the $\mathbf{Y}^{(l)}$ subspace—for a particular subset of the $\mathcal{I}^{(l)}$ indices. In addition to this redundancy-reducing pruning performed prior to applying EM, we also apply a data-sufficiency pruning test—also detailed in Operation (d)—after weighted EM training has been applied at the current lth-order of memory inclusion, in order to ensure suf-

[154] See Footnote 146 regarding the time-dependency of static subspaces.

[155] As noted in Operation (a), the memory inclusion step, τ, was dropped from our initial superscript notation introduced in Section 5.4.2.2 to simplify notation.

[156] See Footnote 60.

ficient data is available to reliably estimate child state densities at the future $(l+1)$th order. Since this latter post-EM pruning condition can only be tested after weighted EM has already been applied, however, we need only consider the aforementioned pre-EM redundancy-reducing condition in the initialization step discussed here.

As summarized in Eq. (5.63), the net result of the pre-EM test for child Gaussian component pruning is that, $\forall i \in \mathcal{I}^{(l)}$, $|\mathcal{J}_i^{(l)}|$ is reduced to one of only two possible values, specifically $|\mathcal{J}_i^{(l)}| \in \{1, J\}$, depending on the value of a distribution flatness measure, ρ_i, calculated based on all incremental data in the $\mathcal{V}_i^{\mathbf{y}^{(l)}}$ subset. Thus, given $\mathcal{G}_{\mathbf{Y}}^{(0)}$, the initialization of our weighted EM algorithm can be summarized as follows:

1. For all $i \in \mathcal{I}^{(l)}$, we estimate ρ_i using Eqs. (5.60)–(5.62) as detailed in Operation (d) below.
2. Given a minimum distribution flatness threshold, $\rho_{\min}$, we apply the pruning condition in Eq. (5.63) to determine $\{i \in \mathcal{I}^{(l)}\colon \rho_i \geq \rho_{\min}\}$—the subset of parent state indices for each of which the incremental $\mathbf{Y}^{(l)}$ data is deemed sufficiently flat to warrant the splitting of the corresponding ith parent state into J child states, whose uni-modal *pdf*s are to be jointly estimated as the Gaussian components of $\mathcal{G}_{\mathbf{Y}_i}^{(l)}$ via weighted EM.
3. Finally, for each of the J-modal $\left\{\mathcal{G}_{\mathbf{Y}_i}^{(l)}\right\}$ GMMs corresponding to the subset of indices obtained above, we use the parameters of $\mathcal{G}_{\mathbf{Y}}^{(0)}$ as the initial 0th-iteration input to our weighted EM algorithm, i.e.,

$$\forall i \in \mathcal{I}^{(l)} | \, \rho_i \geq \rho_{\min}, j \in \mathcal{J}_i^{(l)} = \{1, \ldots, J\}\colon \quad \begin{aligned} \alpha_{ij}^{\mathbf{y}^{(l,0)}} &\xleftarrow{=} \alpha_j^{\mathbf{y}^{(0)}}, && \text{(5.53a)} \\ \boldsymbol{\mu}_{ij}^{\mathbf{y}^{(l,0)}} &\xleftarrow{=} \boldsymbol{\mu}_j^{\mathbf{y}^{(0)}}, && \text{(5.53b)} \\ \mathbf{C}_{ij}^{\mathbf{y}\mathbf{y}^{(l,0)}} &\xleftarrow{=} \mathbf{C}_j^{\mathbf{y}\mathbf{y}^{(0)}}. && \text{(5.53c)} \end{aligned}$$

(2) E-step

Using the $\left\{\left(\alpha_{ij}^{\mathbf{y}^{(l,0)}}, \lambda_{ij}^{\mathbf{y}^{(l,0)}} := (\boldsymbol{\mu}_{ij}^{\mathbf{y}^{(l,0)}}, \mathbf{C}_{ij}^{\mathbf{y}\mathbf{y}^{(l,0)}})\right)\right\}$ estimates of Eq. (5.53) above as initial $\left\{\mathcal{G}_{\mathbf{Y}_i}^{(l,0)}\right\}_{\forall i | \, \rho_i \geq \rho_{\min}}$ model estimates, we now proceed with the first stage of weighted EM training. Replacing the general random vector, $\mathbf{X}$, used in our derivation of weighted EM, by the incremental highband random feature vector, $\mathbf{Y}^{(l)}$, and the Gaussian component index, m, by ij, the E-step deduced from Eq. (5.36) reduces to estimating $P(\lambda_{ij}^{\mathbf{y}^{(l,k)}} | \mathbf{y}_n^{(l)}) \xleftarrow{\hat{=}} P(m | \mathbf{x}_n, \Theta^{(k)})$, for all $n | \, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}$ and all $ij | \, \rho_i \geq \rho_{\min}$—the indices remaining after the application of pre-EM pruning as described above. Thus, using

Bayes' rule, we have,

$$\forall i \in \mathcal{I}^{(l)} | \, \rho_i \geq \rho_{\min}, j \in \mathcal{J}_i^{(l)} = \{1, \ldots, J\}, n | \, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}:$$

$$P(\lambda_{ij}^{\mathbf{y}^{(l,k)}} | \mathbf{y}_n^{(l)}) = \frac{\alpha_{ij}^{\mathbf{y}^{(l,k)}} P(\mathbf{y}_n^{(l)} | \lambda_{ij}^{\mathbf{y}^{(l,k)}})}{\sum\limits_{m \in \mathcal{J}_i^{(l)}} \alpha_{im}^{\mathbf{y}^{(l,k)}} P(\mathbf{y}_n^{(l)} | \lambda_{im}^{\mathbf{y}^{(l,k)}})}. \tag{5.54}$$

(3) M-step

Similarly, by applying the same parameter substitutions noted above to Eqs. (5.39), (5.42), and (5.44), the first-stage M-step is given by

$$\forall i \in \mathcal{I}^{(l)} | \, \rho_i \geq \rho_{\min}, j \in \mathcal{J}_i^{(l)} = \{1, \ldots, J\}:$$

$$\alpha_{ij}^{\mathbf{y}^{(l,k+1)}} = \frac{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} P(\lambda_{ij}^{\mathbf{y}^{(l,k)}} | \mathbf{y}_n^{(l)})}{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)}}, \tag{5.55a}$$

$$\boldsymbol{\mu}_{ij}^{\mathbf{y}^{(l,k+1)}} = \frac{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} P(\lambda_{ij}^{\mathbf{y}^{(l,k)}} | \mathbf{y}_n^{(l)}) \, \mathbf{y}_n^{(l)}}{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} P(\lambda_{ij}^{\mathbf{y}^{(l,k)}} | \mathbf{y}_n^{(l)})}, \tag{5.55b}$$

$$\mathbf{C}_{ij}^{\mathbf{yy}^{(l,k+1)}} = \frac{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} P(\lambda_{ij}^{\mathbf{y}^{(l,k)}} | \mathbf{y}_n^{(l)}) \left[\mathbf{y}_n^{(l)} - \boldsymbol{\mu}_{ij}^{\mathbf{y}^{(l,k+1)}}\right]\left[\mathbf{y}_n^{(l)} - \boldsymbol{\mu}_{ij}^{\mathbf{y}^{(l,k+1)}}\right]^T}{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} P(\lambda_{ij}^{\mathbf{y}^{(l,k)}} | \mathbf{y}_n^{(l)})}. \tag{5.55c}$$

Applied individually over all $i \in \mathcal{I}^{(l)} | \, \rho_i \geq \rho_{\min}$, Eqs. (5.54) and (5.55) are iteratively repeated for each $\mathcal{G}_{\mathbf{Y}_i}^{(l)}$ GMM using the corresponding $\mathcal{V}_i^{\mathbf{y}^{(l)}}$ subset until the relative change in weighted log-likelihood for that ith subset, i.e.,

$$\Delta \mathcal{L}^w \triangleq \frac{\mathcal{L}^w(\Theta^{(k+1)} | \mathcal{V}_i^{\mathbf{y}^{(l)}, w^{(l-1)}}) - \mathcal{L}^w(\Theta^{(k)} | \mathcal{V}_i^{\mathbf{y}^{(l)}, w^{(l-1)}})}{\mathcal{L}^w(\Theta^{(k)} | \mathcal{V}_i^{\mathbf{y}^{(l)}, w^{(l-1)}})} \tag{5.56}$$

where

$$\mathcal{L}^w(\Theta^{(k)} | \mathcal{V}_i^{\mathbf{y}^{(l)}, w^{(l-1)}}) = \sum_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} \log \sum_{j \in \mathcal{J}_i^{(l)}} \alpha_{ij}^{\mathbf{y}^{(l,k)}} P(\mathbf{y}_n^{(l)} | \lambda_{ij}^{\mathbf{y}^{(l,k)}}), \tag{5.57}$$

falls below a particular threshold, $\Delta\mathcal{L}^{w}_{\max}$, thereupon concluding the first stage of our weighted EM-based child state *pdf* estimation.

(4) Final E-step

Finally, through a single weighted EM iteration, we extrapolate the finer child subclasses obtained above in the $\mathbf{Y}^{(l)}$ subspace into the joint-band $\mathbb{Z}^{(l)}$ space. As previously discussed, this extrapolation is achieved by extending the $(l-1)$th-order time-frequency information available in the joint-band $\{\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}$ subsets using the new finer $\mathbf{Y}^{(l)}$-subspace localization information captured into the $\left\{\mathcal{G}^{(l)}_{\mathbf{Y}_i}\right\}$ GMMs corresponding to the non-pruned $\mathcal{I}^{(l)}$ indices. In particular, we first determine the child subclass membership probabilities of the fully-extended lth-order joint-band data in the $\{\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}_{\forall i|\rho_i\geq\rho_{\min}}$ subsets based entirely on the new membership information captured into the $\left\{\mathcal{G}^{(l)}_{\mathbf{Y}_i}\right\}_{\forall i|\rho_i\geq\rho_{\min}}$ GMMs. This effectively augments the information incorporated previously during the construction of the $\{\mathcal{V}_i^{\mathbf{y}^{(l)},w^{(l-1)}}\leftarrow\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}_{\forall i|\rho_i\geq\rho_{\min}}$ subsets—the subsets used to estimate $\left\{\mathcal{G}^{(l)}_{\mathbf{Y}_i}\right\}_{\forall i|\rho_i\geq\rho_{\min}}$ in the first EM stage above—about time-frequency localization in the lower-order $\mathbb{Z}^{(l-1)}$ subspace using the new finer localization information learned through modelling variability in the incremental $\mathbf{Y}^{(l)}$ subspace. Then, in a second step, we estimate the parameters of the joint-band $\left\{\mathcal{G}^{(l)}_{\mathbb{Z}_i}\right\}_{\forall i|\rho_i\geq\rho_{\min}}$ GMMs as those maximizing the weighted log-likelihoods of the corresponding $\{\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}_{\forall i|\rho_i\geq\rho_{\min}}$ subsets given those child subclass memberships determined as described above. These J-modal $\left\{\mathcal{G}^{(l)}_{\mathbb{Z}_i}\right\}_{\forall i|\rho_i\geq\rho_{\min}}$ GMMs, together with the uni-modal $\left\{\mathcal{G}^{(l)}_{\mathbb{Z}_i}\right\}_{\forall i|\rho_i<\rho_{\min}}$ densities estimated in Operation (d) below, represent the densities to be used for future fuzzy clustering in order to obtain the $(l+1)$th-order $\{\mathcal{V}_i^{\mathbb{z}^{(l+1)},w^{(l)}}\}_{i\in\mathcal{I}^{(l+1)}}$ subsets, as described in Operations (a) and (b).

The first step—namely the estimation of child subclass membership probabilities for data in the $\{\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}_{\forall i|\rho_i\geq\rho_{\min}}$ subsets—is simply implemented through an additional E-step using $\{\mathcal{V}_i^{\mathbf{y}^{(l)},w^{(l-1)}}\}_{\forall i|\rho_i\geq\rho_{\min}}$ and the $\left\{\left(\alpha_{ij}^{\mathbf{y}^{(l,k)}},\lambda_{ij}^{\mathbf{y}^{(l,k)}}\right)\right\}$ parameters maximized in the first EM stage; i.e.,

$$\forall i\in\mathcal{I}^{(l)}|\,\rho_i\geq\rho_{\min}, j\in\mathcal{J}_i^{(l)}=\{1,\ldots,J\}, n|\,\mathbb{z}_n^{(l)}\in\mathcal{V}_i^{\mathbb{z}^{(l)}}:$$

$$P(\lambda_{ij}^{\mathbb{z}^{(l,k)}}|\mathbb{z}_n^{(l)})\xleftarrow{=}P(\lambda_{ij}^{\mathbf{y}^{(l,k)}}|\mathbf{y}_n^{(l)})=\frac{\alpha_{ij}^{\mathbf{y}^{(l,k)}}P(\mathbf{y}_n^{(l)}|\lambda_{ij}^{\mathbf{y}^{(l,k)}})}{\sum\limits_{m\in\mathcal{J}_i^{(l)}}\alpha_{im}^{\mathbf{y}^{(l,k)}}P(\mathbf{y}_n^{(l)}|\lambda_{im}^{\mathbf{y}^{(l,k)}})}. \tag{5.58}$$

(5) Final M-step

Similarly, the second step—namely the estimation of the maximum weighted log-likelihood values for the $\left\{\mathcal{G}_{\mathbb{z}_i}^{(l)}\right\}_{\forall i|\rho_i\geq\rho_{\min}}$ model parameters given the $\{P(\lambda_{ij}^{\mathbb{z}^{(l,k)}}|\mathbb{z}_n^{(l)})\}$ posterior probabilities obtained in Eq. (5.58) above—is implemented through a final M-step using the joint-band lth-order data in the $\{\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}_{\forall i|\rho_i\geq\rho_{\min}}$ subsets; i.e.,

$\forall i \in \mathcal{I}^{(l)} | \, \rho_i \geq \rho_{\min}, j \in \mathcal{J}_i^{(l)} = \{1, \ldots, J\}$:

$$\alpha_{ij}^{\mathbb{z}^{(l,k+1)}} = \frac{\sum\limits_{\{n:\,\mathbb{z}_n^{(l)}\in\mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} P(\lambda_{ij}^{\mathbb{z}^{(l,k)}}|\mathbb{z}_n^{(l)})}{\sum\limits_{\{n:\,\mathbb{z}_n^{(l)}\in\mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)}}, \tag{5.59a}$$

$$\boldsymbol{\mu}_{ij}^{\mathbb{z}^{(l,k+1)}} = \frac{\sum\limits_{\{n:\,\mathbb{z}_n^{(l)}\in\mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} P(\lambda_{ij}^{\mathbb{z}^{(l,k)}}|\mathbb{z}_n^{(l)})\,\mathbb{z}_n^{(l)}}{\sum\limits_{\{n:\,\mathbb{z}_n^{(l)}\in\mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} P(\lambda_{ij}^{\mathbb{z}^{(l,k)}}|\mathbb{z}_n^{(l)})}, \tag{5.59b}$$

$$\mathbf{C}_{ij}^{\mathbb{z}\mathbb{z}^{(l,k+1)}} = \frac{\sum\limits_{\{n:\,\mathbb{z}_n^{(l)}\in\mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} P(\lambda_{ij}^{\mathbb{z}^{(l,k)}}|\mathbb{z}_n^{(l)}) \left[\mathbb{z}_n^{(l)} - \boldsymbol{\mu}_{ij}^{\mathbb{z}^{(l,k+1)}}\right]\left[\mathbb{z}_n^{(l)} - \boldsymbol{\mu}_{ij}^{\mathbb{z}^{(l,k+1)}}\right]^T}{\sum\limits_{\{n:\,\mathbb{z}_n^{(l)}\in\mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} P(\lambda_{ij}^{\mathbb{z}^{(l,k)}}|\mathbb{z}_n^{(l)})}. \tag{5.59c}$$

As previously noted, since the $\{\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l-1)}}\}_{\forall i|\rho_i\geq\rho_{\min}}$ subsets also include partial information about the localization of incremental static narrowband data in the $\mathbf{X}^{(l)}$ subspace, maximizing the weighted log-likelihood of these *joint-band* subsets using the finer $\mathbf{Y}^{(l)}$ *highband*-subspace localization information per Eqs. (5.58) and (5.59) implicitly incorporates the important cross-correlation information between data distributions in the $\mathbf{X}^{(l)}$ and $\mathbf{Y}^{(l)}$ subspaces into our lth-order joint-band $\left\{\mathcal{G}_{\mathbb{z}_i}^{(l)}\right\}_{\forall i|\rho_i\geq\rho_{\min}}$ models of child state densities.

v. On the effect of time-frequency localization on computational complexity

To conclude this description of our approach to *pdf* estimation, we note that the overall computational complexity associated with estimating lth-order joint-band densities through our localizing tree-like approach is significantly lower than that required

for global estimation using the conventional EM algorithm whose computational limitations were detailed in Section 5.4.2.1. The reduction in complexity follows directly from localization across time and frequency. In particular:

1. The localization of training data effectively constrains variability across the incremental subspace. Modelling such constrained variability individually within each localized region, in turn, considerably reduces J—the number of Gaussian components needed for mixture modelling, or alternatively, the splitting factor—below what would typically be required to model unconstrained variability across the entire incremental subspace. Indeed, as shown by the results to be detailed in Section 5.4.3.2 based on an initial $\mathcal{G}_{\mathbb{Z}}^{(0)}$ global GMM with $I \coloneqq |\mathcal{J}_1^{(0)}| = 128$, BWE performance saturates at a splitting factor of $J \simeq 4$–6, compared to the ~ 128 components needed for performance saturation when modelling an entire static space.[157]
2. The localization of training data through GMM-based clustering further results in smaller subsets of data. Such reduced subset cardinalities, in turn, translate to fewer operations to be performed at each weighted EM iteration. As detailed in Operation (d) below, we impose an post-EM pruning condition to ensure that the amount of data available for EM training does not fall after fuzzy clustering below the previously-determined threshold of $N_{\text{f/p}} \approxeq 10$.[158]
3. Finally, the localization of training data allows us to estimate the *pdf*s of joint-band data with higher orders of memory inclusion incrementally. This, in turn, allows us to progressively extend our model temporally by modelling variability primarily along the incremental static highband subspaces, $\{\mathbf{Y}^{(l)}\}_{\forall l \geq 0}$, rather than along the fully-extended joint-band spaces, $\{\mathbb{Z}^{(l)}\}_{\forall l \geq 0}$, thereby significantly reducing modelling complexity as a direct result of the difference in dimensionalities—which, in fact, consistently grows with the increase in order of memory inclusion, l.

[157] See Section 3.5.1 and Figure 3.4, in particular, which illustrates static BWE $\overline{d}_{\text{LSD}}$ performance as a function of M, the number of Gaussian components in the global GMM.

[158] See Section 3.5.2 and Figure 3.7, in particular, which illustrates static BWE $\overline{d}_{\text{LSD}}$ and $\overline{Q}_{\text{PESQ}}$ performances as a function of $N_{\text{f/p}}$, the number of data points (frames) available for training per GMM parameter.

(d) Addressing redundancies and potential overfitting by pruning

In introducing our tree-like approach for memory inclusion in Section 5.4.2.2, we emphasized that exploiting the temporal characteristics of speech to achieve a hierarchical time-frequency model represents one of our primary objectives. As detailed in the previous steps, making use of the strong correlation properties between neighbouring frames to carry over time-frequency localization information, from modelling at one memory inclusion index to the next, represents the first means by which temporal characteristics were incorporated in our modelling algorithm. To further incorporate speech temporal characteristics into our model while simultaneously reducing model complexity, we attempt to capture and exploit the redundancies in spectral characteristics that may be present at different temporal sections of the various speech classes underlying our localized time-frequency regions.[159] In particular, similar in concept to maximizing the entropy or information content of a coded speech signal through exploiting the well-known redundancies in speech signals, we measure the extent of spectral variability for the new incremental static highband data in each of the $\{\mathcal{V}_i^{\mathbf{y}^{(l)}} \leftarrow \mathcal{V}_i^{\mathbb{Z}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets. Then, prior to performing weighted EM, we decide accordingly whether such variability warrants *splitting* the ith parent cluster, subclass, or state, for all $i \in \mathcal{I}^{(l)}$, into $|\mathcal{J}_i^{(l)}| = J$ child or daughter states, where J is a splitting factor determined in practice as the number of Gaussian components in the EM initialization GMM, $\mathcal{G}_{\mathbf{Y}}^{(0)}$, as opposed to *pruning* the number of child states to only one, i.e., $|\mathcal{J}_i^{(l)}| = 1$. Our implementation of such pre-EM redundancy-reducing pruning is detailed below.

As discussed in Section 5.4.2.2 and further detailed in Operation (a) above, one of the motivations for our fuzzy clustering approach was to alleviate the risk of overfitting. However, the non-decreasing growth illustrated in Figure 5.8 for the number of the time-frequency states obtained through our tree-like modelling approach motivates us to ensure that sufficient data is always available to reliably estimate those child state densities to be obtained at the future $(l+1)$th-order of memory inclusion based on the lth-order states. As such, we also impose a post-EM pruning condition that directly compares the cardinality of each of the $|\mathcal{I}^{(l+1)} \xleftarrow{=} \mathcal{K}^{(l)}|$ data

[159] See Footnote 139 for examples of the variation in spectral redundancies across time for different sound classes.

subsets, $\{\mathcal{V}_i^{\mathbb{Z}^{(l+1)}}\}_{i\in\mathcal{I}^{(l+1)}} \xleftarrow{\text{Eq.(5.26)}} \{\mathcal{V}_k^{\mathbb{Z}^{(l)}}\}_{k\in\mathcal{K}^{(l)}}$, to a particular threshold determined as a function of the $\mathbf{Y}^{(l+1)}$ subspace dimensionality as well as the number of uni-modal child state densities, J, to be estimated for each parent data subset. We apply this data-sufficiency check after, rather than before, weighted EM training—and thus, potentially pruning some lth-order child state densities despite having been already trained using EM—in order to account for the decrease in $(l+1)$th-order subset cardinalities associated with the edge cases at training audio sample boundaries.[160]

i. Pre-EM pruning

As described in Operation (c), the $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}} \leftarrow \{\mathcal{V}_i^{\mathbb{Z}^{(l)}}\}_{i\in\mathcal{I}^{(l)}} \leftarrow \{\mathcal{V}_k^{\mathbb{Z}^{(l-1)}}\}_{k\in\mathcal{K}^{(l-1)}}$ subsets comprise all previously-obtained information about the distribution of the data in the $\{\mathbb{Z}^{(m)}\}_{m\in\{0,\dots,l-1\}}$ subspaces, including that of time-frequency localization. Hence, these subsets are considered to be reliably and highly localized in time-frequency along the lower-order $\{\mathbb{Z}^{(m)}\}_{m\in\{0,\dots,l-1\}}$ subspaces. In contrast, the $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets contain only partial information about frequency-only localization in the incremental $\mathbf{Y}^{(l)} \leftarrow \mathbf{Z}^{(l)}$ static subspaces added by temporal extension as described in Operation (b). The extent of the correlation of such partial localization information to that in the lower-order subspaces depends entirely on the correlation between the time-dependent $\mathbf{Z}^{(l)} := \mathbf{Z}_{t-l\tau}$ spectra to those of their neighbouring past $\{\mathbf{Z}_{t-m\tau}\}_{m\in\{0,\dots,l-1\}}$ spectra; higher cross-time spectral correlation translates to equally-high frequency localization, and vice versa. Since static $\mathbf{Z}_{t-l\tau}$ spectra that correlate highly with their neighbouring past counterparts add little new information to that already existing in the lower-order $\{\mathcal{V}_i^{\mathbb{Z}^{(m)}}\}_{i\in\mathcal{I}^{(m)}}\big|_{\forall m<l}$ subsets, splitting parent $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets where the $\mathbf{Y}_{t-l\tau} \leftarrow \mathbf{Z}_{t-l\tau}$ data exhibits limited variability in the entire $\mathbf{Y}_{t-l\tau}$ subspace unnecessarily increases our tree-like model's complexity as well as increase the risk of overfitting. Instead, we attempt to maximize the information content of our model by focusing only on those data subsets where the distribution of the incremental $\mathbf{Y}_{t-l\tau}$ data exhibits higher entropy, i.e., where the distribution of the $\mathbf{Y}_{t-l\tau}$ data is *flatter*, rather than *peakier* or more localized, over the entire span of the $\mathbf{Y}_{t-l\tau}$ subspace. To that end, we define a *distribution flatness measure* to quantify the variability of the incremental $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ data in the static $\mathbf{Y}^{(l)}$ subspace, with the flatness esti-

[160] See Eqs. (5.26) and (5.27) for the effect of edge cases on reducing the size of temporally-extended data subsets.

mated based on the variation in the posterior probabilities of the individual Gaussian components of a GMM trained independently to model the entire time-independent $\mathbf{Y} \equiv \mathbf{Y}^{(0)}$ static subspace, given the $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ data. Such a GMM had already been introduced as the reference J-modal $\mathcal{G}_{\mathbf{Y}}^{(0)}$ used for EM initialization.

Similar in concept to the *spectral flatness measure*, introduced in [189] to quantify the *tonality*, or conversely, the *noisiness*, of audio spectra, i.e., their variability across frequency, our distribution flatness measure quantifies the *peakiness*, or conversely, the *whiteness*, of the distribution of incremental static highband data across the frequency-only axis of the $\mathbf{Y}^{(l)}$ subspace. This measure is individually estimated for each of the $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets, based on per-child-state weighted Bayesian occupancies—denoted by $\{\mathcal{O}_{ij}^{\mathbf{y}^{(l)}}\}$, for all $i \in \mathcal{I}^{(l)}$ and $j \in \{1,\ldots,J\}$—which, in turn, are estimated based on the aforementioned posterior probabilities of the J components of $\mathcal{G}_{\mathbf{Y}}^{(0)}$ given the static highband data in each $\mathcal{V}_i^{\mathbf{y}^{(l)}}$ subset.

To estimate $\{\mathcal{O}_{ij}^{\mathbf{y}^{(l)}}\}$, we first define $o_{ij,n}^{(l)}$ representing the hard-decision Bayesian occupancy of the jth initial Gaussian component of $\mathcal{G}_{\mathbf{Y}}^{(0)}$, $(\alpha_j^{\mathbf{y}^{(0)}}, \lambda_j^{\mathbf{y}^{(0)}})$, given the nth data point, $\mathbf{y}_n^{(l)}$, belonging to the ith static highband subset, $\mathcal{V}_i^{\mathbf{y}^{(l)}}$. Then, by adapting our $\overset{*}{\lambda}{}_{ij_k,n}^{\mathbb{z}^{(l)}}$ notation defined in Eq. (5.16) for the kth most-likely Gaussian component, the per-data-point hard-decision occupancies, $\{o_{ij,n}^{(l)}\}_{\forall i,j,n}$, can be written as

$$\forall i \in \mathcal{I}^{(l)}, j \in \{1,\ldots,J\}, n | \mathbf{y}_n^{(l)} \in \mathcal{V}_i^{\mathbf{y}^{(l)}}\text{:} \quad o_{ij,n}^{(l)} = \begin{cases} 1, & \text{if } \overset{*}{\lambda}{}_{j_1,n}^{\mathbf{y}^{(0)}} = \lambda_j^{\mathbf{y}^{(0)}}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.60}$$

Given $\{o_{ij,n}^{(l)}\}_{\forall i,j,n}$, we then estimate the per-child-state weighted Bayesian occupancies per

$$\forall i \in \mathcal{I}^{(l)}, j \in \{1,\ldots,J\}\text{:} \quad \mathcal{O}_{ij}^{\mathbf{y}^{(l)}} = \frac{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} o_{ij,n}^{(l)}}{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)}}, \tag{5.61}$$

using which the distribution flatness, ρ_i, in the $\mathbf{Y}^{(l)}$ subspace, for each of the $|\mathcal{I}^{(l)}|$ $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ subsets, is obtained as the ratio of the geometric mean of the per-child-

state $\{\mathcal{O}_{ij}^{\mathbf{y}^{(l)}}\}_{j\in\{1,\ldots,J\}}$ occupancies to their arithmetic mean; i.e.,

$$\forall i \in \mathcal{I}^{(l)}: \quad \rho_i = \frac{\left(\prod_{j=1}^{J} \mathcal{O}_{ij}^{\mathbf{y}^{(l)}}\right)^{\frac{1}{J}}}{\frac{1}{J}\sum_{j=1}^{J} \mathcal{O}_{ij}^{\mathbf{y}^{(l)}}} \leq 1, \tag{5.62}$$

where lower ρ_i values correspond to peakier, and hence more localized, variability of data in the $\mathbf{Y}^{(l)}$ subspace, and vice versa.

Given a minimum distribution flatness threshold, $\rho_{\min}$, the redundancy-reducing pre-EM pruning test can then be summarized as

$$\forall i \in \mathcal{I}^{(l)}: \quad (\rho_i < \rho_{\min}) \Rightarrow |\mathcal{J}_i^{(l)}| = 1 \ \wedge \ \neg(\rho_i < \rho_{\min}) \Rightarrow |\mathcal{J}_i^{(l)}| = J. \tag{5.63}$$

ii. Post-EM pruning

In addition to the pre-EM redundancy-reducing pruning described above, we also apply a post-EM pruning check to guarantee that the number of data points in the $(l+1)$th-order data subsets to be determined based on the EM-trained lth-order child states—$\{S_{ij}^{(l)} \triangleq (\alpha_{ij}^{\mathbb{z}^{(l)}}, \lambda_{ij}^{\mathbb{z}^{(l)}})\}$, for all $i \in \mathcal{I}^{(l)} | \rho_i \geq \rho_{\min}$ and all $j \in \{1,\ldots,J\}$—is sufficient to reliably estimate finer descendent densities at the future $(l+1)$th-order of memory inclusion. Ensuring a minimum cardinality as such for all subsets obtained through weighted EM is motivated by the progressive decrease in subset cardinalities with increasing memory inclusion index. In particular:

(a) as described in Operation (a), partitioning an arbitrary subset, $\mathcal{V}_i^{\mathbb{z}^{(l)}}$, into J overlapping subsets, $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)}}\}_{j\in\{1,\ldots,J\}}$, based on the K highest soft class memberships of each constituent data point into the J classes underlying a mixture model of the $\mathcal{V}_i^{\mathbb{z}^{(l)}}$ data, results in lower $\mathcal{V}_{ij}^{\mathbb{z}^{(l)}}$ child subset cardinalities—compared to that of the parent $\mathcal{V}_i^{\mathbb{z}^{(l)}}$ subset—for any value of the fuzziness factor satisfying $K < J$;

(b) as suggested by the existence condition for incremental $\mathbf{Z}_{t-(l+1)\tau}$ data in Eq. (5.26), extending an lth-order $\mathcal{V}_k^{\mathbb{z}^{(l)}} \xleftarrow{\text{Eq.(5.24)}} \mathcal{V}_{ij}^{\mathbb{z}^{(l)}}$ child data subset into its $(l+1)$th-order $\mathcal{V}_k^{\mathbb{z}^{(l+1)}} \xleftarrow{\text{Eq. (5.26)}} \mathcal{V}_k^{\mathbb{z}^{(l)}}$ counterpart—by augmenting the $\mathbb{Z}^{(l)}$ feature vectors in $\mathcal{V}_k^{\mathbb{z}^{(l)}}$ with their corresponding incremental $\mathbf{Z}^{(l+1)}$ data—may result in reduced

cardinality, i.e., $|\mathcal{V}_k^{\mathbb{Z}^{(l+1)}}| < |\mathcal{V}_k^{\mathbb{Z}^{(l)}}|$, as a result of the elimination of edge cases at training audio sample boundaries where no $\mathbf{Z}_{t-(l+1)\tau}$ frames exist for the $\mathbf{Z}^{(l)}$ data in $\mathcal{V}_k^{\mathbb{Z}^{(l)}}$.

Let $N_{\min}$ denote the minimum subset cardinality to be ensured for all child subsets derived from weighted EM-based child states. Then, at the conclusion of each $(l < L)$th memory inclusion iteration and for all $i \in \mathcal{I}^{(l)} | \rho_i \geq \rho_{\min}$ and all $j \in \mathcal{J}_i^{(l)} = \{1, \ldots, J\}$, we compare the cardinality of each $\mathcal{V}_k^{\mathbb{Z}^{(l+1)}} \xleftarrow{\text{Eq.(5.26)}} \mathcal{V}_k^{\mathbb{Z}^{(l)}} \xleftarrow{\text{Eq.(5.24)}} \mathcal{V}_{ij}^{\mathbb{Z}^{(l)}}$ child subset—i.e., all $(l+1)$th-order subsets obtained after weighted EM has been applied at order l, followed by fuzzy clustering and the subsequent $\mathbb{Z}^{(l)} \xrightarrow{\text{Eq.(5.26)}} \mathbb{Z}^{(l+1)}$ temporal extension steps—against $N_{\min}$; if the cardinality of one or more of the J $(l+1)$th-order child subsets derived from any particular lth-order $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$ model falls below $N_{\min}$, the underlying lth-order children states—whose *pdf*s have already been jointly estimated as $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$ using weighted EM—are pruned into a single lth-order child state whose uni-modal density is to be re-estimated as shown below.

As shown in Section 3.5.2, the reliable estimation of *pdf*s using full-covariance GMMs is achieved with a minimum of 10 training data points per GMM parameter; i.e., $N_{\text{f/p}} \geq 10$. Thus, using the formula given in Eq. (3.18) for the relation between the number of Gaussian components in a GMM to the number of training data points available per GMM parameter, the minimum cardinality, $N_{\min}$, of a child subset can then be obtained by expressing the cardinality, N, as a function of: (a) J, the number of future Gaussian components—or child states—to be derived from that subset; (b) $N_{\text{f/p}}$, the number of training data points needed per GMM parameter to ensure reliable parameter estimation; and (c) $q := \text{Dim}(\mathbf{Y}^{(l)}) = \text{Dim}(\mathbf{Y})$, the static highband feature vector dimensionality, thus focusing only on highband dimensionality since *pdf* estimation via weighted EM is performed primarily in the incremental highband subspace. In particular,

$$\begin{aligned} N &= N_{\text{f/p}} J \left(1 + q + \frac{q(q+1)}{2}\right) \\ &\geq 10J \left(1 + q + \frac{q(q+1)}{2}\right) \triangleq N_{\min}. \end{aligned} \tag{5.64}$$

Using $N_{\min}$ determined as such and making use of the child subset $ij \to k$ index mapping of Eq. (5.24), the post-EM data-sufficiency pruning condition can then be

summarized as

$$\forall i \in \mathcal{I}^{(l)} | \rho_i \geq \rho_{\min}\text{:}$$

$$|\mathcal{J}_i^{(l)}| = \begin{cases} 1, & \text{if } \exists k = j + \sum_{m<i} |\mathcal{J}_m^{(l)}| \text{: } |\mathcal{V}_k^{\mathbb{z}^{(l+1)}}| < N_{\min},\ \forall j \in \mathcal{J}_i^{(l)} = \{1, \ldots, J\}, \\ J, & \text{otherwise,} \end{cases} \tag{5.65}$$

where, as described above, each $(l+1)$th-order $\mathcal{V}_k^{\mathbb{z}^{(l+1)}}$ subset is obtained from a corresponding lth-order $\mathcal{V}_i^{\mathbb{z}^{(l)}}$ parent subset by performing fuzzy clustering based on $\mathcal{G}_{\mathbb{z}_i}^{(l)}$ followed by temporal extension; i.e., $\mathcal{V}_k^{\mathbb{z}^{(l+1)}} \xleftarrow{\text{Eq.(5.26)}} \mathcal{V}_k^{\mathbb{z}^{(l)}} \xleftarrow{\text{Eq.(5.24)}} \mathcal{V}_{ij}^{\mathbb{z}^{(l)}} \xleftarrow{\text{Eq.(5.18)}} \mathcal{V}_i^{\mathbb{z}^{(l)}}$.

iii. Estimating the parameters of pruned child densities

Finally, the uni-modal densities of those lth-order single-child, or single-component, $\left\{\mathcal{G}_{\mathbb{z}_i}^{(l)}\right\}_{\forall i | \mathcal{J}_i^{(l)} = \{1\}}$ models—i.e., the models corresponding to the pruned $\mathcal{I}^{(l)}$ indices in Eqs. (5.63) and (5.65)—can be straightforwardly estimated by finding the Gaussian *pdf* parameters—i.e., $\{(\boldsymbol{\mu}_{i1}^{\mathbb{z}^{(l)}}, \mathbf{C}_{i1}^{\mathbb{z}\mathbb{z}^{(l)}})\}_{\forall i | \mathcal{J}_i^{(l)} = \{1\}}$, with the $\{\alpha_{i1}^{\mathbb{z}^{(l)}}\}_{\forall i | \mathcal{J}_i^{(l)} = \{1\}}$ priors all reducing to unity—which maximize the weighted log-likelihoods of the corresponding lth-order $\{\mathcal{V}_i^{\mathbb{z}^{(l)}}\}_{\forall i | \mathcal{J}_i^{(l)} = \{1\}}$ parent subsets. In particular, since, for these $\left\{\mathcal{G}_{\mathbb{z}_i}^{(l)}\right\}_{\forall i | \mathcal{J}_i^{(l)} = \{1\}}$ models, $\mathcal{J}_i^{(l)} = \{1\}$, the child subclass memberships of the corresponding data—i.e., the posterior probabilities of the $|\mathcal{J}_i^{(l)}|$ Gaussian components given the data in $\{\mathcal{V}_i^{\mathbb{z}^{(l)}}\}_{\forall i | \mathcal{J}_i^{(l)} = \{1\}}$, or $\{\mathcal{V}_i^{\mathbf{y}^{(l)}}\}_{\forall i | \mathcal{J}_i^{(l)} = \{1\}}$—simply reduce to unity, for all data points. This, in turn, reduces the four weighted EM steps detailed in Operation (c) above for the estimation of $\mathcal{G}_{\mathbb{z}_i}^{(l)}$ models to a single weighted Maximization step similar to the final M-step of Eq. (5.59). As such, the estimation of the pruned child densities is given by

$$\forall i \in \mathcal{I}^{(l)} | \, \mathcal{J}_i^{(l)} = \{1\}\text{:}$$

$$P(\lambda_{ij}^{\mathbb{z}^{(l)}} | \mathbb{z}_n^{(l)}) = P(\lambda_{ij}^{\mathbf{y}^{(l)}} | \mathbf{y}_n^{(l)}) = 1,\ \forall \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}, \mathbf{y}_n^{(l)} \in \mathcal{V}_i^{\mathbf{y}^{(l)}}, \tag{5.66a}$$

$$\Rightarrow \alpha_{i1}^{\mathbb{z}^{(l)}} = 1, \tag{5.66b}$$

$$\boldsymbol{\mu}_{i1}^{\mathbb{Z}^{(l)}} = \frac{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)}\, \mathbb{z}_n^{(l)}}{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)}}, \tag{5.66c}$$

$$\mathbf{C}_{i1}^{\mathbb{Z}\mathbb{Z}^{(l)}} = \frac{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)} \big[\mathbb{z}_n^{(l)} - \boldsymbol{\mu}_{i1}^{\mathbb{Z}^{(l)}}\big]\big[\mathbb{z}_n^{(l)} - \boldsymbol{\mu}_{i1}^{\mathbb{Z}^{(l)}}\big]^T}{\sum\limits_{\{n:\, \mathbb{z}_n^{(l)} \in \mathcal{V}_i^{\mathbb{z}^{(l)}}\}} w_{i,n}^{(l-1)}}. \tag{5.66d}$$

At this point, it is worth noting that, for the pruned uni-modal $\left\{\mathcal{G}_{\mathbb{Z}_i}^{(l)}\right\}_{\forall i|\mathcal{J}_i^{(l)}=\{1\}}$ models estimated as such, performing fuzzy clustering per Operation (a) on the corresponding $\{\mathcal{V}_i^{\mathbb{Z}^{(l)}}\}_{\forall i|\mathcal{J}_i^{(l)}=\{1\}}$ parent data subsets reduces to simply updating the $\{\mathcal{V}_i^{w^{(l-1)}}\}_{\forall i|\mathcal{J}_i^{(l)}=\{1\}}$ parent membership weight subset counterparts into the $\{\mathcal{V}_{i1}^{w^{(l)}}\}_{\forall i|\mathcal{J}_i^{(l)}=\{1\}}$ child subsets with unity lth-order membership weights—per Eqs. (5.19) and (5.20).

(e) Constructing global GMMs

i. Consolidating children pdfs

Given all the $|\mathcal{K}^{(l)}|$ lth-order uni-modal child state densities derived as described above from their respective $|\mathcal{I}^{(l)}|$ lth-order parent states—which are simultaneously the $|\mathcal{K}^{(l-1)}|$ $(l-1)$th-order children states as indicated by Eq. (5.25)—via weighted EM and pruning in Operations (c) and (d), respectively, we conclude the lth increment of our tree-like modelling algorithm by constructing a global GMM, $\mathcal{G}_{\mathbb{Z}}^{(l)}$, modelling the *pdf* over the entire lth-order temporally-extended joint-band space, $\mathbb{Z}^{(l)}$. In order to consolidate all localized $\{\mathcal{G}_{\mathbb{Z}_i}^{(l)}\}_{i\in\mathcal{I}^{(l)}}$ models into a single $\mathcal{G}_{\mathbb{Z}}^{(l)}$ GMM as such, however, the component priors of $\{\mathcal{G}_{\mathbb{Z}_i}^{(l)}\}_{i\in\mathcal{I}^{(l)}}$ must be adjusted. This follows as a result of our approach of breaking down the estimation of a single global *pdf* covering the entire $\mathbb{Z}^{(l)}$ space into the estimation of $|\mathcal{I}^{(l)}|$ localized and independent $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$ *pdf*s, for each of which the $\{\alpha_{ij}^{\mathbb{z}^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$ component priors sum to unity. Since the priors do not thus sum to unity when considering all $\{\mathcal{G}_{\mathbb{Z}_i}^{(l)}\}_{i\in\mathcal{I}^{(l)}}$ *pdf*s, i.e., $\sum_j \alpha_{ij}^{\mathbb{z}^{(l)}} = 1$ for all $i \in \mathcal{I}^{(l)}$ but $\sum_{i,j} \alpha_{ij}^{\mathbb{z}^{(l)}} \neq 1$, combining all the uni-modal child component densities of $\{\mathcal{G}_{\mathbb{Z}_i}^{(l)}\}_{i\in\mathcal{I}^{(l)}}$ into one global $\mathcal{G}_{\mathbb{Z}}^{(l)}$ model requires *weighting* the $|\mathcal{J}_i^{(l)}|$ child densities of each $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$ model in a manner representing the prior probabilities of the corresponding localized time-frequency regions modelled by $\{\mathcal{G}_{\mathbb{Z}_i}^{(l)}\}_{i\in\mathcal{I}^{(l)}}$.

To that end, we model the entire static joint-band $\mathbb{Z}^{(0)}$ space in the first 0th step of our algorithm using a single global GMM, $\mathcal{G}_{\mathbb{Z}}^{(0)}$, with I components; i.e., we do not localize the *pdf* estimation for the initial $\mathbb{Z}^{(0)} \equiv \mathbf{Z} = \begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix}$ space. Per our previous development and indexing notation, this corresponds to modelling a single parent subset, $\mathcal{V}_i^{\mathbb{Z}^{(0)}}$ where $i \in \mathcal{I}^{(0)} = \{1\}$, comprising all static joint-band data available for training, using an $(I \coloneqq |\mathcal{J}_1^{(0)}|)$-modal GMM. By treating the $\{(\alpha_{1j}^{\mathbb{Z}^{(0)}}, \lambda_{1j}^{\mathbb{Z}^{(0)}})\}_{j\in\{1,\ldots,I\}}$ components of $\mathcal{G}_{\mathbb{Z}}^{(0)}$ as I root nodes for all the children states to be estimated in subsequent increments of l, the $\{\alpha_{1j}^{\mathbb{Z}^{(0)}}\}_{j\in\{1,\ldots,I\}}$ priors—corresponding to the prior probabilities of the localized $\{\mathcal{V}_{1j}^{\mathbb{Z}^{(0)}}\}_{j\in\{1,\ldots,I\}}$ time-frequency subsets obtained based on $\mathcal{G}_{\mathbb{Z}}^{(0)}$—can then be progressively updated and passed on to the child states generated along each $(j \in \{1,\ldots,I\})$th branch of the model tree. By using the passed down priors as multiplicative weights to the corresponding descendent $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$ component priors, as shown in Eq. (5.67b) below, we succeed in properly normalizing the $\{\alpha_{ij}^{\mathbb{Z}^{(l)}}\}_{\forall i,j}$ priors of the uni-modal child state densities, obtained at any particular lth order of memory inclusion, such that the relative weights inherited and updated along all I branches from the root states to the child states are taken into account, thereby simultaneously ensuring that $\sum_{i,j}\alpha_{ij}^{\mathbb{Z}^{(l)}} = 1$.

Finally, in a manner similar to that described in Operation (b), we update notation by discarding the ancestry information of the lth-order children subclasses. In particular, we replace the $\mathcal{I}^{(l)}$ and $\{\mathcal{J}_i^{(l)}\}_{i\in\mathcal{I}^{(l)}}$ parent and child index sets enumerating all lth-order Gaussian components—namely, $\{(\alpha_{ij}^{\mathbb{Z}^{(l)}}, \lambda_{ij}^{\mathbb{Z}^{(l)}})\}$ where $i \in \mathcal{I}^{(l)}$ and, $\forall i,\, j \in \mathcal{J}_i^{(l)}$—by the single integer index set, $\mathcal{K}^{(l)} = \{1,\ldots,|\mathcal{K}^{(l)}|\}$. Indexed on $\mathcal{K}^{(l)}$, the parameters of $\mathcal{G}_{\mathbb{Z}}^{(l)}$ can then be easily written as

$$k = j + \sum_{m<i}|\mathcal{J}_m^{(l)}|, \tag{5.67a}$$

$$\forall i \in \mathcal{I}^{(l)}, j \in \mathcal{J}_i^{(l)}\text{:}\quad \alpha_k^{\mathbb{Z}^{(l)}} \xleftarrow{=} \alpha_i^{\mathbb{Z}^{(l)}} \cdot \alpha_{ij}^{\mathbb{Z}^{(l)}}, \tag{5.67b}$$

$$\boldsymbol{\mu}_k^{\mathbb{Z}^{(l)}} \xleftarrow{=} \boldsymbol{\mu}_{ij}^{\mathbb{Z}^{(l)}}, \tag{5.67c}$$

$$\mathbf{C}_k^{\mathbb{Z}\mathbb{Z}^{(l)}} \xleftarrow{=} \mathbf{C}_{ij}^{\mathbb{Z}\mathbb{Z}^{(l)}}, \tag{5.67d}$$

with each lth-order $\alpha_k^{\mathbb{Z}^{(l)}}$ prior obtained via Eq. (5.67b) passed down for the next $(l+1)$th iteration of the algorithm as $\alpha_i^{\mathbb{Z}^{(l+1)}} \xleftarrow{\text{Eq.(5.25)}} \alpha_k^{\mathbb{Z}^{(l)}}$. With $M^{\mathbb{Z}^{(l)}} \coloneqq |\mathcal{K}^{(l)}|$, Eq. (5.67) completely defines all parameters of $\mathcal{G}_{\mathbb{Z}}^{(l)} = \mathcal{G}(\mathbb{Z}^{(l)}; M^{\mathbb{Z}^{(l)}}, \mathrm{A}^{\mathbb{Z}^{(l)}}, \Lambda^{\mathbb{Z}^{(l)}})$, the global

joint-band GMM with lth-order memory inclusion. We note, however, that, for $l < L$, the identification of the $\mathcal{G}_{\mathbb{Z}}^{(l)}$ components using the parent-child ancestry information is required for the fuzzy partitioning of training data in the $\mathbb{Z}^{(l)}$ space as described in Operation (a)—to generate the pairwise-disjoint time-frequency-localized $\{\mathcal{V}_k^{\mathbb{Z}^{(l)},w^{(l)}}\}_{k\in\mathcal{K}^{(l)}}$ subsets in preparation for lth-order post-EM pruning as well as for the next $(l+1)$th modelling iteration. Hence, for $l < L$, we first make use of the $\mathcal{I}^{(l)}$ and $\{\mathcal{J}_i^{(l)}\}_{i\in\mathcal{I}^{(l)}}$ indices in Eqs. (5.21), (5.24), (5.27), and (5.65), prior to discarding that information while constructing $\mathcal{G}_{\mathbb{Z}}^{(l)}$ per Eq. (5.67).

ii. On Markov blankets and the conditional independence properties of the states derived from global GMMs

Given a global $\mathcal{G}_{\mathbb{Z}}^{(l)}$ *pdf* modelling the distribution of training data in entire $\mathbb{Z}^{(l)}$ space as described above, we can now show, as noted in Section 5.4.2.2, that the conditional independence properties of the states represented by the individual Gaussian components of $\mathcal{G}_{\mathbb{Z}}^{(l)}$—i.e., $\{S_k^{(l)} \triangleq (\alpha_k^{\mathbb{Z}^{(l)}}, \lambda_k^{\mathbb{Z}^{(l)}})\}_{k\in\mathcal{K}^{(l)}}$—follow the definition of Markov blankets.[161] In particular, since each kth state corresponds to a uni-modal model of variability in a time-frequency-localized region of the $\mathbb{Z}^{(l)}$ space, the global $\mathbb{Z}^{(l)}$ space can be reduced—from the perspective of that kth state—to a linear vector subspace, $\mathbb{Z}_k^{(l)}$, in which the variability of the $\mathbb{Z}^{(l)}$ data is defined by the uni-modal *pdf*, $p(\mathbb{z}_k^{(l)}) = \alpha_k^{\mathbb{Z}^{(l)}} p(\mathbb{z}_k^{(l)}|\lambda_k^{\mathbb{Z}^{(l)}})$. As such, it is clear from Eq. (5.67) that the likelihood of any realization in $\mathbb{Z}_k^{(l)}$ depends only on the prior probability of the corresponding parent state, as well as that of the kth underlying state itself, but not on any of the *pdf* parameters of the $\{\mathbb{Z}_m^{(l)}\}_{\forall m\neq k}$ vector subspaces underlying the other states of $\mathcal{G}_{\mathbb{Z}}^{(l)}$. Hence, given the parent state, random vector realizations drawn from $p(\mathbb{z}_k^{(l)})$ are conditionally independent of all other realizations drawn from $\{p(\mathbb{z}_m^{(l)})\}_{\forall m\neq k}$, for all $k \in \mathcal{K}^{(l)}$, thereby satisfying the directed local Markov property of directed acyclic graphs. Although we do not make use of this interpretation in our work presented here, it demonstrates the generalization advantage of our tree-like GMM extension approach to other modelling problems.

[161] As defined by Pearl [179], the Markov blanket for a node A in a Bayesian network is the set of nodes $\mathrm{MB}(A)$ composed of A's parents, its children, and its children's other parents. The Markov blanket $\mathrm{MB}(A)$ shields A from the rest of the network; i.e., no other node in the network outside $\mathrm{MB}(A)$ can influence A.

iii. Marginalization

As described in Section 5.4.2.2, performing BWE using our tree-like temporally-extended GMMs requires only the subspace joint-band $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}\}_{l\in\{0,\dots,L\}}$ models. As such, we conclude each $(0<l\leq L)$th iteration of our training algorithm by marginalizing the global $\mathcal{G}_{\mathbb{Z}^{(\tau,l)}} := \mathcal{G}_{\mathbb{Z}}^{(\tau,l)}$ obtained above to $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}$, noting that $\mathcal{G}_{\mathbb{X}^{(\tau,0)}\mathbf{Y}} = \mathcal{G}_{\mathbb{Z}^{(\tau,0)}}$.

Table 5.5: Algorithm for model-based memory inclusion through our tree-like approach to temporally extending GMMs.

---

inputs: $\mathcal{V}^{\mathbf{x}}$ and $\mathcal{V}^{\mathbf{y}}$, the sets of all static narrowband and highband training data, resp.;
τ, memory inclusion step (see definition in Section 5.4.2.2);
L, maximum value for memory inclusion index, l (see definition in Section 5.4.2.2);
I, modality of 0th-order joint-band GMM, $\mathcal{G}_{\mathbb{Z}}^{(0)}$ (see definition in Operation (e));
J, splitting factor, or equivalently, the modality of $\mathcal{G}_{\mathbf{Y}}^{(0)}$, the weighted EM initialization GMM (see definition in Operation (c));
K, fuzziness factor (see definition in Operation (a));
$\rho_{\min}$, distribution flatness threshold (see definition in Operation (d));
$N_{\min}$, child subset cardinality threshold (see definition in Operation (d));
$\Delta\mathcal{L}_{\max}^{w}$, weighted log-likelihood relative change threshold (see definition in Operation (c)).

outputs: $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}\}_{l\in\{0,\dots,L\}}$, the temporally-extended joint-band GMMs to be used for BWE (see illustration in Figure 5.8).

(1) given $\mathcal{V}^{\mathbf{y}}$ and J, construct $\mathcal{G}_{\mathbf{Y}}^{(0)}$ by **conventional EM**;

(2) given $\mathcal{V}^{\mathbf{x}}$ and $\mathcal{V}^{\mathbf{y}}$, construct $\mathcal{V}_1^{\mathbb{Z}^{(0)}}$, the global 0th-order joint-band parent set, by **feature vector concatenation**; i.e., $\mathcal{V}_1^{\mathbb{Z}^{(0)}} = \mathcal{V}^{\mathbf{z}} = \left\{\mathbf{z}_n = \begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix}, \forall \mathbf{x}_n \in \mathcal{V}^{\mathbf{x}}, \mathbf{y}_n \in \mathcal{V}^{\mathbf{y}}\right\}$;

(3) given $\mathcal{V}_1^{\mathbb{Z}^{(0)}}$, construct the I-modal $\mathcal{G}_{\mathbb{X}^{(\tau,0)}\mathbf{Y}} = \mathcal{G}_{\mathbb{Z}}^{(0)} := \mathcal{G}_{\mathbb{Z}_1}^{(0)}$ by **conventional EM** with non-weighted log-likelihood relative change as stopping criterion and $\Delta\mathcal{L}_{\max}^{w}$ as threshold;

for $j = 1$ **to** I **do** (noting that, with $l=0$, $i \in \mathcal{I}^{(0)} = \{1\}$ and $j \in \mathcal{J}_1^{(0)} = \{1,\dots,I\}$)

(4) given $\mathcal{V}_1^{\mathbb{Z}^{(0)}}$, K, and $\mathcal{G}_{\mathbb{Z}_1}^{(0)}$, perform **fuzzy clustering**:
construct $\mathcal{V}_{1j}^{\mathbb{Z}^{(0)},w^{(0)}}$ via Eqs. (5.18), (5.19), and (5.21);

(5) given $\mathcal{V}_{1j}^{\mathbb{Z}^{(0)},w^{(0)}}$ and τ, perform **incremental temporal extension**:
construct $\mathcal{V}_k^{\mathbb{Z}^{(1)},w^{(0)}} \xleftarrow{\text{Eq.(5.27)}} \mathcal{V}_k^{\mathbb{Z}^{(0)},w^{(0)}} \xleftarrow{\text{Eq.(5.24)}} \mathcal{V}_{1j}^{\mathbb{Z}^{(0)},w^{(0)}}$, where $k \in \mathcal{K}^{(0)}$;

(6) perform $\mathcal{K}^{(0)} \xrightarrow{=} \mathcal{I}^{(1)}$ **index mapping** in preparation for next iteration:
$\mathcal{V}_i^{\mathbb{Z}^{(1)},w^{(0)}} \xleftarrow{\text{Eq.(5.25)}} \mathcal{V}_k^{\mathbb{Z}^{(1)},w^{(0)}}$; $\alpha_i^{\mathbb{Z}^{(1)}} \xleftarrow{\text{Eq.(5.25)}} \alpha_k^{\mathbb{Z}^{(0)}} \xleftarrow{\text{Eq.(5.24a)}} \alpha_{1j}^{\mathbb{Z}^{(0)}}$;

(7) construct $\{\mathcal{G}_{\mathbb{Z}}^{(l)}\}_{l\in\{1,\ldots,L\}}$ by iterating over Operations (a)–(e), starting with $\mathcal{G}_{\mathbb{Z}}^{(1)}$ given the parent $\{\mathcal{V}_i^{\mathbb{Z}^{(1)},w^{(0)}}\}_{i\in\mathcal{I}^{(1)}}$ subsets and $\{\alpha_i^{\mathbb{Z}^{(1)}}\}_{i\in\mathcal{I}^{(1)}}$ priors obtained in Step (6):

for $l = 1$ **to** L **do**

for $i = 1$ **to** $|\mathcal{I}^{(l)}|$ **do**

(a) check **pre-EM pruning** condition:
construct $\mathcal{V}_i^{\mathbf{y}^{(l)},w^{(l-1)}} \xleftarrow{\text{Eq.(5.29)}} \mathcal{V}_i^{\mathbb{Z}^{(l)},w^{(l-1)}}$;
given $\mathcal{G}_{\mathbf{Y}}^{(0)}$ and $\mathcal{V}_i^{\mathbf{y}^{(l)},w^{(l-1)}}$, estimate ρ_i per Eqs. (5.60), (5.61), and (5.62);
given ρ_i and $\rho_{\min}$, determine $|\mathcal{J}_i^{(l)}|$ per Eq. (5.63);

(b) given $\mathcal{V}_i^{\mathbf{y}^{(l)},w^{(l-1)}}$, $\Delta\mathcal{L}_{\max}^{w}$, and $\mathcal{G}_{\mathbf{Y}}^{(0)}$, estimate $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$ via **weighted EM** per Operations (c) or (d), depending on $|\mathcal{J}_i^{(l)}|$:

if $|\mathcal{J}_i^{(l)}| = J$ **then**

initialize $\mathcal{G}_{\mathbf{Y}_i}^{(l)}$ as $\mathcal{G}_{\mathbf{Y}_i}^{(l,0)} \xleftarrow{\text{Eq.(5.53)}} \mathcal{G}_{\mathbf{Y}}^{(0)}$;
repeat $k \xleftarrow{=} k+1$
estimate $\mathcal{G}_{\mathbf{Y}_i}^{(l,k+1)} \xleftarrow{\text{Eqs.(5.54),(5.55)}} \mathcal{G}_{\mathbf{Y}_i}^{(l,k)}$ in **first stage** of weighted EM;
calculate $\Delta\mathcal{L}^{w}$ via Eqs. (5.56) and (5.57);
until $(\Delta\mathcal{L}^{w} < \Delta\mathcal{L}_{\max}^{w}) \Rightarrow \mathcal{G}_{\mathbb{Z}_i}^{(l)} \xleftarrow{:=} \mathcal{G}_{\mathbb{Z}_i}^{(l,k)}$;
extrapolate $\mathcal{G}_{\mathbb{Z}_i}^{(l)} \xleftarrow{\text{Eqs.(5.58),(5.59)}} \mathcal{G}_{\mathbf{Y}_i}^{(l)}$ in **second stage** of weighted EM;

else $|\mathcal{J}_i^{(l)}| = 1$

estimate uni-modal $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$ via **reduced weighted EM** per Eq. (5.66);

(c) **if** $l = L$ **then** skip Steps (d)–(g) below $\Rightarrow$ **go to** next i;

for $j = 1$ **to** $|\mathcal{J}_i^{(l)}|$ **do**

(d) given $\mathcal{V}_i^{\mathbb{Z}^{(l)}}$, K, and $\mathcal{G}_{\mathbb{Z}_i}^{(l)}$, perform **fuzzy clustering**:
construct $\mathcal{V}_{ij}^{\mathbb{Z}^{(l)},w^{(l)}}$ via Eqs. (5.18), (5.19), and (5.21);

(e) given $\mathcal{V}_{ij}^{\mathbb{Z}^{(l)},w^{(l)}}$ and τ, perform **incremental temporal extension**:
construct $\mathcal{V}_k^{\mathbb{Z}^{(l+1)},w^{(l)}} \xleftarrow{\text{Eq.(5.27)}} \mathcal{V}_k^{\mathbb{Z}^{(l)},w^{(l)}} \xleftarrow{\text{Eq.(5.24)}} \mathcal{V}_{ij}^{\mathbb{Z}^{(l)},w^{(l)}}$, where $k \in \mathcal{K}^{(l)}$;

(f) check **post-EM pruning** condition for $k \xleftarrow{\text{Eq.(5.24a)}} ij$:
if $(|\mathcal{J}_i^{(l)}| > 1) \wedge (|\mathcal{V}_k^{\mathbb{Z}^{(l+1)}}| < N_{\min})$ **then** $|\mathcal{J}_i^{(l)}| \xleftarrow{=} 1$; **redo** Steps (b)–(e);

(g) perform $\mathcal{K}^{(l)} \xrightarrow{=} \mathcal{I}^{(l+1)}$ **index mapping** to prepare for $(l+1)$th iteration:
$\mathcal{V}_i^{\mathbb{Z}^{(l+1)},w^{(l)}} \xleftarrow{\text{Eq.(5.25)}} \mathcal{V}_k^{\mathbb{Z}^{(l+1)},w^{(l)}}$; $\alpha_i^{\mathbb{Z}^{(l+1)}} \xleftarrow{\text{Eq.(5.25)}} \alpha_k^{\mathbb{Z}^{(l)}} \xleftarrow{\text{Eq.(5.24a)}} \alpha_{ij}^{\mathbb{Z}^{(l)}}$;

(h) given $\{\alpha_i^{\mathbb{Z}^{(l)}}\}_{i\in\mathcal{I}^{(l)}}$ and $\{\mathcal{G}_{\mathbb{Z}_i}^{(l)}\}_{i\in\mathcal{I}^{(l)}}$, **consolidate** all $|\mathcal{K}^{(l)}|$ localized child Gaussian components to construct the lth-order global $\mathcal{G}_{\mathbb{Z}}^{(l)} \xleftarrow{\text{Eq.(5.67)}} \{\mathcal{G}_{\mathbb{Z}_i}^{(l)}\}_{i\in\mathcal{I}^{(l)}}$;

(i) **marginalize** all $M^{\mathbb{Z}^{(l)}} := |\mathcal{K}^{(l)}|$ component densities of $\mathcal{G}_{\mathbb{Z}^{(\tau,l)}} := \mathcal{G}_{\mathbb{Z}}^{(l)}$ to obtain $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}$, the lth-order joint-band subspace GMM to be used for BWE;

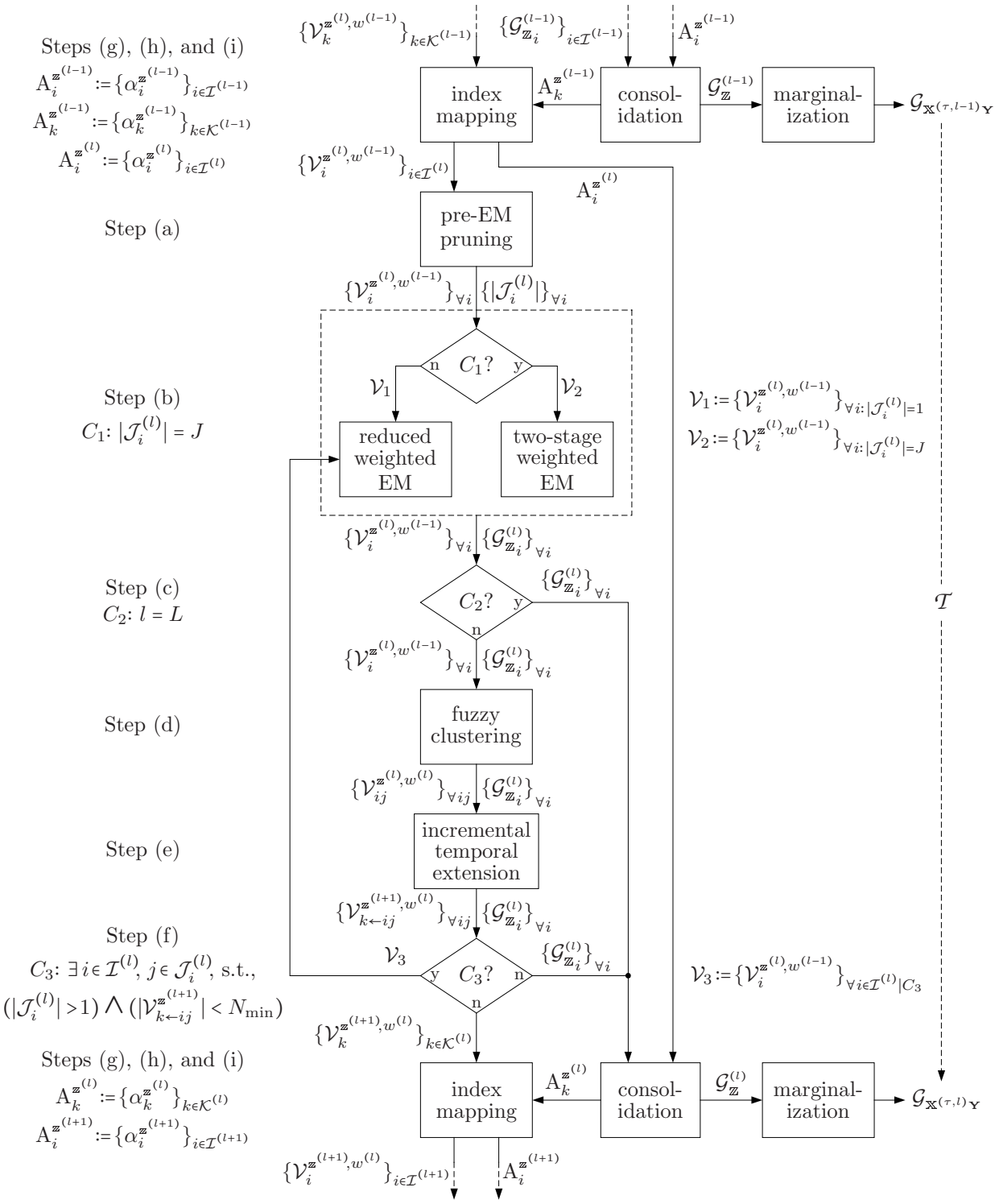


Fig. 5.10: Block diagram of a single ($l > 0$)th-order iteration of our tree-like GMM temporal extension algorithm, with correspondences to the steps of Table 5.5 indicated on the left.

Having detailed the five main operations involved in the implementation of our tree-like GMM extension algorithm, we now summarize the integration of these steps into a complete training procedure. Table 5.5 and Figure 5.10 below provide such a formal synopsis of our model-based memory inclusion algorithm. In particular, Steps (1)–(6) of Table 5.5 describe the operations performed at the $(l = 0)$th iteration to obtain $\mathcal{G}_{\mathbb{X}^{(\tau,0)}\mathbf{Y}}$, as well as the $\mathcal{G}_{\mathbf{Y}}^{(0)}$ initialization GMM and the first-order $\{\mathcal{V}_i^{\mathbb{Z}^{(1)},w^{(0)}}\}_{i\in\mathcal{I}^{(1)}}$ and $\{\alpha_i^{\mathbb{Z}^{(1)}}\}_{i\in\mathcal{I}^{(1)}}$ parent subsets and priors, respectively, representing the inputs to subsequent $(l > 0)$th iterations. The sequence and the integration of operations representing the core of our training algorithm—summarized by the $\mathcal{G}_{\mathbb{X}^{(\tau,l-1)}\mathbf{Y}} \xrightarrow{\mathcal{T}} \mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}$ transformation—are then detailed in Step (7) of Table 5.5, and further illustrated in Figure 5.10.

5.4.2.4 Reliability of temporally-extended GMMs

As described in Section 5.4.2.2, the principles underlying the incremental tree-like design of our GMM temporal extension approach followed from our desire to exploit the information and predictability in speech frames, as well as in the correspondence of GMM-based speech models to underlying acoustic classes, in order to constrain the high degrees of freedom associated with GMM-based modelling of the high-dimensional temporally-extended joint-band spaces. Implemented through time-frequency localization, our approach to constraining the modelling task as such thus aimed to specifically alleviate the detrimental effects of the oversmoothing and overfitting problems comprising the curse of dimensionality in the context of high-dimensional GMM-based modelling. Accordingly, in this section, we assess the reliability of our temporally-extended GMMs in terms of the extent of oversmoothing and overfitting, or lack thereof.

i. Assessing extent of oversmoothing

Oversmoothing was defined and described in Section 5.4.2.1 as the excessive smoothing of MMSE-derived highband spectral characteristics, corresponding to a coarse coverage of the highband spectral space, rather than a continuous one with sufficient spectral variability. It follows from lower source-data contributions in Eq. (3.17) as a result of the tendency of the inter- to intra-band cross-covariance *ratios*[162]—i.e., $\{\mathbf{C}_i^{\mathbf{yx}}\mathbf{C}_i^{\mathbf{xx}^{-1}}\}_{i\in\{1,\ldots,M\}}$ where M is the GMM modality—to decrease with increasing dimensionality. In Section 3.5.1, we made

[162] See Footnote 66.

use of the matrix Frobenius and L_p-norms[163] of these cross-covariance ratios—explicitly representing a joint-band GMM's ability to model information mutual to the disjoint speech frequency bands, rather than band-specific information—to demonstrate the increasing superiority of full-covariance GMMs over diagonal-covariance ones in terms of capturing the sought-after cross-band correlations as GMM modality increases. In a similar manner, we now assess the extent of oversmoothing in our temporally-extended $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}\}_{l\in\{0,\ldots,L\}}$ GMMs by measuring the change in the average Frobenius norms of the corresponding cross-covariance ratios as a function of the memory inclusion index, l, which itself corresponds to the $\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ \mathbf{Y}\end{bmatrix}$ joint-band subspace dimensionality.

As detailed below in Section 5.4.3.1, we temporally extend our static MFCC-based BWE baseline models of Section 5.2.3—represented by the $\boldsymbol{\mathcal{G}} = (\mathcal{G}_{\mathbf{X}\mathbf{C}_y}, \mathcal{G}_{\mathbf{X}G})$ GMM tuple—using our tree-like extension algorithm of Table 5.5, resulting in the memory-inclusive $\{\boldsymbol{\mathcal{G}}^{(\tau,l)} := (\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}, \mathcal{G}_{\mathbb{X}^{(\tau,l)}G})\}_{l\in\{0,\ldots,L\}}$ models. Thus, for the static MFCC-based dimensionalities of $\mathrm{Dim}(\mathbf{X}) = 10$, $\mathrm{Dim}(\mathbf{C}_y) = 6$, and $\mathrm{Dim}(G) = 1$, selected as such per the discussion in Section 5.2.3, the relationship between the lth order of memory inclusion and the dimensionalities of the lth-order temporally-extended $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}$ and $\mathcal{G}_{\mathbb{X}^{(\tau,l)}G}$ GMMs is given by

$$\mathrm{Dim}\left(\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ \mathbf{C}_y\end{bmatrix}\right) = 10(l+1) + 6 \tag{5.68a}$$

$$\mathrm{Dim}\left(\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ G\end{bmatrix}\right) = 10(l+1) + 1. \tag{5.68b}$$

Dropping the fixed memory inclusion step, τ, from superscripts, and focusing only on the higher-dimensional $\{\mathcal{G}_{\mathbb{X}^{(l)}\mathbf{C}_y}\}$ GMMs, we evaluate the average Frobenius norms of the cross-covariance ratios, i.e., $\overline{\left\{\left\|\mathbf{C}_i^{\mathbf{c}_y\mathbb{X}^{(l)}}\left[\mathbf{C}_i^{\mathbb{X}^{(l)}\mathbb{X}^{(l)}}\right]^{-1}\right\|_{\mathrm{F}}\right\}_{i\in\{1,\ldots,M^{\mathbf{x}^{(l)}\mathbf{c}_y}\}}}$, as a function of all $l \in \{0,\ldots,L\}$.[164] We consider only the Frobenius norm rather than also the L_p-norms—where $p \in \{1,2,\infty\}$—previously evaluated in Section 3.5.1, and illustrated in Figure 3.6 in particular, since, per the matrix norm properties detailed in [108, Sections 2.3 and 2.5.3]:

(a) the L_1- and L_∞-norms, $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_\infty$, correspond to the maximum absolute column and row sums of the matrix $\mathbf{A}$, respectively, and hence, are not suitable for comparing norms of matrices with varying dimensionalities—as is the case here for

[163] See Footnote 67 for details on the Frobenius and L_p-norms.

[164] Since the cross-covariance ratio matrices are non-square matrices where determinants are inapplicable, the *weights* represented by such matrices can only be quantified through matrix norms.

the $\left\{\mathbf{C}_i^{\mathbf{c}_y\mathbb{x}^{(l)}}\left[\mathbf{C}_i^{\mathbb{x}^{(l)}\mathbb{x}^{(l)}}\right]^{-1}\right\}$ cross-covariance ratio matrices whose dimensionalities vary with l; and

(b) while both $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_{\mathrm{F}}$ correspond to *weights* along underlying basis vectors by virtue of their relationship with the singular values of $\mathbf{A}$,[165] the Frobenius norm considers all singular values rather than just the largest as is the case for the L_2-norm, and hence, $\left\|\mathbf{C}_i^{\mathbf{c}_y\mathbb{x}^{(l)}}\left[\mathbf{C}_i^{\mathbb{x}^{(l)}\mathbb{x}^{(l)}}\right]^{-1}\right\|_{\mathrm{F}}$ accounts for the scaling applied by the cross-covariance ratios to the source-data contribution along all underlying basis vectors, rather than only along that with the largest scaling.

Figure 5.11(a) illustrates the average Frobenius norm performance obtained, as a function of dimensionality, for several $\mathcal{G}_{\mathbb{x}^{(l)}\mathbf{C}_y}$ GMMs trained at various values for J, K, τ, and $\rho_{\min}$, using the $(I = 128)$-modal $\mathcal{G}_{\mathbf{x}\mathbf{C}_y}$ GMM of our static MFCC-based baseline of Sections 5.2.3 and 5.2.6 as the 0th-iteration model.[166] Except for the temporary slight dip in average norm at initial values of l, the increasing Frobenius norms of Figure 5.11(a) not only indicate the success of our tree-like algorithm in alleviating the oversmoothing concerns associated with GMM-based modelling at high dimensionalities, but they also demonstrate the ability of our algorithm to capture the increasingly-important cross-band correlations as the extent of included memory increases, i.e., as the algorithm incorporates more temporal information from longer causal windows of past narrowband and highband frames, despite the linearly-increasing dimensionality.

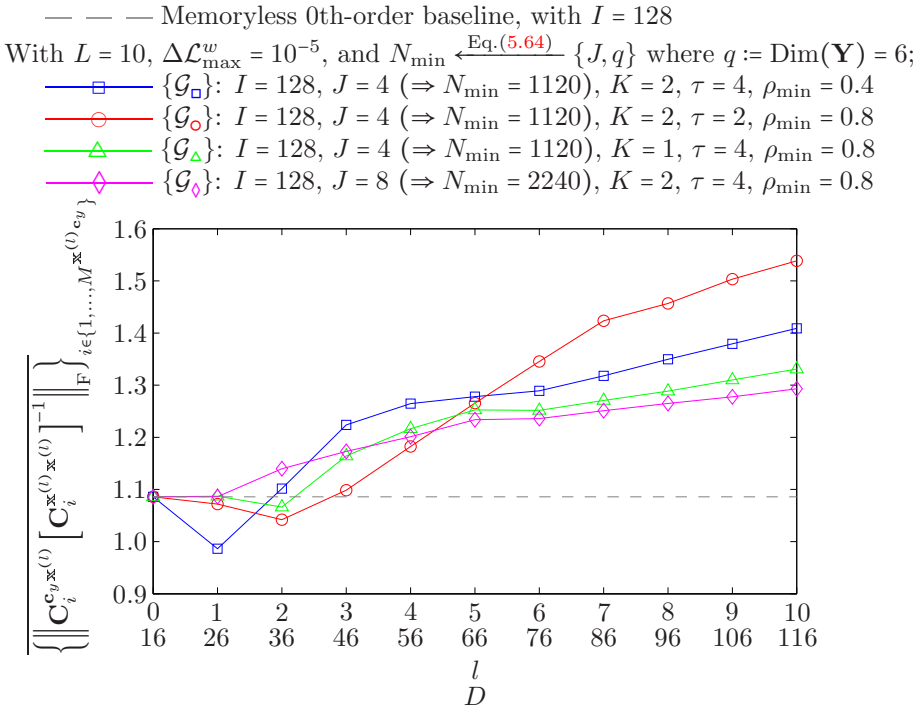
ii. Assessing extent of overfitting

As described in Sections 5.4.2.1 and 5.4.2.2, overfitting is the property whereby the higher sparsity of data associated with modelling distributions in increasingly high-dimensional spaces leads to increasingly suboptimal GMMs with reduced generalization capability. More specifically in our context, as the dimensionality of the $\mathbb{Z}^{(\tau,l)}$ space underlying our temporally-extended GMMs increases with higher orders of memory inclusion, the empty space phenomenon[167] results in increasingly sparse and overlapping densities which, in turn, increases the risk that the available joint-band training data becomes insufficient to reliably estimate the parameters of the temporally-extended GMMs through Expectation-

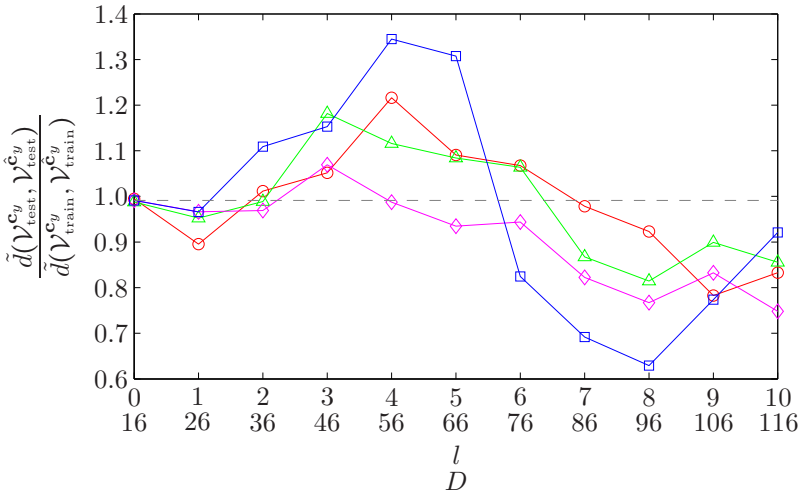
[165] Per [108, Eqs. (2.5.7) and (2.5.8)], the L_2- and Frobenius norms for a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ are related to the singular values of $\mathbf{A}$—$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$, where $p = \min\{m, n\}$—by $\|\mathbf{A}\|_2 = \sigma_1$ and $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{p}\sigma_i^2}$, respectively.

[166] See description of tree-like algorithm inputs in Table 5.5.

[167] See Footnote 140.



(a) Assessing oversmoothing



(b) Assessing overfitting

Fig. 5.11: Assessing oversmoothing and overfitting in the temporally-extended $\{\mathcal{G}_{\mathbb{X}^{(l)}\mathbf{C}_y}\}_{l\in\{0,\ldots,L\}}$ GMMs. Assessed as functions of the memory inclusion index $l \in \{0, \ldots, L\}$ and the associated dimensionality, $D := \text{Dim}\left(\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ \mathbf{C}_y\end{bmatrix}\right)$ as given by Eq. (5.68a), oversmoothing and overfitting are assessed, respectively, through the average Frobenius norms of the inter-band to intra-band cross-covariance ratios—i.e., $\overline{\left\{\left\|\mathbf{C}_i^{\mathbf{c}_y\mathbb{x}^{(l)}}\left[\mathbf{C}_i^{\mathbb{x}^{(l)}\mathbb{x}^{(l)}}\right]^{-1}\right\|_{\text{F}}\right\}_{i\in\{1,\ldots,M^{\mathbb{x}^{(l)}\mathbf{c}_y}\}}}$—and the ratios of the normalized cepstral distance of the test data to that of the training data—i.e., $\frac{\tilde{d}(\mathcal{V}^{\mathbf{c}_y}_{\text{test}}, \mathcal{V}^{\hat{\mathbf{c}}_y}_{\text{test}})}{\tilde{d}(\mathcal{V}^{\mathbf{c}_y}_{\text{train}}, \mathcal{V}^{\hat{\mathbf{c}}_y}_{\text{train}})}$.

Maximization. Such a decrease in EM reliability translates to overfitted models that ultimately lead to poor MMSE-based source-target transformation for test data unseen during training. Accordingly, in order to specifically address the risk of overfitting in our high-dimensional temporally-extended GMMs, we proposed and incorporated the fuzzy GMM-based clustering and the pruning steps of Operations (a) and (d) in Section 5.4.2.3, respectively, into our tree-like GMM construction approach.

To assess the performance of our tree-like algorithm in alleviating the risk of overfitting, we devise a measure to quantify such risk inspired by a cepstral measure proposed in [160]. In order to evaluate the overfitting associated with joint-speaker source-target GMMs obtained through various speaker conversion algorithms, Mesbahi el alia use a *normalized cepstral distance* in [160] defined as the mean Euclidean distance between the target and converted MFCC feature vectors, with the distance between each pair of vectors normalized by the distance between the target and the corresponding source feature vector. The normalization thus corresponds to weighting the distortion in each converted vector based on the difficulty of converting the source speaker vector—conversion distortions for difficult source vectors are given lower weights relative to those obtained for easier-to-convert source speaker vectors. By comparing the mean normalized cepstral distance obtained for a particular test set against that obtained for the corresponding set used to train the conversion GMMs, the generalization capabilities of the GMMs trained using various speaker conversion algorithms can then be quantified. Since conversion distortions for equally-difficult test and training source data are weighted equally, the aforementioned normalization thus ensures that the comparison of performance on the test and training data focuses on GMMs' generalization capabilities, rather than simply the quality of source-target conversion as would be the case by comparing non-normalized mean cepstral distances.

In a similar manner, we quantify the risk of overfitting in our source-target BWE task using normalized cepstral distances between reference highband MFCC feature vectors and their MMSE-based reconstructed counterparts, but with the normalization performed differently from that of [160]. In the speaker conversion task, the source and target feature vectors have the same dimensionality, thus allowing the estimation of cepstral distances between source and target vectors to be then used as normalization as described above. In our BWE task, however, the source and target MFCC-based feature vectors—represented by the temporally-extended narrowband $\mathbb{X}^{(\tau,l)}$ vectors and the static highband $\mathbf{Y}$ vectors, respectively—differ in dimensionality, and hence, estimating source-target cepstral

distances to normalize distortions in reconstructed highband vectors is not feasible. Since the normalization is, in principle, intended to simply weight distortions in a manner accounting for the source-target conversion difficulty, we employ, instead, the likelihood of the source narrowband vectors given our temporally-extended GMMs. As source-data likelihoods represent a measure of GMM-based source-target conversion difficulty, using such likelihoods as multiplicative normalization weights for distortions in the reconstructed target highband vectors enables us to compare distortions in testing and training data in the context of generalization capability; cepstral errors in the reconstructed highband data of equally-likely testing and training source data are weighted equally, with higher emphasis on those errors corresponding to source data with higher likelihoods.

For a particular set of temporally-extended source narrowband data, $\mathcal{V}^{\mathbf{x}^{(\tau,l)}}$, and its corresponding set of reference target highband MFCC vectors, $\mathcal{V}^{\mathbf{y}}$, let $\mathcal{V}^{\hat{\mathbf{y}}} = \{\hat{\mathbf{y}}_n\}$—where $n \in \{n\colon \mathbf{x}_n^{(\tau,l)} \in \mathcal{V}^{\mathbf{x}^{(\tau,l)}}\}$—represent the set of highband feature vectors reconstructed using the joint-band temporally-extended GMM, $\mathcal{G}_{\mathbf{X}^{(\tau,l)}\mathbf{Y}}$, through MMSE-based estimation as detailed in Section 5.4.3.1 below. Then, with the cepstral distances between reference and corresponding MMSE-estimated target highband vectors given by Euclidean distances, i.e.,

$$d^2_{\mathrm{MFCC}}(\mathbf{y}, \hat{\mathbf{y}}) \triangleq \|\mathbf{y} - \hat{\mathbf{y}}\|^2, \tag{5.69}$$

we define the source-data likelihood-based normalized cepstral distance as

$$\tilde{d}(\mathcal{V}^{\mathbf{y}}, \mathcal{V}^{\hat{\mathbf{y}}}) = \frac{\sum\limits_{\{n\colon \mathbf{x}_n^{(\tau,l)} \in \mathcal{V}^{\mathbf{x}^{(\tau,l)}}\}} P(\mathbf{x}_n^{(\tau,l)}|\mathcal{G}_{\mathbf{X}^{(\tau,l)}})\, d_{\mathrm{MFCC}}(\mathbf{y}_n, \hat{\mathbf{y}}_n)}{\sum\limits_{\{n\colon \mathbf{x}_n^{(\tau,l)} \in \mathcal{V}^{\mathbf{x}^{(\tau,l)}}\}} P(\mathbf{x}_n^{(\tau,l)}|\mathcal{G}_{\mathbf{X}^{(\tau,l)}})}, \tag{5.70}$$

where $\mathcal{G}_{\mathbf{X}^{(\tau,l)}} := \mathcal{G}(\mathbf{x}^{(\tau,l)}; M^{\mathbf{x}^{(\tau,l)}}, \mathrm{A}^{\mathbf{x}^{(\tau,l)}}, \Lambda^{\mathbf{x}^{(\tau,l)}})$ is obtained from the joint-band $\mathcal{G}_{\mathbf{X}^{(\tau,l)}\mathbf{Y}}$ by marginalization, and, per Eq. (5.31), the likelihood $P(\mathbf{x}_n^{(\tau,l)}|\mathcal{G}_{\mathbf{X}^{(\tau,l)}})$ is given by

$$P(\mathbf{x}_n^{(\tau,l)}|\mathcal{G}_{\mathbf{X}^{(\tau,l)}}) = \sum_{m=1}^{M^{\mathbf{x}^{(\tau,l)}}} \alpha_m^{\mathbf{x}^{(\tau,l)}} P(\mathbf{x}_n^{(\tau,l)}|\lambda_m^{\mathbf{x}^{(\tau,l)}}). \tag{5.71}$$

Thus, as shown in Eq. (5.70), the cepstral distances between reference and MMSE-estimated highband vectors for a particular $\mathcal{V}^{\mathbf{y}}$ set are normalized by weighting the distance for each

$(\mathbf{y}_n, \hat{\mathbf{y}}_n)$ pair in proportion to the difficulty of converting the corresponding source $\mathbb{x}_n^{(\tau,l)}$ vector, relative to all other vectors in the $\mathcal{V}^{\mathbb{x}^{(\tau,l)}}$ set—where, as described above, conversion difficulty is represented by source-data likelihoods. Cepstral distortions in the target MMSE-estimated highband vectors corresponding to source vectors with higher relative likelihoods are weighted proportionally higher than those target vector distortions of less-likely—i.e., more difficult—source vectors, and vice versa. By normalizing distortions in reconstructed target vectors based on individual data-point likelihoods in relation to the likelihood sum for the whole set, rather than on absolute likelihoods, we ensure that our estimates of overfitting and generalization capability are not biased by the overall likelihood of that particular set.[168] We should also note that, by incorporating the cepstral distances between reference and MMSE-estimated target highband vectors into our $\tilde{d}$ overfitting measure, rather than considering source-data likelihoods alone, we are also accounting for the effect of the cross-band correlation information captured into the joint-band $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}\}$ GMMs on generalization capability, rather than only account for the narrowband-only information in the marginal $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}}\}$ GMMs.

Derived from the TIMIT training and core test sets described in Section 3.2.10, let $\mathcal{V}_{\text{train}}^{\mathbb{x}^{(\tau,l)}}$ and $\mathcal{V}_{\text{test}}^{\mathbb{x}^{(\tau,l)}}$ represent the training and testing sets of lth-order temporally-extended source narrowband data, respectively, with corresponding $\mathcal{V}_{\text{train}}^{\mathbf{c}_y}$, $\mathcal{V}_{\text{train}}^{\hat{\mathbf{c}}_y}$, $\mathcal{V}_{\text{test}}^{\mathbf{c}_y}$, and $\mathcal{V}_{\text{test}}^{\hat{\mathbf{c}}_y}$ sets of MFCC vectors representing the spectral shape of target highband data. Then, by calculating the ratio of the normalized cepstral distance of testing data to that of the training data—i.e., $\frac{\tilde{d}(\mathcal{V}_{\text{test}}^{\mathbf{c}_y}, \mathcal{V}_{\text{test}}^{\hat{\mathbf{c}}_y})}{\tilde{d}(\mathcal{V}_{\text{train}}^{\mathbf{c}_y}, \mathcal{V}_{\text{train}}^{\hat{\mathbf{c}}_y})}$—as a function of the memory inclusion index, l, for all $l \in \{0, \ldots, L\}$, we obtain a measure of potential overfitting in the $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}\}_{l\in\{0,\ldots,L\}}$ GMMs as a function of dimensionality. Given the normalization of per-sample likelihoods by the likelihood sums of the overall testing and training sets as described above, values for $\frac{\tilde{d}(\mathcal{V}_{\text{test}}^{\mathbf{c}_y}, \mathcal{V}_{\text{test}}^{\hat{\mathbf{c}}_y})}{\tilde{d}(\mathcal{V}_{\text{train}}^{\mathbf{c}_y}, \mathcal{V}_{\text{train}}^{\hat{\mathbf{c}}_y})}$ greater than the memoryless baseline at $l = 0$ indicate an increase in likelihood-weighted cepstral

[168] With increasing dimensionalities, source-data likelihoods will typically have a much larger dynamic range than that of the Euclidean $d_{\text{MFCC}}(\mathbf{y}_n, \hat{\mathbf{y}}_n)$ cepstral distances. Consequently, estimates for $\tilde{d}$ can potentially be biased by the overall likelihood sum in the denominator of Eq. (5.70) if this normalizing denominator were to be removed or replaced by a term independent of the source-data likelihoods. Consider, for example, the scenario where we wish to estimate overfitting for a particular set, $\mathcal{V}^{\mathbf{y}}$, with consistently high $\mathcal{V}^{\mathbb{x}^{(\tau,l)}}$ source-data likelihoods, and generally low per-sample $d_{\text{MFCC}}(\mathbf{y}_n, \hat{\mathbf{y}}_n)$ cepstral distortions—which should translate to a generally low value for the normalized cepstral distance, $\tilde{d}$. Replacing the normalizing denominator in Eq. (5.70) by the cardinality of the data—i.e., $|\mathcal{V}^{\mathbb{x}^{(\tau,l)}}|$, effectively transforming $\tilde{d}$ into a mean of likelihood-weighted cepstral distances—would result in a misleadingly high value for $\tilde{d}(\mathcal{V}^{\mathbf{y}}, \mathcal{V}^{\hat{\mathbf{y}}})$, despite the low $d_{\text{MFCC}}(\mathbf{y}_n, \hat{\mathbf{y}}_n)$ cepstral distances.

distances corresponding to decreased generalization capability for the temporally-extended $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{c}_y}\}_{l>0}$ GMMs and, consequently, increased overfitting risk, while lower values for the normalized cepstral distance ratio indicate improved generalization capability.

Figure 5.11(b) illustrates the GMM generalization performance obtained for the example $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{c}_y}\}$ GMMs investigated previously in the context of oversmoothing assessment, with the generalization performance measured in terms of $\frac{\tilde{d}(\mathcal{V}^{\mathbf{c}_y}_{\text{test}},\mathcal{V}^{\hat{\mathbf{c}}_y}_{\text{test}})}{\tilde{d}(\mathcal{V}^{\mathbf{c}_y}_{\text{train}},\mathcal{V}^{\mathbf{c}_y}_{\text{train}})}$—our overfitting measure. With the memoryless $l = 0$ baseline performance nearly at unity, Figure 5.11(b) shows generalization performance to be decreasing to various extents for the different GMMs in the $l = 2$–6 range, before improving consistently for all GMMs. More specifically, the set $\{\mathcal{G}_{\diamond}\}$ exhibits only a slight 7% increase in overfitting at $l = 3$, $\{\mathcal{G}_{\circ}\}$ and $\{\mathcal{G}_{\triangle}\}$ exhibit increased overfitting for $l \in \{3,\ldots,6\}$ with the highest increases reaching $\sim 20\%$ at $l = 3$ and 4, respectively, while $\{\mathcal{G}_{\square}\}$ exhibits the largest degradation in terms of overfitting, reaching $\sim 34\%$ at $l = 4$. Compared to the multiple-fold increases in dimensionality—reaching up to $\frac{116}{16} = 7.25$-fold increase at $l = L = 10$—and noting the fact that no additional data was used for the EM-based training of our temporally-extended GMMs, these performance figures indicate that we have succeeded to a fair extent in avoiding the detrimental effects of increased dimensionality on the generalization capabilities of our high-dimensional temporally-extended GMMs for $l \approxeq 3$–4, being successful to a much larger extent elsewhere.

Reviewing the results of Figure 5.11(b) more closely, we observe that our ability to address the risk of overfitting is closely tied to the effectiveness of our pruning and fuzzy clustering algorithms, as determined by the choices for the distribution flatness threshold and fuzziness parameters, $\rho_{\min}$ and K, respectively. For $\{\mathcal{G}_{\square}\}$ where degradation in generalization performance is highest, $\rho_{\min}$ is lower relative to the value used in constructing the other GMMs. As described in Operation (d) of Section 5.4.2.3, $\rho_{\min}$, corresponding to a threshold on the minimum *whiteness* for the distribution of incremental data, is intended to limit the expansion of the temporally-extended GMM to those localized time-frequency regions where information content is highest. As such, lower $\rho_{\min}$ values translate to less-restrictive distribution flatness thresholds and, consequently, a higher number of Gaussian components in the resulting temporally-extended GMMs. This higher complexity, discussed and illustrated in Section 5.4.3.1 below, naturally increases the risks of overfitting.

To a similar extent, the generalization performances illustrated in Figure 5.11(b) for $\{\mathcal{G}_{\triangle}\}$ and $\{\mathcal{G}_{\diamond}\}$ demonstrate the importance of the fuzziness factor, K, in reducing the risk of overfitting. Despite the two-fold increase in the splitting factor, J (the number of

child states that can potentially be derived at each temporal increment for each parent state), in $\{\mathcal{G}_{\diamond}\}$ compared to $\{\mathcal{G}_{\triangle}\}$, the proportional increase in K for $\{\mathcal{G}_{\diamond}\}$, relative to $\{\mathcal{G}_{\triangle}\}$, completely alleviates any risk of increased overfitting in $\{\mathcal{G}_{\diamond}\}$ as a result of the higher splitting factor. Without such a proportional increase in K, the increased splitting factor would translate into roughly a two-fold reduction in the cardinality of the training data subsets used for the weighted EM-based estimation of child state *pdf*s, and correspondingly, into an equivalent two-fold increase in overfitting risk.

5.4.3 BWE performance using temporally-extended GMMs

Through our tree-like GMM extension algorithm for model-based memory inclusion, we have addressed the drawbacks of our frontend-based approach of Section 5.3—namely, the time-frequency information tradeoff and the non-causality, and associated algorithmic delay, imposed by delta features—while preserving its advantage in terms of the flexibility it provides for the inclusion of memory to varying extents—the primary advantage of delta features and simultaneously the deficiency of first-order HMM-based methods.

In this section, we first describe the modifications to be applied to our static MFCC-based dual-mode BWE system of Section 5.2.2—and illustrated in Figure 5.1—in order for the dual-mode system to be able to exploit the superior cross-band correlation properties of temporally-extend GMMs for improved highband speech reconstruction. Then, we evaluate the memory-inclusive BWE performance obtained using our temporally-extended GMMs, with the static MFCC-based $\boldsymbol{\mathcal{G}} = (\mathcal{G}_{\mathbf{x}\mathbf{C}_y}, \mathcal{G}_{\mathbf{x}_G})$ tuple and results of Section 5.2.6 used as the 0th-iteration model and performance baseline, respectively, for all performance evaluations except those investigating the effect of I—the modality of the 0th-order GMM tuple.

5.4.3.1 System description

As described in Section 5.2, our MFCC-based dual-mode BWE system makes use of two GMMs, represented by the $\boldsymbol{\mathcal{G}} = (\mathcal{G}_{\mathbf{x}\mathbf{C}_y}, \mathcal{G}_{\mathbf{x}_G})$ tuple, to model the joint distributions of the MFCC-parameterized narrowband spectral envelopes with those of the high band, with the shape and gain of the latter modelled independently through $\mathcal{G}_{\mathbf{x}\mathbf{C}_y}$ and $\mathcal{G}_{\mathbf{x}_G}$, respectively. More specifically, the narrowband space modelled in both GMMs is represented by the static MFCC feature vector parameterization given by $\mathbf{x} := \mathbf{c}_x \triangleq [c_{x_1}, \ldots, c_{x_9}, c_{x_0}]^T$ for each frame of the midband-equalized narrowband signal spanning the 0–4 kHz range, while the

highband space in the 4–8 kHz range is represented in $\mathcal{G}_{\mathbf{X}\mathbf{C}_y}$ by the static $\mathbf{c}_y \triangleq [c_{y_1}, \ldots, c_{y_6}]^T$ MFCC feature vectors and in $\mathcal{G}_{\mathbf{X}G}$ by the excitation gain, g. As such, the joint-band dimensionalities for $\mathcal{G}_{\mathbf{X}\mathbf{C}_y}$ and $\mathcal{G}_{\mathbf{X}G}$ are 16 and 11, respectively, with $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{X}\\ \mathbf{C}_y\end{bmatrix}\right) = \begin{bmatrix}10\\6\end{bmatrix}$ and $\mathrm{Dim}\left(\begin{bmatrix}\mathbf{X}\\ G\end{bmatrix}\right) = \begin{bmatrix}10\\1\end{bmatrix}$. Using these static parameterizations and dimensionalities, we construct the lth-order temporally-extended narrowband, highband, and joint-band supervectors—represented by the random feature vector representations $\mathbb{X}^{(\tau,l)}$, $\mathbb{C}_y^{(\tau,l)}$ and $\mathbb{G}^{(\tau,l)}$, and $\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ \mathbb{C}_y^{(\tau,l)}\end{bmatrix}$ and $\begin{bmatrix}\mathbb{X}^{(\tau,l)}\\ \mathbb{G}^{(\tau,l)}\end{bmatrix}$, respectively—by causal concatenation with a frame step of τ as described in Section 5.4.2 above. By constructing lth-order temporally-extended versions of our training data set of Section 3.2.10 as such for all $l \in \{0, \ldots, L\}$, we then proceed to temporally extend the static $\boldsymbol{\mathcal{G}} = (\mathcal{G}_{\mathbf{X}\mathbf{C}_y}, \mathcal{G}_{\mathbf{X}G})$ GMM tuple of the dual-mode BWE system into the memory-inclusive $\{\boldsymbol{\mathcal{G}}^{(\tau,l)} \coloneqq (\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}, \mathcal{G}_{\mathbb{X}^{(\tau,l)}G})\}_{l\in\{0,\ldots,L\}}$ tuples using our tree-like memory inclusion algorithm implemented per Table 5.5.

In addition to a causal concatenation of input static narrowband vectors similar to that discussed above, substituting the static $\boldsymbol{\mathcal{G}}$ tuple in the baseline dual-mode system of Figure 5.1 by the memory-inclusive $\{\boldsymbol{\mathcal{G}}^{(\tau,l)} \coloneqq (\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}, \mathcal{G}_{\mathbb{X}^{(\tau,l)}G})\}_{l\in\{0,\ldots,L\}}$ tuples represents the only modification needed to transform our static BWE system into one that exploits model-based memory inclusion to improve the quality of reconstructed highband speech. In particular, these minor modifications, illustrated in Figure 5.12 below, allow us to perform MMSE-based estimation of highband speech using the same memoryless formulae derived in Section 3.3.1, namely Eqs. (3.12), (3.16) and (3.17), but with the $\mathbf{X}$ input and $\mathcal{G}_{\mathbf{X}\mathbf{Y}}$ GMM parameters replaced by $\mathbb{X}^{(\tau,l)}$ and the parameters of $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}$, respectively.

Figure 5.12 also shows the transient processing required during the initial durations of input speech. For a particular desired memory inclusion index, l, the effective time-dependent order, ℓ, is determined during extension based on the duration of the observed input; for initial speech input where the observed number of input frames at a particular tth frame is insufficient to construct the desired lth-order causal supervectors, the effective order ℓ is determined as $\ell = \left\lfloor \frac{t}{\tau} \right\rfloor$, and set to the desired $\ell = l$ otherwise. Since $\ell < l$ only transiently, namely when $t < l\tau$, the vast majority of our TIMIT test frames are extended at the desired lth memory inclusion index, even at the maximum values of $l = L = 10$ and $\tau = 8$ (corresponding to 800 ms) used in our performance analysis in Section 5.4.3.2 below.[169,170]

[169] See Section 3.2.10 for details on our training and testing data sets derived from the TIMIT corpus.

[170] As described in Section 3.2.8, we parameterize the time-domain speech signal in 20 ms frames with an overlap of 10 ms.

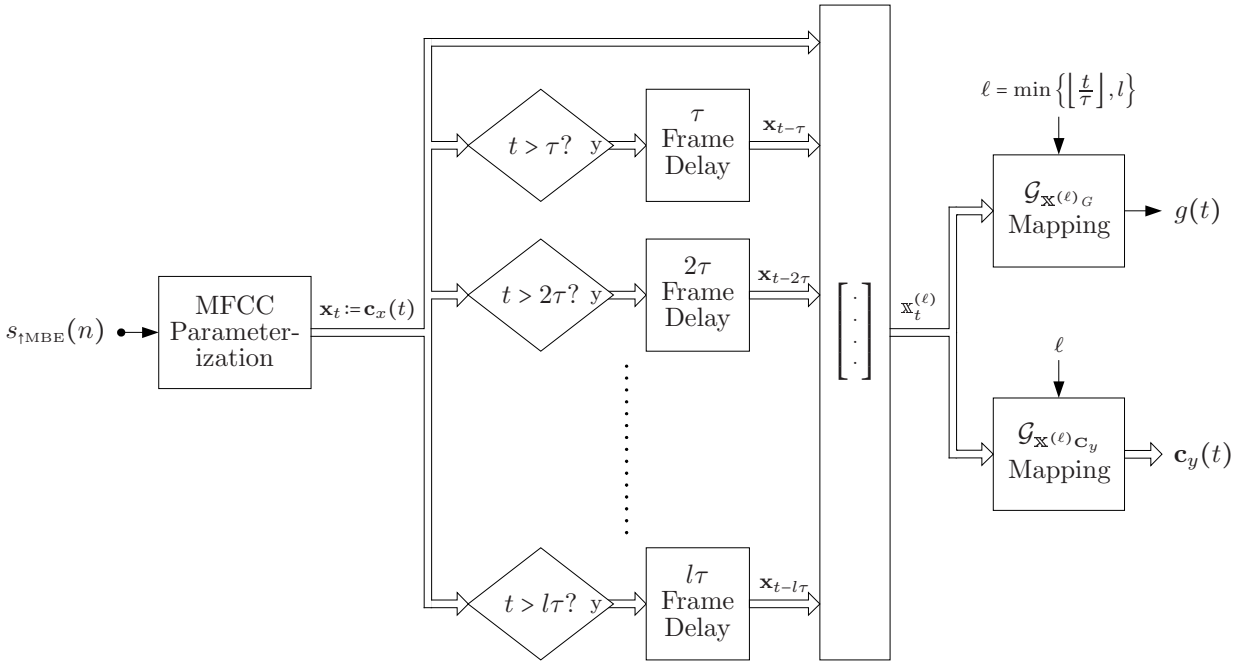


Fig. 5.12: Model-based memory inclusion modifications to the baseline MFCC-based dual-model BWE system of Figure 5.1 to incorporate temporally-extended GMMs. The modifications are applied to the upper-most path of the main processing block in Figure 5.1(b) and to the MMSE estimation block in Figure 5.1(c). With n and t representing the sample and frame time indices, respectively, the input signal, $s_{\uparrow\text{MBE}}(n)$, is that of the midband-equalized and interpolated narrowband speech, while τ and l represent the memory inclusion step and order, respectively.

Given the negligible cost associated with the causal concatenation and GMM substitution modifications described above, the additional computational complexity involved with performing BWE using our temporally-extended $\boldsymbol{\mathcal{G}}^{(\tau,l)} \coloneqq (\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}, \mathcal{G}_{\mathbb{X}^{(\tau,l)}G})$ GMM tuples—relative to the cost of performing BWE using memoryless $\boldsymbol{\mathcal{G}} = (\mathcal{G}_{\mathbf{X}\mathbf{C}_y}, \mathcal{G}_{\mathbf{X}G})$ tuples as described in Section 5.2—is, thus, limited only to the additional cost of performing MMSE-based reconstruction of highband MFCCs using temporally-extended GMMs with higher joint-band dimensionalities and higher modalities compared to the baseline memoryless GMMs. As such, the computational cost of our model-based memory inclusion technique can be easily expressed in terms of the total number of per-frame computations, $N_{\text{FLOPs/f}}$, associated with MMSE estimation per Eqs. (3.12), (3.16) and (3.17), in the same manner previously detailed in Section 3.5.1 for the evaluation of the effect of GMM covariance type on BWE performance and computational complexity. More specifically, for

each of the lth-order $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}$ and $\mathcal{G}_{\mathbb{X}^{(\tau,l)}G}$ GMMs with the modalities $M^{\mathbb{X}^{(\tau,l)}\mathbf{c}_y}$ and $M^{\mathbb{X}^{(\tau,l)}g}$, respectively, we perform the following matrix operations offline prior to extension for all $i \in \{1,\dots,M^{\mathbb{X}^{(\tau,l)}\mathbf{c}_y}\}$ and all $j \in \{1,\dots,M^{\mathbb{X}^{(\tau,l)}g}\}$:

(a) $-\frac{1}{2}\Big[\mathbf{C}_i^{\mathbb{X}^{(\tau,l)}\mathbb{X}^{(\tau,l)}}\Big]^{-1}$ and $-\frac{1}{2}\Big[\mathbf{C}_j^{\mathbb{X}^{(\tau,l)}\mathbb{X}^{(\tau,l)}}\Big]^{-1}$;

(b) $\mathbf{C}_i^{\mathbf{c}_y\mathbb{X}^{(\tau,l)}}\Big[\mathbf{C}_i^{\mathbb{X}^{(\tau,l)}\mathbb{X}^{(\tau,l)}}\Big]^{-1}$ and $\mathbf{C}_j^{g\mathbb{X}^{(\tau,l)}}\Big[\mathbf{C}_j^{\mathbb{X}^{(\tau,l)}\mathbb{X}^{(\tau,l)}}\Big]^{-1}$; and,

(c) $\alpha_i^{\mathbb{X}^{(\tau,l)}\mathbf{c}_y}(2\pi)^{-p/2}\Big|\mathbf{C}_i^{\mathbb{X}^{(\tau,l)}\mathbb{X}^{(\tau,l)}}\Big|^{-1/2}$ and $\alpha_j^{\mathbb{X}^{(\tau,l)}g}(2\pi)^{-p/2}\Big|\mathbf{C}_j^{\mathbb{X}^{(\tau,l)}\mathbb{X}^{(\tau,l)}}\Big|^{-1/2}$;

where $p \coloneqq \mathrm{Dim}\Big(\mathbb{X}^{(\tau,l)}\Big) = 10(l+1)$. Using these pre-computed quantities in the application of the MMSE estimation of Eq. (3.12) for both $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}$ and $\mathcal{G}_{\mathbb{X}^{(\tau,l)}G}$, the total number of the per-frame extension-stage computations—previously given in Eq. (3.34) for a single GMM—can now be calculated for the $(\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}, \mathcal{G}_{\mathbb{X}^{(\tau,l)}G})$ pair as

$$\begin{aligned} N_{\text{FLOPs/f}} &= M^{\mathbb{X}^{(\tau,l)}\mathbf{c}_y}(2p^2+14p+27)+5 \\ &\quad + M^{\mathbb{X}^{(\tau,l)}g}(2p^2+4p+22) \\ &= M^{\mathbb{X}^{(\tau,l)}\mathbf{c}_y}\big(200(l+1)^2+140(l+1)+27\big)+5 \\ &\quad + M^{\mathbb{X}^{(\tau,l)}g}\big(200(l+1)^2+40(l+1)+22\big), \end{aligned} \tag{5.72}$$

where we have substituted the highband parameter dimensionality, q, in Eq. (3.34) by $\mathrm{Dim}(\mathbf{C}_y) = 6$ and $\mathrm{Dim}(G) = 1$ for $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}$ and $\mathcal{G}_{\mathbb{X}^{(\tau,l)}G}$, respectively.

Calculated using Eq. (5.72), the extension-stage computational cost—for the four $\boldsymbol{\mathcal{G}}^{(\tau,l)}$ GMM tuples with the same parameters as those $\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}$ GMMs previously considered in Figure 5.11—is shown in Figure 5.13 below as a function of the memory inclusion index, l, as well as of the combined tuple modality, $M^{\mathbb{X}^{(\tau,l)}\mathbf{c}_y} + M^{\mathbb{X}^{(\tau,l)}g}$. As a result of the increase in both the dimensionalities and modalities of the joint-band $\boldsymbol{\mathcal{G}}^{(\tau,l)}$ GMMs relative to those of our memoryless baseline of Section 5.2, Figure 5.13 shows a corresponding increase in extension-stage computational cost, with the increase at the higher orders of memory inclusion reaching ~ 2–4 orders of magnitude above the cost for our memoryless baseline GMMs. In comparison, previous results in Figure 3.5(b) show an $N_{\text{FLOPs/f}}$ increase of ~2 orders of magnitude when the modality of each of the memoryless $\mathcal{G}_{\mathbf{x}\mathbf{C}_y}$ and $\mathcal{G}_{\mathbf{x}G}$ GMMs is increased from $M^{\text{full}} = 2$ to 256.

To further put the results of Figure 5.13 into context, we compare them against the typical computational capabilities of current personal computers, and more importantly,

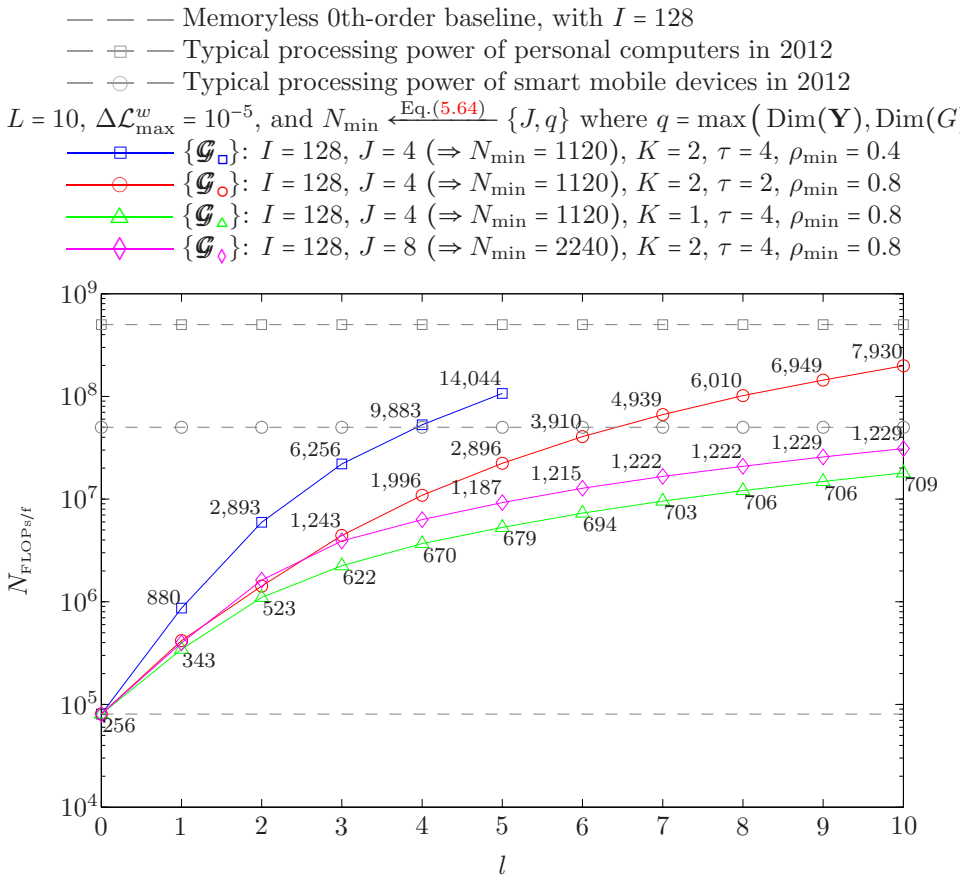


Fig. 5.13: Computational cost of performing MMSE-based estimation of highband MFCCs using temporally-extended GMM tuples. Using Eq. (5.72), the per-frame computational cost represented by $N_{\text{FLOPs/f}}$ is plotted as a function of the memory inclusion index, $l \in \{0, \ldots, L\}$, with the total number of Gaussian components for each lth-order temporally-extended $\mathcal{G}^{(\tau,l)} := (\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}, \mathcal{G}_{\mathbb{X}^{(\tau,l)}G})$ GMM tuple—i.e., $M^{\mathbb{X}^{(\tau,l)}\mathbf{c}_y} + M^{\mathbb{X}^{(\tau,l)}g}$—labelled next to the corresponding data point. Providing a frame of reference for the purpose of practical real-time implementation, the computational capabilities of typical personal computers and smart mobile devices in 2012—calculated in terms of $N_{\text{FLOPs/f}}$ based on figures from [190]—are also shown.

current modern communication devices—e.g., tablets and smart phones. Given the non-causality advantage of our model-based memory inclusion technique, gauging the computational requirements of our BWE technique against the processing capabilities of modern communication devices is important to assess its practicality in terms of real-time implementation. As recently discussed in [190], a standard 2012 laptop has a typical performance of 50 GFLOPs per second, while a typical 2012 tablet or smart phone performs at around

5 GFLOPs per second. Given our 100 frame/s processing rate of the input narrowband speech,[171] these figures correspond to $N_{\text{FLOPs/f}} = 5 \times 10^8$ and 5×10^7 for computers and smart mobile devices, respectively. Based on these latter numbers, Figure 5.13 shows the computational requirements of our model-based memory inclusion technique to be well within the capabilities of laptops and personal computers for all GMMs considered. Relative to the processing power of typical smart mobile devices, however, Figure 5.13 shows that the computational cost of our technique can potentially be too high for real-time implementation at higher orders of memory inclusion, depending on the values chosen for the parameters of our tree-like GMM training algorithm. While the BWE cost using the $\{\mathcal{G}_{\triangle}\}$ and $\{\mathcal{G}_{\diamond}\}$ GMM sets is within the processing power of smart mobile devices up to 400 ms of causal memory inclusion, the cost for $\{\mathcal{G}_{\square}\}$ and $\{\mathcal{G}_{\circ}\}$ reaches the limit of smart mobile device real-time capabilities at 160 and 180 ms of memory inclusion, respectively.

In addition to the observations made in Section 5.4.2.4 regarding the role of the pruning steps of our tree-like GMM extension algorithm in reducing GMM overfitting, the observations above further emphasize the importance of these pruning steps proposed in Operation (d) as an integral component of our algorithm. In particular, we note that, among the $\{\mathcal{G}\}$ sets considered in Figure 5.13, the $\{\mathcal{G}_{\square}\}$ set characterized by having the lowest—and hence, most permissive—value for the distribution flatness threshold, $\rho_{\min}$, is found to be the most computationally demanding, thereby demonstrating the importance of pre-EM pruning in Eq. (5.63) via $\rho_{\min}$. Similarly, the lower computational cost associated with $\{\mathcal{G}_{\diamond}\}$ relative to that associated with $\{\mathcal{G}_{\circ}\}$—noting that both sets share similar values for $\rho_{\min}$ and K, the fuzziness factor, but differ in J, the splitting factor, and consequently differ in $N_{\min}$, the child subset cardinality threshold—demonstrates the effectiveness of post-EM pruning in Eq. (5.65) via $N_{\min}$.

To conclude, we note that the $\{\mathcal{G}_{\triangle}\}$ set, characterized from the other sets in Figure 5.13 by its lower value for the fuzziness factor, K, involves the least computational cost. Per our discussion in Operation (a) regarding our fuzzy clustering approach, this observation is indeed expected since a lower value for K translates into lower cardinalities for the time-frequency-localized child subsets obtained at each iteration of the tree-like algorithm.[172] Per Eq. (5.65), these lower cardinalities result, in turn, in higher likelihoods for post-EM pruning of states in our tree-like training algorithm as a result of the $N_{\min}$ threshold imposed

[171] See Footnote 170.
[172] See Eq. (5.21).

on each child subset's cardinality as a condition for splitting an associated parent state into multiple children states. At the same time, however, we showed through the illustrative example of Figure 5.9, as well as through the overfitting results of Figure 5.11(b), that lower values for K—or, more specifically, lower K/J ratios—correspond to higher overfitting risks in our high-dimensional temporally-extended GMMs. Connecting these various observations together thus emphasizes the importance of choosing a value for K to obtain the compromise—between GMM complexity and generalization capabilities—that is most suitable for the domain in which our model-based BWE technique is implemented. For real-time implementations on smart mobile devices where reducing complexity takes precedence, lower values for the K/J ratio are more suitable. Conversely, for offline BWE implementations where reconstruction quality—and hence, GMM generalization performance—outweighs computational costs, higher values for K/J are more appropriate.

Given the relatively large variability shown above by our model-based approach for memory-inclusive BWE in terms of computational cost, and the practical importance of such cost in general, we include the per-frame computational complexity, $N_{\text{FLOPs/f}}$, as part of the analysis presented below for the BWE performance of our approach.

5.4.3.2 Performance and analysis

Compared to our frontend-based approach of Section 5.3, our algorithm for memory inclusion through the construction of temporally-extended GMMs involves a relatively large number of variables as summarized in the preamble of Table 5.5. As such, performing an exhaustive joint-variable optimization for our temporally-extended $\{\boldsymbol{\mathcal{G}}^{(\tau,l)} = (\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{C}_y}, \mathcal{G}_{\mathbb{X}^{(\tau,l)}G})\}$ GMM tuples in the manner applied for the dynamic $\left\{\overset{\triangle}{\boldsymbol{\mathcal{G}}} = \left(\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}_y}(\overset{\triangle}{\mathbf{x}}, \mathbf{c}_y), \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}(\overset{\triangle}{\mathbf{x}}, g)\right)\right\}$ tuples in Section 5.3.4, is rather prohibitive computationally. Instead, we evaluate and demonstrate the effect of each of our model-based algorithm's parameters on BWE performance individually in order to deduce the parameter ranges corresponding to the best performance achievable within the typical computational capabilities of recent smart mobile devices.

Using the LSD, Itakura-based, and PESQ measures detailed in Section 3.4, we evaluate the performance of our model-based memory-inclusive BWE technique in Figures 5.14–5.18 below, as a function of: $\rho_{\min}$, the distribution flatness threshold; J, the splitting factor; K, the fuzziness factor; τ, the memory inclusion step; and I, the 0th-order GMM modality; respectively. The performance of the memoryless MFCC-based dual-mode BWE system of

Table 5.1—with a modality of 128 for both the static $\mathcal{G}_{\mathbf{X}\mathbf{C}_y}$ and $\mathcal{G}_{\mathbf{X}_G}$ GMMs—represents the memoryless 0th-order baseline for those performances obtained using temporally-extended GMMs in Figures 5.14–5.18. Corresponding to a $\boldsymbol{\mathcal{G}}^{(\tau,l)}$ temporally-extended GMM tuple with $l = 0$ and $I = 128$, we denote the memoryless baseline model by $\boldsymbol{\mathcal{G}}^{(0)}$. For the purpose of further comparing the BWE performance of our model-based memory inclusion technique against that of frontend-based memory inclusion, we also illustrate the performance of our optimized model of Figure 5.7—with $\text{Dim}(\mathbf{X}, \boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{Y}, \boldsymbol{\Delta}_{\mathbf{Y}}, \mathbf{Y}_{\text{ref}}) = (8, 2, 6, 2, 7)$—as a memory-inclusive reference in Figures 5.14–5.18, denoting the optimized frontend-based tuples simply as $\left\{\overset{\triangle}{\boldsymbol{\mathcal{G}}}\right\}$.

To illustrate the effect of memory inclusion on BWE performance, we use the duration of included memory, T, as the abscissa in Figures 5.14–5.18, rather than the memory inclusion index, l, previously used in Figures 5.11 and 5.13.[173] This allows us to: (a) compare performances using the model-based $\{\boldsymbol{\mathcal{G}}^{(\tau,l)}\}$ tuples to those of the frontend-based $\{\overset{\triangle}{\boldsymbol{\mathcal{G}}}\}$ where, per Eq. (4.34), the duration of included memory depends rather on the radius of the delta feature calculation window, L_δ;[174] and (b) make a fair comparison of the performances of various $\{\boldsymbol{\mathcal{G}}^{(\tau,l)}\}$ tuples with different time scales—for tuples with varying values for the memory inclusion step, τ, similar values of l correspond to different extents of memory inclusion. Given our 10 ms frame step,[175] the duration of included memory for $\{\boldsymbol{\mathcal{G}}^{(\tau,l)}\}$ and $\{\overset{\triangle}{\boldsymbol{\mathcal{G}}}\}$ is given by $T = 10 \cdot l \cdot \tau$ and $T = 2 \cdot 10 \cdot L_\delta$, respectively, noting the causality of memory inclusion in the case of $\{\boldsymbol{\mathcal{G}}^{(\tau,l)}\}$ versus its non-causality for $\{\overset{\triangle}{\boldsymbol{\mathcal{G}}}\}$.

Based on the performances shown in Figures 5.14–5.18, we can itemize our findings and conclusions into, first, conclusions based on global performance across all parameters—and their associated ranges—of temporally-extended GMMs, and, second, conclusions based on individual performances as a function of the primary parameters and operations underlying our tree-like GMM extension algorithm—namely those of pruning, splitting factor, fuzzy clustering, memory inclusion step, and initial 0th-order GMM complexity.

[173] To limit the computational complexity associated with our GMM extension algorithm of Table 5.5, training is stopped after completing the lth iteration at which the modality of either of the temporally-extended $\mathcal{G}_{\mathbf{X}^{(\tau,l)}\mathbf{C}_y}$ and $\mathcal{G}_{\mathbf{X}^{(\tau,l)}{}_G}$ GMMs exceeds 10^4.

[174] To differentiate the notation for the radius of the delta feature calculation window in Eq. (4.34) from that of the maximum value of the memory inclusion index used in our tree-like training algorithm in Table 5.5, we denote the former in this section by L_δ.

[175] See Footnote 170.

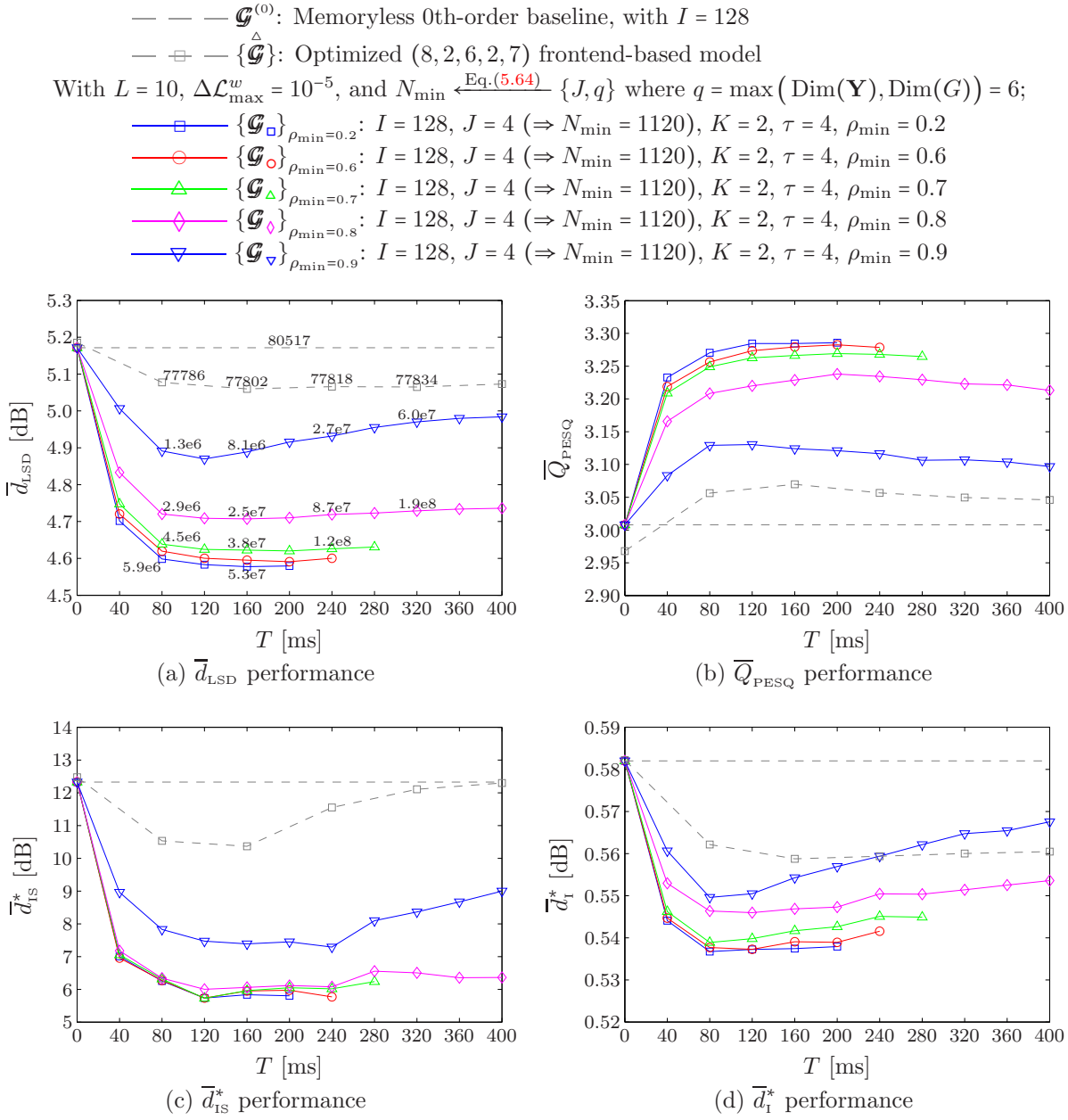


Fig. 5.14: Effect of the distribution flatness threshold, $\rho_{\min}$, on the performance of our model-based memory-inclusive BWE technique. Performances using: (a) the memoryless 0th-order baseline GMM tuple, $\boldsymbol{\mathcal{G}}^{(0)} := (\mathcal{G}_{\mathbf{X}\mathbf{C}_y}, \mathcal{G}_{\mathbf{X}G})$; and (b) the optimized frontend-based GMM tuples, $\Big\{\overset{\triangle}{\boldsymbol{\mathcal{G}}} = \Big(\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{\mathbf{C}}_y}(\overset{\triangle}{\mathbf{x}}, \mathbf{c}_y), \mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}(\overset{\triangle}{\mathbf{x}}, g)\Big)\Big\}$, where $\text{Dim}(\mathbf{X}, \boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{Y}, \boldsymbol{\Delta}_{\mathbf{Y}}, \mathbf{Y}_{\text{ref}}) = (8, 2, 6, 2, 7)$; are shown as references for the performances using temporally-extended $\{\boldsymbol{\mathcal{G}}^{(\tau,l)}\}$ tuples. Performances are plotted as a function of the duration of included memory, T, rather than the memory inclusion index, l, to allow comparison against frontend-based models. In addition to $\overline{d}_{\text{LSD}}$ performance, Subfigure (a) also shows the total per-frame computational cost, $N_{\text{FLOPs/f}}$, for various GMM tuples.

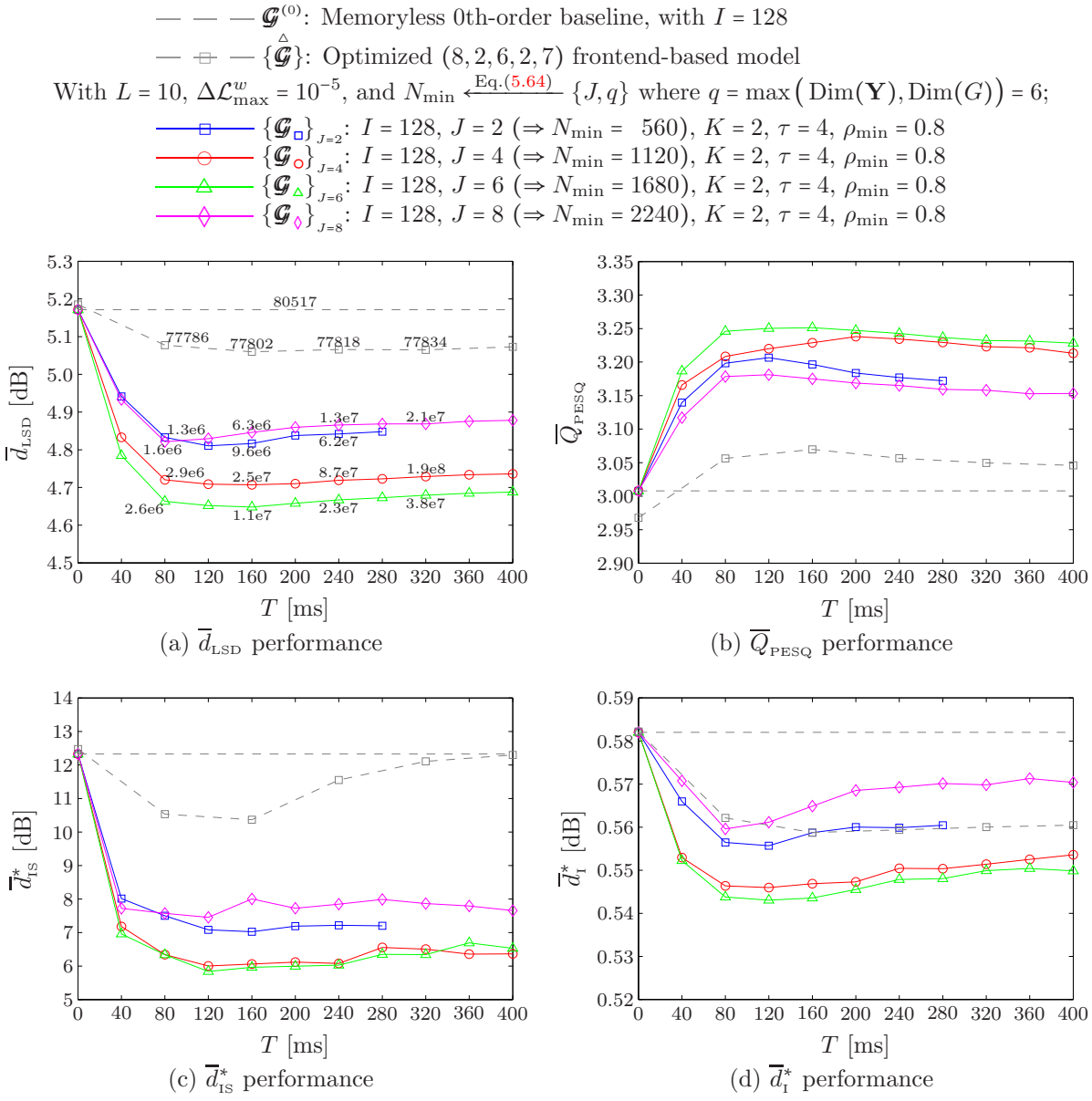


Fig. 5.15: Effect of the splitting factor, J, on the performance of our model-based memory-inclusive BWE technique. Performances using: (a) the memoryless 0th-order baseline GMM tuple, $\mathcal{G}^{(0)}$; and (b) the optimized $(8,2,6,2,7)$ frontend-based GMM tuples, $\left\{\overset{\triangle}{\mathcal{G}}\right\}$; are shown as references for the performances using temporally-extended $\{\mathcal{G}^{(\tau,l)}\}$ tuples. Performances are plotted as a function of the duration of included memory, T, rather than the memory inclusion index, l, to allow comparison against frontend-based models. In addition to $\overline{d}_{\text{LSD}}$ performance, Subfigure (a) also shows the total per-frame computational cost, $N_{\text{FLOPs/f}}$, for various GMM tuples.

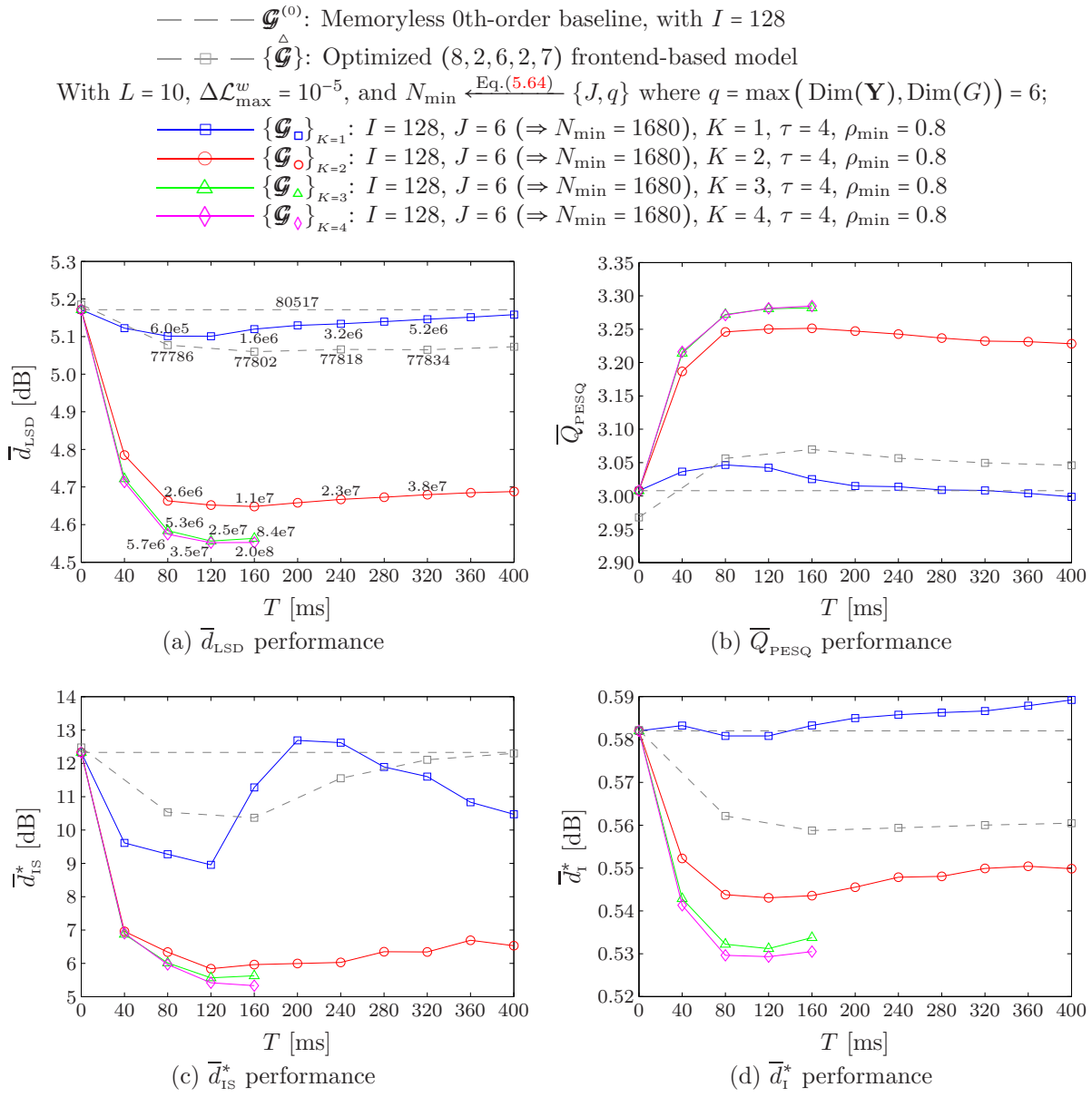
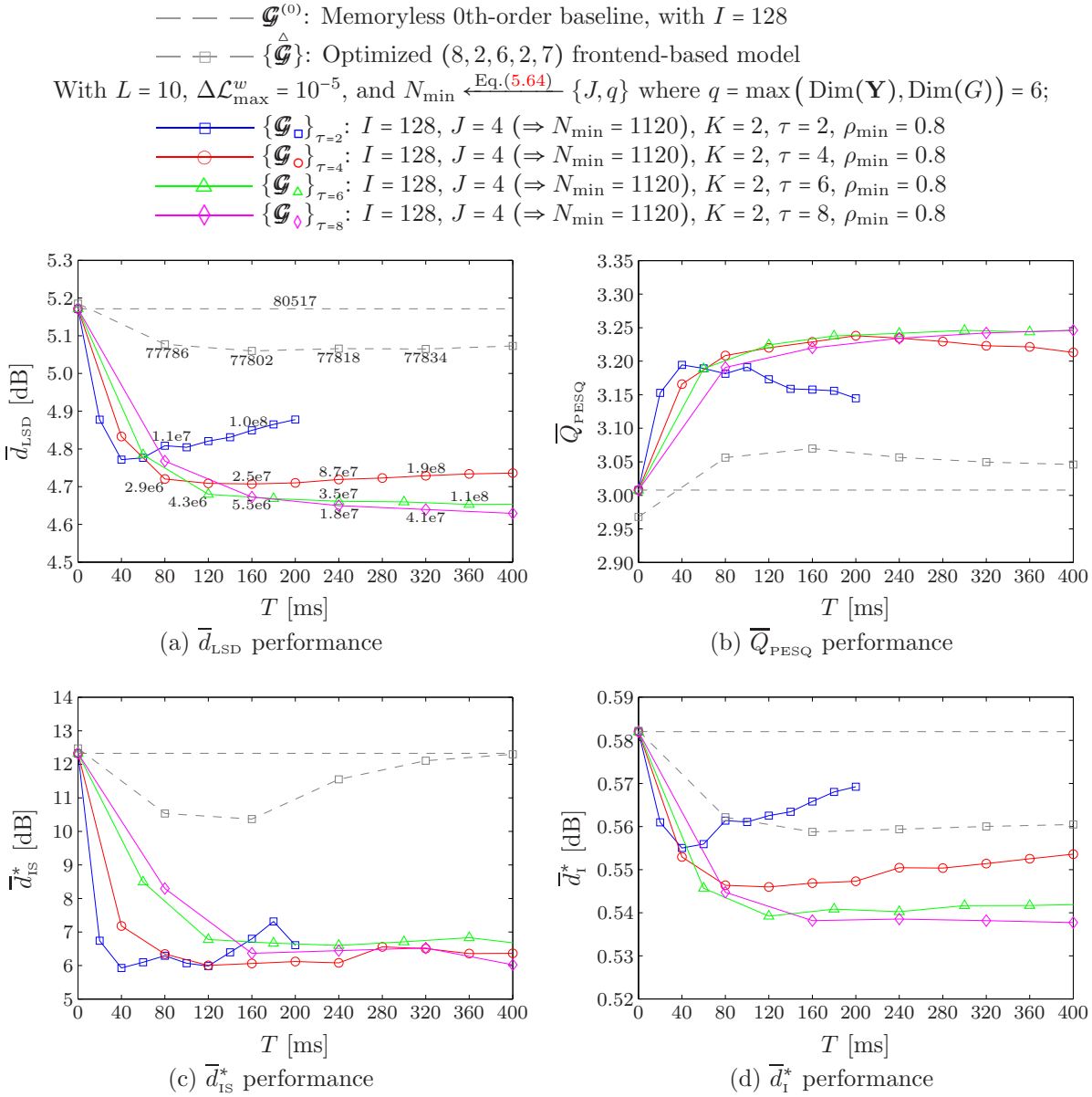


Fig. 5.16: Effect of the fuzziness factor, K, on the performance of our model-based memory-inclusive BWE technique. Performances using: (a) the memoryless 0th-order baseline GMM tuple, $\boldsymbol{\mathcal{G}}^{(0)}$; and (b) the optimized $(8,2,6,2,7)$ frontend-based GMM tuples, $\left\{\overset{\triangle}{\boldsymbol{\mathcal{G}}}\right\}$; are shown as references for the performances using temporally-extended $\{\boldsymbol{\mathcal{G}}^{(\tau,l)}\}$ tuples. Performances are plotted as a function of the duration of included memory, T, rather than the memory inclusion index, l, to allow comparison against frontend-based models. In addition to $\overline{d}_{\text{LSD}}$ performance, Subfigure (a) also shows the total per-frame computational cost, $N_{\text{FLOPs/f}}$, for various GMM tuples.



(a) $\overline{d}_{\text{LSD}}$ performance

(b) $\overline{Q}_{\text{PESQ}}$ performance

(c) $\overline{d}^{\star}_{\text{IS}}$ performance

(d) $\overline{d}^{\star}_{\text{I}}$ performance

Fig. 5.17: Effect of the memory inclusion step, τ, on the performance of our model-based memory-inclusive BWE technique. Performances using: (a) the memoryless 0th-order baseline GMM tuple, $\mathcal{G}^{(0)}$; and (b) the optimized $(8, 2, 6, 2, 7)$ frontend-based GMM tuples, $\left\{\overset{\triangle}{\mathcal{G}}\right\}$; are shown as references for the performances using temporally-extended $\{\mathcal{G}^{(\tau,l)}\}$ tuples. Performances are plotted as a function of the duration of included memory, T, rather than the memory inclusion index, l, to: (a) allow comparison against frontend-based models, and (b) account for the time-scale differences resulting at similar values of l for the different $\mathcal{G}^{(\tau,l)}$ tuples due to the varying value of τ. In addition to $\overline{d}_{\text{LSD}}$ performance, Subfigure (a) also shows the total per-frame computational cost, $N_{\text{FLOPs/f}}$, for various GMM tuples.

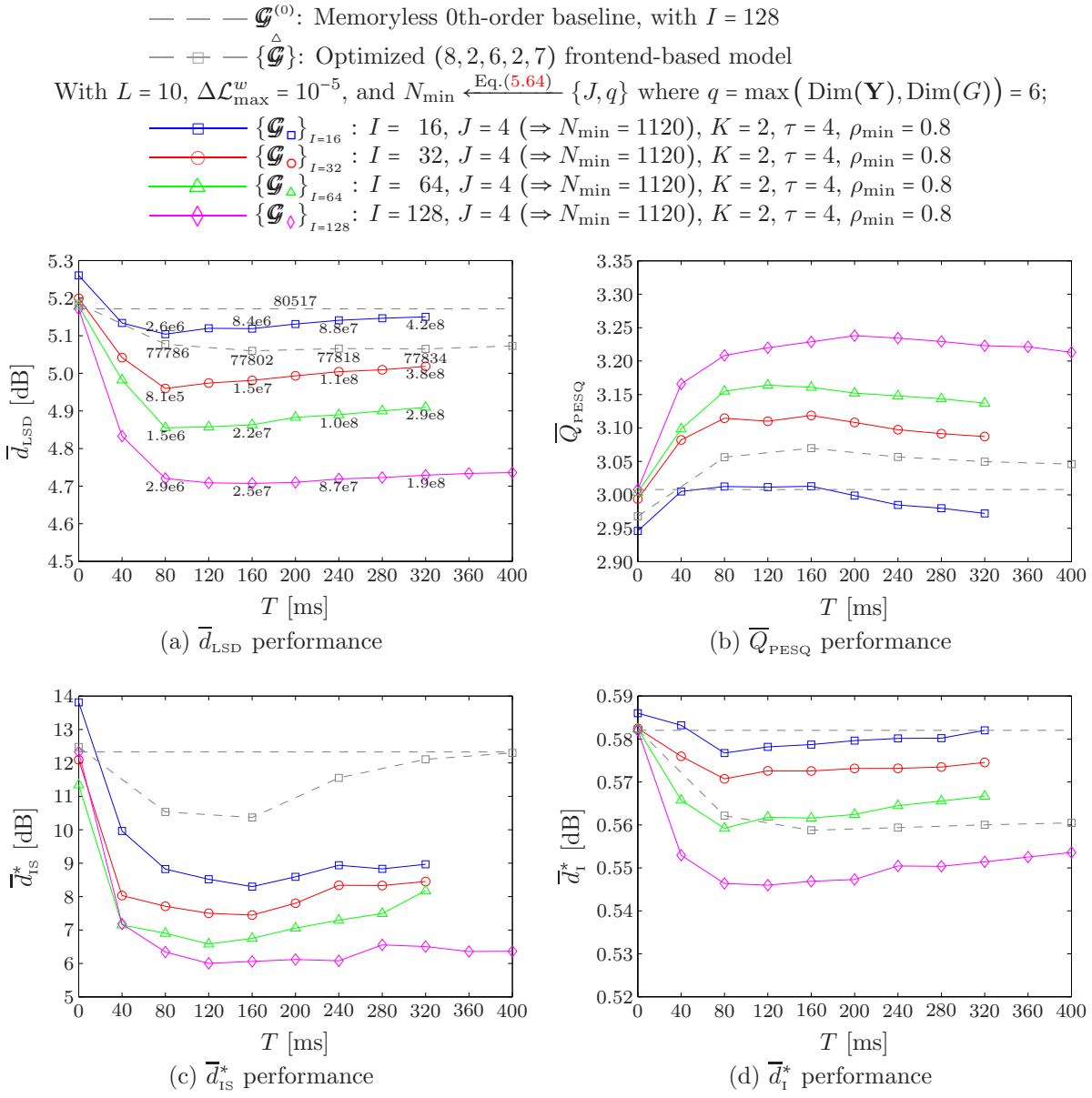


Fig. 5.18: Effect of the 0th-order GMM modality, I, on the performance of our model-based memory-inclusive BWE technique. Performances using: (a) the memoryless 0th-order baseline GMM tuple, $\mathcal{G}^{(0)}$; and (b) the optimized $(8,2,6,2,7)$ frontend-based GMM tuples, $\left\{\overset{\triangle}{\mathcal{G}}\right\}$; are shown as references for the performances using temporally-extended $\{\mathcal{G}^{(\tau,l)}\}$ tuples. Performances are plotted as a function of the duration of included memory, T, rather than the memory inclusion index, l, to allow comparison against frontend-based models. In addition to $\overline{d}_{\text{LSD}}$ performance, Subfigure (a) also shows the total per-frame computational cost, $N_{\text{FLOPs/f}}$, for various GMM tuples.

i. Global performance

- Except for the outlier performance of the excessively-overfitted $\{\mathcal{G}_{\square}\}_{K=1}$ tuples of Figure 5.16 discussed below in more detail, the BWE performances of all temporally-extended GMM tuples with a 0th-order modality of $I = 128$—being thereby comparable to the $(I = 128)$-modal memoryless baseline—are clearly superior to the memoryless performance baseline, across all performance measures, all parameter ranges, and all memory inclusion durations considered—namely, up to 400 ms. This confirms the success of our technique in achieving its basic objective—exploiting the previously-quantified cross-band correlation information in long-term speech to improve BWE performance beyond that achievable by conventional memoryless techniques.

- Except for the excessively-overfitted $\{\mathcal{G}_{\square}\}_{K=1}$ tuples of Figure 5.16 and the excessively-simplified $\{\mathcal{G}_{\square}\}_{I=16}$ tuples of Figure 5.18, all temporally-extended GMM tuples clearly outperform the optimized frontend-based tuples in terms of $\overline{d}_{\text{LSD}}$, $\overline{Q}_{\text{PESQ}}$, and $\overline{d}^{*}_{\text{IS}}$, across all parameter ranges and all memory inclusion durations considered, in some cases by a considerable multiple-fold margin. In terms of the gain-independent $\overline{d}^{*}_{\text{I}}$ performance, however, the improvements obtained via temporally-extended tuples over the frontend-based baseline performance are much less pronounced, and furthermore, are achieved only for particular subsets of the extension algorithm's parameter ranges and memory inclusion durations. Nonetheless, the considerable overall superiority of our model-based approach to memory inclusion is clear from Figures 5.14–5.18, thereby confirming our success in addressing the drawbacks of our frontend-based approach of Section 5.3, and consequently succeeding in translating significantly more of the previously-quantified information-theoretic gains of memory inclusion into measurable BWE performance improvements. A more detailed analysis of the results obtained for the different performance measures, and the implications of these results, is discussed below.

- The BWE performances of all temporally-extended tuples considered reach saturation at various memory inclusion durations of $T \leq 200$ ms, with the majority saturating at ~ 120–160 ms. In other words, the inclusion of causal memory beyond $T = 200$ ms is consistently found to add no further improvements, regardless of the parameter values used in our GMM temporal extension algorithm. This result thus coincides with our previous information-theoretic findings of Section 4.4.3 regarding the saturation of

acoustic-only memory contributions to highband certainty at the syllabic rate.

- By comparing BWE performances against the associated computational costs across T for the $\{\boldsymbol{\mathcal{G}}_{\circ}\}_{J=4}$ and $\{\boldsymbol{\mathcal{G}}_{\triangle}\}_{J=6}$ tuples in Figure 5.15(a), as well as for all tuples in Figures 5.17(a) and 5.18(a), we can conclude that higher GMM complexity does not necessarily translate into higher BWE performance. Indeed, among all the performances considered, the absolute best is achieved with the $\boldsymbol{\mathcal{G}}_{\diamond}\big|_{K=4}$ tuple—shown in Figure 5.16—at $T = 160\,\text{ms}$ with $N_{\text{FLOPs/f}} = 2 \times 10^8$, despite having also considered more complex tuples—such as those with higher memory inclusion in Figure 5.18, for example.[176] This emphasizes the value of the several parameters employed in our GMM extension algorithm in terms of the control and flexibility they provide.

- In terms of the memory inclusion duration at which it is achieved, the highest performance improvement obtained using our model-based memory inclusion approach is consistent with that of our optimized frontend-based approach in Figure 5.7 and Table 5.3. Both approaches achieve the highest improvements at $T = 160\,\text{ms}$.

- As described in Section 5.4.3.1, recent smart mobile devices have a typical processing power equivalent to $N_{\text{FLOPs/f}} \approxeq 5{\times}10^7$. Thus, taking practical real-time implementation into account, the best BWE performance achieved by temporally-extended GMM tuples within the computational capabilities of smart mobile devices is that obtained with $N_{\text{FLOPs/f}} \approxeq 3.5 \times 10^7$ using $\boldsymbol{\mathcal{G}}_{\diamond}\big|_{K=4}$ at $T = 120\,\text{ms}$—shown in Figure 5.16. Nearly identical performance is also achieved at $T = 120\,\text{ms}$ using $\boldsymbol{\mathcal{G}}_{\triangle}\big|_{K=3}$, at the slightly lower computational cost of $N_{\text{FLOPs/f}} \approxeq 2.5 \times 10^7$.

- Table 5.6 details the best performance improvements—absolute as well as computationally-constrained—obtained using temporally-extended tuples. Relative to the memoryless baseline performance of Table 5.1, the improvements achieved using our proposed model-based memory inclusion technique range from ≈ 2.3 times the improvements previously summarized in Table 5.3 using our frontend-based approach for $\overline{d}_{\text{I}}^{\,*}$ performance, to ≈ 5.5 times for $\overline{d}_{\text{LSD}}$ performance. For $\overline{Q}_{\text{PESQ}}$, the measure most

[176] As shown in Table 5.6, the performances of $\{\boldsymbol{\mathcal{G}}_{\diamond}\}_{K=4}$ at $T = 120$ and $160\,\text{ms}$ are virtually identical, with the $\overline{Q}_{\text{PESQ}}$ performance at $T = 160\,\text{ms}$ marginally better. Since PESQ is the measure most subjectively correlated—with an average correlation of 0.935 with subjective MOS scores as described in Section 3.4.3—among the four measures considered, we favour the performance of $\boldsymbol{\mathcal{G}}_{\diamond}\big|_{K=4}$ at $T = 160\,\text{ms}$ as the higher one.

subjectively correlated among the four performance measures considered, the performance improvement resulting from including memory via temporally-extended tuples exceeds that obtained by using dynamic delta coefficient-based tuples by ≈ 4.4 times. As shown in Table 5.6, these significant improvements are achieved at an increase of nearly four orders of magnitude in computational cost.

Table 5.6: Highest BWE performance improvements achieved using model-based memory inclusion via the temporally-extended $\boldsymbol{\mathcal{G}}_{\lozenge}|_{K=4}$ GMM tuple, in comparison to that achieved using the optimal frontend-based $\overset{\triangle}{\boldsymbol{\mathcal{G}}}$ tuple of Table 5.3 with $\text{Dim}(\mathbf{X}, \boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{Y}, \boldsymbol{\Delta}_{\mathbf{Y}}, \mathbf{Y}_{\text{ref}}) = (8, 2, 6, 2, 7)$. Improvements are measured relative to the memoryless MFCC-based dual-mode baseline results of $\boldsymbol{\mathcal{G}}^{(0)}$ in Table 5.1.

	T [ms]	$N_{\text{FLOPs/f}}$	$\overline{d}_{\text{LSD}}$ [dB]	$\overline{Q}_{\text{PESQ}}$	$\overline{d}^{*}_{\text{IS}}$ [dB]	$\overline{d}^{*}_{\text{I}}$ [dB]
$\boldsymbol{\mathcal{G}}^{(0)}$	0	8.1e4	5.17	3.01	12.32	0.5820
$\overset{\triangle}{\boldsymbol{\mathcal{G}}}$	160	7.8e4	5.06	3.07	10.37	0.5588
Improvement	—	—	0.11 (2.2%)	0.06 (2.1%)	1.95 (15.9%)	0.0232 (4.0%)
$\boldsymbol{\mathcal{G}}_{\lozenge}\|_{K=4}$	160	2.0e8	4.55	3.29	5.33	0.5305
Improvement	—	—	0.62 (12.0%)	0.28 (9.2%)	6.99 (56.8%)	0.0515 (8.9%)
$\boldsymbol{\mathcal{G}}_{\lozenge}\|_{K=4}$	120	3.5e7	4.55	3.28	5.42	0.5293
Improvement	—	—	0.62 (12.0%)	0.27 (9.1%)	6.90 (56.1%)	0.0527 (9.1%)

- Similar to the analysis performed in Section 5.3.5.2 by making use of the knowledge about the perceptual principles underlying the four performance measures considered above, we can further analyze the results of Figures 5.14–5.18 and Table 5.6 to better understand the effect of model-based memory inclusion on highband envelope reconstruction accuracy, as follows:

  – Since the $\overline{d}_{\text{LSD}}$ measures weight all deviations in log spectra equally while $\overline{Q}_{\text{PESQ}}$ focuses on over-estimations,[177] then, based on the observation that the $\overline{d}_{\text{LSD}}$ and $\overline{Q}_{\text{PESQ}}$ performances in Figures 5.14–5.18 generally coincide as a function of T, we can conclude that the extent to which the duration of included memory mitigates over- and under-estimations in highband envelopes is consistent for the two types of disturbances across T. In other words, at each particular duration, T, memory inclusion mitigates over- and under-estimations by the same relative extent,

[177] See Sections 3.4.1 and B.1 for details of the $\overline{d}_{\text{LSD}}$ and $\overline{Q}_{\text{PESQ}}$ measures, respectively.

with the duration of included memory having no effect in terms of favouring the alleviation of one type over the other. Coinciding with our previous finding to the same effect in Section 5.3.5.2, this observation confirms the generality of this memory inclusion result. A result in contrast to that of our frontend-based approach, however, is the lower $\overline{Q}_{\text{PESQ}}$ improvement, relative to that of $\overline{d}_{\text{LSD}}$, as shown in Table 5.6 for performances using temporally-extended GMMs. This indicates that our model-based technique is less successful in mitigating over-estimation disturbances in comparison to under-estimations. Nevertheless, as noted above, our model-based approach still outperforms the frontend-based one in terms of overall $\overline{Q}_{\text{PESQ}}$ performance by ≈ 4.4 times.

– In a similar manner, since the symmetrized $\overline{d}^{*}_{\text{IS}}$ and $\overline{d}^{*}_{\text{I}}$ measures weight larger deviations in log spectra more heavily than does the $\overline{d}_{\text{LSD}}$ measure,[178], then, based on the observation that the gain-independent $\overline{d}^{*}_{\text{I}}$ performances generally coincide with those of $\overline{d}_{\text{LSD}}$ in Figures 5.14–5.18 as a function of T, we can conclude that our model-based memory inclusion technique mitigates all degrees of deviations in spectral envelope shapes in a consistent manner across T. In other words, at each particular duration, T, memory inclusion mitigates all deviations by the same relative extent, with the duration of included memory again having no effect in terms of favouring the alleviation of one type over the other. Coinciding with our previous finding to the same effect in Section 5.3.5.2, this observation confirms the generality of this memory inclusion result as well. In an argument similar to that made above for $\overline{Q}_{\text{PESQ}}$, we also note, however, the lower $\overline{d}^{*}_{\text{I}}$ improvement relative to that of $\overline{d}_{\text{LSD}}$, as shown in Table 5.6 for performances using temporally-extended GMMs. This indicates that our model-based technique contrasts with our frontend-based one in that it mitigates the more perceptually-relevant large deviations in highband envelope shape reconstruction less successfully than it does small deviations. In spite of this result, our model-based approach is nevertheless shown to outperform the frontend-based one in terms of overall $\overline{d}^{*}_{\text{I}}$ performance by ≈ 2.3 times.

– For both frontend- and model-based approaches, Figures 5.14–5.18 and Table 5.6 show that improvements in the gain-dependent $\overline{d}^{*}_{\text{IS}}$ performance are relatively

[178] See Section 3.4.2 for details of the $\overline{d}^{*}_{\text{IS}}$ and $\overline{d}^{*}_{\text{I}}$ measures.

higher than those in the similarly-derived but gain-independent $\overline{d}_{\text{I}}^{\star}$ performance, with the discrepancy in performance improvements being higher for our model-based technique. As such, we conclude that our approaches to memory inclusion are generally more successful in translating gain-specific cross-band correlation into measurable BWE performance improvements than they are with cross-band correlations of spectral envelope shapes. More specifically, the $\overline{d}_{\text{IS}}^{\star}$ and $\overline{d}_{\text{I}}^{\star}$ results for $\boldsymbol{\mathcal{G}}_{\lozenge}\big|_{K=4}$ at $T = 160\,\text{ms}$ in Table 5.6 suggest that improvements in the reconstruction of envelope shapes and gains represent $\sim 16\%$ and 84%, respectively, of the overall improvement achieved in the reconstruction of highband envelopes as a result of model-based memory inclusion. For inclusion through the optimized frontend-based $\overset{\triangle}{\boldsymbol{\mathcal{G}}}$ tuples at $T = 160\,\text{ms}$, the improvements in envelope shape and gain reconstruction represent $\sim 25\%$ and 75%, respectively. This observation emphasizes the importance of accurately capturing the cross-band correlations specific to envelope energies, which, in turn, justifies the modelling of energies through: a dedicated GMM, as in our dual-mode BWE system based on that of [55]; through a subband HMM, as in the HMM-based system of [84]; or, through more elaborate schemes, as in the technique of [57] incorporating an asymmetric cost function into the GMM-based MMSE estimation of highband energies.

– To conclude this global performance analysis, we note that the $\overline{d}_{\text{IS}}^{\star}$ performances shown in Figures 5.14–5.18 indicate that our model-based approach further succeeds in alleviating the steep decline suffered with frontend-based tuples for $L_{\delta} > 8$—corresponding to $T > 160\,\text{ms}$. This confirms the superiority of joint-band MMSE estimation using temporally-extended $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}G}\}$ GMMs, rather than the delta coefficient-based $\{\mathcal{G}_{\overset{\triangle}{\mathbf{X}}\overset{\triangle}{G}}(\overset{\triangle}{\mathbf{x}}, g)\}$, in terms of preventing the potentially-detrimental large deviations in highband envelope gain reconstruction.

ii. Individual performance: Effects of pruning

- Figure 5.14 illustrates the effects of the parameters underlying the pre- and post-EM pruning operations on BWE performance. As described in Operation (d) of our tree-like growth algorithm, the purpose of these pruning steps is to reduce model complexity and minimize the risk of overfitting in a manner that maximizes information content in the remaining child *pdf*s generated at each temporal extension step. Indeed, Figure 5.14 demonstrates the direct correlation achieved between the child

distribution flatness threshold, $\rho_{\min}$, and the overall complexity of the resulting GMM tuples, as represented by $N_{\text{FLOPs/f}}$; more restrictive values for $\rho_{\min}$ directly result in lower $N_{\text{FLOPs/f}}$ complexity, and vice versa.

- At the same time, Figure 5.14 also demonstrates the role of the post-EM pruning applied via $N_{\min}$ in ensuring the sufficiency of data points to reliably estimate the *pdf*s of the child states obtained by splitting at each temporal extension step. More specifically, Figure 5.14 shows that the reduction obtained in terms of computational complexity is achieved with minimal overfitting; even with the considerable pre-EM pruning imposed via $\rho_{\min} = 0.9$,[179] the $N_{\min}$ threshold precludes overfitting to the extent that the BWE performance of $\{\mathcal{G}_{\triangledown}\}_{\rho_{\min}=0.9}$ still outperforms that of the memoryless baseline as well as that of the optimized frontend-based tuples.

- Moreover, the observation that performances vary only marginally within the wide $\rho_{\min} = 0.2$–0.7 range indicates that our distribution entropy-based pruning approach does indeed succeed in reducing complexity while preserving most of the information content captured in the tuple with least pruning, $\{\mathcal{G}_{\square}\}_{\rho_{\min}=0.2}$.

- Finally, the lack of rapid decay in performance after reaching saturation for the majority of the temporally-extended tuples considered in Figures 5.14–5.18 indicates the success of our pre- and post-EM pruning steps in constraining the $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}\}$ GMM modality increases associated with progressively-higher memory inclusion indices beyond what is justified by the information content and cardinalities of the time-frequency-localized data subsets.

iii. Individual performance: Effects of the splitting factor

- As introduced in Operation (c), the splitting factor, J, represents the number of the child subclasses that can potentially be inferred from each parent state at any memory inclusion step, based on the cardinalities and distribution flatness—or lack thereof—of the time-frequency-localized data associated by fuzzy clustering with these parent states. In essence, this factor thus corresponds to an extent by which we quantize the variability of data within each of the time-frequency-localized subspace regions

[179] The computational costs associated with $\{\mathcal{G}_{\triangledown}\}_{\rho_{\min}=0.9}$ for $T = 120$–160 ms are ~ 4.5–6.5 times lower than those of the $\{\mathcal{G}_{\square}\}_{\rho_{\min}=0.2}$ tuples where BWE performance improvements are highest.

represented by parent states. As such, higher values for J should translate into higher resolutions for subspace quantization, and consequently, higher performance improvements, up to a point where J disproportionately exceeds the variability of the per-parent time-frequency-localized data, potentially leading to inferior subspace *pdf* modelling and, in turn, degradation in performance. Figure 5.15 indeed confirms this effect of the splitting factor, showing performance improvements saturating for $J \simeq 4$–6, as demonstrated by the $\overline{d}^{*}_{\text{IS}}$ results, in particular, for $\{\mathcal{G}_{\circ}\}_{J=4}$ and $\{\mathcal{G}_{\triangle}\}_{J=6}$, with the performances for J outside this range noticeably inferior, as demonstrated by the results for $\{\mathcal{G}_{\square}\}_{J=2}$ and $\{\mathcal{G}_{\Diamond}\}_{J=8}$.

iv. Individual performance: Effects of fuzzy clustering

- As first discussed in Section 5.4.2.2, the role of our proposed fuzzy GMM-based clustering approach is to alleviate the adverse effects of the empty-space phenomenon associated with *pdf* modelling in high dimensions. By favouring such a soft-decision approach over the conventional hard-decision Bayesian technique to cluster data into time-frequency-localized subsets, and subsequently combining it with a qualitatively-weighted Expectation-Maximization algorithm, we demonstrated—through the illustrative example of Figure 5.9, as well as through the detailed analysis in Section 5.4.2.4—our success in generating excellent time-frequency *pdf* estimates at increasingly-higher dimensionalities, all the while minimizing the risks of both overfitting and oversmoothing. A further examination of the effects of the fuzziness factor, K, on BWE performances in Figure 5.16 confirms our previous findings as follows.

- As described in Operation (a), the fuzziness factor, K, where $1 \leq K \leq J$, corresponds to a qualitative expansion of the localized child data subsets obtained by clustering based on GMM-derived parent states, with the resulting subset cardinalities—as well as overlap—increasing with higher K values. Since higher subset cardinalities translate to lower post-EM pruning likelihoods per Eq. (5.65) when the child subset cardinality threshold, N_{min}, is fixed, higher values of K will thus result in more complex temporally-extended GMMs with higher modalities—i.e., more Gaussian components—at all orders of memory inclusion. This, in turn, results in higher extension-stage computational costs. Figure 5.16 indeed confirms this correlation between K and the $N_{\text{FLOPs/f}}$ complexity.

- On the other hand, as demonstrated by the illustrative example of Figure 5.9, the increased qualitative subset overlap associated with higher K values results in better modelling of the overlap between the underlying time-frequency classes. This, in turn, results in improved time-frequency-localized *pdf* estimates, and consequently, higher-quality global temporally-extended GMMs. This correlation between K and *pdf* estimate quality is indeed confirmed by the higher BWE performance improvements achieved in Figure 5.16 using tuples with higher values for K.

- Moreover, as concluded in the discussion of the aforementioned illustrative example's results, Figure 5.16 further shows that excellent *pdf* estimates can be achieved via our soft-decision approach at relatively low values for K, i.e., where $1 < K \ll J$. Indeed, although $J = 6$, Figure 5.16 shows performances saturating for $\{\boldsymbol{\mathcal{G}}_{\triangle}\}_{K=3}$ and $\{\boldsymbol{\mathcal{G}}_{\Diamond}\}_{K=4}$, i.e., at $K \simeq 3$–4, with the corresponding performance improvements representing the highest among all those achieved in Figures 5.14–5.18.

- Finally, the performances shown in Figure 5.16 for the $\{\boldsymbol{\mathcal{G}}_{\square}\}_{K=1}$ tuples make the modelling advantages of our fuzzy clustering approach quite evident. In particular, these tuples are trained with a fuzziness factor of $K = 1$ where our soft-decision approach reduces to that of conventional hard-decision Bayesian classification. As such, the training of $\{\boldsymbol{\mathcal{G}}_{\square}\}_{K=1}$ using our tree-like algorithm of Table 5.5 takes no advantage of the aforementioned localized subset expansion intended to account for class overlap in high-dimensional spaces. Consequently, the obtained $\{\boldsymbol{\mathcal{G}}_{\square}\}_{K=1}$ tuples exhibit excessive overfitting as clearly indicated by their BWE performances. Further emphasizing the advantages of our fuzzy clustering approach is the observation that, by introducing the slightest possible fuzziness via $K = 2$, significantly superior performances were obtained in Figure 5.16. To conclude, we note that these findings confirm those previously made to the same effect in the illustrative example of Figure 5.9.

v. Individual performance: Effects of the memory inclusion step

- Figure 5.17 illustrates the BWE performances obtained using temporally-extended GMM tuples as a function of the memory inclusion step, τ. As described in Section 5.4.2.3, τ represents the step—in number of frames—between the $l + 1$ static frames used to construct the sequences comprising the lth-order temporally-extended supervectors, such that $\mathbb{X}_t^{(\tau,l)} = [\mathbf{X}_t^T, \mathbf{X}_{t-\tau}^T, \ldots, \mathbf{X}_{t-l\tau}^T]^T$ for the narrow band, for exam-

ple. Effectively, the step, τ, thus allows us to reduce the well-known redundancies between immediately-neighbouring static frames—or, more accurately, to leapfrog such redundancies—when constructing each of our temporally-extended feature vectors, thereby increasing the information content of our temporally-extended data sets as a whole. In essence, this simple memory inclusion step thus mimics the dimensionality-reducing LDA and KLT transforms—previously discussed in Section 4.4.2—in terms of their attempt to maximize the information content of feature vectors.

- The performances shown in Figure 5.17 do indeed reflect the redundancy-reducing effect of τ described above. In particular, the $\overline{d}_{\text{LSD}}$ and $\overline{d}^{*}_{\text{I}}$ performances indicate that overall performance improvements across the range of T generally increase and become more consistent with larger values of τ where feature vectors comprise increasingly-lower redundancies, and hence, increasingly-higher information content. The least improvements, in terms of both value and consistency, are those obtained for $\{\mathcal{G}_{\square}\}_{\tau=2}$—the tuples with the lowest value for τ among all those considered.

- Secondly, we note that, as τ increases, the differences in performances in Figure 5.17 become increasingly smaller, with the improvements in performance reaching saturation for $\tau \cong 6$. Given our 10 ms frame step, these observations coincide with expectations based on the knowledge discussed in Section 1.2 and Appendix A regarding the durations of sounds and phonetic events. More specifically, as the duration of the frame step approaches the average duration of typical phonetic events, roughly around 70 ms, the cross-frame intra-phonetic redundancies that can potentially be reduced through the leapfrogging step become progressively less, until finally reaching saturation when the step equals ~ 70 ms.

- As previously noted, BWE performances are plotted in Figures 5.14–5.18 against T, the duration of included memory, to allow comparison against frontend-based tuples as well as the comparison of model-based tuple performances obtained at different values for τ, the memory inclusion step. Except for the tuples considered in Figure 5.17, however, all temporally-extended tuples in Figures 5.14–5.18 use a fixed value for τ. Noting that $T = 10 \cdot l \cdot \tau$, Figures 5.14–5.16 and 5.18 thus also illustrate performance as a function of l, the memory inclusion index. As such, we observe that, for $\tau = 4$, the performances achieved by temporally-extended tuples consistently reach

saturation for $l \approxeq 2$–3. This optimal range for l is further confirmed for other values of τ by noting that the performances in Figure 5.17 evolve in a consistent manner when viewed as a function of l, rather than T, with the $\square$, $\bigcirc$, $\triangle$, and $\lozenge$ markers denoting performance data points at increasing values of l. Thus, we conclude that it is the extent of memory as represented by inclusion indices, rather than by absolute inclusion durations, that correlates directly with the ability of our tree-like GMM extension algorithm to successfully exploit memory for improved cross-band correlation modelling, and accordingly, improved BWE performances.

- Given that saturation in performance improvements is achieved at $\tau \approxeq 6$ as noted above, the optimal $l \approxeq 2$–3 range translates to $T \approxeq 120$–180 ms, which coincides with the optimal range for memory inclusion duration previously identified in the context of global performance. The observations made above regarding the effects of τ and l thus provide us with a more detailed understanding of how our tree-like GMM extension algorithm achieves its best performance in terms of memory inclusion.

vi. Individual performance: Effects of the initial 0th-order GMM modality

- Per the state space- and subspace clustering-based interpretations introduced in Section 5.4.2.2 for our tree-like extension algorithm, the I component densities of the initial 0th-order GMMs extended via our tree-like algorithm correspond to the initial states or classes—representing projections of L-order temporally-extended classes in the $\begin{bmatrix} \mathbb{X}^{(\tau,L)} \\ \mathbf{Y} \end{bmatrix}$ space onto the static $\begin{bmatrix} \mathbb{X}^{(\tau,0)} \\ \mathbf{Y} \end{bmatrix}$ subspace—from which all the finer and higher-order time-frequency states, or subclasses, in the $\left\{\begin{bmatrix} \mathbb{X}^{(\tau,l)} \\ \mathbf{Y} \end{bmatrix}\right\}_{l\in\{1,\ldots,L\}}$ subspaces progressively descend. For a fixed splitting factor, J, applied to all such I initial states, this results in a close correlation between the 0th-order modality, I, and the total number, $M^{(l)}$, of child states obtained at low orders of memory inclusion. In other words, higher initial I modalities are more likely to translate into higher modalities for the $(l << L)$th-order temporally-extended $\{\mathcal{G}_{\mathbb{X}^{(\tau,l)}\mathbf{Y}}\}$ GMMs, which, in turn, translate into finer time-frequency localization, and hence, improved temporally-extended joint-band modelling. This correlation between I and $M^{(l)}$ is expected for low orders of memory inclusion up to the point where the variability and/or cardinality of the time-frequency-localized data subsets obtained via fuzzy clustering become sufficiently low such that no further localization is allowed by pre- and post-EM pruning.

- Illustrating the effects of I on the performance of temporally-extended tuples, Figure 5.18 indeed confirms the behaviour described above. In particular, the performance improvements achieved at any particular memory inclusion duration are shown to directly correlate with the initial 0th-order modality. This correlation is observed not only for low orders of memory inclusion, but also for higher orders of l.

- Since the correlation of I with improved performances is observed across all orders of memory inclusion, Figure 5.18 thus indicates that the 0th-order modality, I, also directly affects the ability of our tree-like algorithm in achieving reliable time-frequency localization at higher orders of memory inclusion as well. To elaborate, we note that, despite differences in initial 0th-order modality, all tuples will converge to comparable complexities at higher orders of memory inclusion. This follows as a result of: (a) applying a fixed threshold, N_{min}, for subset cardinalities throughout, and (b) using the same amount of training data to train each of the tuples considered in Figure 5.18. More specifically, although the temporal extension of tuples with lower 0th-order modalities results in tuples with $(0 < l << L)$th-order modalities that are lower compared to those obtained for tuples with higher 0th-order modalities, these differences in modalities continue as a function of T, or l, only until the effects of pruning lead to the convergence of modalities for both sets of tuples. Indeed, Figure 5.18 illustrates this behaviour via the $N_{\text{FLOPs/f}}$ complexity; tuples with lower 0th-order modalities have lower complexities—relative to those with higher initial modalities—for the same values of l for $0 < l << L$, until all sets of tuples eventually converge to comparable complexities as l increases.

  Despite the convergence in complexity, and thereby, in the extent of time-frequency localization as well, the performances of tuples with lower 0th-order modalities do not eventually catch up with those of the tuples with higher initial modalities. Rather, the performances for all tuples saturate at roughly the same order of memory inclusion regardless of initial modalities. This indicates that the information content and quality of the time-frequency-localized *pdf*s estimated at all $(l \in \{1, \ldots, L\})$th orders of memory inclusion correlate strongly with the initial 0th-order modality, to the extent that any two sets of equally-complex tuples—and hence, with similar degrees of time-frequency localization—can vary considerably in terms of quality and the associated BWE performance as a result of performing temporal extension using

0th-order GMMs with different modalities. In other words, the lower quality of static *pdf* estimates associated with coarser 0th-order frequency localization is, in fact, inherited by the descendent temporally-extended GMMs obtained at all $(l > 0)$th orders of memory inclusion, with these lower modelling qualities not being compensated by subsequent increases in time-frequency localization.

5.4.3.3 Comparisons to relevant model-based memory inclusion approaches

Through the detailed analysis presented in Sections 5.4.3.1 and 5.4.3.2 above, we have shown that our model-based approach to memory inclusion clearly outperforms that of Section 5.3 based on incorporating delta features. Although this superior performance is achieved at an increase of ~ 3–4 orders of magnitude in extension-stage computational cost, we have shown that such costs are within the typical computational capabilities of modern smart mobile devices. In addition to thus translating the previously-shown cross-band correlation into tangible BWE performance improvements more successfully, our tree-like temporal extension technique also outperforms the delta coefficient-based approach in that involves no algorithmic delay, all the while preserving the latter's advantage in terms of the ability to incorporate varying extents of long-term memory—up to and exceeding syllabic durations—into the joint-band model. Having thus compared our two techniques in detail, we now compare our tree-like memory inclusion approach to relevant works in the literature, focusing specifically on the model-based techniques reviewed in Section 5.4.1.

As discussed in Section 5.4.1, the use of GMMs as the primary means to statistically model joint-band correlations for the purpose of BWE has been restricted to memoryless implementations due to the dimensionality-related limitations detailed in Section 5.4.2.1. Similarly, the use of neural networks in BWE has also been restricted to memoryless implementations, and even then achieving only mixed and inconclusive performances. As such, among the five general model-based approaches discussed in Section 5.4.1, only those incorporating memory through codebook mapping, HMMs, and non-HMM state space techniques, can in practice be compared against our model-based memory inclusion technique. As in Section 5.3.5.3, we simplify this comparison by assuming that the test sets used by the cited techniques are sufficiently diverse such that the results reported therein can be considered general enough for direct comparison against each other, as well as against our results in Table 5.6. In other words, we preclude any effects that the test set differences—

relative to the TIMIT core test set described in Section 3.2.10—may have on the generality, and hence the comparability, of performances.

In the context of codebook mapping, we noted in Section 5.4.1.4 that the works of [130] and [131] represent the exceptions to the generally memoryless implementations of codebook-based BWE. Having then described both of these techniques in detail, we noted that the three-step quantization technique of [130] is quite limited in its use of memory in that it only incorporates information from immediately preceding frames into codevector interpolation.[180] More importantly for the purpose of comparison at hand, however, is that only informal subjective results are reported in [130]. In contrast, the 256-codeword predictive VQ[181]-based BWE approach of [131] is reported to achieve a highband $\overline{d}_{\text{LSD}}$ improvement of 0.45 dB relative to conventional memoryless VQ of equal codebook size, while also incorporating memory at the limited interframe level as in [130]. To put these results into the same frame of reference as the $\overline{d}_{\text{LSD}}$ improvements of Table 5.6 achieved by our model-based state space approach to memory inclusion, we note that:

1. The 0.62 dB $\overline{d}_{\text{LSD}}$ improvement reported for $\boldsymbol{\mathcal{G}}_{\lozenge}\big|_{K=4}$ in Table 5.6 is calculated relative to the performance of $\boldsymbol{\mathcal{G}}^{(0)}$, our MFCC-based memoryless baseline implementation of the dual-mode BWE system, detailed in Section 5.2.

2. As shown in Section 5.2.6 by comparing the results of Table 5.1 to those of 3.1, our MFCC-based implementation of the memoryless dual-mode BWE system achieves a BWE performance that is nearly similar—lower by a $\overline{d}_{\text{LSD}}$ difference of 0.06 dB, to be exact—to that obtained using the LSF-based system detailed in Section 3.2, which, in turn, is itself based on the reference system of [55].

3. The dual-model system of [55] is, in fact, an improvement over the earlier system of [54] employing GMM-based statistical modelling only, with no midband equalization.[182] This latter system itself achieves a $\overline{d}_{\text{LSD}}$ improvement of 0.96 dB over the split VQ-based technique of [69] which uses three separate 32-word codebooks to map voiced, unvoiced, and mixed narrowband sounds into their highband counterparts.

[180] See Section 2.3.3.2 for details on codebook mapping with interpolation, or fuzzy VQ.
[181] See Footnote 23.
[182] As discussed in Sections 2.3.2.4, 3.2.3, and 3.2.4, the dual-mode system of [55] improves upon the GMM-only system of [54] by using midband equalization to extend the narrowband signal into the 3.4–4 kHz range, which in turn allows the use of the signal across the 3–4 kHz range—rather than in the 2–3 kHz range—to generate the 4–8 kHz highband excitation signal by full-wave rectification.

4. The voicing-based three-way split codebook mapping technique of [69], using a total of 96 codewords, is quite similar to the two-way split codebook approach of [63]. With a total of 128 voicing-based codewords, this latter technique is shown in [63] to marginally outperform similarly-size conventional codebook-based mapping by a $\overline{d}_{\text{LSD}}$ difference of 0.07 dB.

Thus, notwithstanding the relatively minor effects of differences in reference codebook sizes or in the frequency ranges of the highband content reconstructed by BWE,[183] aggregating the $\overline{d}_{\text{LSD}}$ improvements listed above for our dual-model BWE system with model-based memory inclusion results in an overall improvement of ≈ 1.59 dB, corresponding to ~ 3.5 times that achieved by the predictive VQ-based approach of [131], relative to conventional memoryless codebook-based baselines. This demonstrates the clear superiority of our model-based approach to memory inclusion over that of [131]. For illustration, the differences among the performances of the BWE techniques listed above, as well as those to be further discussed below, are shown in Figure 5.19. With the performance differences plotted to scale, Figure 5.19 thus puts the BWE performances cited throughout this section into an informative relative perspective.

Focusing next on the more successful HMM-based memory inclusion techniques for BWE, we noted in Section 5.4.1.3 that, except for the more recent work of [163] and the early approach of [84], all HMM-based approaches proposed in the literature share the same idea underlying the work in [39] and [87]. In our detailed description of these four first-order HMM-based techniques in Sections 2.3.3.4 and 5.4.1.3, we also noted, however, that no comparisons are reported in either [39] or [84] for the performances of their techniques relative to those of other BWE techniques. In contrast, the HMM-based techniques of [87] and [163] are compared, respectively, to the piecewise-linear and codebook-based mapping techniques, described in Sections 2.3.3.1 and 2.3.3.2, respectively, in terms of $\overline{d}_{\text{LSD}}$ and $\overline{Q}_{\text{PESQ}}$ performances. Before comparing the performances achieved by these two HMM-based techniques to those in Table 5.6 for our temporally-extended GMM-based approach, however, we further note that, in addition to incorporating memory through first-order HMMs, the techniques of [87] and [163] also use delta and delta-delta features as a secondary

[183] The 0.45 dB $\overline{d}_{\text{LSD}}$ improvement reported in [131] is calculated over the 4–7 kHz range, while the performances reported for the GMM-only BWE system in [54] are calculated for the 3.5–7 kHz range. As discussed in Section 3.4.1, however, the $\overline{d}_{\text{LSD}}$ performances calculated throughout our work presented herein are estimated over the wider 4–8 kHz range. This latter range is also used in [63] to compare the $\overline{d}_{\text{LSD}}$ performances of several linear- and codebook-based mapping techniques.

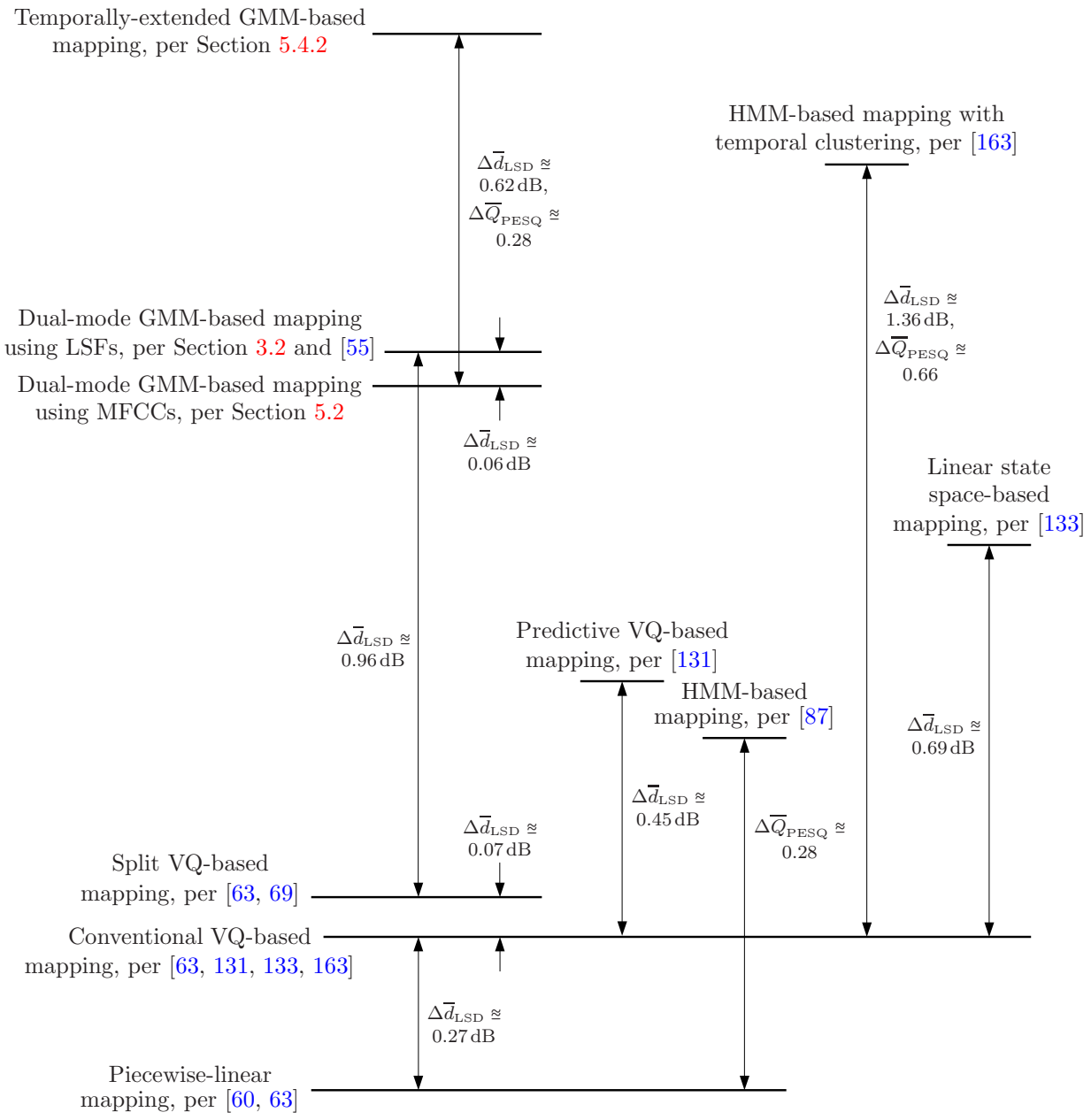


Fig. 5.19: Illustrating differences among the performances of the relevant model-based BWE techniques cited throughout this section. Although we discard minor differences between the performances of those techniques using the same model-based approach (several works use conventional VQ-based mapping as a performance baseline, for example), all distances are plotted to scale based on the results reported in the cited techniques, with higher levels corresponding to better BWE performances. As suggested by the $\overline{d}_{\text{LSD}}$ and $\overline{Q}_{\text{PESQ}}$ results of Table 5.6 and [163], a linear relation between the two measures can be assumed for the purpose of this illustration, with the relationship's parameters estimated based on those results.

means for the inclusion of memory. Incorporating longer-term memory into joint-band modelling as such, the BWE techniques of [87] and [163] thus also contrast with other HMM-based techniques by attempting to mitigate the 20–40 ms limitations imposed on the extent of memory that can be incorporated by the use of first-order HMMs.

As previously discussed in Sections 2.3.3.4, 5.3.5.3, and 5.4.1.3, the 64-state HMM-based BWE technique of [87] achieves only a $\overline{Q}_{\text{PESQ}}$ improvement of 0.28 relative to the quite-unsophisticated 4-partition piecewise-linear mapping technique of [60]. While the comparative performance illustration in Figure 5.19 shows that this oft-cited HMM-based approach—of both [87] and [39]—thus outperforms those techniques based on conventional and split codebook mapping, it also shows it to be quite inferior to almost all the other advanced model-based approaches considered in this section.[184] Our proposed temporally-extended GMM-based BWE technique, in particular, is shown to outperform this HMM-based approach by ~ 1.24 dB, corresponding to ~ 2 times the improvement achieved relative to the baseline based on piecewise-linear mapping.

In the supervised HMM-based approach of [39] and [87] evaluated above, HMM chains statistically model the spectra of narrowband-only features as well as their first-order dynamics, with the cross-band correlations with highband envelopes modelled through a tied codebook. In contrast, the more recent technique of [163] trains joint-band HMMs in an unsupervised manner, effectively clustering—or segmenting—joint-band data into separate neighbourhoods, each of which comprising a set of joint-band data with high spectral and first-order temporal correlations. Using sequences of such temporally-clustered data, wideband spectral envelopes are then estimated in the extension stage by linear prediction, rather than by codebook mapping as in [39] and [87]. Having already discussed this BWE technique in Section 5.4.1.3 with detail, we repeat its underlying idea here in order to note the similarities with our proposed temporally-extended GMM-based technique in terms of the time-frequency localization used to improve the modelling of joint-band dynamics. Despite these similarities, Figure 5.19 shows that our model-based approach also outperforms this first-order HMM-based technique, albeit by a much smaller margin—$\Delta\overline{d}_{\text{LSD}} \simeq 0.23$ dB—

[184] Given the approximations assumed in illustrating Figure 5.19, namely: (a) the linear relationship between $\overline{d}_{\text{LSD}}$ and $\overline{Q}_{\text{PESQ}}$, (b) equating performances for different techniques based on the same model-based approach, (c) discarding the effects of differences among the various techniques in terms of training and testing conditions as well as in terms of the ranges of the highband frequency content reconstructed by BWE, the ~ 0.1 dB difference between the $\overline{d}_{\text{LSD}}$ performances of the HMM-based technique of [87] and the predictive VQ-based one of [131] is too close to clearly favour one technique over the other.

than those obtained relative to the other comparable techniques. More specifically, a $\overline{d}_{\text{LSD}}$ improvement of $\approx 1.36\,\text{dB}$ is reported in [163] for the proposed HMM-based technique relative to conventional 128-codeword VQ. In comparison, our temporally-extended $\mathcal{G}_{\lozenge}\big|_{K=4}$ tuples of Table 5.6 achieve a cumulative $\overline{d}_{\text{LSD}}$ improvement of $\approx 1.59\,\text{dB}$, relative to a similar VQ-based baseline. In addition to thus achieving a superior performance, our proposed technique also contrasts with that of [163] in that in involves no algorithmic delay—as incurred by delta features and the Viterbi algorithm employed in [163] for HMM state sequence decoding during extension.

Reiterating our arguments from previous sections, we attribute this lower success of first-order HMM-based techniques in general, relative to our temporally-extended GMM-based approach, to the 20–40 ms limitation imposed on the extent of cross-band information that can be incorporated by first-order-only HMMs. As shown in Section 4.4.3, such short-term information represents only a minor portion of the maximum mutual information achievable at syllabic durations. We also note that, for the approach of [87] in particular, the delta and delta-delta features are incorporated only for the narrow band. As such, dynamic cross-band correlations are captured in the manner modelled by Scenario 1 in Section 4.4.3, for which we showed the information gains achievable to be rather minimal in comparison to those achievable by the inclusion of delta features in the parameterizations of both frequency bands, as modelled by Scenario 2. In contrast, the technique of [163] incorporates delta and delta-delta features per Scenario 2, which partially contributes to the superior performance achieved by this technique relative to that of [87].

Finally, Figure 5.19 compares the performance achieved by the dynamic linear state space approach of [133] to that of our model-based technique, as well as to those of the techniques discussed above. As described in Section 5.4.1.5, this approach achieves its best BWE performance with $\approx 300\,\text{ms}$ of memory inclusion. Except for the HMM-based technique of [163] using temporal clustering, this state space technique outperforms all HMM- and codebook-based BWE techniques reviewed above, albeit it at a higher computational cost. It does possess an advantage over the technique of [163], however, in that it involves no algorithmic delay. In addition to the similarity it thus has with our model-based technique in terms of real-time processing capability, this approach shares the state space concept underlying the time-frequency localization central to our technique—illustrated in Figure 5.8. On the other hand, this approach of [133] models temporal and spectral joint-band correlations through a dynamic linear model, rather than through GMMs. Despite

using a large number of linear *modes*, the linear assumption expectedly limits the ability to model complex joint-band correlations. As such, our temporally-extended GMM-based techniques is found to outperform this linear state space-based approach by a considerable $\Delta\overline{d}_{\text{LSD}} \simeq 0.9\,\text{dB}$, corresponding to a performance improvement difference of ~ 2.3 times relative to conventional VQ-based performance.

5.5 Summary

Since an extended summary of our work in this chapter is presented next in Section 6.1.5, we conclude here by providing a rather brief summary.

Building upon our information-theoretic findings of Chapter 4, this chapter detailed our proposed approaches for improving BWE performance via memory inclusion. First, an MFCC-implementation of our baseline LSF-based dual-mode memoryless BWE system was proposed. Despite the non-invertible partial loss of information associated with MFCC parameterization, we showed that high-quality highband speech can indeed be reconstructed from GMM-based MMSE estimates of highband MFCCs. We then presented two novel memory-inclusive BWE techniques. The first mimics the methodology used in Chapter 4 by using delta features to extend the acoustic classes underlying the GMM-based joint-band models along long-term temporal axes. Although this frontend-based approach to memory inclusion succeeds in achieving only modest BWE performance improvements, it requires minimal modifications to the baseline memoryless system, involves no increases in extension-stage computational cost nor in training data requirements, and hence, provides an easy and convenient means for exploiting memory to improve BWE performance. In our second approach, we focus instead on modelling the high-dimensional distributions underlying sequences of joint-band feature vectors. In particular, we made use of the correspondence of GMM component densities to underlying classes and the strong correlation between neighbouring frames in order to devise a GMM training algorithm that effectively breaks down the infeasible task of modelling high-dimensional *pdf*s into a series of progressive tree-like time-frequency-localized estimation operations. Incorporating the temporally-extended GMMs obtained as such into our dual-mode BWE system results in substantial performance improvements that exceed not only those of our first delta coefficient-based approach, but also other comparable model-based memory-inclusive BWE techniques, notably those based on HMMs.

Chapter 6

Conclusion

For the purposes of a quick review, we conclude this thesis by first presenting an extended summary of all content presented throughout the thesis, with particular emphasis on our findings and contributions. Representing future work, we then discuss the potential avenues for improving the techniques and approaches presented herein, followed by a brief discussion of their applicability to BWE in general, as well as to contexts other than that of BWE.

6.1 Extended Summary

6.1.1 Motivation

This thesis presents our work on improving the artificial *bandwidth extension* (BWE) of narrowband speech—the bandlimited speech of traditional telephony. To introduce BWE and show its relevance, we started in Chapter 1 by providing a historical background of traditional telephony. We noted, in particular, that, while traditional telephony has undergone many advances since its inception in 1876, the bandwidth of telephony speech has always been rather limited relative to the full spectrum of speech. This followed as a result of technological limitations as well as the necessity to balance quality and intelligibility with economic viability. During the early twentieth century, for example, telephony speech was bandlimited to as low as 2.5 kHz [5]. Subsequently standardized in the 1960s, the bandwidth of telephony speech has been limited ever since to the 0.3–3.4 kHz *narrowband* range [8, 9]. As illustrated in Figure 5.2, however, the frequency spectra of speech can extend to over 20 kHz. More importantly, many of the distinctive acoustic features of several classes of sounds—mainly, fricatives, stops, and affricates—were shown

in Section 1.1.3.1 to lie outside the narrowband range. As such, narrowband speech exhibits not only a *toll* quality that is noticeably inferior to its wideband counterpart extending up to 7–8 kHz, but also reduced intelligibility especially for consonant sounds.

As an alternative to the costly-prohibitive complete wideband digitization of the now-ubiquitous traditional telephone network, wideband speech reconstruction through BWE attempts to regenerate the *highband* frequency content above 3.4 kHz—and, optionally, the *lowband* content below 300 Hz—lost during the filtering processes employed in traditional networks. Applied in the receiving end, BWE thus provides backward compatibility with existing networks. Based on the assumption that the missing highband spectral content to be reconstructed is sufficiently correlated with that of the available narrowband input, BWE has been the subject of considerable research where the objective is to learn as much of such *cross-band correlation* as possible in a training stage. The work in [109, 124, 125] presented evidence, however, that this cross-band correlation is, in fact, rather low when the joint-band information being modelled is limited to only that of conventionally-parameterized speech signals—i.e., when only the information from quasi-stationary 10–30 ms segments of narrowband and highband speech is considered. Despite this low cross-band dependence, the majority of BWE techniques proposed in the literature have relied—and continue to rely—on *memoryless mapping* between the spectra of both bands, thereby making no use of the significant information carried by the dynamic spectral and temporal events in long-term speech segments. Quantifying and exploiting such information—referred to herein as speech *memory*—for the purpose of improving BWE performance represents the focus of our work presented here. To illustrate their scope and importance for perception, the spectral and temporal characteristics comprising speech memory were discussed in Section 1.2 and are further detailed in Appendix A. Among the observations made in these discussions, the most notable is that phoneme perception is likely accomplished by analyzing dynamic acoustic patterns over segments corresponding roughly to syllables, and hence, to improve the perceived quality of extended speech, BWE systems should in turn exploit long-term information extending up to syllabic durations.

6.1.2 Reviewing BWE techniques and principles

To allow for a detailed and comprehensive description of the joint-band modelling approaches used in Chapters 3–5 as well as to put our BWE techniques proposed therein into

perspective, we followed up on the introduction of Chapter 1 above by presenting a broad review of previous BWE work and underlying principles in Chapter 2. In particular, the early non-model-based approaches to BWE were first discussed in brief, focusing thereafter on the prevalent state-of-the-art *model-based* approaches. Using primarily the source-filter speech model, these latter approaches reduce the BWE problem of highband speech reconstruction to two separate tasks—generating a highband excitation signal and a highband spectral envelope [49, 50]. These two elements of the highband signal can then be combined in a linear prediction (LP) synthesis filter to reconstruct highband speech, which, in turn, is added to the narrowband input in order to generate wideband speech.

Since it is the quality of reconstructed highband envelopes, rather than that of the excitation signal, that was shown—in, e.g., [39, 58, 88]—to be far more important for the subjective quality of extended speech, we devoted the greater part of our review in Chapter 2 to those approaches concerned with modelling the cross-band correlations of spectral envelopes. Ranging in complexity from simple linear mapping to the advanced statistical modelling approaches based on *Gaussian mixture models* (GMMs) where highband speech is reconstructed by minimum mean-square error (MMSE) estimation given the narrowband input, the surveyed techniques were shown to vary greatly in their ability to model the complex and non-linear cross-band correlations. With hidden Markov models (HMMs) being the basis of memory-inclusive exceptions to the mostly-memoryless approach to BWE in the literature, the advantages and drawbacks of HMM-based BWE techniques were also discussed. We noted, in particular, that, while HMMs exploit interframe dependencies for joint-band modelling, their use of speech memory is rather limited to the short-term 20–40 ms durations. This follows as a result of typically using first-order-only HMMs to mitigate the higher complexity and data requirements associated with more general HMMs.

By using an illustrative example in Section 2.3.3.5, we then showed that GMMs, in particular, represent the tool most suited to our purpose—investigating the role of speech memory in improving BWE performance through apt cross-band correlation modelling. More specifically, not only do GMM-based BWE techniques outperform those based on the common codebook mapping approach at comparable or slightly higher complexity, but they also contrast with other techniques in that GMMs—as multi-modal density representations—have an intuitive correspondence with the acoustic classes underlying the joint-band feature vector distributions being modelled. Since these classes are shared to varying extents by the representations of both frequency bands, joint-band GMMs inherently learn their cross-

band statistical properties, thereby improving the ability of MMSE estimation to generate perceptually-relevant highband spectral envelopes. Indeed, it is this very correspondence that inspires the two approaches we propose in Chapter 5 for the inclusion of memory into the joint-band modelling paradigm.

6.1.3 Dual-mode BWE and the GMM framework

Having discussed the principles underlying BWE in Chapter 2, we then continued our presentation in Chapter 3 by describing the details of our GMM-based BWE technique. Based on the system proposed in [55], our BWE implementation—illustrated in Figure 3.1—is a *dual-mode* technique that exploits equalization in addition to GMM-based statistical modelling. Equalization is used to extend the bandwidth of narrowband speech up to approximately 4 kHz at the high end. The 0.3–4 kHz midband-equalized narrowband signal is then used for the GMM-based MMSE estimation of the complementary highband spectrum in the 4–8 kHz range, with the equalized signal in the 3–4 kHz range further processed to generate an enhanced excitation signal.

Briefly discussed in [55], the motivation for parameterizing spectral envelopes in the dual-mode BWE system using line spectral frequencies (LSFs) was also discussed in detail in Section 3.2.2. We showed, in particular, that LSFs guarantee LP synthesis filter stability, improve the robustness of BWE to estimation errors, and improve the ability of GMMs to capture perceptually-significant events in spectral envelopes.

Given the central role of GMMs in our work presented in Chapters 4 and 5 on the inclusion of speech memory, as well as in BWE in general, the GMM framework was studied further in more detail in Section 3.3. First, the derivation of the MMSE estimation of target features given those of the source and using joint-density GMMs was presented. Then, by using the obtained formulae to derive the exact per-frame *extension-stage* computational and memory costs of performing MMSE estimation using full- as well as diagonal-covariance GMMs, we showed that full-covariance GMM are, in fact, more computationally efficient than those with diagonal covariances for the purpose of achieving similar BWE performances. By investigating the finer cross- and auto-covariance matrix properties of the joint-band GMM components, we further illustrated a tight correlation between source-target conversion performance and full-covariance GMMs. Representing one of the contributions in this thesis, this analysis and subsequent conclusion challenge the

assumption commonly stated and used in the source-target conversion literature in general, e.g., in [40], whereby the performance obtained using a GMM with a particular number of full-covariance Gaussians can be obtained by a corresponding GMM with a larger set of diagonal-covariance Gaussians in a manner that nevertheless preserves, or even reduces, overall computational or memory costs, or both. Indeed, this very assumption has led to the predominant use of diagonal-covariance GMMs in GMM-based BWE research and implementation, despite the fact that, with the continuous advances in offline processing capabilities, the computational cost of the offline maximum likelihood (ML) GMM training stage has become rather less important than the cost of online real-time MMSE estimation. Based on this analysis, we thenceforth focused only on the use of full-covariance GMMs in the remainder of our work.

After describing the measures used to evaluate the BWE performances discussed throughout our work, the memoryless LSF-based dual-mode BWE performance baseline was then presented. Unlike previous BWE works where only a single performance measure is typically used, we chose an ensemble of objective measures that collectively ensure our reported results are: (a) comparable to those of previous works (via log-spectral distortion, or $\overline{d}_{\text{LSD}}$), (b) strongly correlated with subjective measures (via the perceptual evaluation measure of speech quality, or $\overline{Q}_{\text{PESQ}}$), and (c) sufficiently detailed to allow the individual evaluation of gain-related and spectral shape-related BWE performance improvements (via the symmetrized Itakura-Saito and Itakura distortion measures, or $\overline{d}^{*}_{\text{IS}}$ and $\overline{d}^{*}_{\text{I}}$, respectively).

6.1.4 Modelling speech memory and quantifying its role in improving cross-band correlation

Although the few memory-inclusive BWE techniques proposed in the literature report performances that are superior to those of the conventional memoryless approach, none of these works has explicitly quantified the cross-band correlation gains associated with the use of speech memory. In fact, as noted in Section 6.1.1 above, only a handful of works have even attempted to verify and quantify the cross-band correlation assumption itself. As such, Chapter 4 was devoted to modelling speech memory and quantifying its effect to determine the value and potential of the inclusion of such memory in terms of improving BWE performance.

Building on the work of [109] where the *certainty* about the high band given the narrow

band was quantified as the ratio of the mutual information (MI) between the two bands to the discrete entropy of the high band, we estimated and compared highband certainties in both the memoryless and memory-inclusive conditions. With the MI estimated numerically as in [109] using stochastic integration of test features vectors over the marginal and joint narrowband and highband *pdf*s modelled by GMMs, our contributions in terms of highband certainty estimation are four-fold:

(a) First, we estimate discrete highband entropies through resolution-constrained vector quantization (VQ) in steps of increasing resolution such that the spectral distortion associated with our entropy estimates is guaranteed to fall below the 1 dB $\overline{d}_{\text{LSD}}$ spectral transparency threshold proposed in [115]. By using VQ as such rather than first estimating differential highband entropy followed by entropy-constrained scalar quantization (SQ) as proposed in [109], we make use of the space filling, shape, and memory advantages of VQ to obtain superior estimates for the discrete highband entropy, which, in turn, lead to more accurate highband certainty estimates.

(b) Secondly, unlike the SQ-based approach of [109], our proposed VQ-based technique does not require any correspondence between the quantization mean-square error and $\overline{d}_{\text{LSD}}$ spectral error. This allows the estimation of highband certainties for any form of spectral envelope parameterization as long as $\overline{d}_{\text{LSD}}$ can be calculated from the quantized feature vectors.

(c) Thirdly and most importantly, we quantify the cross-band correlation gains attainable by memory inclusion by explicitly incorporating speech memory into the feature vector spaces used for highband certainty estimation. The ability to estimate highband certainty with memory incorporated in the parameterization frontend as such follows as a result of the ability provided by our proposed technique to estimate the spectral error associated with quantization over any arbitrary subspace of the entire vector-quantized highband feature vector space.

(d) Finally, our last contribution, detailed in Sections 4.3.5 and 4.4.3.2, is the adaptation of the $\overline{d}_{\text{LSD(RMS)}}$ lower bound proposed in [125] to our context of quantifying the role of memory inclusion. Derived as a function of the aforementioned information-theoretic measures, this bound effectively translates highband certainty estimates into an upper bound on achievable BWE performance.

Taking advantage of the parameterization-independence property of our VQ-based technique discussed above, we compared two different parameterizations in terms of their ability

to retain the mutual cross-band information relevant to BWE. In addition to the LSFs used by our dual-mode BWE system, we also chose *mel-frequency cepstral coefficients* (MFCCs) for our information-theoretic investigation specifically for their superior MI and class separability properties relative to several common speech parameterizations. These properties, demonstrated in [126, 135], suggest the superior aptitude of MFCCs for the particular task of capturing the cross-band correlation information crucial for BWE.

To incorporate memory into our information-theoretic investigation of cross-band correlation, we used *delta features* as a means by which to explicitly parameterize long-term speech dynamics in each of the two frequency bands. Detailed in Sections 4.4.1 and 4.4.2, delta features can be calculated for any form of conventional static parametrization, and, more importantly, allow us to to model long-term information in speech segments extending up to 600 ms.

Having detailed our methodology for highband certainty estimation as well as our frontend-based approach to modelling speech memory as summarized above, we then proceeded to quantify the role of speech memory in Section 4.4.3 by estimating highband certainties in the multiple scenarios and contexts in which the *dynamic* (static+delta) representation of one or both frequency bands can be applied. This investigation led us to several conclusions which can be itemized as follows:

(a) Incorporating the long-term speech dynamics of only one of the two frequency bands into the joint-band model achieves marginal cross-band correlation gains. We showed, in particular, that narrowband spectral dynamics provide minimal information about the properties of static highband spectra; appending delta features to the static narrowband parameterization—without any truncation in the dimensionalities of the static or delta features—resulted in, roughly, a mere 2% relative increase in highband certainty when using MFCCs, 5% when using LSFs.

(b) In contrast, the inclusion of memory via delta features into the parameterizations of both frequency bands was shown to result in considerable cross-band correlation gains, and hence, considerably higher certainty about the dynamic representation of highband spectra. More specifically, the addition of delta features to the static narrowband and highband MFCC-based parameterizations resulted, roughly, in a relative increase of 99% in terms of dynamic highband certainty, 115% for LSF-based parameterizations. Under the constraint of fixed dimensionality where the reference per-band dimensionalities are preserved by substituting—rather than appending—delta

features in lieu of high-order static features, the relative gains in dynamic highband certainty are reduced to ~ 78% and ~ 10% for MFCCs and LSFs, respectively.

(c) Incorporating memory into the modelling frontend via delta features involves a *time-frequency information tradeoff*. Resulting from the non-invertibility of delta features, this tradeoff was demonstrated by comparing the highband certainties achieved when delta features are appended to their static counterparts relative to those certainties achieved in the substitution scenario. As noted in Item (b) above, the net effect of such tradeoff on highband certainty is a maximum relative increase of, roughly, 78% rather than 99% for MFCCs, or only ~ 10% rather than ~ 115% for LSFs. These figures were summarized in Table 4.4.

(d) The information-theoretic gains achieved by memory inclusion reach saturation at long-term durations of ~ 200 ms. Corresponding to the syllabic 4–5 Hz rate, our results were thus found to coincide with earlier findings regarding the acoustic-only information content in the long-term speech signal.

(e) MFCCs were found to consistently outperform LSFs in capturing the cross-band correlation information central to BWE. The considerable difference in performance is reflected in the highband certainties measured in both the memoryless and memory-inclusive conditions summarized in Tables 4.2 and 4.4, respectively. With the MFCC-based highband certainties reaching double those based on LSFs in many cases, we note in particular the certainties measured in the memory-inclusive scenario where the delta features of low-order static features replace an equal number of high-order parameters in the reference memoryless static feature vectors—36.5% for MFCCs compared to 17.5% for LSFs. These performance differences were attributed to the improved class separability associated with MFCCs, as well as the lower spectral error associated with vector-quantizing truncated MFCC feature vectors. By being less susceptible as such to the adverse effects of the time-frequency information tradeoff, MFCC-based implementations of BWE were concluded to be potentially superior to those based on LSFs, particularly under constraints of fixed dimensionality.

Finally, we note that the practical significance of these information-theoretic gains was further demonstrated by making use of the aforementioned bounding relation between achievable MFCC-based $\overline{d}_{\mathrm{LSD(RMS)}}$ performance and the estimated information-theoretic measures. In particular, we showed that the ~ 99% and ~ 78% relative highband certainty gains measured, respectively, in the appending and substitution scenarios, correspond, respectively,

to 1.66 and 0.82 dB decreases in the best achievable $\overline{d}_{\text{LSD(RMS)}}$ performance of BWE. By comparing these potential improvements to those reported in earlier BWE works, we confirmed that memory inclusion can indeed result in BWE performance improvements that are, at least, comparable to those of oft-cited BWE techniques.

6.1.5 Incorporating speech memory into the BWE paradigm

Using the conclusions of Chapter 4 as the basis for subsequent work, we then focused in Chapter 5 on converting the information-theoretical gains quantified as discussed above into tangible BWE performance improvements.

First, we started by investigating the *reconstruction of speech from MFCCs* in order to exploit the superior highband certainties demonstrated in Chapter 4 for the inclusion of memory using MFCCs, rather than LSFs. Such reconstruction has been quite limited in the speech processing literature, in general, due to the non-invertibility of several steps employed in MFCC parameterization—namely, using the magnitude of the complex spectrum, the mel-scale filterbank binning, and the possible higher-order cepstral coefficient truncation. Indeed, this difficulty of synthesizing speech from MFCCs has effectively precluded their use in the context of BWE, despite their superior MI and class separability properties previously demonstrated in [126, 135]. Using *high-resolution* inverse discrete cosine transform (DCT) per [151], followed by LP analysis on the resulting high-resolution power spectra, we showed, however, that fine spectral detail can be obtained from the GMM-based MMSE estimates of highband MFCCs, with the DCT cosine functions acting as interpolation functions. As shown in Figure 5.1 and Tables 3.1 and 5.1, incorporating this MFCC inversion scheme into our memoryless dual-model BWE system enabled us to reconstruct highband speech with a quality that is nearly identical to that obtained using LSFs, despite the partial loss of information associated with MFCC parameterization.

Given the ability provided by our proposed MFCC-based BWE system to potentially exploit the superior certainty advantages associated with MFCC-based memory inclusion, we then proceeded by presenting the first of two distinct and novel approaches for memory-inclusive BWE. In particular, we followed the same methodology used to quantify the information-theoretic effects of memory in Chapter 4 by incorporating such memory exclusively into the MFCC parameterization frontend in the form of delta features. Despite the fact that, in practice, only the MMSE-estimated static highband features can be used for

the reconstruction of highband spectral envelopes, we showed that the certainty achievable for static-only highband MFCCs can nevertheless be improved by the inclusion of highband delta features—in addition to those of the narrow band—into the joint-band GMM-based model. Illustrated in Figure 5.4, this finding was confirmed by demonstrating the effect of the strong correlation between the delta parameterizations of both bands in improving the ability of the overall dynamic joint-band GMM to model the underlying phonemic classes, specifically in the static highband subspace that is, in fact, the only highband space actually needed for extension.

Given the aforementioned time-frequency information tradeoff imposed by the non-invertibility of delta features, we then performed an *empirical optimization* of dimensionalities in order to determine the optimal allocation of the available degrees of freedom among the static and delta features of both frequency bands such that static highband certainty is maximized. Integrating *frontend-based memory inclusion* optimized as such into our MFCC-based BWE system, as shown in Figure 5.6, resulted in relative performance improvements ranging from 2.1% in terms of $\overline{Q}_{\text{PESQ}}$ to 15.9% for $\overline{d}^{\star}_{\text{IS}}$, with a BWE algorithmic delay of 80 ms resulting from the non-causality of delta feature calculation. Although modest, these improvements were shown to coincide with the highband certainty gains measured when only the static highband subspace is considered. Moreover, they were achieved with no increases in run-time computational cost nor in training data requirements, and required only minimal modifications to the memoryless BWE system. As such, our proposed frontend-based memory inclusion approach provides a simple, inexpensive, and convenient means by which to realize some of the BWE performance improvements achievable by the inclusion of memory.

As an alternative to using a frontend dimensionality-reducing transform as the means for incorporating memory into the joint-band BWE model as discussed above, we focused instead in our second proposed approach on *modelling the high-dimensional distributions* underlying sequences of joint-band feature vectors. In addition to addressing the delta feature drawback of non-causality as well as the time-frequency information tradeoff associated with frontend dimensionality-reducing transforms in general, transferring the memory inclusion task from the frontend to the modelling space allows us to exploit prior knowledge about the properties of GMMs and speech to improve our models of the underlying classes along spectral and temporal axes. Indeed, by using: (a) the correspondence of GMM component densities to underlying classes, and (b) the strong correlations between neighbouring

speech frames, we showed that the problem of modelling high-dimensional GMM-based *pdf*s can be transformed into a time-frequency state space modelling task where the complexities associated with high-dimensional GMM parameter estimation can be circumvented. More specifically, we used sequences of past frames to grow high-dimensional GMMs in a *progressive tree-like* fashion, with the GMM component densities treated as states, or classes, corresponding individually to *time-frequency-localized* regions—regions that collectively span the full space underlying the modelled feature vector sequences. At each step of this tree-like progression, previously-estimated component densities are viewed as *parent* states from which finer *child* states can be estimated by incorporating the incremental information obtained by causally extending the input training data sequences—i.e., extending the sequences of static feature vectors further into the past. Illustrated in Figure 5.8, this progressive tree-like approach to the inclusion of memory into joint-band GMMs thus effectively breaks down the infeasible task of modelling such high-dimensional *temporally-extended* GMMs into a series of localized modelling operations with considerably lower complexity and fewer degrees of freedom.

In formulating our tree-like model-based approach to memory inclusion, we further presented two novel techniques intended to to ensure the robustness of the obtained temporally-extended GMMs to the oversmoothing and overfitting risks associated with GMM parameter estimation in high-dimensional settings in general:

(a) Since dimensionalities increase progressively with each step of our tree-like modelling technique, the overlap between the classes underlying the temporally-extended GMMs under training also increases progressively. This, in turn, increases the risk of overfitting. In contrast to the conventional Bayesian clustering approach where the risk of overfitting is compounded by hard-decision classification, our proposed *fuzzy GMM-based clustering* technique uses soft decisions to partition training data into fuzzy time-frequency child clusters, which are then used to estimate the parameters of the densities underlying the aforementioned time-frequency-localized regions as discussed below. Through an illustrative example, we showed this approach to be quite successful in alleviating the risk of overfitting, while simultaneously precluding any oversmoothing that can potentially result from relaxing the conventional hard-decision classification of training data.

(b) To incorporate the soft membership weights of the data subsets obtained by fuzzy clustering into the aforementioned estimation of localized *pdf*s, we also proposed and

derived a *weighted* implementation of the conventional Expectation-Maximization (EM) GMM-training algorithm. In particular, new iterative EM update formulae were derived such that a weighted log-likelihood function that takes account of the soft membership weights is maximized. The convergence of our iterative weighted algorithm was then proved.

In addition to these two algorithms, a third fundamental component of our tree-like GMM temporal extension algorithm was also formulated in order to maximize the information content of the resulting GMMs. Similar in concept to maximizing the entropy of a coded speech signal by exploiting the well-known redundancies in speech signals, this proposed *pruning* algorithm first measures the spectral variability of the incrementally-localized child data subsets—obtained by fuzzy clustering then used to train child state *pdf*s—using a *distribution flatness measure* in order to decide if the variability is sufficiently high to warrant splitting the parent state into multiple children states, prior to performing weighted EM. In a second post-EM step, we also apply a *cardinality test* to ensure that descendent child states—to be estimated in the future increment of the tree-like algorithm—can be reliably estimated without the risk of overfitting. Summarizing the overall tree-like GMM training algorithm, Table 5.5 and Figure 5.10 concisely illustrate how these component techniques are all melded together.

By formulating novel measures based on covariance matrix norms and normalized cepstral distances, respectively, we were then able to demonstrate the reliability of our high-dimensional temporally-extended GMMs in terms of robustness to both oversmoothing and overfitting. We thereafter described the modifications to be applied to our memoryless MFCC-based BWE system such that the dual-mode system can exploit the superior cross-band correlation properties of temporally-extend GMMs for improved highband speech reconstruction. Illustrated in Figure 5.12, these model-based modifications address the drawbacks of our frontend-based approach—namely, the time-frequency information trade-off and the non-causality, and associated algorithmic delay, imposed by delta features—while preserving its advantage in terms of the flexibility it provides for the inclusion of memory to varying extents—the primary advantage of delta features and simultaneously the deficiency of the oft-cited first-order HMM-based methods.

Our temporally-extended GMM-based BWE technique was then evaluated extensively in terms of both BWE performance and extension-stage computational costs. Relative to the memoryless baseline, results showed that our model-based approach to memory in-

clusion achieves considerable performance improvements across all performance measures, with the best improvements ranging from a relative 9.1% in terms of $\overline{Q}_{\text{PESQ}}$ to 56.1% for $\overline{d}^{*}_{\text{IS}}$, at a causal memory inclusion of 120 ms. Compared to the performance results achieved using our delta coefficient-based BWE technique, these results also showed that our second proposed technique significantly outperforms the frontend-based approach in terms of successfully translating the previously-quantified information-theoretic gains of memory inclusion into measurable BWE performance improvements. Although the advantages of model-based memory inclusion in terms of performance and real-time practicality were achieved at a run-time computational cost increase of nearly four orders of magnitude, relative to the memoryless baseline as well as to the computationally equally-inexpensive frontend-based approach, we nonetheless showed that these computational costs are within the typical capabilities of modern communication devices—e.g., tablets and smart phones.

Finally, through a detailed performance comparison, our temporally-extended GMM-based BWE technique was also shown to outperform comparable techniques incorporating model-based memory inclusion, in some cases by a wide margin. The techniques compared ranged from those based on predictive VQ, e.g., [131], to the HMM-based techniques often cited as being more successful, e.g., [87]. By illustrating this comparison, Figure 5.19 provides a rather informative and concise perspective of the relative success of current state-of-the-art BWE landscape.

6.2 Potential Avenues of Improvement and Future Work

In addition to ideas that can potentially improve the performance and generalization of our proposed BWE techniques, we now discuss relevant research avenues unaddressed in this thesis due to scope, time, and space limitations. These ideas and topics of interest can be categorized by context as follows.

6.2.1 Dual-mode BWE and statistical modelling

(a) As detailed in Section 3.2.3, the dual-mode technique of [55] upon which our BWE implementation is based uses equalization to *recover*—rather than reconstruct by GMM-based statistical mapping—the lowband and midband content in the 100–300 Hz and 3.4–4 kHz ranges, respectively. This approach followed from the higher likelihood for improved speech reconstruction with equalization given the knowledge available

about the filter response characteristics of the G.712 telephone channel. Although our focus in this thesis has been the reconstruction of content above 4 kHz, the perceptual importance of the lowband and midband ranges presents a motivation for further research. We noted in Section 1.1.3.1 that the lowband content adds *naturalness* to the speech signal as well as improve the perception of nasals and voicing in fricatives, stops, and affricates. Similarly, we showed in Section 1.1.3.3 that the 0.8 bark 3400–3889 Hz subband was found in [27] to be more perceptually important than many other subbands outside the 300–3400 Hz range. Among the ideas to be investigated to improve the recovery of speech in both these ranges, augmenting equalization with statistical modelling is of particular interest. More specifically, by statistically modelling narrowband speech jointly with the true gain in the bands to be equalized, the reconstruction of lowband and midband speech can be separated into signal *shape recovery* via equalization in conjunction with signal *gain reconstruction* via GMMs. Alternatively, the statistical estimation of equalization gain can be performed as a corrective post-equalization step where a gain ratio—rather than absolute gain—is estimated via GMMs. This latter approach would, in essence, be similar to that used for the estimation of the highband excitation gain—calculated per Eq. (3.3) as the square root of the ratio of energy in the original highband signal to the energy in the reconstructed signal—as described in Section 3.2.5.

(b) Throughout our work, our approach to statistical modelling has been exclusively speaker-independent. Notwithstanding the additional training and testing data requirements in terms of size and labelling, performing joint-band modelling in a speaker-dependent manner, however, has the potential to considerably improve the MMSE-based estimation of highband speech. Indeed, as noted in Section 4.4.3.2, the speaker-dependent HMM-based BWE technique of [39], for example, was shown to outperform the corresponding speaker-independent implementation by an average of $\overline{d}_{\text{LSD(RMS)}} \simeq 1$ dB. Given the observation that $\overline{d}_{\text{LSD(RMS)}}$ performance improvements are, in general, only slightly higher than the corresponding $\overline{d}_{\text{LSD}}$ improvements, similar improvements achievable by introducing speaker dependence would thus potentially be comparable to those achieved by the best performing BWE techniques. This projection follows directly from a comparison to the ranges illustrated in Figure 5.19 for the $\overline{d}_{\text{LSD}}$ performance improvements achieved by state-of-the-art BWE techniques.

6.2.2 Frontend-based memory inclusion

(a) In contrast to our frontend-based approach to memory inclusion where only the first-order regression of long-term dynamics was captured via delta features, the HMM-based techniques of [87] and [163] additionally use delta-delta features to parameterize the second-order regression. As discussed in Section 5.3.5.3, however, these techniques rely primarily on the first-order HMM state transition probabilities to model the cross-band correlation of speech dynamics. This minor role of dynamic parameterization in these techniques is emphasized by the absence of any information regarding: (a) the durations used to calculate the first- and second-order delta features, and more importantly, (b) the contribution of such features to the overall BWE performance improvements reported therein. Although the improvements achieved by our delta coefficient-based BWE technique were rather modest, the improvements achieved by the additional inclusion of delta-delta features in the field of automatic speech recognition (ASR) motivates us to investigate their benefits, or lack thereof, in the context of BWE. Worthy of note in this context is that, in addition to their effect in capturing speech dynamics, first- and second-order delta features applied in the spectral domain—rather than in the typical cepstral domain—were recently shown to improve robustness to additive noise and reverberation [191].

(b) As discussed in Section 4.4.2, the differential transform used to generate delta features can be viewed as as a special case of dimensionality-reducing transforms that compact the temporal information from sequences of static feature vectors into a single vector of dynamic features. From this perspective, we noted that other transforms can then be equally applied for the purpose of memory inclusion, most notably those of linear discriminant analysis (LDA) and the Karhunen-Loève transform (KLT). In comparison to the differential transform of Eq. (4.34), LDA is characterized by its superior discriminative ability of the underlying classes while the KLT is known for its superior decorrelating properties. As a result of their advantages, LDA and the KLT were both shown in [149] to outperform delta features in terms of encoding temporal information from sequences of static feature vectors. Since that comparison was performed in the context of a digit recognition task, however, it does not account for the BWE-specific effects of time-frequency information tradeoff imposed by the non-invertibility of all three transforms. Nevertheless, the superior temporal

compaction demonstrated in [149] for LDA and the KLT suggests a time-frequency information tradeoff that is potentially more favourable for the purpose of cross-band correlation modelling than that associated with delta features. As such, the inclusion of memory via the addition of LDA- or KLT-based dynamic features to the static features necessary for speech reconstruction represents a research topic of interest.

6.2.3 Tree-like GMM temporal extension

Despite the superior BWE performance achieved using our tree-like memory inclusion algorithm, we believe that the generalization and modelling performance of our algorithm can be further improved through some modifications. Indeed, comparing the best results associated with the $\boldsymbol{\mathcal{G}}_{\lozenge}\big|_{K=4}$ tuple in Table 5.6, to the information-theoretic gains reported in Table 4.4 for the memory-inclusive scenario, suggests that there are additional performance gains to be yet achieved. More specifically, Table 4.4 indicates that the highband certainty gains associated with memory inclusion translate to a range of 0.82–1.62 dB of absolute improvement in terms of the aforementioned lower bound of BWE $\overline{d}_{\text{LSD(RMS)}}$ performance. In comparison, the maximum 0.62 dB $\overline{d}_{\text{LSD}}$ improvement reported in Table 5.6 corresponds to only 0.73 dB in terms of $\overline{d}_{\text{LSD(RMS)}}$. Hence, the improvements attained by our model-based memory inclusion approach can theoretically be doubled. To realize these potential gains, we list the following modifications to our algorithm as future avenues of research:

(a) As described in Operation (c) of Section 5.4.2.3, the splitting factor, J, controls the branching complexity of our tree-like training algorithm by defining the number of child states to be derived from each lth-order parent state. To minimize overfitting while maximizing the information content of these lth-order child state *pdf*s, the branching complexity was subsequently moderated by the pruning described in Operation (d), with the result that the effective number of child states to be derived for each lth-order parent state is a binary number given by $|\mathcal{J}_i^{(l)}| \in \{1, J\}$. Rather than constrain the progressive generation of time-frequency states in such a binary hard-decision manner, however, a gradual pruning approach may be more beneficial. In particular, the pre-EM pruning condition of Eq. (5.63) can be relaxed such that the distribution flatness, ρ_i, of each $(i \in \mathcal{I}^{(l)})$th time-frequency-localized data subset is repeatedly estimated based on $\mathcal{G}_{\mathbf{Y}}^{(0)}$ initialization GMMs with decreasing complexity—in terms of the number of Gaussian components, J—until the distribution flatness

exceeds the specified flatness threshold, $\rho_{\min}$, or the minimum number of child states, i.e., $|\mathcal{J}_i^{(l)}| = 1$, is reached. In other words, Eqs. (5.60), (5.61), (5.62) are repeated with descending values for J until the maximum value for $1 \leq |\mathcal{J}_i^{(l)}| \leq J$ is found such that the right-hand side condition of Eq. (5.63) is satisfied. As a result of the higher resolution used as such to model the lth-order time-frequency-localized distributions, this gradual pruning approach should, in theory, result in an improved global model for the entire temporally-extended joint-band space at memory inclusion order l.

(b) As described in Operation (a) of Section 5.4.2.3, our fuzzy clustering algorithm was proposed in order to account for the overlap between the lth-order child classes when partitioning the associated lth-order parent data into corresponding time-frequency-localized child subsets (which, in turn, become the $(l+1)$th-order parent subsets). To control the *softness* of this classification, the fuzziness factor, K, was introduced. The extent of the expansion of the partitioned child subsets is determined by using normalized posterior probabilities in Eq. (5.19) to calculate K membership weights—and hence, K different destination child subsets—for each data point in the parent subset. In subsequently implementing this fuzzy clustering approach within our tree-like GMM training algorithm, a fixed value for K was used.

Although the normalization used in Eq. (5.19) allows subset expansion to account for the actual extent of class overlap (represented by overlap in the tails of the Gaussian *pdf*s corresponding to the underlying time-frequency classes), using a fixed value for K for all classes spanning the entire lth-order time-frequency space results in the same expansion complexity for all classes, regardless of differences in terms of the extent of overlap. However, time-frequency regions where there is minimal class overlap do not require the same high values for K otherwise needed in regions with high overlap in order to achieve the same modelling accuracy. Accordingly, using dynamic overlap-dependent values for K—rather than quantitatively expand all subsets equally through a uniform value—allows us to make more efficient use of the available training resources, and hence, achieve a potentially better overall $(l+1)$th-order GMM-based model. To that end, K can be optimized dynamically during the training algorithm of Table 5.5 as a function of the areas under the overlapping tails of the Gaussian densities representing lth-order child classes. Alternatively, fuzzy clustering can be performed in an iterative manner—independently for each parent

data subset—with the value of K incrementally increased at each iteration until a stopping criterion associated with the change in child subset mean and/or variance is reached. Similar in concept to the stopping criteria used in iterative EM or VQ training (where the change in training data likelihood, or mean-square error in the case of VQ, is compared to a particular threshold after each EM iteration), a stopping criterion based on the change in the parameters of child subset distributions corresponds to the convergence of the iterative fuzzy clustering towards a particular classification accuracy.

(c) As detailed in Operation (e) of Section 5.4.2.3, the time-frequency-localized states obtained via our tree-like GMM growth algorithm have the conditional independence properties of Markov blankets as defined in [179].[185] For the localized subsets obtained by fuzzy clustering, in particular, these conditional independence properties are rather evident for all memory inclusion orders of $l > 0$. For each lth-order time-frequency-localized parent data subset, $\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l)}}$, our fuzzy clustering algorithm exclusively uses the corresponding $|\mathcal{J}_i^{(l)}|$-modal *pdf* given by the GMM, $\mathcal{G}_{\mathbb{z}_i}^{(l)}$, to cluster the data into $|\mathcal{J}_i^{(l)}|$ lth-order child subsets, $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)},w^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$, per Eqs. (5.18)–(5.21). Since $\mathcal{G}_{\mathbb{z}_i}^{(l)}$ is itself estimated exclusively for the $\mathcal{V}_i^{\mathbb{z}^{(l)},w^{(l)}}$ parent subset using weighted EM, it is clear that the $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)},w^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$ child subsets are thus conditionally independent of all other lth-order parent subsets, $\{\mathcal{V}_m^{\mathbb{z}^{(l)},w^{(l)}}\}_{\forall m\neq i}$, and hence, are also conditionally independent of all lth-order child subsets descending from these parent subsets, i.e., $\left\{\{\mathcal{V}_{mj}^{\mathbb{z}^{(l)},w^{(l)}}\}_{j\in\mathcal{J}_m^{(l)}}\right\}_{\forall m\neq i}$. Using the arguments given in Operation (e) regarding the correspondence of Eq. (5.67b) to the conditional independence of all child states descending from the same parent state, the $\{\mathcal{V}_{ij}^{\mathbb{z}^{(l)},w^{(l)}}\}_{j\in\mathcal{J}_i^{(l)}}$ child subsets can then be shown to be conditionally independent themselves.

Although these conditional independence properties considerably simplify the overall training algorithm in Table 5.5 as well as improve its interpretation intuitively, in reality the time-frequency-localized states underlying all subsets—parent as well as child subsets—do overlap, and hence, conditional independence among states is, in fact, rather unlikely. As discussed in Item (b) above, our fuzzy clustering approach accounts for such overlap via the soft membership weights. However, per Eqs. (5.16)

[185] See Footnote 161 for the formal definition of Markov blankets.

and (5.19), it only does so for *sibling* states—states descended from the same parent state and which correspond to the component densities of one particular GMM modelling a unique parent data subset. In other words, our current implementation of fuzzy clustering restricts the modelling of class overlap to only that between sibling classes for all $l > 0$. Accordingly, it follows that extending the input domain of fuzzy clustering to all lth-order child states should result in higher-quality localized subsets as a result of modelling class overlap across the entire lth-order temporally-extended joint-band space.

Corresponding to a more realistic relaxation of the aforementioned conditional independence properties, this modification to fuzzy clustering can be implemented by substituting all references to the priors, $\mathrm{A}_i^{\mathbb{z}^{(l)}} = \{\alpha_{ij}^{\mathbb{z}^{(l)}} := P(\lambda_{ij}^{\mathbb{z}^{(l)}})\}_{j \in \mathcal{J}_i^{(l)}}$, and densities, $\Lambda_i^{\mathbb{z}^{(l)}} = \{\lambda_{ij}^{\mathbb{z}^{(l)}} := (\boldsymbol{\mu}_{ij}^{\mathbb{z}^{(l)}}, \mathbf{C}_{ij}^{\mathbb{z}\mathbb{z}^{(l)}})\}_{j \in \mathcal{J}_i^{(l)}}$, of $\mathcal{G}_{\mathbb{z}_i}^{(l)}$, $\forall i \in \mathcal{I}^{(l)}$, in Eqs. (5.16)–(5.21), by the corresponding priors and densities of the global lth-order temporally-extended GMM, $\mathcal{G}_{\mathbb{z}}^{(l)}$, given by Eq. (5.67). In the context of the overall algorithm of Table 5.5, the modification can be implemented by moving steps (d)–(g) to succeed, rather than precede, step (h).

(d) In a manner similar to that performed above for the splitting and fuzziness factors, J and K, respectively, the memory inclusion step, τ, can also be modified to be dynamic, rather than fixed as illustrated in Figure 5.8. As discussed in Item v of Section 5.4.3.2, τ indirectly allows us to increase the information content of the temporally-extended data by leapfrogging redundancies between immediately-neighbouring static frames when constructing temporally-extended feature vectors. Setting τ dynamically in a manner dependent on the information content of the concatenated static feature vectors should thus further increase the overall information content of our tree-like algorithm. To that end, we could make use of the distribution flatness measure already introduced in Operation (d) of Section 5.4.2.3 as a means by which to measure the self-information of the child data subsets obtained by fuzzy clustering. More specifically, the self-information of child data subsets obtained at a particular memory inclusion index, l, can be estimated as a function of τ prior to the application of pre-EM pruning, then used to optimize τ accordingly. It should be noted, however, that, since the same value of τ must be used for all child subsets at the same lth order of memory inclusion, such a dynamic information-dependent optimization of

τ can only be performed globally at each step of our tree-like GMM training algorithm. In other words, the previously-fixed τ now becomes the order-dependent $\tau(l)$. This modification thus contrasts with those discussed above for J and K which can be dynamically modified on a per-parent-state basis, rather than globally at each lth order. This modification, to be applied during the temporally-extended GMM training stage, requires a corresponding—but rather straightforward—change in the reconstruction of temporally-extended supervectors during the extension stage—more specifically, replacing $n\tau$ in Figure 5.12 by $\sum_{m=1}^{n}\tau(m)$, for all $n \in \{1,\ldots,l\}$.

(e) As first described in Section 5.4.2.2 then later detailed in Operation (d) of Section 5.4.2.3, the variability obtained by pruning in terms of the number of child states that can potentially be estimated for each parent state, not only increases the model's information content, but is also intended to model the large variability among different speech classes—as well as among different realizations of the same classes—in terms of the rate of change of spectral properties across time.[186] It is this particular time-dependent variability that HMMs are known to model well through intra- and inter-state transitions. While using a dynamic information-dependent memory inclusion step, $\tau(l)$, during training and extension-stage mapping—as described in Item (d) above—should alone improve the ability of our tree-like algorithm to model such variability, employing a more sophisticated dynamic approach for the reconstruction of temporally-extended supervectors during the mapping stage should further improve our ability to account for temporal variations in long-term dynamics. One such approach is to use *dynamic time warping* (DTW) [10, Section 10.6.2] to dynamically determine the optimal sequence of $l+1$ input narrowband feature vectors—among all the paths by which $l+1$ frames can be chosen from the $l\tau+1$ consecutive input vectors resulting from the static frontend at order l—such that the likelihood of the lth-order sequence constructed by concatenating these vectors, given the lth-order temporally-extended narrowband GMM obtained by marginalizing its joint-band counterpart, is maximized. In other words, rather than construct temporally-extended narrowband supervectors from input static feature vectors via $\mathbb{X}_t^{(\tau,l)} = [\mathbf{X}_t^T, \mathbf{X}_{t-\tau}^T, \ldots, \mathbf{X}_{t-l\tau}^T]^T$ when using a fixed τ, or via $\mathbb{X}_t^{(\tau,l)} =$

[186] See Footnote 139 for a clarifying example of the differences among classes in terms of spectral variability as a function of time.

$[\mathbf{X}_t^T, \mathbf{X}_{t-\tau(1)}^T, \mathbf{X}_{t-\tau(1)-\tau(2)}^T, \ldots, \mathbf{X}_{t-\sum_{n=1}^{l}\tau(n)}^T]^T$ when using an order-dependent $\tau(l)$, we instead construct $\mathbb{X}_t^{(\tau,l)}$ as $\mathbb{X}_t^{(\tau,l)} = [\mathbf{X}_t^T, \mathbf{X}_{t-\tau+\varepsilon(1)}^T, \mathbf{X}_{t-2\tau+\varepsilon(2)}^T, \ldots, \mathbf{X}_{t-l\tau+\varepsilon(l)}^T]^T$, or as $\mathbb{X}_t^{(\tau,l)} = [\mathbf{X}_t^T, \mathbf{X}_{t-\tau(1)+\varepsilon(1)}^T, \mathbf{X}_{t-\tau(1)-\tau(2)+\varepsilon(2)}^T, \ldots, \mathbf{X}_{t-\sum_{n=1}^{l}\tau(n)+\varepsilon(l)}^T]^T$, respectively, with the additive time index deviations, $\{\varepsilon(n)\}_{n\in\{1,\ldots,l\}}$, determined online during mapping by DTW such that the likelihood $P(\mathbb{x}_t^{(\tau,l)}|\mathcal{G}_{\mathbb{X}^{(\tau,l)}}) = \sum_{m=1}^{M}\alpha_m^{\mathbb{x}^{(\tau,l)}} P(\mathbb{x}_t^{(\tau,l)}|\lambda_m^{\mathbb{x}^{(\tau,l)}})$ is maximized individually for each input $\mathbb{x}_t^{(\tau,l)}$ supervector. As typically done in the application of DTW, constraints on the maximum values $\{\varepsilon(n)\}_{n\in\{1,\ldots,l\}}$ can attain should be imposed to limit increases in computational complexity as well as to ensure that a reasonable degree of local spectral continuity is preserved.

6.3 Applicability of our Research and Contributions

As repeatedly noted throughout the thesis, we have attempted to present our research as generally as possible to emphasize and widen the potential for its application. As such, we now conclude by briefly discussing the applicability of our work to BWE in general, as well as to non-BWE contexts.

Despite exclusively using the dual-mode BWE technique of [55] as the vehicle for our research, it is clear that the approaches proposed in Chapter 5 for the purpose of improving BWE are easily transferable to other BWE techniques based on the statistical modelling of cross-band correlation via GMMs. Our frontend- and model-based techniques for the inclusion of memory can indeed be applied to any BWE technique using the GMM-based mapping approach of [82], regardless of the type of features used to parameterize speech in the narrow and high frequency bands. Similarly, we have shown that our GMM- and VQ-based information-theoretic approach proposed in Chapter 4 for the purpose of quantifying long-term speech dynamics can be applied to any form of parameterization, as long as spectral errors can be calculated from that parameterization.

Although these approaches noted above were proposed for the purpose of quantifying and exploiting speech memory in the context of BWE, they are, in fact, also equally applicable to other contexts where source-target transformation is performed via GMMs. Among such contexts, the field of speaker conversion, e.g., [40, 159–161], is most notable. Indeed, many of the similarities between BWE and speaker conversion were discussed throughout the thesis. Other examples of related GMM-based fields that were not previously discussed, however, include conversion in the context of text-to-speech (TTS) synthesis,

e.g., [78], speaker de-identification, e.g., [192], and articulatory-to-acoustic—and the corresponding acoustic-to-articulatory inverse—mapping, e.g., [193]. Since the majority of these works typically use diagonal-covariance GMMs for source-target transformation, our investigation and subsequent conclusions in Chapter 3 on the role of GMM covariance type—where common assumptions about the effects of covariance type on the performance and computational costs of MMSE-based transformation were challenged—also gain particular importance.

In addition to these domains, we note that our work on quantifying the information content of long-term speech can also be beneficial to those of speech coding and enhancement. In particular, our proposed information-theoretic approach can be used to quantify the relative importance of long-term speech in arbitrary frequency bands—rather than only in the 0.3–4 and 4–8 kHz bands of midband-equalized narrowband and highband speech, respectively—for the purposes of determining the optimal allocation of coding or robustness resources. In essence, this application would be similar in concept to the work of [27] where subjective—rather than objective—evaluations were used to determine the relative importance of memoryless—rather than long-term—content in several frequency bands within the 50–7000 Hz range. In addition to quantifying the relative importance of different frequency bands, our information-theoretic technique can similarly be used to also evaluate the long-term information retention capabilities of different speech parameterizations.

Last but not least, we note the relevance of our tree-like GMM training technique to the machine learning contexts of mixture model-based density estimation and clustering. As discussed in Section 5.4.2.1, addressing the oversmoothing and overfitting problems associated with modelling in high-dimensional settings is among the topics these fields are most concerned with, e.g., [154–158, 174]. Given the success of our tree-like algorithm in mitigating such dimensionality-related problems in the context of long-term speech modelling, we can project a comparable success for its application—as a whole algorithm as well as in terms of its individual fuzzy clustering and weighted EM algorithms—to the general machine learning contexts of density estimation or clustering. This success is, however, conditional on the requirement that the high-dimensional data to be modelled, or clustered, has the same properties of long-term speech that made our time-frequency localization approach possible—namely, (a) a strong correlation across the dimensions of the feature vectors of the data, and (b) an underlying multi-modal distribution where densities can intuitively converge to model individual generative classes.

Appendix A

Dynamic and Temporal Properties of Speech

A.1 Temporal Cues

In addition to the short-term spectral characteristics of Section 1.1.3.1 which act as cues to voicing, manner and place of articulation (and their longer-term dynamic variants discussed in Section A.2 below), the perception of speech also exploits many temporal cues that complement and, in many cases, supersede spectral cues. Indeed, temporal cues have been shown sufficient to achieve 90% correct identification of words when spectral detail is severely degraded through substitution by only three broad bands of noise [167]. Moreover, some languages, e.g., Swedish and Japanese, use duration directly as a phonemic cue, in the sense that some phonemes differ only by duration and not spectrally [10, Section 5.6.1]. Generally, however, duration is a secondary phonemic cue utilized when a primary cue is ambiguous; e.g., the /b/ closure duration in the word *rabid* is normally short; if the closure is prolonged, *rapid* is heard. Thus, cues for voicing may be found in the durational balance between a stop and a preceding vowel. Another example where duration influences perception is in fricative+sonorant clusters; normally, a short interval (about 10 ms) intervenes between the cessation of frication and the onset of voicing, when this duration exceeds about 70 ms, listeners tend to perceive a stop phoneme in the interval despite the lack of the burst associated with stops. Place of articulation in stops is also affected by closure duration in some cases; stop closures tend to be longer for labials than for alveolars and velars. As such, longer stops bias perception towards labials.

Other temporal cues include voice onset time (VOT) in stop+sonorant clusters—the time from stop release (the start of the resulting sound burst) to the start of vocal fold

periodicity—and the timing and duration of pitch and formant transitions before and after sonorants. These temporal cues are particularly dependent on context as described next.

A.2 Coarticulation and the Inherent Variability in Speech

While short-term spectral features, such as those described in Section 1.1.3.1, provide distinctive cues for most *phones* (the physical sounds produced when a phoneme is articulated), speech does not simply consist of a concatenation of discrete phones with ideal *steady-state* characteristics. Rather, vocal tract articulators move gradually from one phoneme's articulatory gestures to those corresponding to the next—a property called *coarticulation*[187]. Thus, through coarticulation, phonemes' acoustic features affect those of several preceding and ensuing phones, often across syllable and syntactic boundaries. For example, lip rounding for a vowel usually commences during preceding nonlabial consonants by lowering their formants in anticipation of the rounded vowel. While such formant lowering does not cause the consonants to be perceived differently when spoken in context, it does affect their spectral properties. Coarticulation thus results in diffusing perceptually-important phonemic information across time at the expense of phonemic spectral distinctiveness. In fact, classical steady-state positions and formant frequency targets for many phonemes are rarely achieved in natural coarticulated speech.

In addition to coarticulation, speech exhibits inherent variability. Repeated pronunciations of the same phoneme by a single speaker differ from one another, with versions from different speakers differing to an even higher extent. Comparing segments in identical phonetic contexts, a speaker produces standard deviation variations on the order of 5–10 ms in phone durations and 50–100 Hz in F1–F3 [10, Section 3.7.1]. Variations in different contexts beyond these amounts are attributed to coarticulation. Consequently, coarticulation and the inherent variability of speech result in phonemes with infinite variations of phones that are rather viewed as consisting of transient and highly context-dependent initial and final segments, with a steady-state segment in between that is less affected by phonetic context.

A detailed description of the dynamic effects of coarticulation on the spectral and temporal cues of speech is beyond the scope of this work.[188] In the following, however, we

[187] See Footnote 19.

[188] A detailed and thorough review of coarticulation and its effects on speech perception is provided in [10, Section 3.7 and Chapter 5].

demonstrate the significance of such dynamic coarticulation effects on perception:

Vowel identification

When vowels are produced in contexts, i.e., not in isolation, formants undershoot their targets. Perception of such vowels depends on a complex auditory analysis of formant movements before, during, and after the vowel. In CVC (consonant-vowel-consonant) syllables, listeners perform worse in vowel identification when the middle 50–65% of the vowel is excised and played to listeners in isolation, than if the CV and VC transitions (containing the other 35–50% of the vowel) are heard instead [10, Section 5.4.3]. Short portions of the CV and VC transitions often permit identification of the vowel when a large part of the vowel is removed, indicating the importance of dynamic spectral transitions for vowel intelligibility.

While spectra dominate in regards to vowel perception, temporal coarticulation factors affect phone identification; e.g., lax and tense vowels tend to be heard when formant transitions are slow and fast, respectively.

Perception of consonant voicing

As many syllable-final voiced obstruents have weak vocal cord vibrations, the primary cues may be durational [10, Section 5.5.3.1]: voicing is perceived more often when the prior vowel is long and has a higher durational proportion of formant steady state to final formant transition. In French vowel+stop sequences, the duration of the closure, the duration and intensity of voicing, and the intensity of the release burst, as well as the preceding vowel duration, all interact to affect voicing perception. In English VC contexts, the glottal vibration in the vowel usually continues into the initial part of a voiced stop, whereas voicing terminates abruptly with oral tract close in unvoiced stops. This difference appears to be the primary cue to voicing perception in final English stops.

For voicing in syllable-initial stops, VOT seems to be the primary cue [10, Section 5.5.3.2]; a rapid voicing onset after stop release leads to voiced stop perception, while a long VOT cues an unvoiced stop. A secondary cue is the value of F1 at voicing onset, where lower values cue voiced stops. This follows from the fact that F1 rises in CV transitions as the oral cavity opens from stop constriction to vowel articulation. The duration and extent of the F1 rising transition significantly affects stop voicing perception.

In consonant clusters within a syllable, only certain sequences of consonants are permissible. English, for example, requires that obstruents within a cluster have common voicing, i.e., all voiced or all unvoiced (e.g.,*steps*, te*xts*, do*gs*).

Perception of consonant manner of articulation

The timing of transitions to and from vocal tract constrictions associated with consonants influences perception of the consonants; e.g., when steady formants are preceded by linearly rising formants, /b/ is heard if the transition is short and /w/ if more than 40 ms. With very long transitions (> 100 ms), a sequence of vowels beginning with /u/ is heard instead. In contrast, if falling formants are used, /g/, /j/, and /i/ are successively heard as the transition duration increases [10, Section 5.5.1.1].

Perception of consonant place of articulation

Weak continuant consonants (continuant consonants are all consonants except for stops) are primarily distinguished by spectral transitions at phoneme boundaries. Spectral transitions are also more reliable for the perception of consonant place than steady-state spectra for stops and forward fricatives [10, Section 5.5.2]. In stop+sonorant sequences, for example, transitions are more important than burst amplitude for the perception of /b/ than for /d/. Similarly, transitions are more reliable place cues before nonfront vowels. In the case of unreleased plosives in VC syllables, spectral transitions provide the sole place cues. For CV stimuli from natural speech, stop bursts and ensuing formant transitions have equivalent perceptual weights. In stressed CV contexts and synthetic CV stimuli, however, VOT and amplitude also play a role when formant transitions give ambiguous cues; VOT duration distinguishes labial from alveolar stops (labial stops have the shortest VOTs, while velars have the longest, with bigger differences for unvoiced stops), and spectrum amplitude changes at high frequencies (F4 and higher formants) can also reliably separate labial and alveolar stops: when high-frequency amplitude is lower at stop release than in the ensuing vowel, labials are perceived [10, Section 5.5.2.1].

Among the phonological constraints of syllable contexts is that consonants in final nasal+unvoiced stop clusters must have the same place of articulation (e.g., li*mp*, li*nt*, li*nk*).

A.3 Prosody: Suprasegmental and Syntactic Information

The analysis above illustrates the importance of the dynamic properties of speech as cues integral to speech perception, but only at the segmental phonological level (i.e., at the segmental level of sequences of one to three phones at most, and without the aid of linguistic or syntactic information). In particular, we observe that the mapping from phones (with their varied acoustic correlates) to individual phonemes is likely accomplished by analyzing dynamic acoustic patterns—both spectral and temporal—over sections of speech corresponding roughly to syllables [10, Section 5.4.2]. Meaningful speech, however, also incorporates language-dependent *prosody*—suprasegmental and syntactic information that extends beyond phone boundaries into syllables, words, phrases, and sentences. Prosody concerns the relationships of duration, amplitude, and F0 of sound sequences. A such, suprasegmental and syntactic information manifests into recognizable long-term acoustic patterns of rhythm and intonation that assist in recognizing and identifying speech units smaller than the entire sentence. Prosody, for example, assists in word recognition, especially in tonal languages, e.g., Japanese, where different F0 patterns superimposed on identical segment sequences cue different words [10, Section 3.8]. In fact, prosody contains sufficient information such that speech communication can still be achieved with severely degraded spectra; Blesser shows in [168] that subjects can converse by exploiting F0, duration, and amplitude, with spectral segmental information effectively destroyed through spectral rotation at 4 kHz (replacing low-frequency content by that at high frequency and vice versa).

Lexical (word) stress is an example of suprasegmental intonation features that is as important to the identification of a spoken word as the use of the proper sequence of phonemes [10, Section 5.7.1]. In English, F0 is the most important acoustic correlate of stress, duration secondary, and amplitude least important. At a higher level, prosody shifts from the syllable- and word-highlighting effects of stress to highlighting syntactic features. The primary function is to aid in segmenting utterances into small phrasal groups and syntactic structures in order to facilitate the transfer of information; monotonic speech, i.e., speech lacking F0 variation, without pauses usually contains enough segmental information for message intelligibility, but it is also fatiguing to listen to.

Appendix B

The PESQ Algorithm

B.1 Description

As noted in Section 3.4.3, the calculation of the PESQ score is rather complex as it involves many time- and frequency-domain processing steps over the length of a test speech signal—assumed to be a few seconds long.[189] Indeed, as stated in [120, Section 10], a description of the PESQ algorithm—illustrated in Figure B.1—can not be easily expressed in mathematical formulae, but is rather textual in nature. As such, based on [119–122], we describe the algorithm as follows:

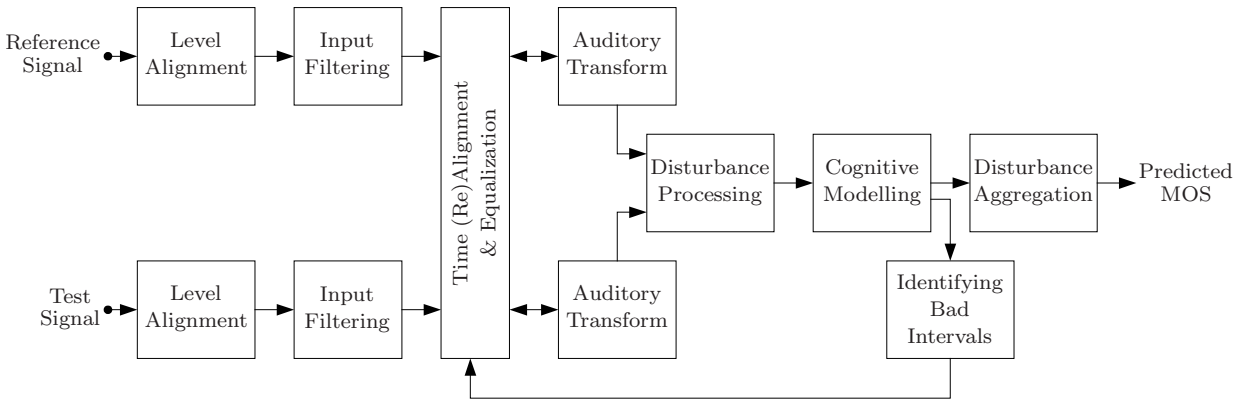


Fig. B.1: The PESQ algorithm. See [121, Figures 2, 3; 122, Figure 1].

Level alignment The reference and test signals are first aligned to a standard listening level.

[189] Most of the experiments used in calibrating and validating PESQ contained recordings of 2–4 sentences separated by silence, totalling 8–12 s in duration [120, Section 8.1.2].

Input filtering Signals are filtered (using an FFT) with an input filter to model the narrowband characteristic of a standard telephone handset in the case of P.862—extended later in P.862.2 to allow PESQ evaluation for wideband (50–7000 Hz) speech signals.

Time alignment Assuming piecewise constant delays between the reference and test signals, both signals are time-aligned through a series of steps:

- envelope-based delay estimation using the entire original and degraded signals,
- dividing both signals into utterances,
- envelope-based delay estimation per utterance,
- fine correlation/histogram-based identification of delay per utterance,
- utterance splitting and realignment to test for delay changes during speech.

These steps provide a delay estimate for each utterance, which is then used to find a per-frame delay for use in the auditory transform.

Auditory transform A psychoacoustic model maps the signals into a representation of perceived loudness in time and frequency as follows:

- **Perceptual frequency warping:** FFT coefficients in each 32 ms frame (with 50% overlap) are grouped into 42 bins that are equally spaced on a modified Bark scale.[190]
- **Frequency equalization:** Since severe filtering is disturbing to listeners while mild filtering effects have minimal influence on overall perceived quality (especially when no reference is available to the subject), partial compensation is used to provide PESQ score robustness to such imperceptible filtering effects in the test signal. The mean Bark spectrum for active speech frames is calculated using only the time-frequency cells whose power is more than 30 dB above the absolute hearing threshold. Per modified Bark bin, a partial compensation factor is calculated from the ratio of the test signal spectrum to that of the original signal, bounded to 20 dB, and then used to equalize the reference signal to the test signal. Compensation is applied to the original signal since the degraded test signal is the one judged by subjects in an ACR experiment.

[190] See Section 4.2.1 for more details on perceptual frequency mapping.

- **Equalization of gain variations:** Imperceptible short-term gain variations are partially compensated by processing per-frame Bark spectra. The ratio between the audible powers—i.e., where spectra exceed the absolute hearing threshold—of the reference and test signals in each frame is used to identify gain variations. This ratio is filtered with a first-order lowpass filter and then used to equalize the degraded signal to the reference.
- **Loudness mapping:** The equalized Bark spectrum is then mapped to a Sone loudness scale, resulting in loudness densities—the perceived loudness in each time-frequency cell.

Disturbance processing *Disturbances* are computed as the signed difference between the test and reference loudness in each time-frequency cell. Positive disturbances indicate noise addition while negative ones indicate signal attenuation. Reference and test frames where time alignment results in negative delays longer than half a frame, are discarded.

Cognitive modelling In addition to the perceptual processing described above, two important cognitive effects are modelled into the per-time-frequency cell disturbances:

- **Masking:** Masking is the perceptual property where small intensity differences are inaudible in the presence of stronger intensities within—as well as in neighbouring—time-frequency cells. Within-cell masking is applied in the PESQ model by generating a *deadzone* in each time-frequency cell using a simple threshold below which disturbances are inaudible. The threshold is set to the lesser of the loudness of the reference and test signals, divided by four. The threshold is then subtracted from the absolute loudness difference, and values less than zero are set to zero. The net effect is that disturbances are pulled towards zero, thereby generating the masking deadzone where only those time-frequency cells with disturbance values outside the zone are perceived as distorted. Methods for applying masking across time-frequency cells were examined with earlier perceptual models but did not improve overall performance, and thus, were not used in PESQ.
- **Asymmetry in disturbance perception:** The perception of disturbances is generally asymmetric in the sense that a reference signal distorted additively can be decomposed into two different percepts—the original signal and the additive

distortion—with such distortions being clearly audible. In contrast, an attenuated or omitted time-frequency component can not be similarly decomposed and the distortion is less objectionable to listeners. This effect is modelled in PESQ by calculating an asymmetrical disturbance per time-frequency cell by multiplying the cell disturbance with an asymmetry factor. The PESQ asymmetry factor is calculated as the ratio of the Bark spectral densities of the test and reference signals in each time-frequency cell, raised to the power of 1.2, and bounded with an upper limit of 12. Values smaller than 3 are set to zero such that only those time-frequency cells for which the distorted Bark spectral density exceeds that of the reference by the corresponding amount, remain as nonzero values.

Identifying and realigning bad intervals The time alignment pre-processing described above may fail to correctly identify delays, resulting in intervals of consecutive frames with disturbances above a trained threshold. Bad intervals identified as thus are realigned; new delay values are estimated by locating the maximum cross-correlations between the absolute reference and test signals pre-compensated with the delays observed during pre-processing. Disturbances for the bad intervals are recomputed and, if smaller, replace the original disturbances.

Disturbance aggregation In the last processing step of the PESQ algorithm, symmetric and asymmetric per-cell disturbances are aggregated separately in time and frequency and then linearly combined to calculate the perceived overall speech quality for the entire test speech file:

- **Aggregation in frequency:** Symmetric and asymmetric disturbances are first integrated along the frequency axis using two different L_p-norms, giving a per-frame measure of perceived distortion. A series of constants proportional to the width of the modified Bark bins are used such that a bin's disturbance is weighted in the L_p-norm by the bin's width on the perceptual modified Bark scale. The two weighted L_p-norms—symmetric and asymmetric—thus obtained are then multiplied by a factor inversely proportional to the power of the reference signal frame such that disturbances for low-intensity reference frames are emphasized.
- **Aggregation in time:** To model the property whereby temporally localized errors dominate perception, the symmetric and asymmetric frame disturbances obtained above are aggregated in time on two different time scales. First, frame

disturbances are aggregated over split-second intervals of approximately 320 ms using L_6-norms. The obtained split-second disturbances are then aggregated over the active interval of the speech file using L_2-norms. The value of p is higher for aggregation over the shorter split-second intervals to give higher weight to localized distortions.

- **PESQ score calculation:** Finally, the average symmetric and asymmetric disturbance values are linearly combined to calculate a PESQ score whose range is −0.5 to 4.5.

A reference ANSI-C implementation for the PESQ algorithm above is provided in Annex A of the ITU-T P.862 Recommendation [120].

B.2 Training and Optimization

The various PESQ model parameters employed in the auditory transform and disturbance processing were optimized on a large set of subjective experiments such that the highest average correlation coefficient is achieved with subjective MOS scores. In particular, as described in [121, Section 4; 122, Section 2.7], 30 subjective tests covering a wide range of conditions were used in the final training of the model. Starting with a large number of symmetric and asymmetric disturbance parameters calculated for each of the subjective test conditions, training was performed in an iterative manner in order to jointly optimize the various components of the model—i.e., maximize the correlation of the final PESQ scores with subjective quality—while minimizing the risk of over-training associated with training using a large set of separate parameters.

As noted in [120], the PESQ model parameters obtained by optimization as such lead to MOS-like PESQ scores between 1.0 (bad) and 4.5 (no distortion) in most cases. With extremely high distortions, however, PESQ scores may fall below 1.0, although this is very uncommon [121, 122].

References

[1] A. Gabrielsson, B. N. Schenkman, and B. Hagerman, "The effects of different frequency responses on sound quality judgments and speech intelligibility," *J. Speech Hear. Res.*, vol. 31, no. 2, pp. 166–177, June 1988. [Cited on pages 2, 12, and 82]

[2] J. Rodman, "The effect of bandwidth on speech intelligibility." White paper, Polycom®, January 2003. Available online at http://support.polycom.com/global/documents/support/technical/products/voice/soundstation_vtx1000_wp_effect_bandwidth_speech_intelligibility.pdf. [Cited on page 2]

[3] A. G. Bell, "Improvement in Telegraphy." U.S. Patent 174,465, March 1876. [Cited on page 2]

[4] B. M. Oliver, J. R. Pierce, and C. E. Shannon, "The philosophy of PCM," *Proc. IRE*, vol. 36, no. 11, pp. 1324–1331, November 1948. [Cited on pages 3 and 11]

[5] W. H. Martin, "Transmitted frequency range for telephone message circuits," *Bell Sys. Tech. J.*, vol. 9, no. 3, pp. 483–486, July 1930. [Cited on pages 3, 4, and 285]

[6] G. Wilkinson, "The new audiometry," *J. Laryngology & Otology*, vol. 40, no. 8, pp. 538–548, August 1925. [Cited on page 3]

[7] A. H. Inglis, "Transmission features of the new telephone sets," *Bell Sys. Tech. J.*, vol. 17, no. 3, pp. 358–380, July 1938. [Cited on page 4]

[8] ITU-T Recommendation G.232, "12-channel terminal equipments," November 1988. [Cited on pages 4, 5, and 285]

[9] ITU-T Recommendation G.712, "Transmission performance characteristics of pulse code modulation channels," November 2001. [Cited on pages 4, 65, and 285]

[10] D. O'Shaughnessy, *Speech Communications: Human and Machine*. Piscataway, NJ, USA: Wiley-IEEE Press, second ed., 1999. [Cited on pages 5, 6, 8, 9, 12, 13, 14, 16, 48, 67, 70, 83, 100, 102, 103, 140, 194, 216, 304, 307, 308, 309, 310, and 311]

[11] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, January 1947. [Cited on pages 7, 10, and 67]

[12] I. B. Crandall, "The composition of speech," *Phys. Rev.*, vol. 10, no. 1, pp. 74–76, July 1917. [Cited on pages 7 and 10]

[13] A. M. A. Ali, J. van der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of fricatives," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2217–2235, May 2001. [Cited on pages 7 and 8]

[14] G. E. Peterson and H. L. Barney, "Control methods used in a study of vowels," *J. Acoust. Soc. Am.*, vol. 24, no. 2, pp. 175–184, March 1952. [Cited on page 9]

[15] G. A. Campbell, "Telephonic intelligibility," *Phil. Mag.*, vol. 19, no. 6, pp. 152–159, January 1910. [Cited on page 10]

[16] H. Fletcher, *Speech and Hearing*. New York, NY, USA: D. Van Nostrand Company, Inc., 1929. [Cited on page 10]

[17] J. B. Allen, "How do humans process and recognize speech?," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 567–577, October 1994. [Cited on page 10]

[18] H. Fletcher, "The nature of speech and its interpretation," *Bell Sys. Tech. J.*, vol. 1, no. 1, pp. 129–144, July 1922. [Cited on page 10]

[19] H. Fletcher and R. H. Galt, "The perception of speech and its relation to telephony," *J. Acoust. Soc. Am.*, vol. 22, no. 2, pp. 89–151, March 1950. [Cited on page 10]

[20] H. Fletcher, "Hearing, the determining factor for high-fidelity transmission," *Proc. IRE*, vol. 30, no. 6, pp. 266–277, June 1942. [Cited on page 10]

[21] J. D. Harris, H. L. Haines, and C. K. Myers, "The importance of hearing at 3 KC for understanding speeded speech," *Laryngoscope*, vol. 70, no. 2, pp. 131–146, February 1960. [Cited on page 10]

[22] R. A. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey, "The contribution of consonants versus vowels to word recognition in fluent speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Atlanta, GA, USA, pp. II 853–856, May 1996. [Cited on page 10]

[23] ANSI S3.5-1969, "American national standard: Methods for the calculation of the Articulation Index," 1969. [Cited on page 10]

[24] ANSI S3.5-1997, "American national standard: Methods for the calculation of the Speech Intelligibility Index," 1997. [Cited on page 10]

[25] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, January 1949. [Cited on page 11]

[26] E. Meijering, "A chronology of interpolation: From ancient astronomy to modern signal and image processing," *Proc. IEEE*, vol. 90, no. 3, pp. 319–342, March 2002. [Cited on page 11]

[27] S. Voran, "Listener ratings of speech passbands," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Pocono Manor, PA, USA, pp. 81–82, September 1997. [Cited on pages 11, 12, 65, 66, 298, and 306]

[28] ITU-T Recommendation G.722, "7 kHz audio-coding within 64 kbit/s," November 1988. [Cited on pages 11 and 14]

[29] M. Oshikiri, H. Ehara, and K. Yoshida, "A scalable coder designed for 10-kHz bandwidth speech," in *Proc. IEEE Workshop on Speech Coding*, Tsukuba City, Japan, pp. 111–113, October 2002. [Cited on pages 12 and 14]

[30] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," August 1996. [Cited on pages 12 and 183]

[31] M. Oshikiri, H. Ehara, and K. Yoshida, "Efficient spectrum coding for super-wideband speech and its application to 7/10/15 kHz bandwidth scalable coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Montréal, QC, Canada, pp. I 481–484, May 2004. [Cited on pages 12 and 14]

[32] R. V. Cox, "Three new speech coders from the ITU cover a range of applications," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 40–47, September 1997. [Cited on page 13]

[33] 3GPP Recommendation TS 26.290, "Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions," September 2004. [Cited on page 14]

[34] ITU-T Recommendation G.722.2, "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)," July 2003. [Cited on pages 14 and 25]

[35] M. Yong, G. Davidson, and A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, New York, NY, USA, vol. 1, pp. 402–405, April 1988. [Cited on page 16]

[36] M. R. Zad-Issa and P. Kabal, "Smoothing the evolution of the spectral parameters in linear prediction of speech using target matching," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Munich, Gemrany, vol. 3, pp. 1699–1702, April 1997. [Cited on page 16]

[37] J. Samuelsson and P. Hedelin, "Recursive coding of spectrum parameters," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 492–503, July 2001. [Cited on pages 16 and 188]

[38] T. Eriksson and F. Norden, "Memory vector quantization by power series expansion [in speech coding]," in *Proc. IEEE Workshop on Speech Coding*, Tsukuba City, Japan, pp. 141–143, October 2002. [Cited on page 16]

[39] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Process.*, vol. 83, no. 8, pp. 1707–1719, August 2003. [Cited on pages 17, 33, 34, 49, 50, 56, 83, 100, 139, 149, 186, 187, 219, 279, 281, 287, and 298]

[40] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, January 1995. [Cited on pages 21, 45, 78, 79, 289, and 305]

[41] B. Iser and G. Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Proc. European Conf. Speech, Commun. Tech., EUROSPEECH*, Geneva, Switzerland, pp. 565–568, September 2003. [Cited on pages 25, 41, 83, and 185]

[42] H. Yasukawa, "Quality enhancement of band limited speech by filtering and multi-rate techniques.," in *Proc. Int. Conf. Spoken Language Process., ICSLP*, Yokohama, Japan, pp. 1607–1610, September 1994. [Cited on page 27]

[43] L. Laaksonen, J. Kontio, and P. Alku, "Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Philadelphia, PA, USA, pp. I 809–812, March 2005. [Cited on pages 27 and 183]

[44] E. Hänsler and G. Schmidt, *Speech and Audio Processing in Adverse Environments*. Berlin, Germany: Springer, 2008. [Cited on pages 27 and 28]

[45] H. Yasukawa, "Signal restoration of broad band speech using nonlinear processing," in *Proc. European Signal Process. Conf., EUSIPCO*, Trieste, Italy, pp. 987–990, September 1996. [Cited on page 28]

[46] G. Fant, *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton, 1960. [Cited on page 29]

[47] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*. Upper Saddle River, NJ, USA: Pearson-Prentice Hall, fourth ed., 2007. [Cited on pages 29 and 156]

[48] H.-M. Zhang and P. Duhamel, "On the methods for solving Yule-Walker equations," *IEEE Trans. Signal Process.*, vol. 40, no. 12, pp. 2987–3000, December 1992. [Cited on page 30]

[49] Y. Yoshida and M. Abe, "An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping," in *Proc. Int. Conf. Spoken Language Process., ICSLP*, Yokohama, Japan, pp. 1591–1594, September 1994. [Cited on pages 30, 37, and 287]

[50] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proc. European Signal Process. Conf., EUSIPCO*, Edinburgh, UK, pp. 1178–11181, September 1994. [Cited on pages 30, 32, 37, 86, and 287]

[51] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Washington, DC, USA, vol. 4, pp. 428–431, April 1979. [Cited on pages 31 and 32]

[52] C. K. Un and D. T. Magill, "The residual-excited linear prediction vocoder with transmission rate below 9.6 kbits/s," *IEEE Trans. Commun.*, vol. 23, no. 12, pp. 1466–1474, December 1975. [Cited on page 31]

[53] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Tampa, FL, USA, pp. 937–940, March 1985. [Cited on page 31]

[54] Y. Qian and P. Kabal, "Dual-mode wideband speech recovery from narrowband speech," in *Proc. European Conf. Speech, Commun. Tech., EUROSPEECH*, Geneva, Switzerland, pp. 1433–1437, September 2003. [Cited on pages 32, 35, 46, 48, 66, 67, 68, 139, 278, and 279]

[55] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Montréal, QC, Canada, pp. I 713–716, May 2004. [Cited on pages 32, 35, 46, 50, 55, 59, 64, 65, 66, 67, 72, 83, 162, 270, 278, 280, 288, 297, and 305]

[56] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech using piecewise linear mapping," in *Proc. European Conf. Speech, Commun. Tech., EUROSPEECH*, Rhodes, Greece, pp. 1643–1646, September 1997. [Cited on pages 32, 36, 38, 40, 83, and 185]

[57] M. Nilsson and W. B. Kleijn, "Avoiding over-estimation in bandwidth extension of telephony speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Salt Lake City, UT, USA, vol. 2, pp. 869–872, May 2001. [Cited on pages 32, 56, and 270]

[58] C. Avendano, H. Hermansky, and E. A. Wan, "Beyond Nyquist: Towards the recovery of broad-bandwidth speech from narrow-bandwidth speech," in *Proc. European Conf. Speech, Commun. Tech., EUROSPEECH*, Madrid, Spain, pp. 165–168, September 1995. [Cited on pages 32, 36, 56, and 287]

[59] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *Proc. IEEE Workshop on Speech Coding*, Porvoo, Finland, pp. 171–173, June 1999. [Cited on pages 37, 38, 64, 149, 150, and 189]

[60] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Salt Lake City, UT, USA, vol. 1, pp. 665–668, May 2001. [Cited on pages 32, 36, 51, 64, 183, 186, 280, and 281]

[61] J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP J. Appl. Signal Process.*, vol. 2001, no. 4, pp. 266–274, December 2001. [Cited on pages 33 and 55]

[62] S. Vaseghi, E. Zavarehei, and Q. Yan, "Speech bandwidth extension: Extrapolations of spectral envelope and harmonicity quality of excitation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Toulouse, France, pp. III 844–847, May 2006. [Cited on pages 34, 35, and 83]

[63] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *Proc. IEEE Workshop on Speech Coding*, Porvoo, Finland, pp. 174–176, June 1999. [Cited on pages 36, 37, 38, 52, 83, 150, 279, and 280]

[64] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley-Interscience, second ed., 2006. [Cited on pages 37, 108, 109, 130, and 166]

[65] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, January 1980. [Cited on page 37]

[66] C.-F. Chan and W.-K. Hui, "Wideband re-synthesis of narrowband CELP-coded speech using multiband excitation model," in *Proc. Int. Conf. Spoken Language Process., ICSLP*, Philadelphia, PA, USA, vol. 1, pp. 322–325, October 1996. [Cited on pages 37, 57, 58, and 64]

[67] C.-F. Chan and W.-K. Hui, "Quality enhancement of narrowband CELP-coded speech via wideband harmonic re-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Munich, Germany, vol. 2, pp. 1187–1191, April 1997. [Cited on pages 37 and 83]

[68] I. Y. Soon and C. K. Yeo, "Bandwidth extension of narrowband speech using soft-decision vector quantization," in *Proc. IEEE Int. Conf. Inform., Commun., Signal Process., ICICS*, Bangkok, Thailand, pp. 734–738, December 2005. [Cited on pages 38, 52, and 83]

[69] Y. Qian and P. Kabal, "Wideband speech recovery from narrowband speech using classified codebook mapping," in *Proc. Australian Int. Conf. Speech Science, Tech.*, Melbourne, Australia, pp. 106–111, December 2002. [Cited on pages 39, 139, 278, 279, and 280]

[70] Y. Tanaka and N. Hatazoe, "Reconstruction of wideband speech from telephone-band speech by multi-layer neural networks," *Spring meeting of ASJ (Acoustical Society of Japan)*, pp. 255–256, March 1995. In Japanese. [Cited on pages 39 and 185]

[71] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: Wiley-Interscience, second ed., 2001. [Cited on pages 39, 40, 76, 106, 127, 200, and 221]

[72] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice Hall, second ed., 1999. [Cited on pages 39 and 41]

[73] A. E. Bryson and Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control*. Waltham, MA, USA: Blaisdell, 1969. [Cited on page 39]

[74] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 544–548, October 1994. [Cited on pages 42 and 44]

[75] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, no. 1, pp. 67–72, February 1975. [Cited on pages 43 and 86]

[76] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via EM algorithm," *J. Royal Stat. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977. [Cited on pages 43 and 69]

[77] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, February 1989. [Cited on pages 45 and 48]

[78] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Seattle, WA, USA, vol. 1, pp. 285–288, May 1998. [Cited on pages 45, 46, and 306]

[79] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," in *Proc. European Conf. Speech, Commun. Tech., EUROSPEECH*, Madrid, Spain, pp. 447–450, September 1995. [Cited on pages 45 and 47]

[80] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465 – 479, July 1971. [Cited on pages 45 and 46]

[81] H. Fischer, *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. New York, NY: Springer, 2010. [Cited on page 46]

[82] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Istanbul, Turkey, vol. 3, pp. 1843–1846, June 2000. [Cited on pages 46, 47, 50, 55, 76, 83, 162, 184, 189, and 305]

[83] A. H. Nour-Eldin, "Robust automatic recognition of bluetooth speech," Master's thesis, INRS-ÉMT, Université du Québec, 2003. [Cited on page 48]

[84] M. Hosoki, T. Nagai, and A. Kurematsu, "Speech signal band width extension and noise removal using subband HMM," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Orlando, FL, USA, pp. I 245–248, May 2002. [Cited on pages 48, 49, 51, 55, 100, 186, 270, and 279]

[85] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, no. 1, pp. 164–171, February 1970. [Cited on page 48]

[86] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 13, no. 2, pp. 260–269, April 1967. [Cited on pages 49 and 186]

[87] G. Chen and V. Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Montréal, QC, Canada, pp. I 709–712, May 2004. [Cited on pages 49, 50, 64, 100, 160, 161, 183, 186, 187, 279, 280, 281, 282, 297, and 299]

[88] J.-M. Valin and R. Lefebvre, "Bandwidth extension of narrowband speech for low bit-rate wideband coding," in *Proc. IEEE Workshop on Speech Coding*, Delavan, WI, USA, pp. 130–132, September 2000. [Cited on pages 56, 66, and 287]

[89] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), ch. 4, pp. 121–173, Amsterdam, Netherlands: Elsevier, 1995. [Cited on page 57]

[90] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 8, pp. 1223–1235, August 1988. [Cited on page 57]

[91] J. Epps and W. H. Holmes, "Speech enhancement using STC-based bandwidth extension," in *Proc. Int. Conf. Spoken Language Process., ICSLP*, Sydney, Australia, vol. 2, pp. 519–522, December 1998. [Cited on pages 57, 58, and 150]

[92] P. Kabal, "Linear-phase FIR filter design tools." MATLAB® Central File Exchange: File 24662, July 2009. Available online at http://www.mathworks.com/matlabcentral/fileexchange/24662. [Cited on page 60]

[93] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. 57, no. Supplement 1, pp. S35–S35, April 1975. [Cited on pages 61 and 101]

[94] F. K. Soong and B.-W. Juang, "Line spectrum pair LSP and speech data compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, San Diego, CA, USA, pp. 1.10.1–1.10.4, March 1984. [Cited on page 63]

[95] H. W. Schüssler, "A stability theorem for discrete systems," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 1, pp. 87–89, February 1976. [Cited on page 63]

[96] T. Bäckström and C. Magi, "Properties of line spectrum pair polynomials–A review," *Signal Process.*, vol. 86, no. 11, pp. 3286–3298, November 2006. [Cited on page 63]

[97] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. 28, no. 2, pp. 129–137, March 1982. [Cited on pages 69 and 110]

[98] P. Kabal, "Time windows for linear prediction of speech." Technical report, Department of Electrical & Computer Engineering, McGill University, November 2009. Available online at http://www-mmsp.ece.mcgill.ca/Documents/Reports/2009/KabalR2009b.pdf. [Cited on pages 70, 71, and 72]

[99] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975. [Cited on pages 71 and 113]

[100] Y. Tohkura, F. Itakura, and S. Hashimoto, "Spectral smoothing technique in PARCOR speech analysis-synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 6, pp. 587–596, December 1978. [Cited on page 71]

[101] Y. Tohkura and F. Itakura, "Spectral sensitivity analysis of PARCOR parameters for speech data compression," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 3, pp. 273–280, June 1979. [Cited on page 71]

[102] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, no. 3, pp. 309–321, June 1975. [Cited on pages 71 and 73]

[103] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized linear prediction of speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 65–73, January 2008. [Cited on page 72]

[104] P. Kabal, "Ill-conditioning and bandwidth expansion in linear prediction of speech." Technical report, Department of Electrical & Computer Engineering, McGill University, February 2003. Available online at http://www-mmsp.ece.mcgill.ca/Documents/Reports/2003/KabalR2003a.pdf. [Cited on pages 72 and 73]

[105] P. Kabal, "Ill-conditioning and bandwidth expansion in linear prediction of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Hong Kong, Hong Kong, pp. I 824–827, April 2003. [Cited on pages 72 and 73]

[106] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research data base: Specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, Palo Alto, CA, USA, pp. 93–99, February 1986. [Cited on page 73]

[107] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. Hoboken, NJ, USA: Wiley-Interscience, 1982. [Cited on page 76]

[108] G. H. Golub and C. F. van Loan, *Matrix Computations*. Baltimore, MD, USA: The Johns Hopkins University Press, third ed., 1996. [Cited on pages 79, 94, 246, and 247]

[109] M. Nilsson, H. Gustafsson, S. V. Andersen, and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Orlando, FL, USA, pp. I 525–528, May 2002. [Cited on pages 81, 85, 97, 99, 101, 104, 107, 108, 109, 112, 115, 286, 289, and 290]

[110] A. H. Gray, Jr. and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 380–391, October 1976. [Cited on pages 84, 85, 86, and 87]

[111] P. Hedelin and J. Skoglund, "Vector quantization based on Gaussian mixture models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 385–401, July 2000. [Cited on pages 85 and 108]

[112] W. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Hartford, CT, USA, vol. 2, pp. 204–207, May 1977. [Cited on page 85]

[113] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ, USA: Prentice Hall, 1988. [Cited on pages 85 and 86]

[114] J. L. Flanagan, "Difference limen for the intensity of a vowel sound," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1223–1225, November 1955. [Cited on page 85]

[115] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 1, pp. 3–14, January 1993. [Cited on pages 85, 86, 101, 109, 110, 116, and 290]

[116] F. ltakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. Japan*, vol. 53-A, no. 1, pp. 36–43, 1970. [Cited on page 86]

[117] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y.Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 367–376, August 1980. [Cited on page 86]

[118] G. Chen, S. N. Koh, and I. Y. Soon, "Enhanced Itakura measure incorporating masking properties of human auditory system," *Signal Process.*, vol. 83, no. 7, pp. 1445–1456, July 2003. [Cited on page 86]

[119] ITU-T Recommendation P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," November 2005. [Cited on pages 88 and 313]

[120] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," February 2001. [Cited on pages 88, 89, 90, 313, and 317]

[121] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II – Psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 765–778, October 2002. [Cited on pages 89, 90, 313, and 317]

[122] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Salt Lake City, UT, USA, vol. 2, pp. 749–752, May 2001. [Cited on pages 90, 313, and 317]

[123] T. Minka, "Lightspeed toolbox: Efficient operations for MATLAB® programming," May 2011. Version 2.6. © Microsoft Corporation. Available online at http://research.microsoft.com/en-us/um/people/minka/software/lightspeed. [Cited on page 94]

[124] M. Nilsson, S. V. Andersen, and W. B. Kleijn, "On the mutual information between frequency bands in speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Istanbul, Turkey, vol. 3, pp. 1327–1330, June 2000. [Cited on pages 99 and 286]

[125] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Orlando, FL, USA, pp. I 237–240, May 2002. [Cited on pages 99, 101, 108, 115, 120, 121, 122, 286, and 290]

[126] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Montréal, QC, Canada, pp. I 697–700, May 2004. [Cited on pages 99, 101, 106, 191, 291, and 293]

[127] H. Hermansky and S. Sharma, "TRAPS — Classifiers of temporal patterns," in *Proc. Int. Conf. Spoken Language Process., ICSLP*, Sydney, Australia, vol. 3, pp. 1003–1006, December 1998. [Cited on page 100]

[128] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Munich, Germany, vol. 3, pp. 1647–1650, April 1997. [Cited on pages 100 and 140]

[129] H. Pulakka, V. Myllylä, L. Laaksonen, and P. Alku, "Bandwidth extension of telephone speech using a filter bank implementation for highband mel spectrum," in *Proc. European Signal Process. Conf., EUSIPCO*, Aalborg, Denmark, pp. 979–983, August 2010. [Cited on pages 100, 160, 183, and 185]

[130] U. Kornagel, "Spectral widening of telephone speech using an extended classification approach," in *Proc. European Signal Process. Conf., EUSIPCO*, Toulouse, France, pp. 339–342, September 2002. [Cited on pages 187, 188, and 278]

[131] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Philadelphia, PA, USA, pp. I 805–808, March 2005. [Cited on pages 187, 188, 278, 279, 280, 281, and 297]

[132] K.-T. Kim, M.-K. Lee, and H.-G. Kang, "Speech bandwidth extension using temporal envelope modeling," *IEEE Signal Process. Lett.*, vol. 15, pp. 429–432, 2008. [Cited on pages 100, 160, 161, 162, and 184]

[133] S. Yao and C.-F. Chan, "Speech bandwidth enhancement using state space speech dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Toulouse, France, pp. I 489–492, May 2006. [Cited on pages 100, 188, 189, 280, and 282]

[134] A. H. Nour-Eldin, T. Z. Shabestary, and P. Kabal, "The effect of memory inclusion on mutual information between speech frequency bands," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Toulouse, France, pp. III 53–56, May 2006. [Cited on page 100]

[135] A. H. Nour-Eldin and P. Kabal, "Objective analysis of the effect of memory inclusion on bandwidth extension of narrowband speech," in *Proc. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, Antwerp, Belgium, pp. 2489–2492, August 2007. [Cited on pages 100, 291, and 293]

[136] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 254–272, April 1981. [Cited on pages 100, 125, and 126]

[137] P. Mermelstein, "Distance measures for speech recognition—psychological and instrumental," in *Pattern Recognition and Artificial Intelligence* (C. H. Chen, ed.), pp. 374–388, New York, NY, USA: Academic, 1976. [Cited on pages 101 and 103]

[138] S. S. Stevens and J. Volkmann, "The relation of pitch to frequency: A revised scale," *Am. J. Psych.*, vol. 53, no. 3, pp. 329–353, July 1940. [Cited on page 102]

[139] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical band width in loudness summation," *J. Acoust. Soc. Am.*, vol. 29, no. 5, pp. 548–557, May 1957. [Cited on pages 102 and 104]

[140] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *J. Acoust. Soc. Am.*, vol. 33, no. 2, pp. 248–248, February 1961. [Cited on page 103]

[141] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, January 1974. [Cited on page 105]

[142] P. E. Pfeiffer, *Concepts of Probability Theory*. Mineola, NY, USA: Dover Publications, Inc., second ed., 1978. [Cited on page 107]

[143] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997. [Cited on page 109]

[144] W. B. Kleijn, "A basis for source coding." Lecture notes, KTH (Royal Institute of Technology) Stockholm, July 2004. [Cited on page 109]

[145] W. R. Bennett, "Spectra of quantized signals," *Bell Sys. Tech. J.*, vol. 27, no. 3, pp. 446–472, July 1948. [Cited on page 110]

[146] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization and the vector quantizer advantage," *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1020–1033, September 1989. [Cited on page 112]

[147] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall, 1993. [Cited on page 121]

[148] R. Hagen, "Spectral quantization of cepstral coefficients," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Adelaide, Australia, pp. I 509–512, April 1994. [Cited on page 121]

[149] B. Milner, "Inclusion of temporal information into features for speech recognition," in *Proc. Int. Conf. Spoken Language Process., ICSLP*, Philadelphia, PA, USA, vol. 1, pp. 256–259, October 1996. [Cited on pages 127, 128, 191, 299, and 300]

[150] A. H. Nour-Eldin and P. Kabal, "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech," in *Proc. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, Brisbane, Australia, pp. 53–56, September 2008. [Cited on page 145]

[151] T. Ramabadran, J. Meunier, M. Jasiuk, and B. Kushner, "Enhancing distributed speech recognition with back-end speech reconstruction," in *Proc. European Conf. Speech, Commun. Tech., EUROSPEECH*, Aalborg, Denmark, pp. 1859–1862, September 2001. [Cited on pages 145, 146, 150, 156, 157, and 293]

[152] B. Milner and X. Shao, "Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model," in *Proc. Int. Conf. Spoken Language Process., ICSLP*, Denver, CO, USA, pp. 2421–2424, October 2002. [Cited on pages 145 and 156]

[153] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch frequency," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Istanbul, Turkey, vol. 3, pp. 1299–1302, June 2000. [Cited on pages 146, 150, and 157]

[154] W. Pan and X. Shen, "Penalized model-based clustering with application to variable selection," *J. Mach. Learn. Res.*, vol. 8, pp. 1145–1164, May 2007. [Cited on pages 147, 191, 204, and 306]

[155] D. L. Elliot, "Covariance regularization in mixture of gaussians for high-dimensional image classification," Master's thesis, Department of Computer Science, Colorado State University, 2009. [Cited on page 191]

[156] A. Krishnamurthy, "High-dimensional clustering with sparse Gaussian mixture models." Unpublished paper, 2011. Available online at `www.cs.cmu.edu/~akshaykr/files/sgmm_paper.pdf`. [Cited on pages 191, 192, and 204]

[157] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Series B*, vol. 58, no. 1, pp. 267–288, 1996. [Cited on page 192]

[158] C. Bouveyron, S. Girard, and C. Schmid, "High-dimensional data clustering," *Comput. Stat. Data Anal.*, vol. 52, no. 1, pp. 502–519, 2007. [Cited on pages 147, 192, 198, and 306]

[159] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," in *Proc. European Conf. Speech, Commun. Tech., EUROSPEECH*, Geneva, Switzerland, pp. 2413–2416, September 2003. [Cited on pages 147, 190, 191, and 305]

[160] L. Mesbahi, V. Barreaud, and O. Boëffard, "Comparing GMM-based speech transformation systems," in *Proc. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, Antwerp, Belgium, pp. 1989–1992, August 2007. [Cited on pages 191 and 249]

[161] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Philadelphia, PA, USA, pp. I 9–12, March 2005. [Cited on pages 147, 191, and 305]

[162] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE-ASSP*, vol. 30, no. 4, pp. 679–681, August 1982. [Cited on page 158]

[163] C. Yağli and E. Erzin, "Artifical bandwidth extension of spectral envelope with temporal clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Prague, Czech Republic, pp. 5096–5099, May 2011. [Cited on pages 160, 161, 183, 186, 187, 279, 280, 281, 282, and 299]

[164] K.-T. Kim, J.-Y. Choi, and H.-G. Kang, "Perceptual relevance of the temporal envelope to the speech signal in the 4–7 kHz band," *J. Acoust. Soc. Am.*, vol. 122, no. 3, pp. EL88–EL94, August 2007. [Cited on pages 161 and 162]

[165] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," Juanuary 2003. [Cited on page 162]

[166] D. L. Clark, "High-resolution subjective testing using a double-blind comparator," *J. Audio Eng. Soc.*, vol. 30, no. 5, pp. 330–338, May 1982. [Cited on page 162]

[167] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, October 1995. [Cited on pages 170 and 307]

[168] B. Blesser, "Speech perception under conditions of spectral transformation: I. Phonetic characteristics," *J. Speech Hear. Res.*, vol. 15, no. 1, pp. 5–41, March 1972. [Cited on pages 170 and 311]

[169] J. Herre and M. Lutzky, "Perceptual audio coding of speech signals," in *Springer Handbook of Speech Processing* (J. Benesty, M. M. Sondhi, and Y. Huang, eds.), ch. 18, pp. 393–412, Berlin, Germany: Springer, 2008. [Cited on page 177]

[170] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ, USA: Prentice Hall, fourth ed., 2002. [Cited on page 189]

[171] S. Yao and C.-F. Chan, "Block-based bandwidth extension of narrowband speech signal by using CDHMM," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Philadelphia, PA, USA, pp. I 793–796, March 2005. [Cited on page 189]

[172] R. E. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton University Press, 1957. [Cited on page 190]

[173] K. P. Murphy, "An introduction to graphical models." Unpublished paper, May 2001. Available online at http://www.cs.ubc.ca/~murphyk/Papers/intro_gm.pdf. [Cited on page 191]

[174] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY, USA: Springer, second ed., 2009. [Cited on pages 191, 193, 198, 199, and 306]

[175] J. Hadamard, *Four Lectures on Mathematics: Delivered at Columbia University in 1911*. Columbia University Press, 1915. [Cited on page 192]

[176] L. Parsons, E. Haque, and H. Liu, "Evaluating subspace clustering algorithms," in *Proc. Workshop on Clustering High Dimensional Data and its Applications, SIAM Int. Conf. Data Mining*, pp. 48–56, April 2004. [Cited on page 192]

[177] A. H. Nour-Eldin and P. Kabal, "Memory-based approximation of the Gaussian mixture model framework for bandwidth extension of narrowband speech," in *Proc. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, Florence, Italy, pp. 1185–1188, August 2011. [Cited on page 193]

[178] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, March 2011. [Cited on page 193]

[179] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers, Inc., 1988. [Cited on pages 193, 202, 241, and 302]

[180] D. W. Scott and J. R. Thompson, "Probability density estimation in higher dimensions," in *Computer Science and Statistics: Proceeedings of the Fifteenth Symposium in the Interface* (J. E. Gentle, ed.), pp. 173–179, Amsterdam, New York: North Holland-Elsevier Science Publishers, 1983. [Cited on page 200]

[181] A. Kandel, *Fuzzy Techniques in Pattern Recognition*. New York, NY, USA: Wiley-Interscience, 1982. [Cited on page 200]

[182] A. Baraldi and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition—Part I," *IEEE Trans. Sys., Man, and Cybern., B*, vol. 29, no. 6, pp. 778–785, December 1999. [Cited on page 200]

[183] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, NY, USA: Plenum Press, 1981. [Cited on page 200]

[184] J. Zhang, M. A. Anastasio, X. Pan, and L. V. Wang, "Weighted expectation maximization reconstruction algorithms for thermoacoustic tomography," *IEEE Trans. Med. Imag.*, vol. 24, no. 6, pp. 817–820, June 2005. [Cited on page 201]

[185] Y. Matsuyama, "The α-EM algorithm and its basic properties," *Systems and Computers in Japan*, vol. 31, no. 11, pp. 12–23, October 2000. [Cited on page 201]

[186] R. M. Golden, *Mathematical Methods for Neural Network Analysis and Design*. Cambridge, MA, USA: MIT Press, 1996. [Cited on page 206]

[187] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models." Technical report TR-97-021, International Computer Science Institute, 1997. Available online at http://ssli.ee.washington.edu/~bilmes/mypubs/bilmes1997-em.pdf. [Cited on pages 219, 221, and 223]

[188] S. Borman, "The Expectation Maximization algorithm: A short tutorial." Unpublished paper, July 2004. Available online at http://www.cs.utah.edu/~piyush/teaching/EM_algorithm.pdf. [Cited on pages 219, 221, 223, 224, and 225]

[189] A. H. Gray, Jr. and J. D. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 3, pp. 207–217, June 1974. [Cited on page 235]

[190] F. Wray, "A brief future of computing." Featured article, PlanetHPC, Edinburgh Parallel Computing Centre, University of Edinburgh, November 2012. Available online at http://www.planethpc.eu/index.php?option=com_content&view=article&id=66:a-brief-future-of-computing. [Cited on page 257]

[191] K. Kumar, C. Kim, and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Prague, Czech Republic, pp. 4784–4787, May 2011. [Cited on page 299]

[192] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Voice convergin [*sic*]: Speaker de-identification by voice transformation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, Taipei, Taiwan, pp. 3909–3912, April 2009. [Cited on page 306]

[193] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Commun.*, vol. 50, no. 3, pp. 215–227, March 2008. [Cited on page 306]