Robust Bandwidth Extension of Narrowband Speech

Wei-shou Hsu

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

November 2004

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering.

Abstract

Telephone speech often sounds muffled and thin due to its narrowband characteristics. With the increased availability of terminals capable of receiving wideband signals, extending the bandwidth of narrowband telephone speech at the receiver has drawn much research interest. Currently, there exist many methods that can provide good reconstructions of the wideband spectra from narrowband speech; however, they often lack robustness to different channel conditions, and their performances degrade when they operate in unknown environments.

This thesis presents a bandwidth extension algorithm that mitigates the effects of adverse conditions. The proposed system is designed to work with noisy input speech and unknown channel frequency response. To maximize the naturalness of the reconstructed speech, the algorithm estimates the channel and applies equalization to recover the attenuated bands. Artifacts are reduced by employing an adaptive and a fixed postfilters.

Subjective test results suggest that the proposed scheme is not affected by channel conditions and is able to produce speech with enhanced quality in adverse environments.

Sommaire

Les conversations téléphoniques paraissent souvent étouffée, à cause de leurs caractéristiques de bande étroite. Avec une plus grande disponibilité des bornes téléphoniques capables de recevoir les signaux à large bande, prolonger la largeur de bande à bande étroite du signal téléphonique au récepteur a attiré beaucoup d'intérêt de recherches. Actuellement, il existe beaucoup de méthodes qui peuvent fournir de bonnes reconstructions des spectres à bande large à partir du discours à bande étroite; cependant, elles manquent souvent de robustesse aux variations du canal, et leurs performances se dégradent quand elles fonctionnent dans des environnements aux caractéristiques inconnues.

Cette thèse présente un algorithme de prolongation de la largeur de bande qui atténue les effets des conditions défavorables. Le système proposé est conçu pour fonctionner avec un signal de la parole en entrée bruyant et une réponse en fréquence du canal inconnue. Pour maximiser le naturel du discours reconstruit, l'algorithme fait une estimation du canal et applique l'égalisation pour récupérer les bandes atténuées. Les artéfacts sont réduits en utilisant un filtre adaptatif et un postfiltre fixe.

Les résultats de tests subjectifs suggèrent que l'arrangement proposé n'est pas affecté par les variations entre des canaux et est capable de produire le discours avec une qualité accrue dans les environnements défavorables.

Acknowledgments

I would like to thank my supervisor, Prof. Peter Kabal, for his invaluable advice and constant motivations. I am grateful to Prof. Kabal for providing financial support to carry on the research. I would also like to thank Mr. Yasheng Qian, whose bandwidth extension algorithm forms the basis of this thesis work, for his patient explanations on numerous occasions throughout my studies.

I wish to thank my fellow Telecommunications and Signal Processing Laboratory graduate students for the wonderful experiences I have had at the lab. In addition, I would like to acknowledge Alex for his help on various matters from general computer knowledge to questions on signal processing, Sien for the many tips he gave me regarding thesis formatting, and Denis for the French translation of the abstract. Furthermore, I would like to thank the volunteers who participated in the subjective listening tests.

I am indebted to my parents for their continual encouragements and unwavering support. I want to thank my family for being with me throughout the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human speech generally contains a perceptually significant amount of energy from 50 Hz up to 8 kHz. However, the current public switched telephone network (PSTN) transmits only a narrowband portion of the signal, in the frequency range of approximately between 300 Hz and 3.4 kHz. The result of this reduction in bandwidth is the familiar telephone speech, which sounds thin and muffled.

The pursuit of better speech quality has led to advancement in speech coding technology and made the sending and receiving of wideband speech possible. However, the transmission of such signal requires the construction of a new network that supports higher bitrates and larger bandwidth, which is expensive and time-consuming. With the public telephone networks already very well established, it would seem logical to utilize its infrastructure in some way so that cost of wideband speech can be reduced. Furthermore, since the PSTN is still the most widely used network with its users unlikely to abandon it anytime soon, the wideband terminals will have to be compatible with the current narrowband network to be useful for the foreseeable future.

Since the signals will be bandpass filtered when they pass through the narrowband network, the advantage offered by a wideband terminal would be lost when the terminal is connected to the PSTN. One possible solution is to artificially extend the signal bandwidth at the receiving end by reconstructing the wideband speech given the narrowband PSTN output. This process of regenerating wideband spectra from narrowband signals is referred to as bandwidth extension.

1.1 Basic Requirements of a Bandwidth Extension System

The ultimate goal of bandwidth extension is to generate a more natural sounding speech signal. To do so, the narrowband telephone speech should be retained as much as possible, as it is the true voice of the speaker.

Furthermore, in order to effectively integrate bandwidth extension with the current network, the algorithm should use no additional information other than the waveform of the narrowband signal. Therefore, one primary constraint for bandwidth extension is that the algorithm has to be able to predict the wideband spectrum solely based on the correlation between the known narrowband and the missing bands. Another practical constraint is that a user should not be able to notice any additional delay introduced by the extra processing in the receiving end. Therefore, a bandwidth extension algorithm should be able to work in real time.

1.2 History of Bandwidth Extension

The inferiority of telephone speech becomes very apparent when it can immediately be compared with a higher-quality broadcast speech recorded at a studio, such as the case when a listener listens to the host of a radio or TV show talking to a correspondent over telephone. The difference in quality can be particularly annoying for the listener, as different levels of concentration are needed to understand the two different types of speech. In the early 1970s, complaints from listeners and desire of providing better broadcast telephone speech led the BBC to one of the early attempts at recovering wideband speech [1].

However, due to the knowledge and technology available at that time, little attention was given to bandwidth extension implemented only at the receiving end, and no satisfactory results were reported. Most researchers would instead focus on altering both the transmitter and the receiver.

Although not much effort was made toward bandwidth extension during the 1970s and the 1980s, there were some works that had similar goals. For example, from the speech coding perspective, the goal of reducing the number of bits allocated to the high-frequency band naturally led to attempts, e.g. [2], in coding only partial information of the highband spectrum while recovering the rest from the narrowband signal at the receiving end. Another approach with similar objective is described in [3], where Patrick et al.

apply frequency mapping at the transmitter to send the high-frequency band of unvoiced phonemes over the telephone network. The signal would then be reconstructed at the receiver by the inverse frequency mapping operation.

Interest in bandwidth extension began to grow in the late eighties and early nineties thanks to increase in processing power and improvements in wideband speech coding technology. Since then, many new methods have been presented. Several methods [4–9] used codebook mapping between narrowband and wideband spectral parameter vectors for wideband reconstruction. Linear mappings using one or more matrices [10–12] have also been proposed. Methods based on statistical mapping were first discussed in [13] and later improved using Gaussian mixture model (GMM) [14–16] and Hidden Markov model (HMM) [17, 18]. Other attempts include methods based on multi-rate processing proposed by Yasukawa [19, 20], which have the advantage of being computationally simple.

1.3 Problems with Existing Bandwidth Extension Systems

Although much research has been done on bandwidth extension, there is a common drawback in most algorithms proposed so far — the systems are inflexible to different channel conditions.

Shown in Fig. 1.1 is a simple linear model of the transmission channel. As can be seen in the diagram, there are two main ways by which a transmitted speech signal can be distorted. One of the distortions come as a result of the bandlimited nature of the channel, which acts as a linear bandpass filter and behaves in a convolutive manner with respect to the signal. The other distortions appear in the form of additive noise, which can come from sources such as background noise at the transmitting end or quantization noise due to coding of the signal.

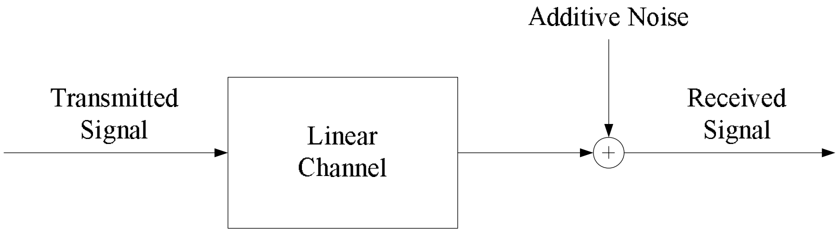


Fig. 1.1 A simple model of the transmission channel

Most algorithms so far, however, have made the assumptions that the frequency response of the channel is known and the speech is not corrupted by noise. Because of these assumptions, these systems are susceptible to channel mismatch and thus are prone to suboptimal and unpredictable performance. The remainder of this section will describe how different channel conditions might affect a bandwidth extension system.

1.3.1 Problem of Additive Noise

The primary effect of noise is obvious as it can drastically reduce the intelligibility and quality of the incoming speech. Fortunately, because of the significance and prevalence of the problem in various speech processing applications, noise reduction for speech enhancement has already been widely researched, and there exist systems that can provide satisfactory results for narrowband speech.

However, because noise is random, its values cannot be exactly determined and complete recovery of the clean speech is impossible. No matter how good the noise suppressor is, some information about the narrowband signal is still going to be lost or altered. If a bandwidth extension system does not take noise into consideration and does not attempt to salvage the lost information, the likely result is that the wideband spectrum of the noise, rather than that of the clean speech, will be reconstructed, thus causing further degradation.

1.3.2 Problem of Unknown Channel Response

Many bandwidth extension algorithms (e.g. [5, 10, 14, 21]) assume that the telephone channel passes the speech component between 300 and 3400 Hz while completely suppressing anything outside of this band. Some systems assume that the entire region below 4000 Hz is available (e.g. [8, 12, 18]). Other bandwidths that have been used include 300 – 3200 Hz [6] and 300 – 3500 Hz [7].

Systems that are based on having the entire 0 – 4000 Hz band available will have to work with a preprocessor that can reconstruct the frequency regions below 4 kHz but outside the passband of the channel. For systems that work directly with telephone-band speech, the problem is that not all networks have the same frequency response. For example, older networks might cut off the signal at 300 Hz and 3.2 kHz, while some other networks can transmit up to 3.6 kHz.

Since the channel response is not known, parameters extracted from the input speech might differ from those extracted under the designed condition, and the performance of the system might be degraded. Furthermore, there are risks of discarding available content if the assumed bandwidth of the telephone band is narrower than the actual bandwidth of the channel.

1.4 Thesis Contribution

As mentioned in the previous section, the lack of robustness to unknown channel conditions is a major problem to existing bandwidth extension systems and has yet to receive much attention. The objective of this thesis is to provide a robust bandwidth extension scheme that can handle various channel conditions.

Specifically, the system described in [22] will be used as the core in this thesis work, whose aim is to improve the performance of the system when the input speech is corrupted by noise and the channel response is unknown. The focus will be on robustness towards high-frequency extension, and the improvement will mainly come in the form of a preprocessor which can (1) suppress noise for quality enhancement and better parameter extraction and (2) estimate the channel frequency response and choose between two modes of recovery based on the amount of high-frequency content available below 4 kHz. Two postfilters are also designed to reduce possible artifacts produced by the system.

1.5 Thesis Organization

This thesis contains five chapters. Fundamentals of bandwidth extension are given in Chapter 2, including discussions on some of the existing methods and previous works in the area.

Chapter 3 describes the core bandwidth extension system, which comes from the algorithm used in [22]. This system serves as the starting point for this thesis work.

Chapter 4 discusses the effect of channel mismatch and presents the proposed scheme that aims at improving the performance of the core system under adverse conditions.

Chapter 5 evaluates the performance of the proposed system and presents test results and discussions.

Chapter 6 gives the conclusion and provides ideas for further improvement and future research.

Chapter 2

Bandwidth Extension

Due to the bandlimited nature of the conventional telephone networks, telephone speech carries only a narrowband portion of the original speech. The loss of the low- and high-frequency spectra not only degrades speech quality but also increases the difficulty in distinguishing fricatives, such as /s/ and /f/, since their major discriminating characteristics are in the lost high-frequency band.

The study of bandwidth extension endeavors to enhance the quality of narrowband speech by reconstructing the missing frequency bands outside the passband of conventional telephone networks. The reconstruction of wideband spectra has to be based solely on the information available in the telephone-band signal, so that it can be implemented only at the receiver and be compatible with existing infrastructure.

This chapter will describe the fundamentals of bandwidth extension and outline some methods that have been previously applied. It starts by presenting an important speech model, the linear separable source-filter model, which is followed by a discussion of bandwidth extension methods based on this model. Finally, the previous publications regarding practicability of bandwidth extension will be examined.

2.1 Linear Separable Source-Filter Model

The linear separable source-filter model, shown in Fig. 2.1, is widely used in various speech processing applications to model human speech production, and it is the basis of most existing bandwidth extension systems.

According to this model, speech is produced by a spectrally flat excitation source mod-

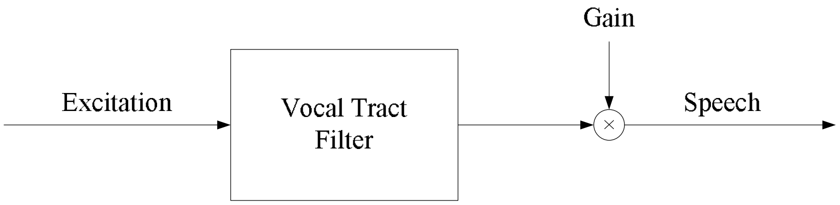
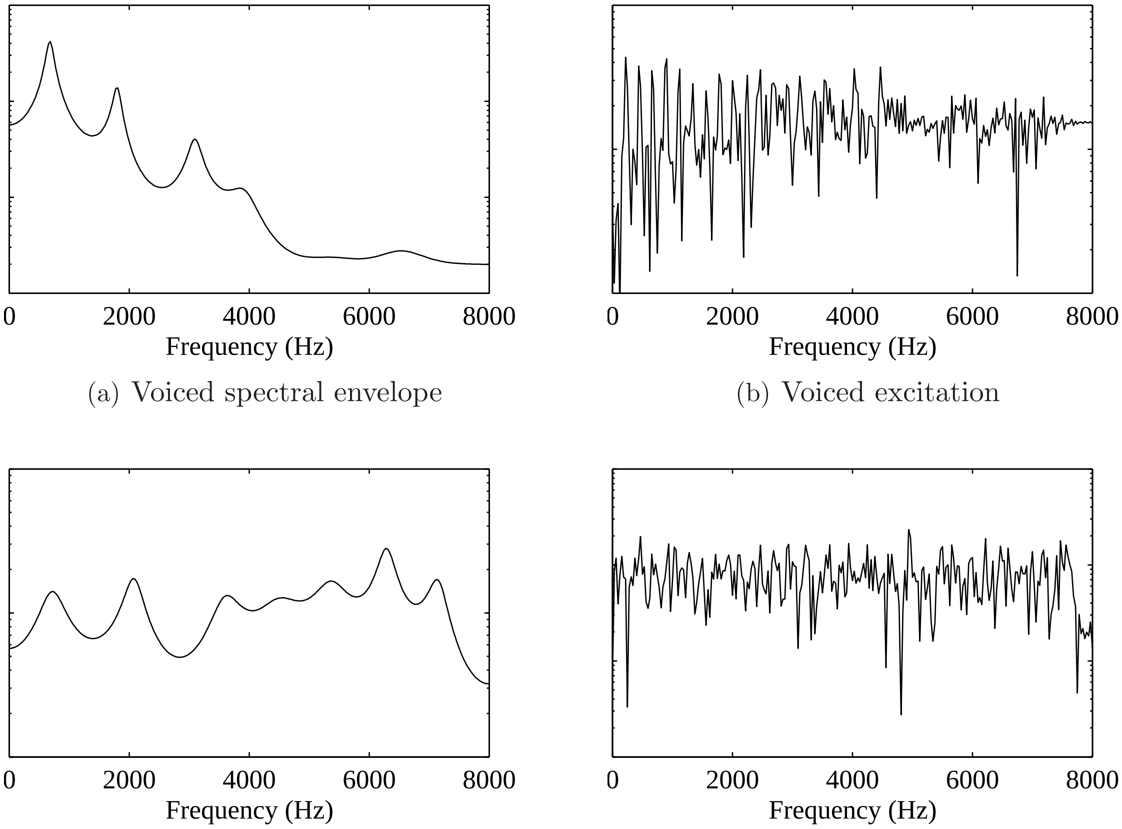


Fig. 2.1 Source-filter model of human speech production



(a) Voiced spectral envelope

(b) Voiced excitation

(c) Unvoiced spectral envelope

(d) Unvoiced excitation

Fig. 2.2 Examples of spectral envelopes and excitations

ulated by a linear spectral shaping filter corresponding to the vocal tract. The spectral shaping filter is commonly modelled as an all-pole filter using linear prediction (LP) analysis. The excitation signal, which accounts for the fine structure of the spectrum, can contain a pulse train, representing the glottal pulses during voiced sounds, and white noise, which comes from the turbulent air flow in unvoiced phonemes. Figure 2.2 shows examples of the frequency spectra of these two main components in the source-filter model obtained by a 16^{th}-order LP analysis.

One of the fundamental assumptions in the source-filter model is that the sound source and the spectral shaping filter are independent. Using this model, bandwidth extension can be separated into the tasks of recovering the wideband spectral envelope and the wideband excitation signal, the reconstruction of which can be further broken down to the generation of the excitation waveform and the estimation of its gain.

2.2 High-Frequency Band Spectral Envelope Regeneration

Highband spectral envelope regeneration can be accomplished by mapping narrowband features to highband parameters using some predefined transformations. The narrowband feature vector can include any parameter that can be derived from the narrowband signal. On the other hand, when modelled using LP analysis, the highband vector consists of the LP coefficients or other equivalent forms, such as the line spectral frequencies (LSF), which can later be used to reconstruct the highband spectrum by LP synthesis. In mathematical form, the highband spectral envelope estimation can be expressed as a transformation f mapping narrowband vector $\boldsymbol{x}$ to highband vector $\boldsymbol{y}$, i.e.,

$$\boldsymbol{y} = f(\boldsymbol{x}).$$

This section outlines some of the existing methods used for the transformations from narrowband to highband spectral envelope parameters.

2.2.1 Codebook Mapping

Many early bandwidth extension algorithms (e.g. [4–6]) are based on codebook mapping, which originates from the vector quantization techniques in signal compression. In vector quantization, a pre-trained codebook is used to encode an input vector by replacing it with the code vector that is closest to the input as determined by a predefined distance measure.

The codebook mapping method for bandwidth extension adds one extra step to the vector quantizer. Whereas a vector quantizer uses only one codebook, the codebook mapping method uses two codebooks, one for narrowband and the other for highband feature vectors. The two codebooks are trained together and have a one-to-one correspondence between their entries. The process of codebook mapping, as shown in Fig. 2.3, starts the same way as a vector quantizer by searching through the narrowband codebook for the code vector closest to the input narrowband feature vector. The algorithm then maps the optimal narrowband entry to the highband codebook and uses the corresponding highband feature vector as the estimate for the missing spectral envelope.

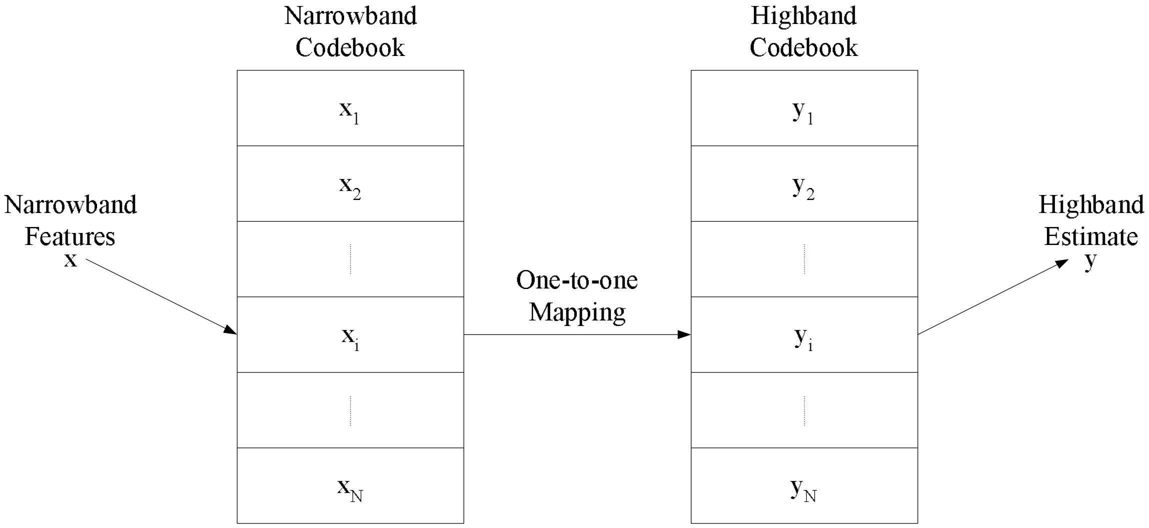


Fig. 2.3 Codebook mapping

The basic codebook mapping method can be improved by employing multiple codebooks. Because voiced and unvoiced sounds have very distinctive spectral shapes, different codebooks specialized for different voicing degrees are often used, as in [8, 9].

Another improvement that can be made to codebook mapping is to interpolate by taking a weighted average between K highband candidates that correspond to the K best narrowband vectors, where $K > 1$ [8]. By averaging the highband vectors, the algorithm can increase the number of possible outputs and reduce the probability of large distortions.

2.2.2 Linear Mapping

Another way of mapping from the narrowband feature vector to an estimate of the highband vector is by using a linear transformation, represented as matrix multiplications as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x},$$

where $\boldsymbol{x}$ is a vector of parameters extracted from the narrowband signal, $\boldsymbol{y}$ is the highband vector, and $\boldsymbol{A}$ is a matrix obtained during training.

Since the narrowband-highband correspondence is highly non-linear, linear mapping can be improved by modifying it to a piecewise operation [10, 12]. Similar to multiple codebook mapping, a piecewise linear mapping can be realized by employing more than one matrices, with each one being used for different types of phonemes.

2.2.3 Statistical Mapping

Codebook mapping and linear mapping can estimate the highband spectral envelope with little computational complexity. However, the deterministic nature of both methods makes them inflexible. With the increase in processing power, more complex methods based on statistical mapping can be employed.

Statistical mapping estimates highband parameters based on probabilistic measures. Given the input narrowband vector, an algorithm using statistical mapping can calculate the output based on measures such as maximum likelihood.

The pioneering work of statistical mapping is described in [13]. In [13], spectral envelopes are modelled as combinations of signals emitted from different random sources. Each source generates autoregressive spectral envelope parameters based on a Gaussian distribution. Furthermore, narrowband and highband sources are correlated according to a transition probability matrix

$$\boldsymbol{P} = [p_{ij}],$$

where p_{ij} is the probability that the highband speech is generated by the j^{th} highband source given that the narrowband portion is generated by the i^{th} narrowband source. The statistical model of random spectral envelope generators is illustrated in Fig. 2.4 with an example using three narrowband and two highband sources.

As pointed out in [23], statistical methods can be seen as generalizations of codebook

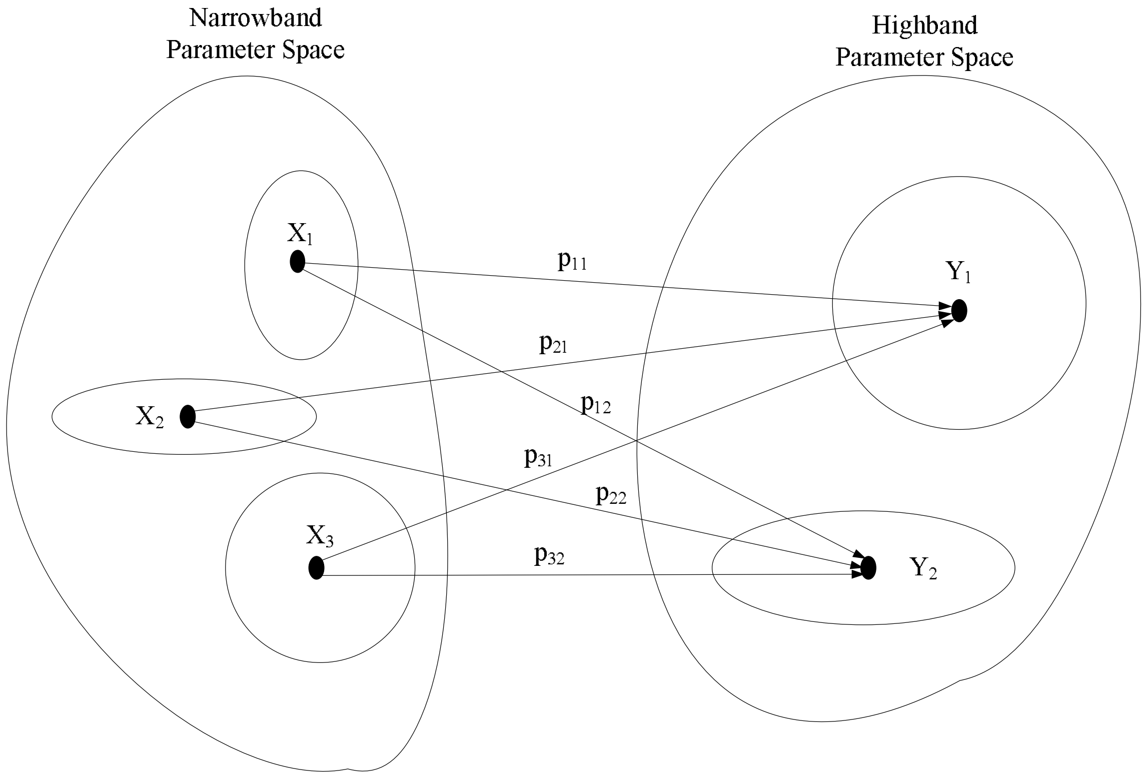


Fig. 2.4 Statistical model

mapping with interpolation. Whereas codebook mapping with interpolation generates highband estimates by averaging K vectors using fixed weights, statistical methods average every vector in the codebook using weights calculated according to a probability distribution that is determined during training. From this perspective, codebook mapping can be regarded as a statistical model whose random sources have zero variance and whose narrowband-highband transition is governed by an identity matrix.

Recently, the GMM has been used to better model the probability distribution of the vectors ([14–16]). The HMM also has been applied to take advantage of the interframe dependence inherent in speech signals ([17, 18]).

2.3 High-Frequency Band Excitation Regeneration

Excitation source is often considered less important than spectral envelope. Nonetheless, it contains significant information regarding the harmonic structure of the speech, and it needs to be estimated to be able to synthesize the highband spectrum. This section describes some existing approaches to excitation regeneration.

2.3.1 Pulse and Noise Excitations

As mentioned in Chapter 1, the goal of reducing transmission rate had naturally led researchers in speech coding to look into reconstructing speech spectra with limited information at the receiver. Because the vocal tract shape largely determines the phonemes being produced, the spectral envelope contributes considerably more to speech intelligibility than the excitation source does. As a result, efforts were made to regenerate the fine structure of the speech with as little transmitted information as possible. One early attempt in this direction is the use of pulse and noise excitations.

In the most rudimentary form, speech source can be modelled as a pulse train for voiced sounds and white noise for unvoiced sounds, and successive pulses are separated by an amount of time equal to the fundamental period of the speech. Under this simple model, the excitation source can be determined by first making a voicing decision on the input speech frame. If the frame is determined to be unvoiced, white noise is generated as excitation. Otherwise, the fundamental frequency is estimated and then used to generate the excitation pulses. Figure 2.5 shows a block diagram depicting the basic excitation model.

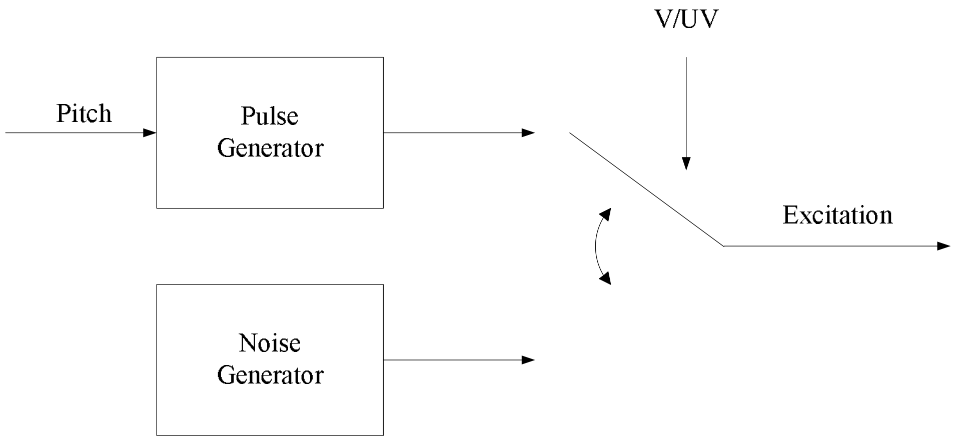


Fig. 2.5 Pulse and noise excitation

This simple excitation model was first used in the traditional pitch-excited LPC vocoder, where the pitch and the degree of voicing were the only excitation parameters transmitted.

The same method can be easily applied, as in [5, 14], to bandwidth extension by estimating the voicing degree and the fundamental frequency from the input narrowband speech.

2.3.2 Non-Linear Distortion

Another method that has its roots in speech coding is the use of non-linear distortions, which were employed in voice-excited vocoders, or baseband vocoders. Voice-excited vocoders were proposed as a middle ground between the low bit-rate analysis-synthesis coders and the high-quality waveform coders [24]. They make the tradeoffs between bit-rate and quality by transmitting the waveform of a low-frequency band, the baseband, which is then used to regenerate high-frequency excitation at the receiver. A conceptually similar coding scheme, employed in the residual-excited linear prediction (RELP) vocoder [25], is based on linear prediction and transmits the low-frequency parts of the LP residual signal as the baseband.

High-frequency regeneration methods based on non-linear distortions reconstruct the excitation signal by applying a zero-memory non-linear transformation, such as a full-wave rectifier or a square operation, to the baseband waveform. The basic concept of this method is depicted in Fig. 2.6. The non-linear transformation creates high-frequency components that have continuous harmonic structures with the baseband. The resulting signal is then spectrally flattened so that the excitation does not affect the overall spectral shape.

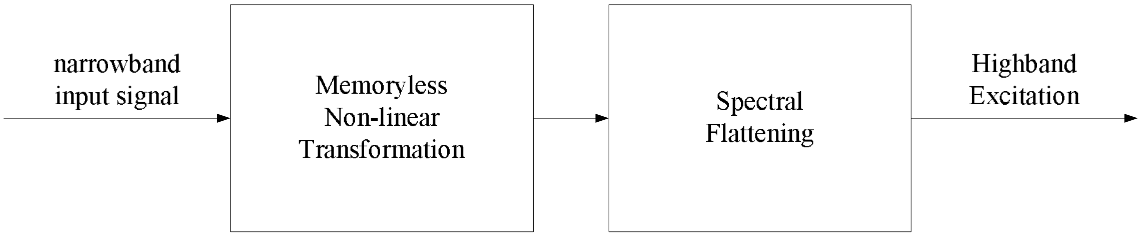


Fig. 2.6 Non-linear transformation

Although the voice-excited and the RELP vocoders were originally used to transmit speech sampled at 8 kHz and the baseband was often below 1 kHz, it is obvious that the concept of high-frequency regeneration in the vocoders is exactly the same as bandwidth extension of telephone speech. Non-linearities have been applied to bandwidth extension in, for example, [26], where the absolute value function is used, and [21], which employs the cubic operation.

2.3.3 Spectral Folding and Spectral Translation

Also originated from the research in baseband coders were methods based on spectral folding and spectral translation. First proposed in [2], spectral folding has quickly gained popularity since then and is perhaps the most widely used method for high-frequency source regeneration in bandwidth extension today (e.g., [4, 7, 17, 27]).

The strength of the spectral folding method lies in its simplicity. Assuming the bandwidth of wideband speech is an integer multiple of that of the baseband, folded images of the baseband spectrum can be generated by inserting zeros between each sample. To extend the bandwidth from 4 kHz to 8 kHz, therefore, an upsampling factor of two can be applied. The resulting upsampled signal can then be used to synthesize highband speech. The upsampling operation is generally applied to the linear prediction residual signal, so that only the fine structures are replicated in the highband.

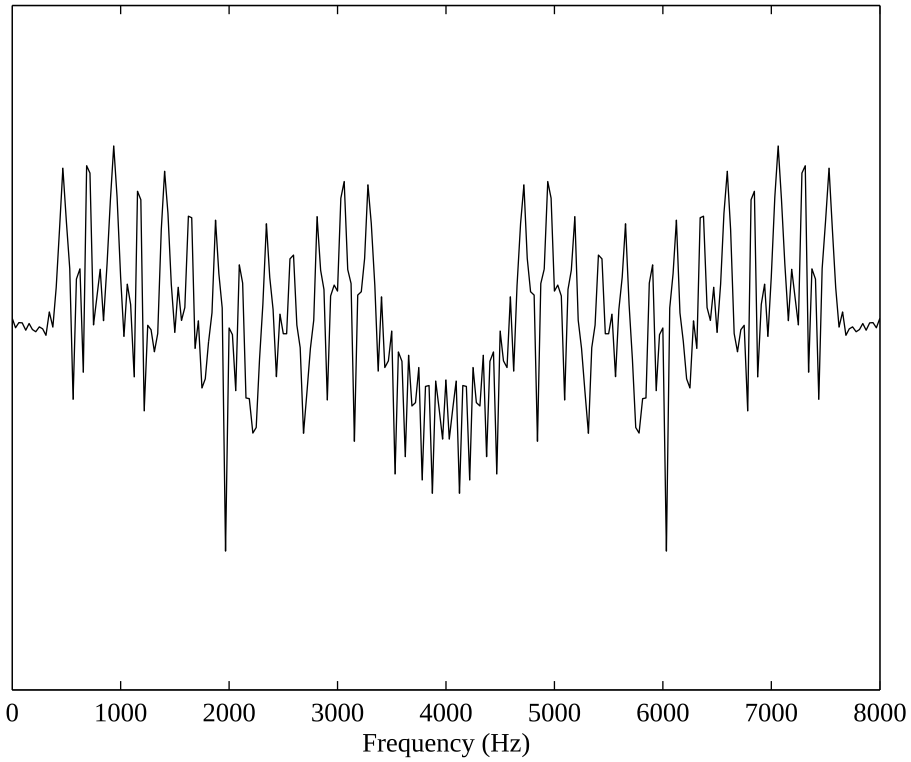


Fig. 2.7 Spectral folding

One possible problem of applying spectral folding to telephone speech is the fact that telephone speech does not have the entire frequency band below 4 kHz available. Specifically, although a sampling rate of 8 kHz can theoretically preserve the spectrum up to 4 kHz, there are always various degrees of attenuation in telephone speech at the higher frequencies above 3 kHz. Upsampling by two will fold the spectrum around 4 kHz and thus create a hole in the middle, as shown in Fig. 2.7. One way to work around the problem, as discussed in [19], is to first downsample the input signal to 7 kHz, thus eliminating the band above 3.5 kHz. Spectral folding is then applied to the resulting signal to obtain the wideband excitation. Finally, a 7-to-8 sampling rate conversion is used to obtain the desired sampling rate of 16 kHz.

Spectral translation is a method conceptually similar to spectral folding and was also proposed in [2]. Whereas spectral folding generates mirrored images of the baseband spectrum, spectral translation simply creates exact copies in the high-frequency bands.

The basic form of spectral folding and spectral translation methods as presented in [2] contain one common problem. If additional care is not taken, the pitch structure will be destroyed since the harmonics in the highband replica will not be aligned with the true harmonics in the narrowband. Therefore, several algorithms (e.g., [15, 28, 29]) use modified frequency shifting methods that retain harmonic continuity in the highband.

2.3.4 Modulated Noise Excitation

Human ears gradually lose their abilities to resolve the fine structures in speech spectra as the frequency increases. In addition, the speech signal itself often loses its harmonic structures and becomes more noise-like at high frequencies. These facts have motivated the use of noise modulation for highband source regeneration [30].

The noise modulation method generates the highband excitation signal by taking the time-domain envelope of the 3 – 4 kHz band and using it to modulate a noise signal to obtain the high-frequency components. A block diagram depicting this process is shown in Fig. 2.8.

2.4 High-Frequency Band Energy Estimation

In the two previous sections, several methods of regenerating the highband spectral envelope and excitation source were described. However, the focus was on the general shape of the signal waveform. Since the estimated highband spectrum has to be recombined with the telephone-band speech, it is important that the highband energy is adjusted according to

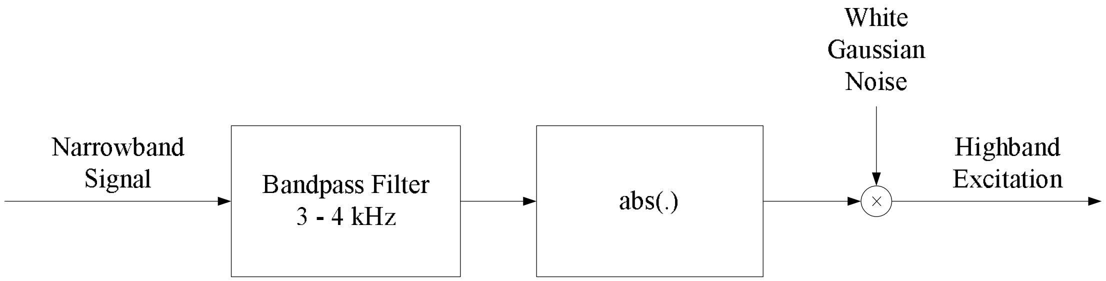


Fig. 2.8 Noise modulation

the amount of energy present in the narrowband. The highband energy adjustment is the focus of this section.

2.4.1 Narrowband Energy Matching

Many bandwidth extension algorithms reconstruct the highband spectrum by first generating the entire wideband spectrum, which is then highpass filtered. This can be done simply by using the parameters of the wideband spectral envelope rather than those of the highband during training. In this case, energy can be adjusted by scaling the reconstructed signal so that its telephone-band portion has equal amount of energy as the input narrowband signal.

2.4.2 Gain Ratio Estimation

Another way to adjust the highband energy is to estimate the ratio $\frac{E_h}{E_n}$, where E_h is the energy of the highband, and E_n is an energy value that can be derived from known information, e.g., the energy of the input narrowband signal. The gain ratio can be estimated in the same manner as spectral envelope by considering it as a missing parameter to be recovered. For example, any one of the aforementioned mapping methods can be used to compute the gain ratio estimate.

2.5 Low-Frequency Band Regeneration

Although the focus of this thesis work is on robustness towards high-frequency reconstruction, the extension towards the low-frequency region is briefly discussed in this section.

The main difference between lowband and highband spectra is that while speech often contains a large percentage of noise-like components at high frequencies, the lowband generally contains strong harmonic structures during voiced sounds. Therefore, recovering the missing harmonics is the most important task in lowband regeneration.

2.5.1 Harmonic Recovery

In [26], it is assumed that the fundamental frequency is always greater than 100 Hz and that the telephone band is cut off at 300 Hz. Under these assumptions, there are at most two harmonics that need to be recovered. These harmonics are reconstructed by a sinusoidal oscillator whose frequencies are controlled by pitch analysis. The harmonics are then scaled according to the output of a multi-layer perceptron.

A similar scheme is described in [11], which uses the additional step of injecting noise into the harmonics. The combination of noise and harmonics are used to excite a spectral shaping filter estimated using codebook mapping or linear mapping.

2.5.2 Equalization

If sufficient lowband content is available, estimation errors associated with harmonics recovery can be avoided by deterministically boosting the lowband spectrum using an equalizer. This method assumes that channel response is known and thus can be compensated by a filter that has approximately the inverse response of the channel. Examples of lowband equalization can be found in [31] and [16]. Equalization has the advantage of being able to maintain the lowband spectral structure. The drawback is that it is designed for only one kind of channel response and it can increase the noise level at low frequencies.

2.6 Feasibility of Bandwidth Extension

Because of the physical nature of human speech production, there exists an inherent mutual dependency between different frequency bands in the spectrum. However, exactly how much correlation exists is still an open question, and it is natural to ask whether there is enough correlation to ensure satisfactory reconstruction of the missing bands.

2.6.1 Information Theoretic Perspective

Several works attempt to answer the question from an information theoretic point of view. In [32], it is reported that a lower bound on mutual information exists between lower-band (0 – 4 kHz) spectral parameters and highband (above 4 kHz) spectral slope and gain. This result confirms the existence of inter-band correlations.

In [33], an attempt is made at determining whether the dependency between the telephone band and the highband is strong enough for bandwidth extension. Parameterizing the spectral envelopes as mel-frequency cepstral coefficients (MFCC) and modelling the probability distribution using GMM, the authors of [33] observe that the mutual information between different bands is low, especially for fricative sounds, and the reconstructed highbands generally contain perceptible spectral distortions. They conclude that a memoryless mapping of spectral envelope parameters can perform reasonably well because the resulting signal sounds pleasant and not because the mapping is accurate.

Another experiment regarding the information theoretical background of bandwidth extension is reported in [34], where a lower bound on the mean log spectral distortion of the highband spectral envelope is found. The lower bound is expressed in terms of the highband entropy and the mutual information between the narrowband and the highband, and it poses a limit on any memoryless mapping methods.

2.6.2 Speech Coding Perspective

The practicability of bandwidth extension has also been tested from the speech coding point of view, where the objective is to code the highband with as few bits as possible with the ultimate goal of not having to allocate any bits for the highband, i.e., bandwidth extension with no side information.

In [26] and [35], it is claimed that highband reconstruction without side information produces unsatisfactory results, and the reason given in [35] is that the narrowband to highband transformation is a one-to-many mapping and as a result correct mapping cannot be ensured without side information. Therefore, as claimed in both papers, some highband information needs to be transmitted for adequate reconstructions at the receiver end.

This conclusion, nonetheless, is a consequence of the speech coding approach of bandwidth extension. Speech coding aims to find the best trade-off between bit-rates and quality. From this perspective, bandwidth extension with no side information serves as a lower bound in performance since it represents the lowest bit-rate possible.

Furthermore, the research in speech coding is concerned with reconstructing the exact copy of the original signal. However, as can be inferred from [33], a bandwidth extension system can make the speech sound pleasant without having a good approximation of the original spectrum. Therefore, these results show that the focus of bandwidth extension should be on improving the speech quality in general rather than estimating the original wideband signal.

2.7 Chapter Summary

This chapter described various methods of bandwidth extension based on the linear source-filter model of speech production. First, the basic concept of the source-filter model was introduced. Codebook mapping, linear mapping, and statistical mapping were then discussed as methods for spectral envelope reconstruction. For high-frequency excitation regeneration, the methods of pulse excitation, non-linear transformation, spectral folding, and noise modulation were described. Highband energy adjustment was then examined, and the problem of low-frequency extension was also briefly treated. Finally, previous works on determining the feasibility of bandwidth extension were discussed.

Chapter 3

Core Bandwidth Extension System

As mentioned in Chapter 1, the goal of this thesis work is to provide a robust bandwidth extension scheme that can operate under different conditions. The basis of the proposed scheme comes from the system described in [22], a block diagram of which is shown in Fig. 3.1.

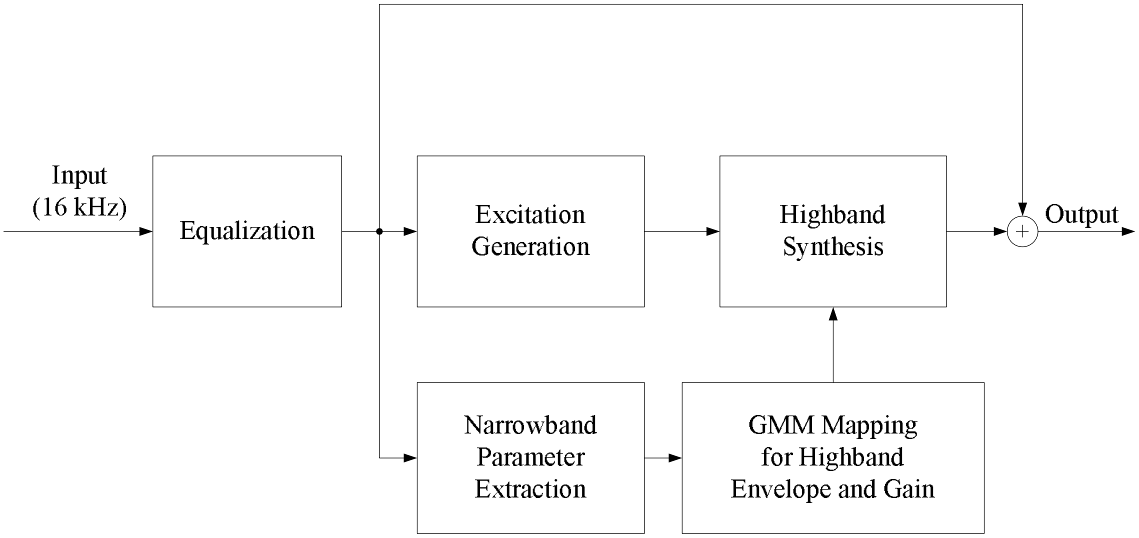

Input
(16 kHz)
Equalization
Excitation Generation
Highband Synthesis
+
Output
Narrowband Parameter Extraction
GMM Mapping for Highband Envelope and Gain

Fig. 3.1 Core system

As can be seen in the diagram, the algorithm consists of three main components, namely

- equalization of the telephone-band signal for recovery of attenuated components,

- generation of highband excitation, and
- reconstruction of highband spectral envelope and excitation gain

The core bandwidth extension system will be the focus of this chapter, and functions of the main components will be described.

3.1 Equalization

According to the ITU-T G.712 standard [36], an analogue-to-analogue channel between 2-wire ports should have an approximately flat frequency response between 300 and 3400 Hz, with various degrees of attenuation outside this region. Figure 3.2 shows the exact specification on the allowable frequency distortions. Although the standard does not set an upper bound on the attenuation applied outside the passband between 300 and 3400 Hz, practical implementations are not likely to completely suppress the attenuation band. If the components outside the telephone band are available, recovery through equalization can better preserve the naturalness than reconstruction methods that discard those components.

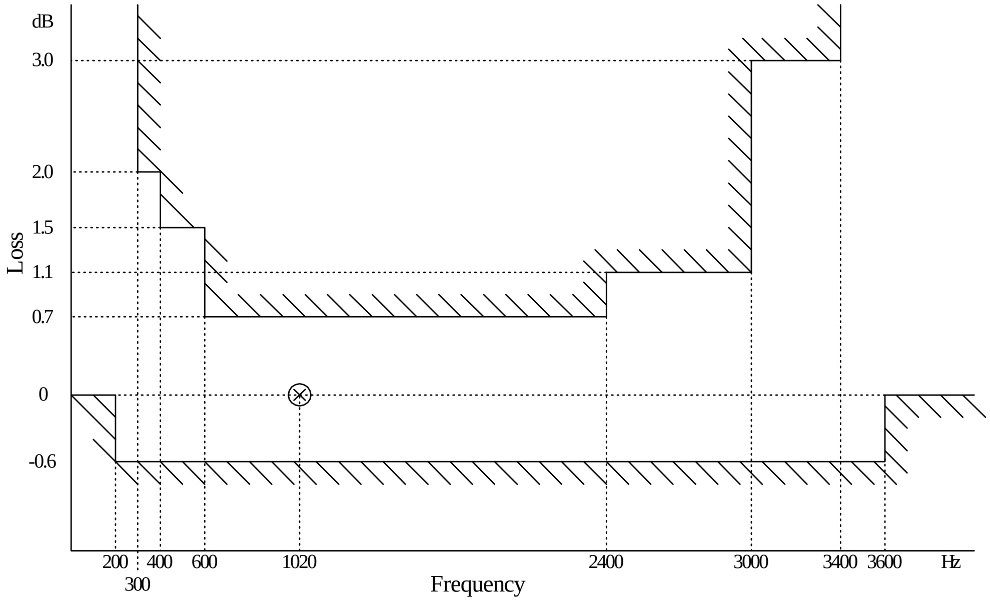


Fig. 3.2 ITU-T G.712 specification for attenuation for analogue-to-analogue channels between 2-wire ports

Motivated by the aforementioned reason, equalization was used in [22] to compensate the attenuation below 4 kHz. Through experiments, it was observed that an ITU-T G.712 channel filter can be characterized as having an attenuation up to 18 dB as indicated in Table 3.1. Two equalizers were designed to recover the attenuated components below 300 Hz and between 3400 and 4000 Hz. The approximate combined frequency response of the two equalizers are shown in Fig. 3.3.

Table 3.1 Characterizations of channel frequency response in [22]

Frequency (Hz)	100 – 300	300 – 3400	3400 – 4000
Attenuation (dB)	0 – 10	0	0 – 18

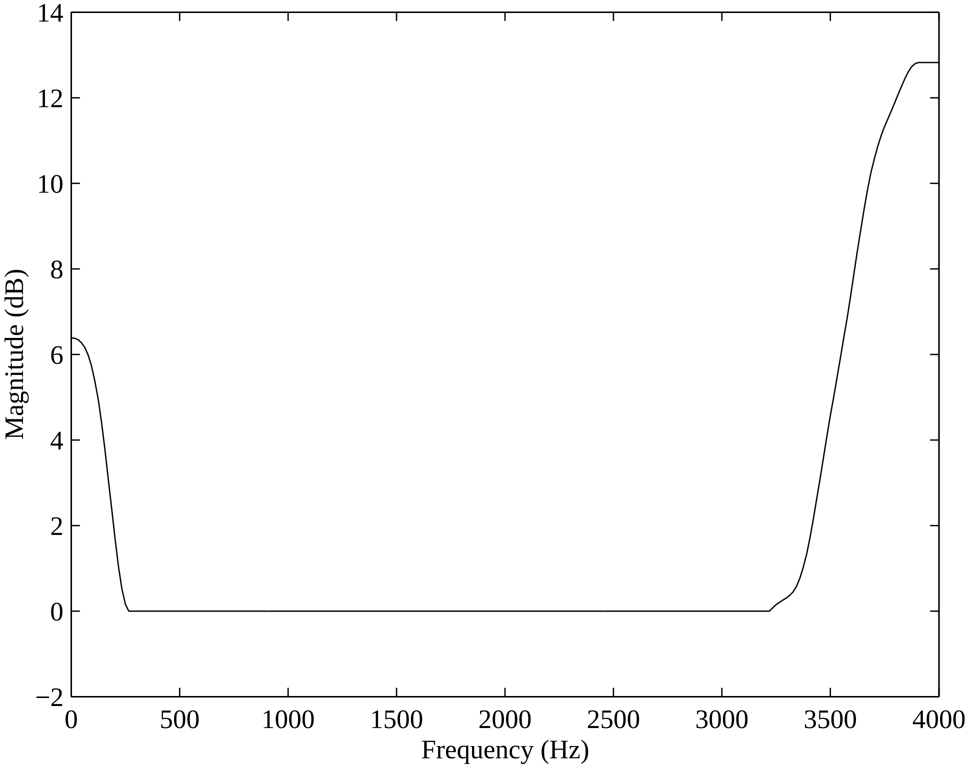
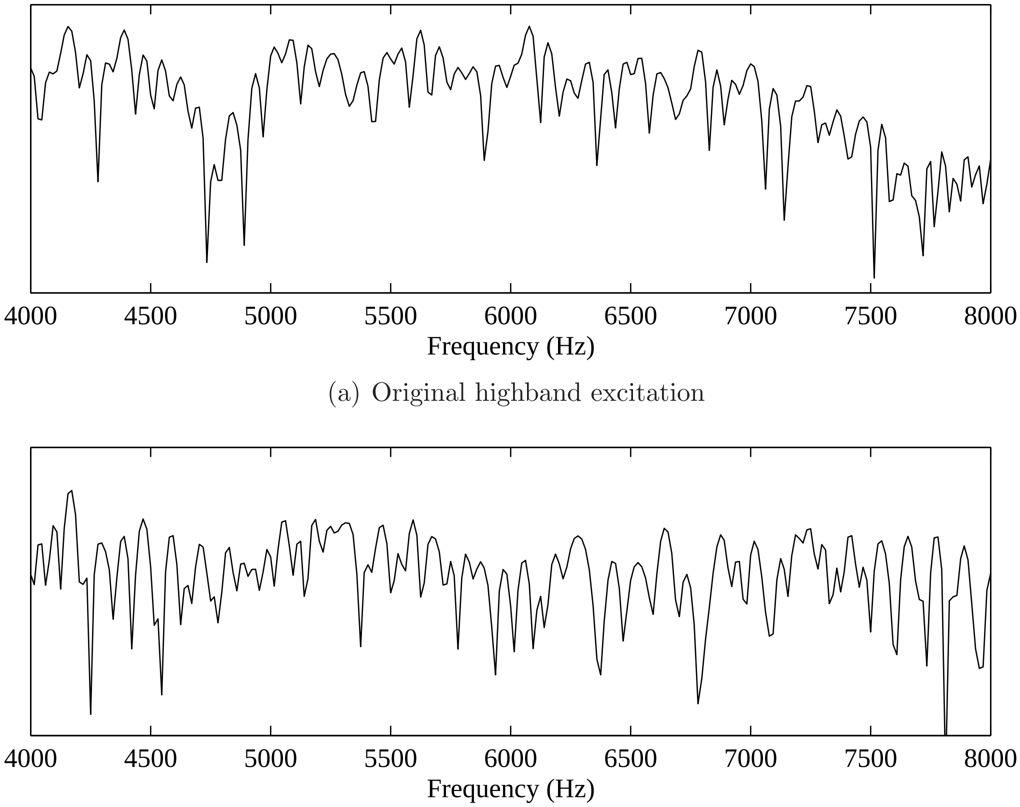


Fig. 3.3 Equalization used in the core system

3.2 Highband Excitation Generation

Highband excitation signal in the core system is generated using the noise modulation method described in Chapter 2. First, the band between 3 and 4 kHz is extracted by passing the input signal through a bandpass filter. The high-frequency components are then generated by taking the time envelope of the bandpass filtered signal. Finally, the spectrum is approximately flattened by modulation of white Gaussian noise. Figure 3.4 shows an example of the original highband LP residual and the highband excitation signal reconstructed by noise modulation. Through informal listening tests, it was determined that the reconstructed speech using noise modulated excitation incurs little loss of quality compared to the original wideband signal.



(a) Original highband excitation

(b) Reconstructed highband excitation

Fig. 3.4 A comparison of (a) the highband LP residual of the original wideband signal and (b) the reconstructed excitation signal by bandpass white Gaussian noise modulation

One concern regarding the use of bandpass modulated noise is whether enough content is available in the bandpass region of telephone speech from 3 to 4 kHz. In [22], the effect of the bandlimited nature of telephone speech was alleviated through equalization. Therefore, equalization not only improves the speech quality in the telephone band, but it also contributes to the reconstruction of the highband excitation signal.

3.3 Highband Spectral Envelope and Excitation Gain Estimation

For the reconstruction of highband spectral envelopes, statistical mapping based on GMM is employed.

The narrowband and highband spectral envelopes are represented as the LSF obtained from a 14^{th}-order and a 10^{th}-order LP analysis, respectively. In addition to the LSF, the pitch filter gain is also included in the narrowband feature vector to better capture the voicing characteristics of the telephone-band speech.

Modelling the joint probability distribution as Gaussian mixtures, the highband features are obtained from narrowband features using the minimum mean square error (MMSE) estimator.

This section starts by introducing the concept of GMM. The processes of obtaining the GMM parameters and estimating the highband feature vectors are then described.

In this section, random variables are denoted by capital letters while letters in boldface indicate vectors.

3.3.1 Gaussian Mixture Model

A GMM with M mixture components and L dimensions can be seen as a combination of M random sources, each having an L-variate Gaussian distribution. The probability density function (PDF) of a GMM can be expressed as

$$f(\boldsymbol{z}|\theta) = \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{z}), \tag{3.1}$$

where $\alpha_m \in (0,1)$ is the mixing proportion of the m^{th} mixture component, $f_m(z)$ is an L-variate Gaussian PDF with mean vector μ_m and covariance matrix Σ_m, and $\theta = \{\alpha_m, \mu_m, \Sigma_m : m = 1, 2, ..., M\}$ is the parameter set of the GMM.

From a speech production perspective, different mixture components in a GMM can conceptually be regarded as representing sound sources that generate different classes of speech, such as vowels, fricatives, and stop consonants. Each source contributes to the output according to the parameter set θ. Specifically, the parameters μ_m and Σ_m determine how each random source generates an instance of the sound class it represents, and the mixing proportions α_m govern how much effect each mixture component has on the output. Mixture components with higher values of α_m will have greater effects than those with lower mixing proportions.

3.3.2 GMM Parameters

In [22], two GMM's with 128 mixture components each are used. The first GMM has 25 dimensions and specifies the joint probability distribution of the narrowband feature vector and the highband LSF, whereas the other GMM has 16 dimensions and determines the relationship between the narrowband features and the highband excitation gain. For both GMM, the covariance matrices $\{\Sigma_m\}$ are modelled as diagonal matrices. In other words, different dimensions in the same mixture component are assumed to be independent.

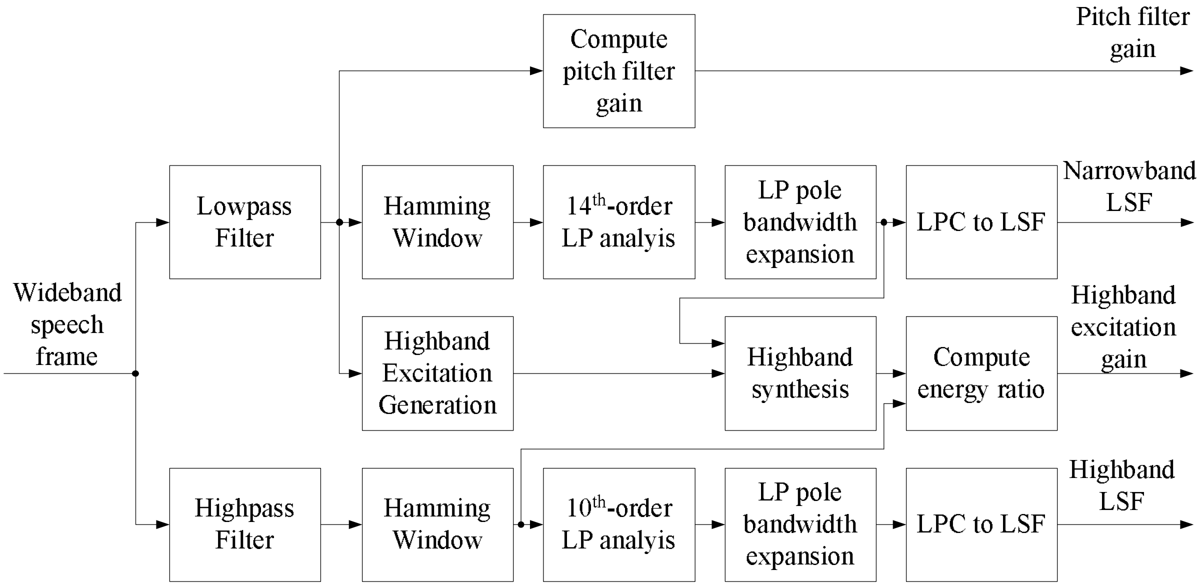


Fig. 3.5 Procedure for calculating training data

To obtain the GMM parameters $\{\alpha_m, \mu_m, \Sigma_m\}$, the expectation-maximization algorithm is employed to iteratively search for the maximum likelihood estimators. The training data are collected from a speech database comprising approximately 50 minutes of studio-quality wideband speech. Figure 3.5 illustrates the process of collecting the training data. There are four parameters that need to be calculated, i.e.,

- narrowband LSF,
- narrowband pitch filter gain,
- highband LSF, and
- highband excitation gain

The procedure takes as input non-overlapping 20-ms wideband speech frames and filters them into two subbands, the telephone band and the highband. The LSF vectors of the two subbands are obtained from LP analyses on Hamming-windowed speech frames with 60-Hz LP pole bandwidth expansion, which is performed by multiplying the LP coefficients with a truncated decaying exponential sequence. The highband excitation gain is computed as

$$g = 10 \log \frac{E_h}{E_s},$$

where E_h is the energy of the highpass filtered frame, and E_s is the energy of the synthesized highband frame.

3.3.3 GMM Mapping

Figure 3.6 shows a block diagram of highband spectral envelope and excitation gain estimation. The highband features are estimated by mapping narrowband parameters to the MMSE estimator according to the Gaussian mixture PDF obtained during training.

Denoting the narrowband features and the highband LSF as random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, the MMSE estimate of the highband LSF given the observed narrowband feature vector $\boldsymbol{x}$ is the vector $\hat{\boldsymbol{y}}$ such that

$$\epsilon^2(\hat{\boldsymbol{y}}) = \min_{\boldsymbol{y}} \epsilon^2(\boldsymbol{y}), \tag{3.2}$$

where

$$\epsilon^2(\boldsymbol{y}) = E[(\boldsymbol{y} - \boldsymbol{Y})^2 | \boldsymbol{X} = \boldsymbol{x}] \tag{3.3}$$

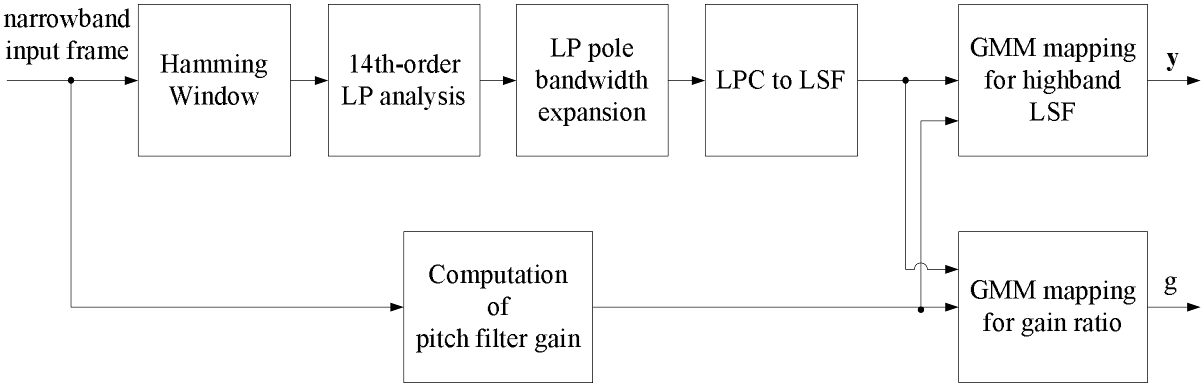

narrowband input frame
Hamming Window
14th-order LP analysis
LP pole bandwidth expansion
LPC to LSF
GMM mapping for highband LSF
y
Computation of pitch filter gain
GMM mapping for gain ratio
g

Fig. 3.6 Estimation of highband spectral envelope and excitation gain

is the mean square error of a highband features estimate $\boldsymbol{y}$.

It is known from probability that the MMSE estimate of a random variable V given another random variable U is the conditional expectation

$$\hat{v} = E[V|U]. \tag{3.4}$$

Therefore, the highband features estimate that is optimal in the mean square sense is simply the expected value of the highband features given the narrowband parameters $\boldsymbol{x}$,

$$\hat{\boldsymbol{y}} = E[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]. \tag{3.5}$$

The MMSE estimate of the highband spectral envelope parameters can be derived as follows:

$$\begin{aligned}
\hat{\boldsymbol{y}} &= E[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}] \\
&= \int_{\Omega_y} \boldsymbol{y} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})\, d\boldsymbol{y} \\
&= \int_{\Omega_y} \frac{\boldsymbol{y} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{f_{\boldsymbol{X}}(\boldsymbol{x})}\, d\boldsymbol{y} \\
&= \frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x})} \int_{\Omega_y} \boldsymbol{y} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{y},
\end{aligned} \tag{3.6}$$

where Ω_y denotes the sample space of $\boldsymbol{Y}$, $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$ the conditional PDF of the highband LSF given the narrowband vector, $f_{\boldsymbol{X}}(\boldsymbol{x})$ the PDF of the narrowband feature vector, and

$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ the joint PDF of the narrowband and highband parameters.

Modelling the probability distribution using GMM, the PDF of the narrowband vector $f_{\boldsymbol{X}}(\boldsymbol{x})$ is equal to

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}), \tag{3.7}$$

and similarly the joint PDF $f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ of narrowband features and highband LSF may be expressed as

$$\begin{aligned} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) &= \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x},\boldsymbol{y}) \\ &= \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}) f_m(\boldsymbol{y}), \end{aligned} \tag{3.8}$$

where the last line follows from the use of diagonal covariance matrices in the GMM.

Substituting with Eq. (3.8), the integral in Eq. (3.6) can then be computed as

$$\begin{aligned} \int_{\Omega_y} \boldsymbol{y} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{y} &= \int_{\Omega_y} \boldsymbol{y} \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}) f_m(\boldsymbol{y})\, d\boldsymbol{y} \\ &= \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}) \int_{\Omega_y} \boldsymbol{y} f_m(\boldsymbol{y})\, d\boldsymbol{y} \\ &= \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}) \boldsymbol{\mu}_{m,y}, \end{aligned} \tag{3.9}$$

where $\boldsymbol{\mu}_{m,y}$ denotes the mean vector of $\boldsymbol{Y}$ in the m^{th} mixture component of the GMM.

Finally, substituting Eq. (3.7) and Eq. (3.9) in Eq. (3.6), the MMSE estimate of the highband LSF can be obtained:

$$\hat{\boldsymbol{y}} = \frac{\sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}) \boldsymbol{\mu}_{m,y}}{\sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x})}. \tag{3.10}$$

with

$$f_m(\boldsymbol{x}) = \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{14}\frac{(\boldsymbol{x}_i - \mu_{m,x_i})^2}{\sigma^2_{m,x_i}}\right\}}{(2\pi)^7\prod_{i=1}^{14}\sigma_{m,x_i}}, \tag{3.11}$$

where μ_{m,x_i} and σ^2_{m,x_i} denote the mean and variance, respectively, of the i^{th} narrowband LSF in the m^{th} mixture component of the GMM.

The MMSE estimate of the highband excitation gain can be calculated in the same fashion and is equal to

$$\hat{g} = \frac{\sum_{m=1}^{M}\alpha_m f_m(\boldsymbol{x})\mu_{m,g}}{\sum_{m=1}^{M}\alpha_m f_m(\boldsymbol{x})}. \tag{3.12}$$

3.4 Chapter Summary

This chapter described the system proposed in [22], which serves as the basis of this thesis work. The three main components of the system — equalization, highband excitation generation, and highband spectral envelope and gain estimation — were presented. Equalization is used to recover attenuated telephone-band components below 4 kHz. For the high-frequency band above 4 kHz, the excitation signal is generated by bandpass-modulated white Gaussian noise, while the spectral envelope and gain factor were reconstructed using GMM mapping. The concept of GMM was introduced, and finally the MMSE estimator for GMM mapping was derived.

Chapter 4

Robust Bandwidth Extension

In Chapter 2, the backgrounds of bandwidth extension were introduced. Several methods that had been previously applied were described. All of the algorithms discussed in Chapter 2 have been employed with various degrees of success. However, they have the common weakness of being inflexible to different channel conditions.

The algorithms are generally trained and tested only for clean, studio-quality speech. Any deviation from the ideal condition will result in degradations in performance and possibly annoying artifacts. Since no telephone channel is ideal and each channel has its unique characteristics, drastic reduction in reconstructed speech quality is bound to occur if no safeguard measures are taken.

In Chapter 3, the bandwidth extension algorithm proposed in [22] was described in more details. Similar to many other systems, it too suffers from lack of robustness to mismatch in training and testing conditions.

This chapter will describe the effects of non-ideal operating conditions and present the proposed scheme that uses the algorithm in [22] as the core bandwidth extension system for robustness against additive noise and channel response mismatch.

4.1 Robustness Against Additive Noise

Noise has always been a major problem for speech processing applications. It can reduce intelligibility, degrade quality, and affect the accuracy of parameter estimation.

Noise can come from many sources. In this section, the focus will be on sources that can be modelled as additive noise. Examples of additive noise sources include

- background noise, such as street noise at a public pay phone and machine noise in a factory,
- quantization noise in analogue-to-digital signal conversion,
- coding noise due to data loss incurred through speech compression, and
- channel noise that results from data corruption during transmission through an imperfect channel.

The effects additive noise has on the core bandwidth extension system are fourfold:

- additive noise decreases the intelligibility and the quality of the telephone-band speech;
- the noise level is increased along with speech during equalization, thus compromising the effectiveness of the equalizers;
- if the band between 3 and 4 kHz is noisy, the reconstructed highband will likely be noisy too, since high-frequency excitation generation depends on this band;
- noise can cause the computations of narrowband LSF and pitch filter gain to be inaccurate, leading to undesirable results in GMM mapping for highband LSF and gain estimation.

The first problem in the list involves only the telephone-band speech, whereas the other three are concerned with the recovery of the attenuated and the missing bands. In the following, the proposed ways to counter these problems are presented.

4.1.1 Noise Reduction for Enhancement of Noisy Narrowband Speech

Noise reduction for speech enhancement is a well-studied area in speech processing. Due to the importance of the problem, much effort has already been devoted to improving the quality of speech corrupted by noise. Although the search for better noise suppressors that can work in more adverse conditions will be ever-ongoing, there exist nowadays many systems that can provide satisfactory enhancement.

In this thesis work, the noise suppression preprocessor used in the Enhanced Variable Rate Codec (EVRC) [37] is employed for telephone-band speech enhancement. Figure 4.1 shows the block diagram of the noise suppressor.

The EVRC noise suppressor is based on the concept of short-time spectral weighting. The input 80-sample frame is combined with 24 samples of the previous frame and, after the application of a smoothed trapezoid window, is transformed to the frequency domain via a 128-point Fast Fourier Transform. In the frequency domain, the FFT coefficients are divided into 16 non-overlapping channels, with the channel bandwidth gradually increasing towards higher frequency to model the critical bands in human hearing. An adaptive filter with constant gain in each channel is calculated based on the signal-to-noise ratio (SNR) estimates. For low-SNR regions, stronger suppression is applied. If no noise is detected, the filter is set to unit gain. Finally, the filtered signal is transformed back to the time domain through overlap-add inverse FFT.

One of the features the EVRC noise suppressor has is the use of a noise floor which limits the maximum attenuation applied by the filter to 13 dB. The noise floor allows the algorithm to avoid severe distortion and leave noise residual that can help mask artifacts. In addition, speech with some amount of background noise can sound more natural by giving the listener a sense of the speaker's environment.

4.1.2 Equalization

Due to the masking phenomenon in human hearing, a weak signal cannot be heard by the human ears if a stronger signal is present at the same frequency. Therefore, noise is more perceptible at spectral valleys than spectral peaks. The most common situation is when there is quantization or coding noise between harmonics during voiced sounds. In this case, the equalizers can have the harmful effect of increasing the perceptual noise level in the spectral valleys.

To counter this effect, an adaptive postfilter in the form of a one-tap comb filter is utilized. The transfer function of the filter can be expressed as

$$h(z) = G(1 + \gamma z^{-p}), \tag{4.1}$$

where p is the pitch period of the speech frame, $\gamma > 0$ controls the strength of the filter, and G is a scaling factor that normalizes the energy of the filter output.

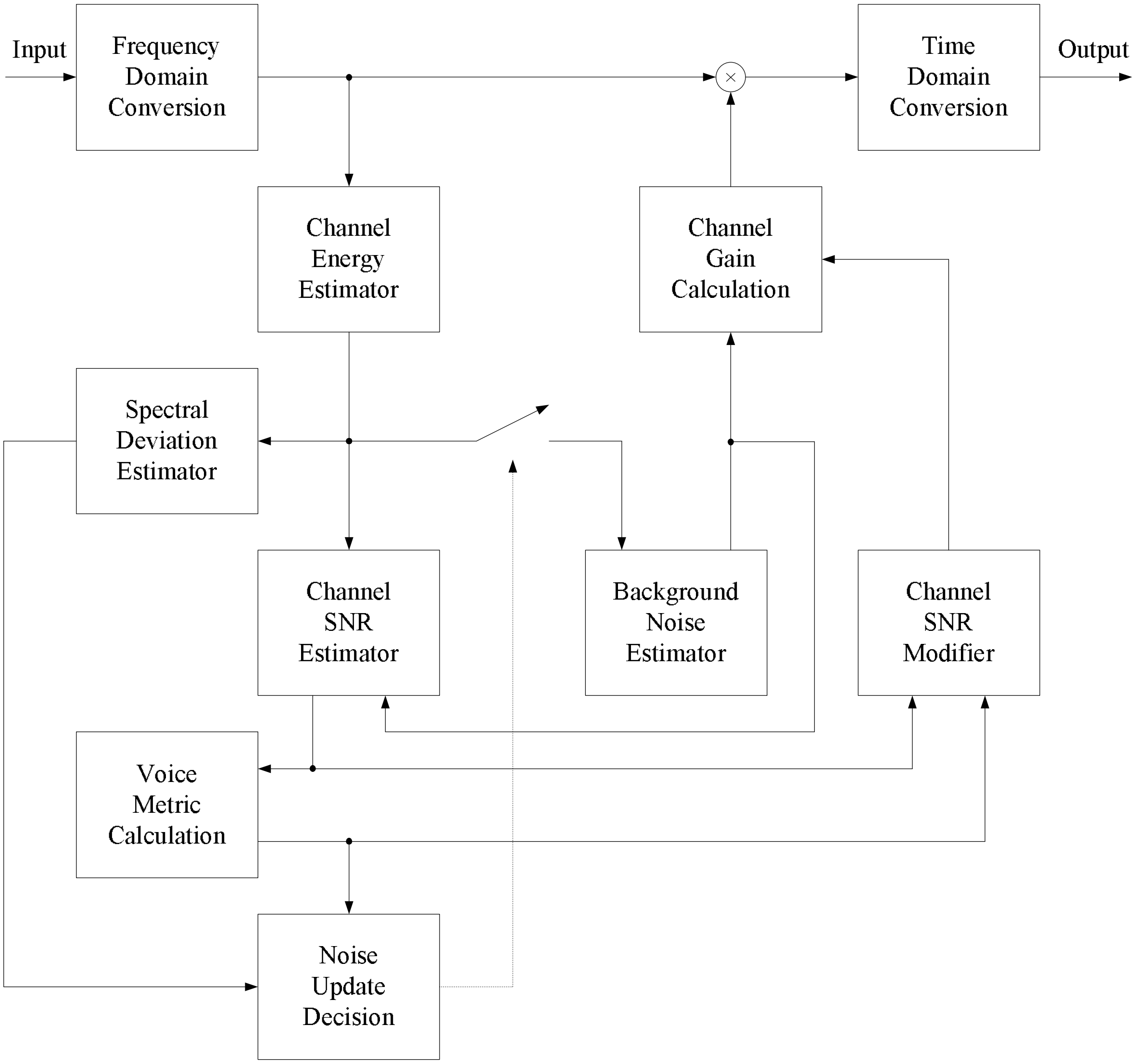


Fig. 4.1 Noise suppressor employed in EVRC

The pitch analysis is performed on 20-ms frames, and the fundamental frequency is assumed to be between 50 and 400 Hz. The pitch period is determined as the position of the first peak greater than 0.6 in the normalized autocorrelation function

$$c[k] = \frac{\displaystyle\sum_{n=0}^{N-1} s[n]s[n-k]}{\sqrt{\displaystyle\sum_{n=0}^{N-1} s^2[n] \sum_{n=0}^{N-1} s^2[n-k]}}, \tag{4.2}$$

where $s[n]$ is the input speech frame, with negative indices corresponding to samples from the previous frame.

The choices of γ and G follow from the method proposed in [38]. The factor γ is selected based on the voicing degree obtained during pitch analysis. Specifically, γ is obtained from the pitch filter gain β according to the relationship

$$\gamma = \begin{cases} 0 & \text{if } \beta < 0.6, \\ 0.9\beta & \text{if } 0.6 \le \beta \le 1, \\ 0.9 & \text{if } \beta > 1. \end{cases} \tag{4.3}$$

As can be seen, the comb filter is turned off during unvoiced phonemes. This avoids introducing unnecessary harmonic structures into unvoiced sounds, which generally have noisy spectra even under ideal conditions. For voiced sounds, the sharpness of the comb filter is determined by the voicing degree, with stronger suppression applying to more strongly voiced sounds. Finally, an upper bound of 0.9 is used for γ to avoid too much suppression that might produce artifacts.

The scaling factor G has to be chosen so that the comb filter output has approximately equal energy as the input. Otherwise the voiced sounds will be amplified by different factors while the unvoiced sounds remain unaltered. To avoid unwanted amplification, the scaling factor is set as

$$G = \begin{cases} \dfrac{1}{1+\gamma} & \text{if } \beta \le 1, \\ \dfrac{1}{1+\dfrac{\gamma}{\beta}} & \text{if } \beta > 1, \end{cases} \tag{4.4}$$

A large pitch filter gain likely indicates a voice onset frame, where the current frame has significantly higher energy than the previous frame. Therefore, the factor of $\frac{1}{\beta}$ is included in the denominator in the case $\beta > 1$ to avoid over-compensation by the scale factor.

Substituting for γ and G in Eq. (4.1), the transfer function of the comb filter can now be expressed as

$$h(z) = \begin{cases} 1 & \text{if } \beta < 0.6, \\ \dfrac{1 + 0.9\beta z^{-p}}{1 + 0.9\beta} & \text{if } 0.6 \le \beta \le 1, \\ \dfrac{1 + 0.9 z^{-p}}{1 + 0.9/\beta} & \text{if } \beta > 1 \end{cases} \tag{4.5}$$

The filtering is applied to 20-ms frames every 10 ms. The filter output is multiplied by a Hamming window to smooth the frame boundaries.

Fig. 4.2 shows the frequency responses of a comb filter with fixed pitch period p and varying pitch filter gain β. A 20-dB and a 40-dB offset is added to the spectra for $\beta = 1$ and $\beta = 0.6$, respectively, to facilitate viewing.

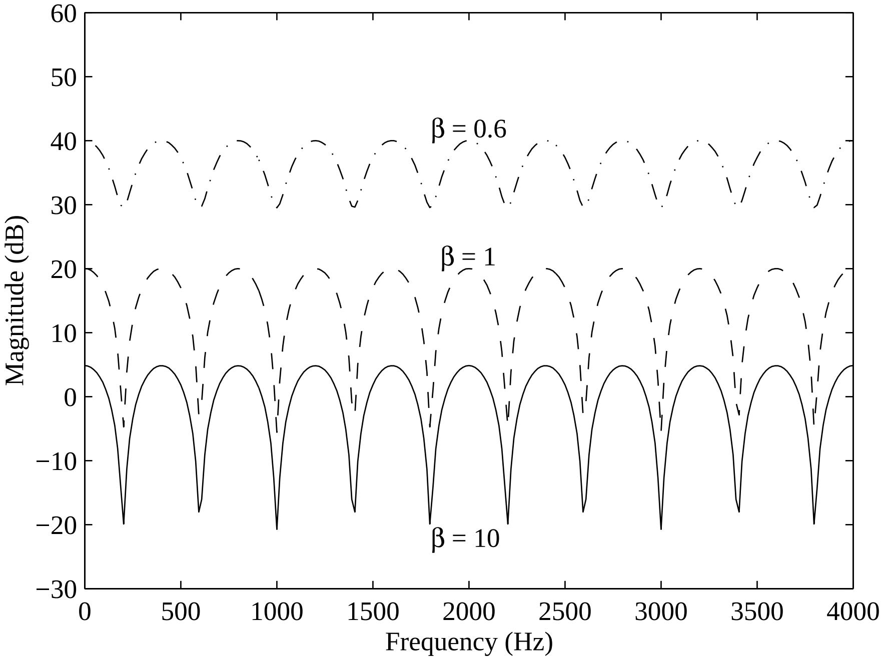


Fig. 4.2 Frequency responses of comb filters. Adjacent graphs are separated by 20-dB offsets to facilitate viewing.

In the case when the attenuated bands are completely submerged in background noise, the algorithm turns off the equalizers automatically since equalization can only further reduce the speech quality. The criterion used to turn on and off equalization is based on the SNR estimates obtained in the EVRC noise suppressor. Equalization is used if the two highest channels, spanning the frequency range between 3000 and 4000 Hz, both are estimated to have SNR greater than 10 dB. On the other hand, if at least one of the two channels has an SNR less than 10 dB, no equalization is applied. Nevertheless, the comb filter is applied in all cases.

4.1.3 Excitation Generation

Under the scheme of bandpass modulated noise, the highband excitation is derived from the narrowband spectrum between 3 and 4 kHz. If this band contains noise, the distortions will be carried onto the highband reconstruction. This problem was investigated in [39], where an adaptive smoothing function of the form

$$e_{sm}[n] = \alpha * e_{sm}[n-1] + (1-\alpha) * e[n] \tag{4.6}$$

was applied to the modulating envelope signal prior to the multiplication with white noise, where $e[n]$ is the modulating envelope and α is between 0 and 1.

To better understand the effects of noise on the excitation signal, tests were conducted by injecting bandpass filtered white Gaussian noise into the 3 – 4 kHz band prior to calculating the highband excitation. The original narrowband signal was not altered, so the result of the GMM mapping would remain unchanged. Through informal listening tests, it was found that additive noise affects the generation of highband excitation in two basic manners:

- noise in the 3 – 4 kHz band causes distortions in the spectral shape of the reconstructed highband excitation signal;
- noise adds extra energy to the 3 – 4 kHz band, and the extra energy will propagates through the process of excitation regeneration and in the end cause over-estimation of the highband spectrum.

To further isolate the two effects, additional tests were performed. First, white noise spectra were used as the highband excitation signal, while the highband energy was scaled back to what would have been obtained with clean narrowband speech. The results of this test showed that distortions to the excitation spectrum introduced only a small amount of degradation when the highband energy could be accurately estimated. On the other hand, too much energy in the highband could be very annoying even with the correct spectral shape.

The reason that the distortions to highband excitation spectrum do not have a great effect on the reconstructed speech quality may be attributed to the fact that the human ears are insensitive to high-frequency spectral fine structures. It can thus be concluded that perceptual artifacts due to errors in highband excitation reconstruction often do not originate from distortions in the excitation spectrum, but rather from over-estimation of excitation energy. In light of this observation, the proposed scheme focuses on the second effect on the list.

In the proposed scheme, the use of the EVRC noise suppressor can automatically reduce the energy between 3 and 4 kHz when noise is present. As noted previously, however, the noise suppressor applies at most a 13-dB attenuation, which still leaves a large amount of noise in low-SNR situation. Excessive energy, coupled with an already highly distorted spectrum, ultimately leads to noisy artifacts in the highband.

The solution is to consider the highly noisy bands as lost bands. If the highest frequency channel, between 3500 and 4000 Hz, is estimated to have an SNR less than 5 dB, the band is completely suppressed and treated as if it has been filtered out due to sharp cutoff in the channel frequency response. The details will be presented in the next section, when robustness against channel response variations is discussed.

4.1.4 Noise Reduction for Better GMM Mapping

In addition to reducing the intelligibility and quality of speech, noise can also affect speech processing applications by making accurate feature estimations more difficult. It has been a major problem in applications such as speech recognition, speaker verification, and speech coding, where accurate modelling of speech is essential. In bandwidth extension applications, the narrowband features are needed to estimate the highband parameters. It is important that the narrowband-to-highband mapping is robust to noise.

As mentioned in Section 4.1.1, the EVRC noise suppressor is employed in the proposed scheme to reduce noise. Like many other noise suppression systems, the EVRC noise suppressor aims to enhance the quality of the speech. However, an improved perceptual quality does not necessarily lead to more accurate feature estimations. To be able to better recover the original parameters from noise-corrupted speech, a different scheme specifically designed for feature extraction should be employed.

The EVRC noise suppressor sets the maximum attenuation at 13 dB to avoid severe spectral distortions. Although the upper bound on attenuation can have good perceptual effects, it changes the overall shape of the spectral envelope by leaving noise in the spectral valleys, thus making LP analysis unreliable. In [40] and [41], it was concluded that a high upper bound on attenuation can be beneficial for coding the LSF in low-SNR situations, and a scheme that adaptively sets the upper bound according to the SNR was proposed.

In this thesis work, the simple solution of increasing the maximum attenuation level to 25 dB is employed. This is implemented by modifying the EVRC noise suppression algorithm to include a second branch that outputs the more strongly suppressed signal. The original output with 13-dB noise floor is applied to the narrowband component of the reconstructed speech, while the more strongly suppressed output is used for parameter estimation in GMM mapping. Little additional complexity is incurred and no extra algorithmic delay is introduced using this method.

The reason this simple method is used is the observation that an accurate estimate of speech features for highband spectrum generation may not be as crucial as it is for speech coding and speech recognition applications. The extracted features in a bandwidth extension system are not used to reconstruct the original spectrum as in a speech coder, nor do they have to accurately represent the speech at the phoneme level as in a speech recognizer.

For a bandwidth extension system, the most important issue is to correctly identify the input speech frames as belonging to one of the different broad classes of speech, such as vowels, fricatives, and plosive consonants. Speech of different classes generally have very different spectral shapes. For example, vowels normally have falling or flat spectra with low energy in the highband, whereas the high-frequency spectra of fricatives are usually flat or rising with high energy. Obviously, a misclassification can easily cause perceptual artifacts. Spectra of speech within the same class, on the other hand, are not so distinctly different from each other, and a misidentification within the same class will likely have little perceptual effect.

The GMM implicitly classifies the speech through soft-decision probabilistic measures and is inherently more robust than hard-decision classifiers such as codebook mapping. By strongly attenuating the low-SNR regions, only the prominent portions of the frequency spectrum remain. For example, a vowel sound corrupted by white noise might end up having a highly lowpass-tilted spectral envelope due to strong suppression at high frequencies, thus reducing the chance of being misidentified as a consonant.

4.2 Robustness Against Mismatch in Channel Response

The ITU-T G.712 standard has a set of specifications for the amount of attenuation the telephone networks can impose on the transmitted signals. In the core system, operations of the equalizers rely on the assumptions that the telephone channels conform to the ITU-T G.712 standard and their frequency responses can be characterized as in Table 3.1.

However, older channels often are not compliant with the ITU-T G.712 recommendation. Furthermore, there can be significant variations even among channels that conform to the standard. Variations in channel response can affect the core bandwidth extension system in two ways:

- mismatch of training and testing channel conditions can cause errors in the GMM mapping for highband parameters.
- difference in channel response in the attenuated bands reduce the effectiveness of the equalizers, as they are designed for only one channel condition;

This section presents the proposed scheme for robust bandwidth extension against channel response mismatch.

4.2.1 Speech Parameter Extraction

Variations in channel frequency response do not usually result in drastic degradation in quality as noise does; however, they can still affect speech processing applications by altering the speech parameters.

In this thesis work, it is assumed that the channel frequency response is approximately flat within the band between 300 and 3200 Hz [1]. Outside of the passband, the channel

[1] If the speech signal is digitized at the source (as would be the case for cell phones), the channel frequency response can be expected to follow the G.712 mask, which is flat to within ±0.5 dB over most of the speech spectrum.

filter can have anything from a sharply cut off to an all-pass frequency response.

Under the above assumption, the agreement between training and testing conditions can be ensured by limiting the speech to between 300 and 3200 Hz. Figure 4.3 illustrates the preprocessor used before speech feature extraction. During both training and testing, the input narrowband speech is passed through the same bandpass filter while being downsampled to 6400 Hz. The bandpass filter is an 200^{th}-order finite impulse response (FIR) filter with the magnitude response shown in Fig. 4.4.

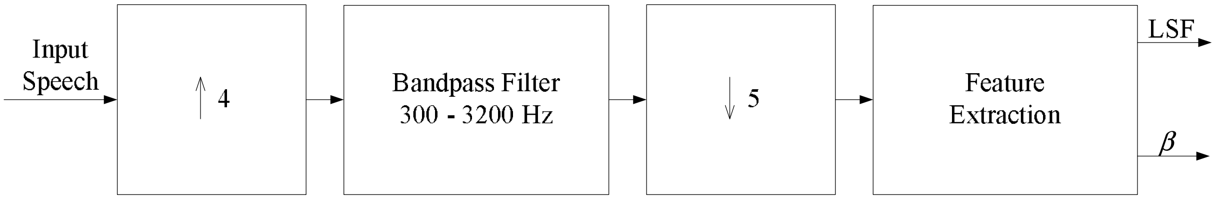


Fig. 4.3 Bandwidth adjustment before feature extraction

By downsampling to 6400 Hz, it can be ensured that the components above 3200 Hz are completely removed and do not affect the narrowband LSF computations. However, the lowband below 300 Hz is not removed in a similar fashion because there is rarely any LSF in the narrow stopband and thus the complexity required for such operation might not be warranted.

One concern for not using the entire telephone-band speech for high-frequency band estimation is the possible reduction of narrowband-highband correlation that plays an important role in bandwidth extension. However, discarding parts of the spectrum before highband feature estimation can be justified. In [32], experiments were conducted to determine the mutual information between narrowband and highband spectral envelopes as a function of the bandwidth of the narrowband. It was observed that the mutual information did not increase much after the narrowband was expanded over 2 kHz. Furthermore, although the content above 3200 is not used in the GMM mapping, it is still needed to generate highband excitation and, as discussed in the previous section, plays an important part in determining the highband energy.

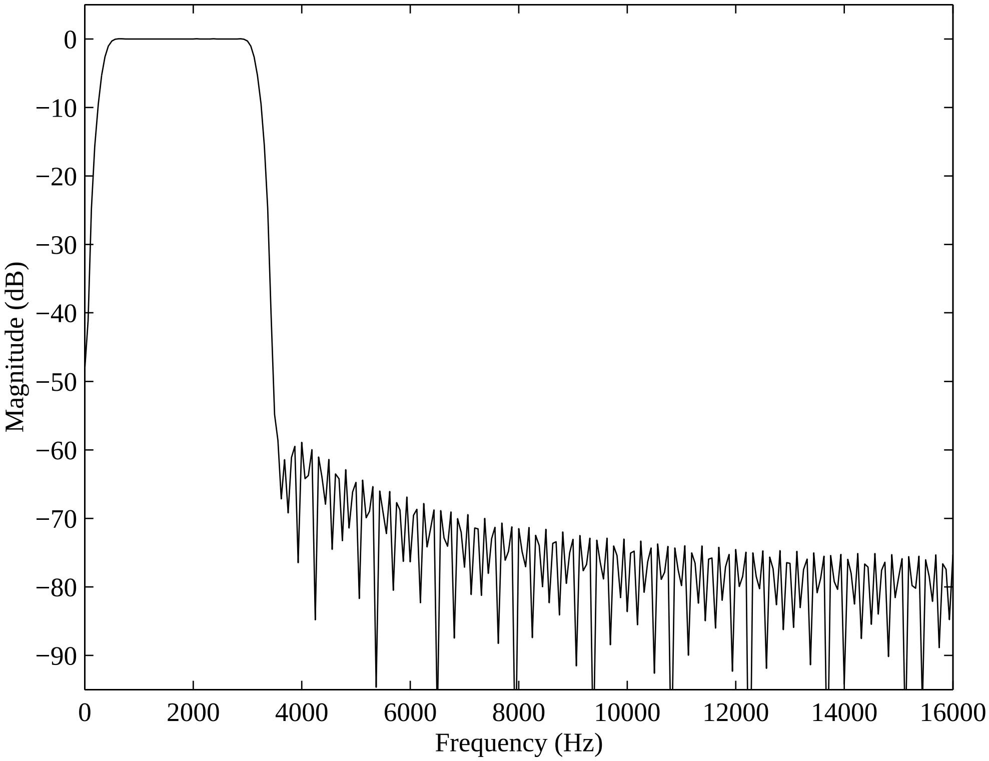


Fig. 4.4 300 – 3200 Hz Bandpass Filter

4.2.2 Recovery of Attenuated Band

Since the equalizers in [22] are designed for a specific channel condition, their performances degrade when different kinds of channels are encountered. If more energy than expected is available in the attenuated bands, the equalizers can produce artifacts by providing too much boost. On the other hand, if the channel attenuation is too strong, equalization might not be enough to recover the content.

One simple solution would be to filter out the content above 3200 Hz, as done for feature extraction, and recover the 3200 – 4000 Hz band as part of the statistical mapping, thereby foregoing the use of an equalizer. However, this method would discard available natural speech and replace it with synthetic content.

A better way would be to include equalization as part of the scheme for the recovery of attenuated band. For equalization to work, there are two conditions:

- the channel response can be estimated.

- enough speech content is available in the attenuated band for a satisfactory recovery using equalization;

The proposed scheme for attenuated band recovery is shown in Fig. 4.5. It can select between two different modes of recovery depending on the amount of available frequency content.

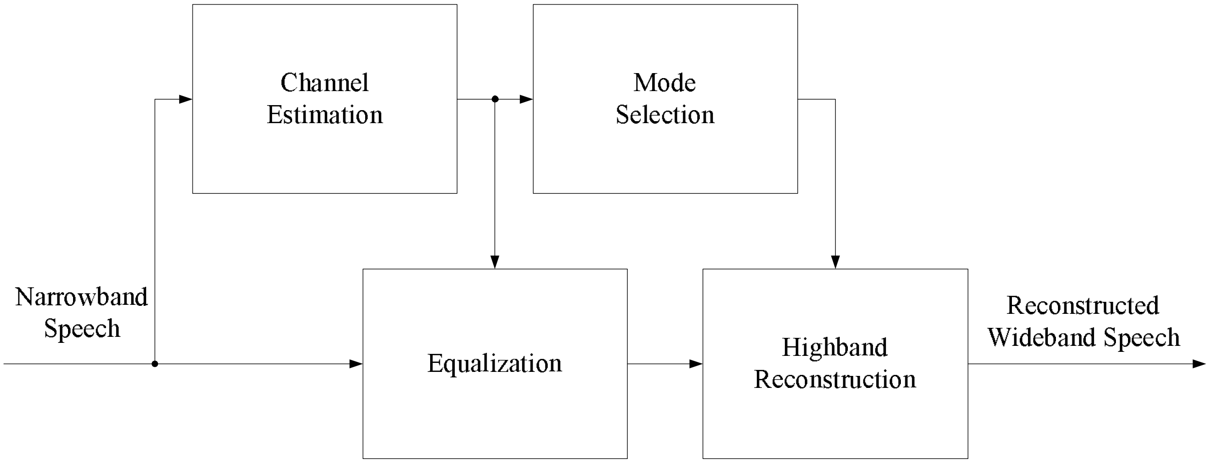


Fig. 4.5 Attenuated band recovery

Channel Response Estimation

First, the channel response in the attenuated band is estimated. Because speech spectra of vowels often are lowpass tilted and have little high-frequency content, it is difficult to make accurately estimate of the channel response using vowel speech frames. Therefore, the estimation procedure is performed only when the input is a consonant with enough high-frequency spectrum. The criterion is based on a threshold of the zero crossing count of the input frame:

$$estimate\_flag = \begin{cases} \text{on} & \text{if } zcc \geq zcc_{th} \\ \text{off} & \text{if } zcc < zcc_{th} \end{cases}, \tag{4.7}$$

where the value $zcc_{th} = 32$ has been found to be a good threshold for a 80-sample frame sampled at 8 kHz. To avoid being affected by low-level noise, a zero crossing is counted only when there is a sign change and the magnitude of the current sample is greater than 24 for 16-bit data, i.e., at a full scale of 32768.

The value of the flag is determined on a frame-by-frame basis. If the flag is switched on, the channel response is estimated in the frequency domain via a K-point FFT. Denoting the magnitudes of the FFT coefficients of the channel filter as $H[k]$, the procedure is as follows:

- Compute the magnitudes of the K-point FFT of the speech frame, $S[k]$, $k = 0...K-1$;
- Keep the coefficients from $k = \frac{5K}{16}$ to $\frac{K}{2} - 1$(i.e., from 2500 Hz to 4000 Hz), and normalize the coefficients so that the maximum in that range equals 0 dB;
- Set $H\left[\frac{K}{2}\right] = S\left[\frac{K}{2}\right]$, and label $\frac{K}{2}$ (4000 Hz) as a peak;
- Iteratively search the FFT coefficients from $i = \frac{K}{2} - 1$ down to $\frac{5K}{16}$:
  - If $S[i] > \max\{S[\text{previous peak}], S[i-1], S[i+1]\}$, then set $H[i] = S[i]$ and label i as a peak;
- After the search is done, $H[k]$ is obtained by linearly interpolating between the peaks. However, if the resulting $H[k]$ is less than $S[k]$, set $H[k] = S[k]$.
- Filter coefficients to the left of the leftmost peak (which is also the highest peak and has a magnitude of 0 dB) are set to 0 dB.

Since the EVRC noise suppressor computes an 128-point FFT for each frame, the channel response estimation and equalization can be conveniently implemented as part of the noise suppressor. Figure 4.6 shows the 128-point FFT magnitude spectrum of a speech frame and the channel response estimate for that frame. Also shown in the figure is the frequency response of the lowpass filter used to obtain the speech. As can be seen, the estimated and the true channel response are close even though only one frame is used in the estimation.

The channel response obtained for each frame is continuously accumulated and averaged. Assuming a time-invariant channel, the averaged estimate should converge close to the true characteristics of the channel.

Equalization

Once an estimate of the channel response is obtained, an equalizer can be constructed as the inverse response of the channel. However, in the case of severe attenuation, care must

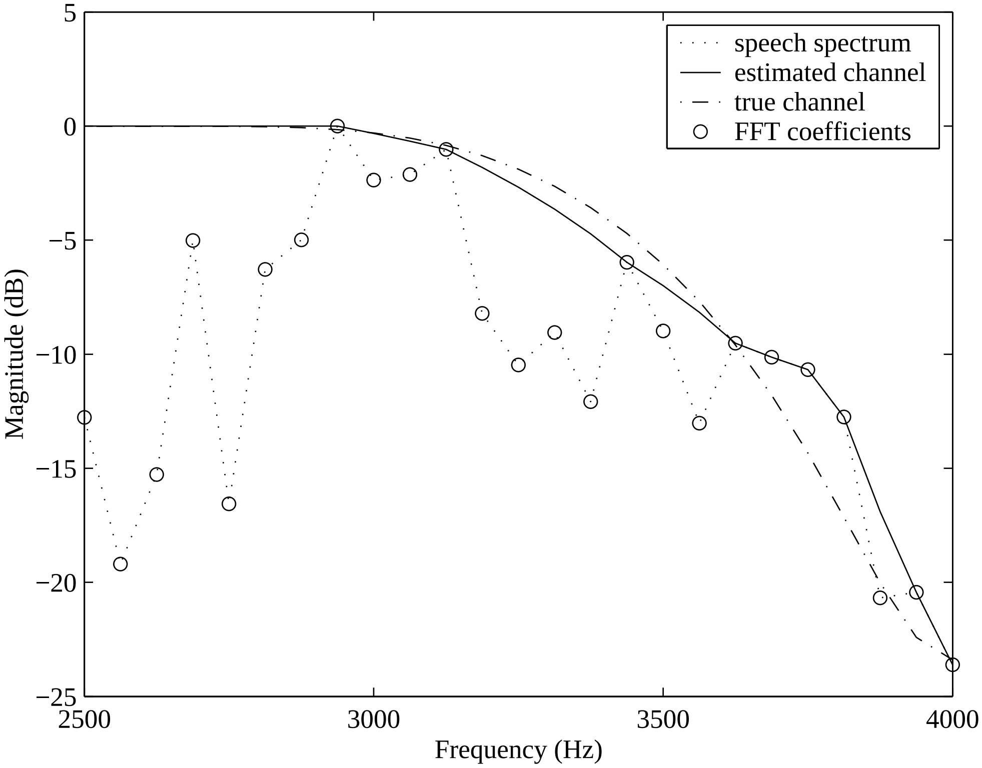


Fig. 4.6 Estimate of channel frequency response using an utterance of /ʃ/

be taken to ensure that the amplification is not overly applied. In the proposed scheme, the maximum gain of the equalizer is set at 12 dB to prevent raising too much noise. Figure 4.7 shows the equalizer response calculated from the channel estimate in Fig. 4.6.

As mentioned in the previous section, the equalizer is turned off when the SNR in the attenuated band is low. In addition, equalization is applied only after 10 instances of channel estimate have been taken and averaged because the estimate can be unreliable at the early stage.

Mode Selection and Highband Reconstruction

It is possible that the channel causes heavy distortion on the speech signal such that recovering the attenuated band via equalization is not feasible. Furthermore, the performance of highband excitation generation will degrade drastically if the 3 – 4 kHz band is too severely distorted.

To mitigate the effects of high attenuation, the proposed scheme introduces two modes

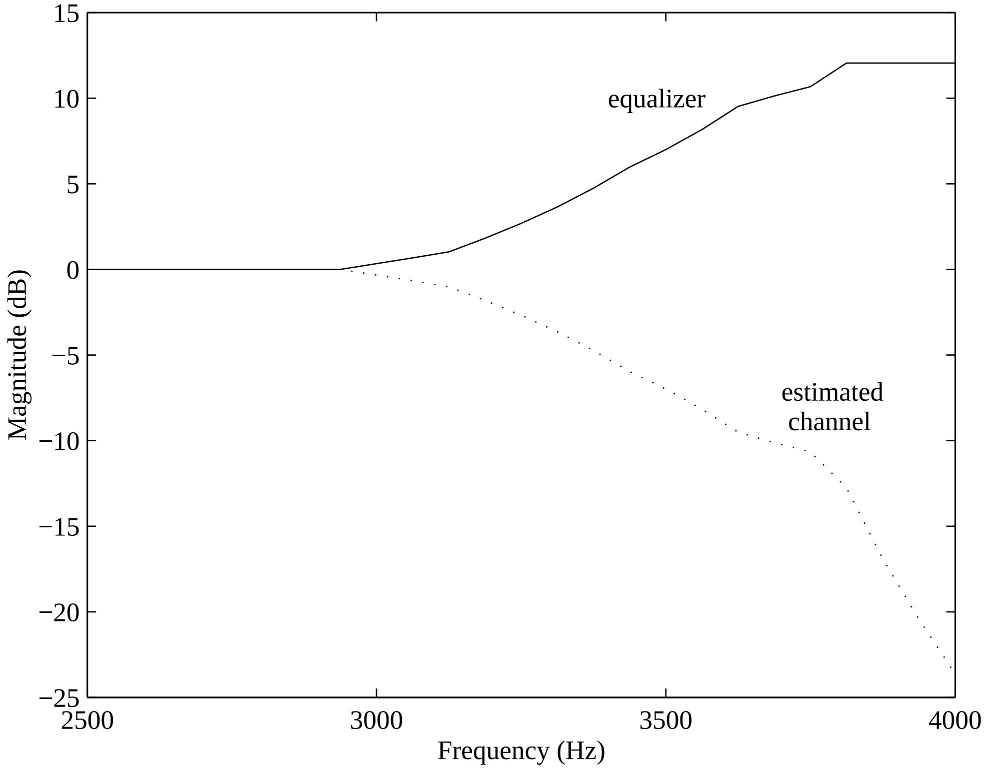


Fig. 4.7 Magnitude response of the equalizer and the estimated channel

of operation whose use depends on whether the attenuated band can be recovered by an equalizer.

The first mode is the normal operation, where equalization is applied to the speech spectrum below 4 kHz and the missing content above 4 kHz is reconstructed as in the core system. To use this mode of operation, the channel estimate of at least part of the 3500 – 4000 Hz band has to be greater than -20 dB.

On the other hand, if the frequency response of the channel filter is estimated to be less than -20 dB between 3500 and 4000 Hz, this band is considered as too severely attenuated and will be reconstructed as part of the missing highband using GMM mapping. As discussed in the previous section, this band is also considered to be lost if its estimated SNR is less than 5 dB. Different GMM's are used for the two modes; nonetheless, the methods are the same. Noise modulation is also employed for highband excitation generation in the second mode; the only difference is that the band between 2500 and 3500 Hz is used in the modulation.

4.3 Other Features

In addition to what was described in the previous two sections, the proposed scheme also contains a few other features for improving the robustness of the system. These additional features are presented in this section.

4.3.1 Spectral Shaping

In [42], experiments were conducted to find the best speech passbands between 50 and 7000 Hz. The results suggested that most of the improvement made in highband extension could be obtained by recovering the spectrum below 5000 Hz. Furthermore, the speech content at very high frequencies are more difficult to reconstruct and can easily cause artifacts if overestimated. Therefore, in the proposed scheme, highband spectral shaping is employed to taper the highband spectrum towards higher frequencies.

Spectral shaping is performed at two places. The first instance of spectral shaping is done during training. Instead of using a highpass filter to obtain the highband portion of the wideband speech, a bandpass filter with gradually decreasing high-end cutoff is employed. Figure 4.8 shows the filters used during training.

Spectral shaping is also explicitly applied at the end of bandwidth extension operations, when a postfilter is used to suppress high-frequency content. The postfilter is a 10^{th}-order FIR filter with the magnitude responses shown in Fig. 4.9.

4.3.2 Energy Adjustment

In the original bandwidth extension system, a constant 3-dB attenuation is applied to the reconstructed highband spectrum to reduce the chance of overestimation. With the use of spectral shaping postfilter, the need for the constant attenuation is diminished. However, overestimation still occurs from time to time for unvoiced sounds, especially in noisy environment. Therefore, in the proposed system, the 3-dB attenuation is applied when the narrowband speech contains significant high-frequency content. A threshold of zero crossing count is again used as the criterion because it can be a good indicator of the presence of high-frequency noise.

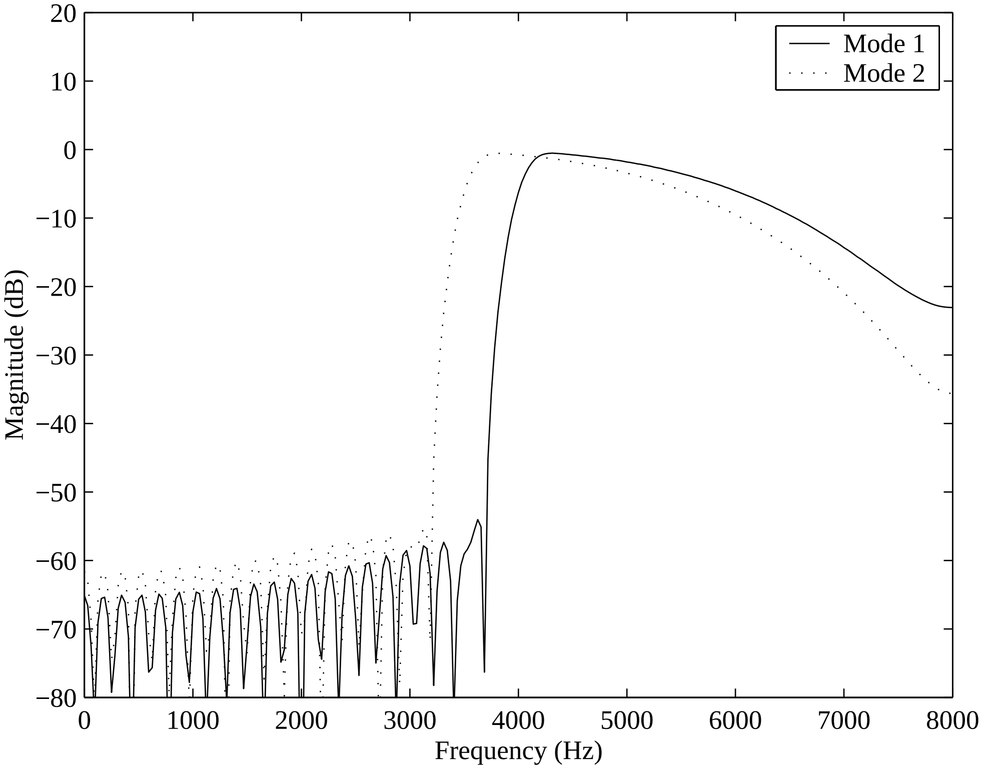


Fig. 4.8 Filters used to obtain highband speech in training

4.3.3 Parameter Smoothing

The bandpass filtering prior to parameter extraction sometimes causes fluctuations in LSF computations when the speech contains a formant close to 3200 Hz, the cutoff frequency of the bandpass filter. Because LSF exhibits high interframe correlation, the solution of smoothing the narrowband LSF before GMM mapping is employed. Specifically, a weighted average of LSF is applied as follows:

$$LSF_{sm}[i] = 0.5 * LSF[i] + 0.3 * LSF[i-1] + 0.2 * LSF[i-2], \tag{4.8}$$

where i is the index of the current frame, and LSF_{sm} is the resulting smoothed LSF vector. To avoid unnecessary smoothing, the weighted average is taken only when an LSF coefficient is greater than three radians.

In the core system, the same smoothing function is employed for the estimated highband LSF to reduce rapid spectral variations. To further ensure robustness, the weighted average

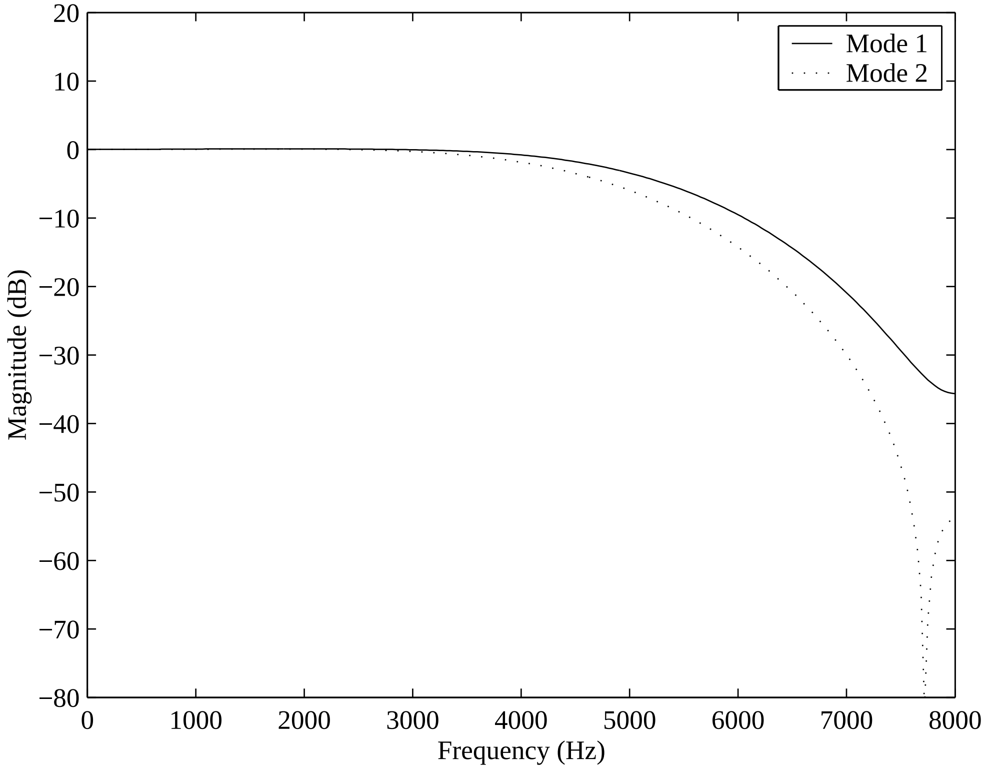


Fig. 4.9 Postfilter

is also taken conditionally for highband energy. Smoothing is applied if the current frame has higher energy than the two previous frames. Otherwise, the highband energy is not adjusted.

4.3.4 Frame Overlap

During training of the GMM, Hamming windows are applied to the highband speech before the calculation of LSF. Therefore, the reconstructed highband spectrum can have discontinuities at the frame boundaries. Smoother transitions between successive frames can be obtained by overlapping speech frames.

Frame overlap is one of the upgrades added to the core system after its publication in [22]. In the original upgrade, different degrees of overlap can be chosen. In this thesis work, an 50-percent frame overlap is employed to take into account the Hamming windows used during training. In other words, the input signal is processed as 20-ms frames with 10-ms frame advance.

4.4 Chapter Summary

In this section, the proposed scheme for robust bandwidth extension was presented. The effects of additive noise and channel variations were described and the solutions were discussed. Figure 4.10 shows the overall diagram of the proposed robust scheme.

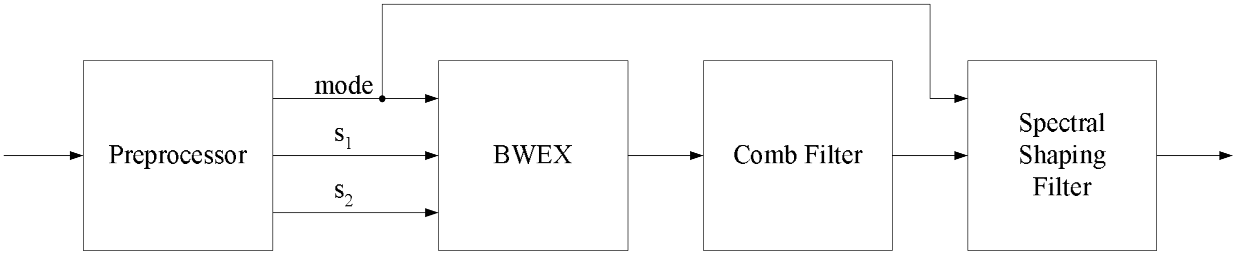


Fig. 4.10 Overall system

A preprocessor, employing the EVRC noise suppressor and an equalizer is applied to the input speech to enhance the narrowband speech prior to bandwidth extension. The preprocessor outputs two versions of enhanced speech, one enhanced for improved perceptual quality and the other for better feature extraction.

The range of the reconstructed highband can be chosen to be above either 3500 Hz or 4000 Hz depending on whether enough content is available in the input speech between 3500 and 4000 Hz.

To ensure robustness, the reconstructed wideband speech is passed through two filters, a comb filter that suppresses noise between harmonics below 4 kHz and a spectral shaping lowpass filter that tapers the highband spectrum to avoid overestimation.

Chapter 5

System Evaluation

In Chapter 4, the proposed scheme for robust bandwidth extension was presented. To evaluate the performance of the system, subjective tests have been conducted. This chapter describes the processes and the results of the subjective tests.

5.1 Test Data

The speech files used in the tests are taken from the files supplied by the 2002 IEEE Workshop on Speech Coding [43]. The original speech data are recordings of two male and two female utterances sampled at 16 kHz. Each file contains an utterance of two unrelated sentences taken from the phonetically balanced Harvard list [44]. The speech signals were recorded with 0.3 second of silence at both the beginning and the end of each file and 0.6 second of silence between the two sentences. Table 5.1 shows the sentences spoken in the test files.

Table 5.1 Test speech utterances

Speaker	Sentences
Female 1	The small pup gnawed a hole in the sock. The fish twisted and turned on the bent hook.
Female 2	Weave the carpet on the right hand side. Hemp is a weed found in parts of the tropics.
Male 1	His wide grin earned many friends. Flax makes a fine brand of paper.
Male 2	Hats are worn to tea and not to dinner. The ramp led up to the wide highway.

The signals are filtered and downsampled to 8 kHz for input to the bandwidth extension system. To test the proposed algorithm, two different filters are employed and their magnitude responses are shown in Fig. 5.1. The first filter has a sharp cutoff at 3400 Hz while the second filter has a gradually decreasing frequency response at high frequencies. Because the focus of this thesis work is on highband extension, the filters do not cut off the lowband spectra during testing.

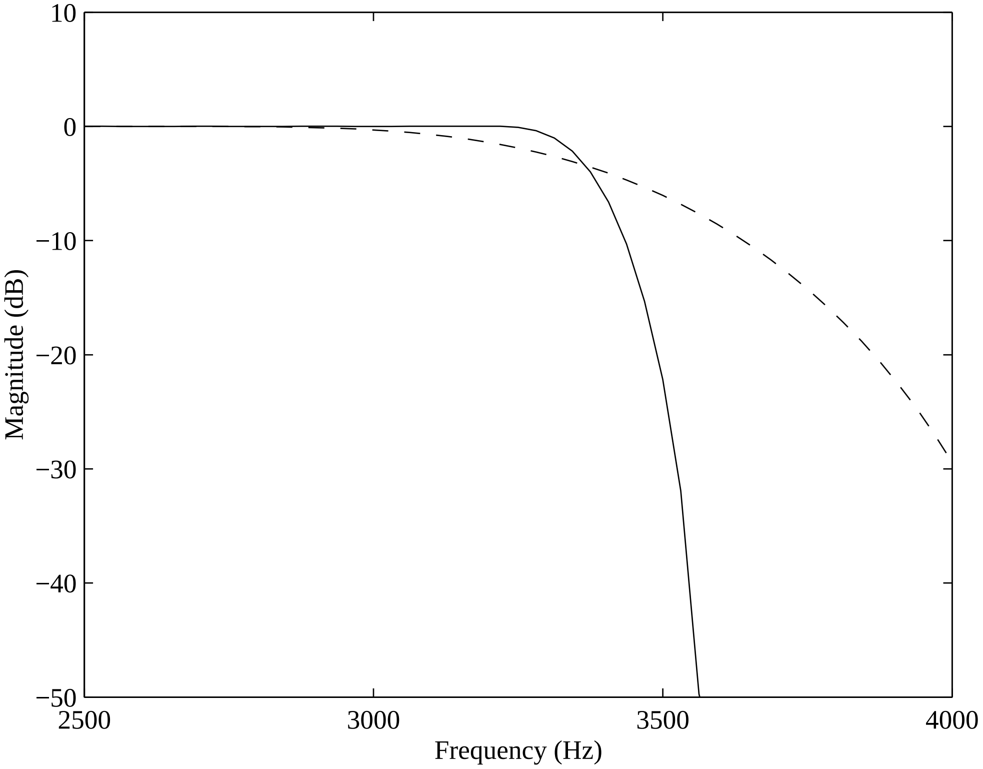


Fig. 5.1 Filters used in the subjective tests

To test for robustness against additive noise, five different kinds of environmental noise are artificially injected into the speech. The five noise sources are babble, car noise, factory noise, Hoth noise and street noise. Samples of babble and factory noise are taken from the Noisex database [45], while car, factory, and street noise are taken from the noise database used in the characterization tests for the ITU-T G.729 Codec [46]. During the tests, the noise is scaled such that the corrupted speech signals have an SNR of 10 dB, where the SNR is defined as the ratio of the active speech level to the root mean square (RMS) of the noise,

$$SNR_{dB} = 20\log_{10}\left(\frac{\text{Active speech level}}{\text{RMS of noise}}\right),$$

and it is calculated after the noisy speech has passed through the filters. The active speech level is computed according to Method B in ITU-T Recommendation P.56 [47].

Finally, to simulate the transmission of coded speech, all speech data are coded using the ITU-T G.729 speech codec prior to bandwidth extension. The procedure for generating the test data can be summarized as in Fig. 5.2.

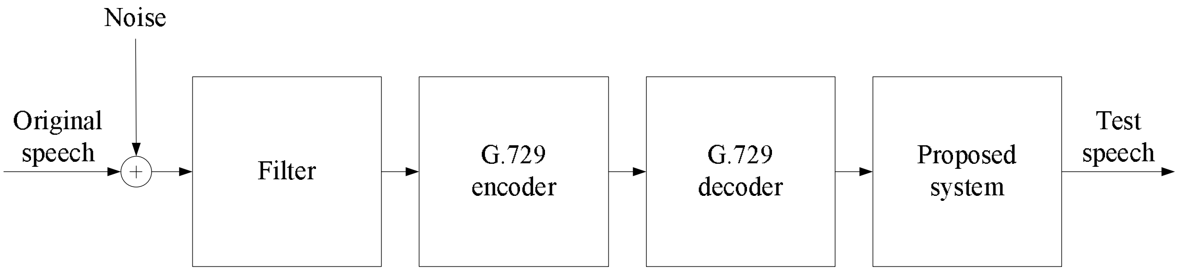


Fig. 5.2 Procedure for generating the test data

5.2 Test Procedure

A-B comparison tests were employed to determine the effectiveness of the proposed system under various conditions. There were a total of 36 tests covering nine different conditions. The sequence of the tests was determined randomly and the same for all participants. For each test, a narrowband signal and the corresponding reconstructed wideband signal are played to the test subject once, but the participants had the option of hearing the files in the same order for a second time before making their decisions. The playing order in each test was also determined randomly; however, the orders were set up such that two tests in each testing condition started with the narrowband signal and the other two with the reconstructed signal. At the end of each test, the participant would write down which of the two speech they preferred or if they had no preference. The exact wordings of the choices were

- A is better,
- B is better, and

- no preference,

where it was explained to the participants that A and B denoted the first and the second files, respectively, played in each test.

Five test subjects participated in the listening test. All of the participants were students of the Electrical and Computer Engineering department at McGill University.

Table 5.2 shows the conditions that were tested, where h_1 refers to the filter with sharp cutoff and h_2 the filter with gradually decreasing magnitude response. As a reference for the best performance possible, the tests include the situation when the input signal is clean and the channel has a flat frequency response. In the cases of noisy speech, the output of the EVRC noise suppressor was used as the narrowband signal in the test so that its effects could be excluded from the results.

Table 5.2 Tested conditions

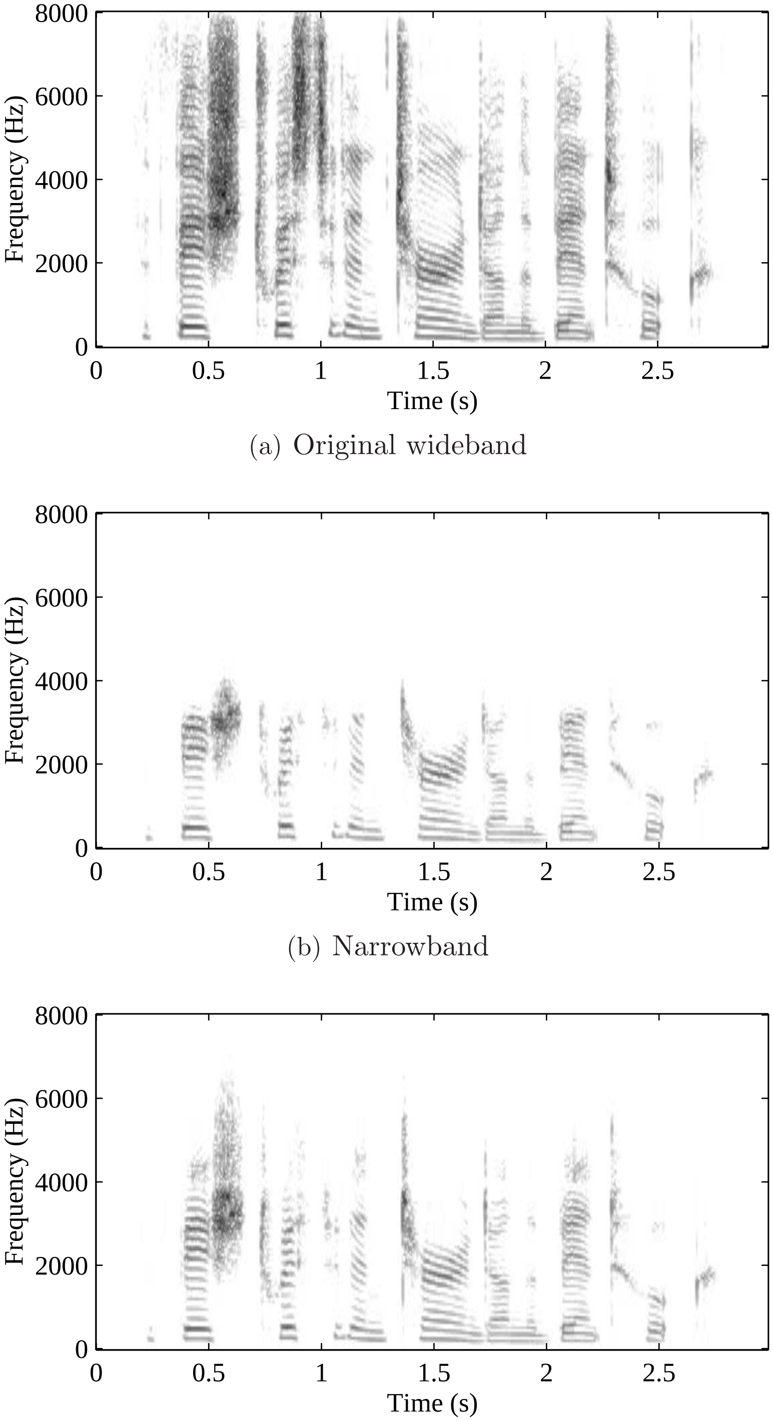
Channel	Noise
flat	none
h_1	none car
h_2	none babble car factory Hoth street

5.3 Results and Discussions

Fig. 5.3 and Fig. 5.4 depict the spectrograms of the first female utterance in the test, showing the evolution of the speech spectra for the speech under different conditions.

It can be seen from the spectrograms that the algorithm can estimate the highband spectrum with reasonable success, although the content at very high frequencies are suppressed by the spectral shaping postfilter. To further evaluate the system, results from the subjective tests are analyzed.

Table 5.3 shows the results of the subjective tests. The results for a given test are expressed in the format u_1-u_2-u_3, where u_1 refers to the number of participants that preferred

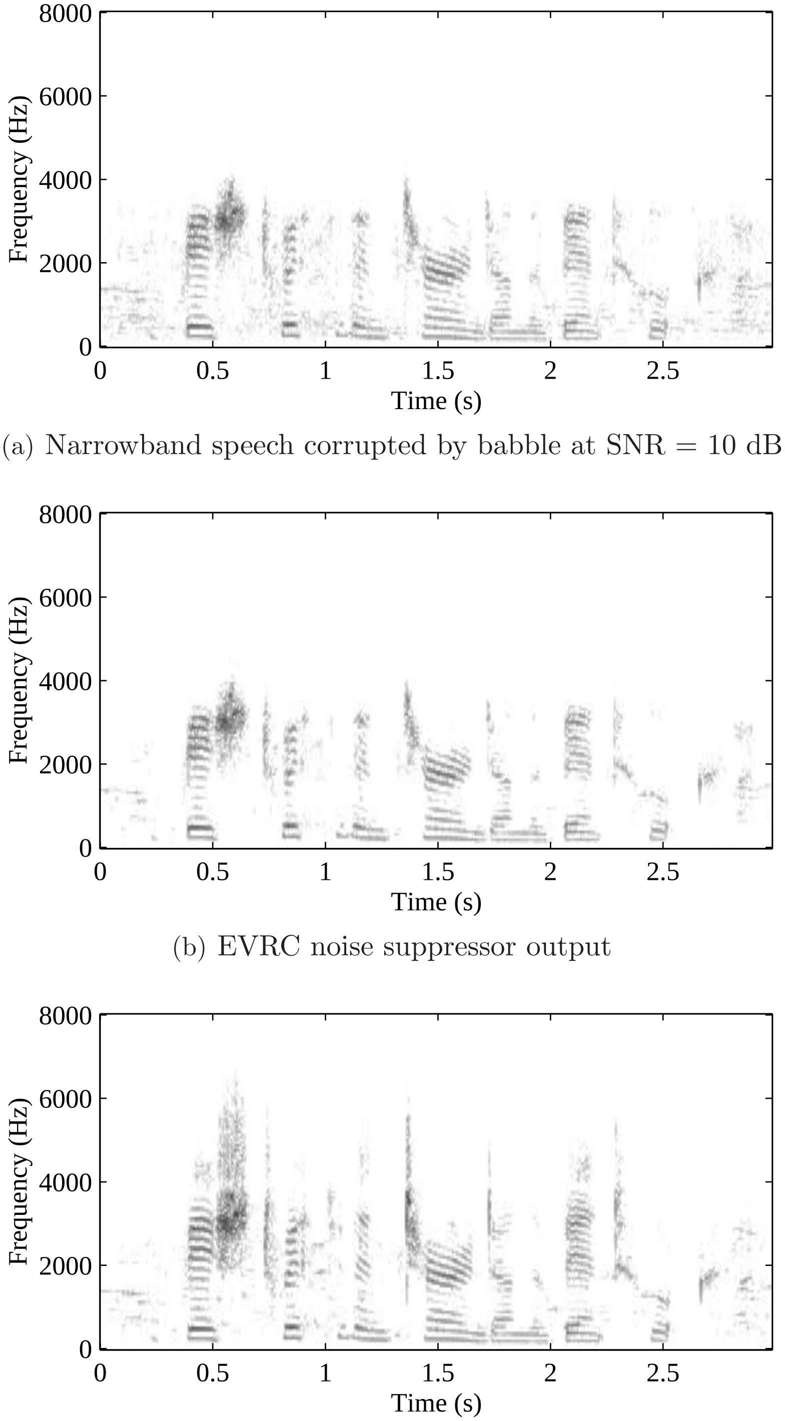


(a) Original wideband

(b) Narrowband

(c) Reconstructed wideband

Fig. 5.3 Spectrograms of speech under clean condition



(a) Narrowband speech corrupted by babble at SNR = 10 dB

(b) EVRC noise suppressor output

(c) Reconstructed wideband

Fig. 5.4 Spectrograms of speech under noisy condition

the speech reconstructed by the proposed algorithm, u_2 the number of no preferences, and u_3 the number of participants that chose the narrowband speech.

Table 5.3 Results of subjective tests. The scores u_1-u_2-u_3 denote in order (u_1) the number of participants that preferred the bandwidth extended speech, (u_2) the number of no preferences, and (u_3) the number of participants that chose the narrowband speech.

Channel	Noise	Speech				Total
		Female 1	Female 2	Male 1	Male 2	
flat	none	4 - 0 - 1	4 - 0 - 1	3 - 1 - 1	2 - 1 - 2	13 - 2 - 5
h_1	none	4 - 0 - 1	3 - 0 - 2	5 - 0 - 0	4 - 0 - 1	16 - 0 - 4
	car	4 - 0 - 1	4 - 0 - 1	2 - 2 - 1	3 - 2 - 0	13 - 4 - 3
h_2	none	4 - 0 - 1	3 - 0 - 2	3 - 0 - 2	4 - 0 - 1	14 - 0 - 6
	babble	5 - 0 - 0	4 - 0 - 1	4 - 0 - 1	2 - 1 - 2	15 - 1 - 4
	car	3 - 0 - 2	5 - 0 - 0	4 - 1 - 0	2 - 1 - 2	14 - 2 - 4
	factory	3 - 1 - 1	5 - 0 - 0	4 - 0 - 1	2 - 1 - 2	14 - 2 - 4
	street	5 - 0 - 0	5 - 0 - 0	5 - 0 - 0	3 - 1 - 1	18 - 1 - 1
Total		32 - 1 - 7	33 - 0 - 7	30 - 4 - 6	22 - 7 -11	117 -12 -31

Several observations can be drawn from the test results. First, the tests confirm that the proposed scheme can work under various channel conditions. Overall, the bandwidth extended speech was preferred over 70% of the time while being rejected by the listeners in less than 20% of the tests. The scores are approximately uniform for all testing conditions, showing that improvements are still made despite the reduction in narrowband-highband correlation due to adverse conditions. In fact, the best scores were obtained when the speech was corrupted by street noise or sharply lowpass filtered at 3400 Hz. On the other hand, input with the "perfect" characteristics, i.e. clean and non-filtered, results in one of the lowest scores. This may be attributed to the fact that there is less room for quality improvement when the input contains all the information below 4 kHz. With the narrowband signal already having close to wideband quality, the listeners become more aware of the distortions introduced by the bandwidth extension algorithm during comparison.

A breakdown of the test results by speaker's gender also gives an important insight. The combined test score for the two female speech files is 65-1-14, whereas the two male speech files have a total score of 52-11-17. In other words, bandwidth extension achieved a quality gain for female speech in over 80% of the tests but had only a 65% success rate for male speech. The number of rejections are approximately equal for male and female speech, and the discrepancy mostly shows up in the number of no preferences. This is mainly due to the fact that the frequency spectra of male speech have low energies between 3 and 4 kHz and are much more lowpass tilted than those of female voice. Consequently, the estimated excitation gain is likely to be low and the reconstructed excitation signal is too weak to make a perceptible difference.

Comments from the participants after the tests indicated that most of the time a distinctive difference in speech clarity could be heard. In addition, the reconstructed speech gave the feeling that the speaker was just nearby. However, sometimes the quality gain of reconstructed highband was offset by the extra perceived noise introduced by bandwidth extension. When the participants opted for the more muffled narrowband speech, the reasons were generally the temporal smoothness, which was maintained in the original narrowband signal but disturbed when the highband was added due to fluctuations in the reconstructed high-frequency spectrum.

Originally, Hoth noise was included as one of the test conditions. However, it was determined that the proposed system did not have much effect when the input speech was corrupted by Hoth noise at 10-dB SNR and was subsequently taken out of the tests for the final two participants. Because Hoth noise contained significant high-frequency content, much of the speech component above 2.5 kHz was destroyed at 10-dB SNR, leaving little information for highband excitation generation. The result was virtually no highband reconstruction, and it was reflected in the test score of 3-0-9 for the three test subjects that evaluated it, including no preferences for all six tests using male speech.

5.4 Algorithmic Delays

As mentioned in Chapter 1, it is important that a bandwidth extension system does not cause any noticeable delays. The proposed algorithm has several inherent processing overheads:

- 3 ms from the EVRC noise suppressor,
- 20 ms due to the frame length used, and
- 3 ms as a result of the interpolation filter.

Therefore, a total of 26 ms of algorithmic delays is introduced by the system.

5.5 Chapter Summary

In this chapter, performance of the proposed robust bandwidth extension system was examined. A-B comparison tests were conducted and the results were analyzed. Finally, the delays caused by the algorithm were calculated.

Chapter 6

Conclusions

This thesis has focused on the design of a robust bandwidth extension system. The goal was to implement a system that can enhance the quality of narrowband telephone speech by reconstructing the high-frequency spectrum while at the same time is able to operate under various channel conditions. The proposed algorithm is based on the system presented in [22] and improves upon it by including a preprocessor and two postfilters for safeguard against adverse environments.

This chapter gives a summary of this thesis work and discusses possible future directions for research.

6.1 Thesis Summary

In Chapter 1, the concept of bandwidth extension is introduced and a brief overview of the history of bandwidth extension is presented. The motivation for bandwidth extension is humans' innate preference for more natural speech quality, which the narrowband telephone speech lacks due to the loss of low- and high-frequency content. Several existing methods are mentioned in the chapter, and the need for robustness is discussed.

The existing methods are described in more details in Chapter 2. First, the linear source-filter model for human speech production is introduced. Under this model, bandwidth extension can be separated into the tasks of reconstructing the wideband spectral envelope and excitation signal, which include the generation of the excitation waveform and the estimation of its gain. Some methods that have previously been applied to each task are explained. At the end of the chapter, previous works regarding the feasibility and possible limitation of a bandwidth extension system are discussed.

Chapter 3 presents the core bandwidth extension system used in this thesis work. The system employs the method of bandpass envelope modulated noise for highband excitation signal generation and GMM mapping for the reconstruction of highband spectral envelopes and excitation gain. One important aspect of the core system is the use of equalizers to recover the attenuated spectra below 4 kHz. If the attenuated band is not too strongly suppressed, equalization can recover the content while maintaining the spectral structure of the original speech.

The effects that additive noise and channel variations have on the core bandwidth extension system are discussed in Chapter 4, and solutions to counter these effects are proposed. In Section 4.1, the proposed scheme to combat additive environmental and coding noise is presented. The EVRC noise suppressor is employed for quality improvement in the telephone band. In addition, it is also modified to produce a second output used for GMM mapping, obtained by employing a higher suppression limit. Increasing the suppression limit allows more noise removal in the spectral valleys and reduces the chance of misclassifying the input speech signal during GMM mapping. Afterwards, a one-tap adaptive comb filter is designed to reduce noise between harmonics.

Section 4.2 presents the solutions for the problem of channel variations. For equalization to achieve its optimal performance, the channel frequency response must be estimated. The proposed scheme estimates the channel response in the frequency domain using a 128-point FFT, and an equalizer is constructed as the inverse channel filter with a maximum gain of 12 dB. However, if the attenuation is too severe or the noise too strong at the high frequencies, the band between 3500 and 4000 Hz is considered missing, and the recovery mode is switched from equalization to statistical recovery using GMM mapping. Furthermore, to reduce the effects of channel mismatch between training and testing conditions, the input speech is passed through a $300 - 3200$ Hz bandpass filter and then downsampled to 6400 Hz prior to GMM mapping. This procedure allows the estimations of highband spectral envelopes and excitation gain to be unaffected by how the channel behaves between 3200 and 4000 Hz.

Section 4.3 describes other components of the proposed system. A spectral shaping postfilter is employed to suppressed the very high-frequency content because that part of the speech spectrum is highly noisy and its reconstruction tends to create artifacts. A constant 3 dB attenuation is applied to the highband when the input is considered to have significant high-frequency components. To reduce possible temporal fluctuations in the highband, successive frames are overlapped and the parameters are smoothed by a weighted average.

Chapter 5 evaluates the performance of the proposed system. Spectrograms were used as preliminary tests to see the reconstructed spectra. Listening tests were then performed to compare the subjective quality between narrowband and reconstructed wideband speech. The test results show that the reconstructed speech is usually preferred by the listeners in all tested conditions.

6.2 Future Research Directions

The subjective tests have shown that the listeners sometimes still prefer the original narrowband voice over the reconstructed wideband speech. This result indicates that there is still much room for improvement. Moreover, many other real world situations exist and are not covered by the proposed scheme. This section provides some general ideas for further improvement and research.

First, the system is designed for stationary environments and is not equipped for sudden changes in channel conditions. For example, an abrupt increase in high-frequency noise, such as a car honk, may cause unacceptable degradation to the reconstructed speech. The time-varying channel characteristics of various mobile applications can also pose problems. The system can be improved by expanding it to cover more different conditions.

In addition, the highband reconstruction in effect shuts off when the noise becomes too strong, as evident from the test results of speech corrupted by Hoth noise. The system might be able to benefit from a more aggressive approach in both narrowband noise reduction and highband estimation. A noise suppressor specifically designed to fit the needs of a bandwidth extension system could be a first step in this direction.

Finally, as the subjective tests showed, the system performs more poorly with male speech than female speech. Therefore, extra performance gain might be obtained by employing a gender-dependent algorithm.

References

[1] M. G. Croll, “Sound-quality improvement of broadcast telephone calls,” Tech. Rep. 1972/26, BBC, Aug. 1972.

[2] J. Makhoul and M. Berouti, “High-frequency regeneration in speech coding systems,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Washington, DC), pp. 428–431, Apr. 1979.

[3] P. J. Patrick, R. Steele, and C. S. Xydeas, “Frequency compression of 7.6 kHz speech into 3.3 kHz bandwidth,” *IEEE Trans. Communications*, vol. 31, pp. 692–701, May 1983.

[4] H. Carl and U. Heute, “Bandwidth enhancement of narrow-band speech signals,” in *Proc. European Signal Processing Conf.*, (Edinburgh, Scotland), pp. 1178–1181, Sept. 1994.

[5] Y. Yoshida and M. Abe, “An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping,” in *Proc. Int. Conf. Spoken Language Processing*, (Yokohama, Japan), pp. 1591–1594, Sept. 1994.

[6] V. Iyengar, R. Rabipour, P. Mermelstein, and B. R. Shelton, “Speech bandwidth extension method and apparatus.” U.S. Patent 5455888, Oct. 1995.

[7] N. Enbom and W. B. Kleijn, “Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients,” in *IEEE Speech Coding Workshop*, (Porvoo, Finland), pp. 171–173, June 1999.

[8] J. Epps and W. H. Holmes, “A new technique for wideband enhancement of coded narrowband speech,” in *IEEE Speech Coding Workshop*, (Porvoo, Finland), pp. 172–174, June 1999.

[9] Y. Qian and P. Kabal, “Wideband speech recovery from narrowband speech using classified codebook mapping,” in *Proc. Australian Int. Conf. Speech Science, Technol.*, (Melbourne, Australia), pp. 106–111, Dec. 2002.

[10] Y. Nakatoh, M. Tsushima, and T. Norimatsu, “Generation of broadband speech from narrowband speech using piecewise linear mapping,” in *Proc. European Conf. Speech Commun., Technol.*, (Rhodes, Greece), pp. 1643–1646, Sept. 1997.

[11] G. Miet, A. Gerrits, and J. C. Valière, “Low-band extension of telephone-band speech,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Istanbul, Turkey), pp. 1851–1854, June 2000.

[12] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, “Speech enhancement via frequency bandwidth extension using line spectral frequencies,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Salt Lake City, UT), pp. 665–668, May 2001.

[13] Y. M. Cheng, D. O’Shaughnessy, and P. Mermelstein, “Statistical recovery of wideband speech from narrowband speech,” *IEEE Trans. Speech, Audio Processing*, vol. 2, pp. 544–548, Oct. 1994.

[14] K.-Y. Park and H. S. Kim, “Narrowband to wideband conversion of speech using GMM based transformation,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Istanbul, Turkey), pp. 1843–1846, June 2000.

[15] M. Nilsson and W. B. Kleijn, “Avoiding over-estimation in bandwidth extension of telephony speech,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Salt Lake City, UT), pp. 869–872, May 2001.

[16] Y. Qian and P. Kabal, “Dual-mode wideband speech recovery from narrowband speech,” in *Proc. European Conf. Speech Commun., Technol.*, (Geneva, Switzerland), pp. 1433–1436, Sept. 2003.

[17] P. Jax and P. Vary, “Wideband extension of telephone speech using a Hidden Markov Model,” in *IEEE Speech Coding Workshop*, (Delavan, WI), pp. 133–135, Sept. 2000.

[18] G. Chen and V. Parsa, “HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Montreal, QC), pp. 709–712, May 2004.

[19] H. Yasukawa, “Spectrum broadening of telephone band signals using multirate processing for speech quality enhancement,” *IEICE Trans. Fundamentals*, vol. E78-A, pp. 996–998, Aug. 1995.

[20] H. Yasukawa, “Restoration of wide band signal from telephone speech using linear prediction error processing,” in *Proc. Int. Conf. Spoken Language Processing*, (Philadelphia, PA), pp. 901–904, Oct. 1996.

[21] B. Iser and G. Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signal," in *Proc. European Conf. Speech Commun., Technol.*, (Geneva, Switzerland), pp. 565–568, Sept. 2003.

[22] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Montreal, QC), pp. 713–716, May 2004.

[23] J. Epps, *Wideband Extension of Narrowband Speech for Enhancement and Coding*. Ph.D. thesis, University of South Wales, School of Electrical Engineering and Telecommunications, Sept. 2000.

[24] E. E. David, M. R. Schroeder, B. F. Logan, and A. J. Prestigiacomo, "Voice-excited vocoders for practical speech bandwidth reduction," *IEEE Trans. Inform. Theory*, vol. 8, pp. 101–105, Sept. 1962.

[25] C. K. Un and D. T. Magill, "The residual-excited linear prediction vocoder with transmission rate below 9.6 kbits/s," *IEEE Trans. Communications*, vol. 23, pp. 1466–1474, Dec. 1975.

[26] J.-M. Valin and R. Lefebvre, "Bandwidth extension of narrowband speech for low bit-rate wideband coding," in *IEEE Speech Coding Workshop*, (Delavan, WI), pp. 130–132, Sept. 2000.

[27] 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., Technol.*, (Madrid, Spain), pp. 165–168, Sept. 1995.

[28] U. Kornagel, "Spectral widening of the excitation signal for telephone-band speech enhancement," in *Proc. Int. Workshop Acoustic Echo, Noise Control*, (Darmstadt, Germany), pp. 215–218, Sept. 2001.

[29] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, pp. 1707–1719, Aug. 2003.

[30] A. McCree, "A 14 kb/s wideband speech coder with a parametric highband model," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Istanbul, Turkey), pp. 1153–1156, June 2000.

[31] D. O. Bowker, J. T. Ganley, and J. H. James, "Telephone network speech signal enhancement." U.S. Patent 5195132, Mar. 1993.

[32] M. Nilsson, S. V. Andersen, and W. B. Kleijn, "On the mutual information between frequency bands in speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Istanbul, Turkey), pp. 1327–1330, May 2000.

[33] 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. Acoustics, Speech, Signal Processing*, (Orlando, FL), pp. 525–528, May 2002.

[34] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Orlando, FL), pp. 237–240, May 2002.

[35] Y. Agiomyrgiannakis and Y. Stylianou, "Combined estimation/coding of highband spectral envelopes for speech spectrum expansion," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Montreal, QC), pp. 469–472, May 2004.

[36] "Transmission performance characteristics of pulse code modulation channels." ITU-T Recommendation G.712, June 1996.

[37] "Enhanced variable rate codec (EVRC)." 3GPP2 Specification C.S0014-0 v1.0, Dec. 1999.

[38] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 59–70, Jan. 1995.

[39] A. McCree, T. Unno, A. Anandakumar, A. Bernard, and E. Paksoy, "An embedded adaptive multi-rate wideband speech coder," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Salt Lake City, UT), pp. 761–764, May 2001.

[40] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," in *IEEE Speech Coding Workshop*, (Porvoo, Finland), pp. 165–167, June 1999.

[41] R. Martin, I. Wittke, and P. Jax, "Optimized estimation of spectral parameters for the coding of noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Istanbul, Turkey), pp. 1479–1482, June 2000.

[42] S. Voran, "Listener ratings of speech passbands," in *IEEE Workshop on Speech Coding for Telecommunications*, (Pocono Manor, PA).

[43] 2002 IEEE Workshop on Speech Coding. [Online]. Available: http://kt-lab.ics.nitech.ac.jp/~sako/scw/web/official/index.php.

[44] IEEE Standards Publication No. 297, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoustics*, vol. AU-17, pp. 225–246, Sept. 1969.

[45] Signal Processing Information Base. [Online]. Available: http://spib.rice.edu/spib.html.

[46] “ITU-T coded speech database.” Supplement 23 to ITU-T P-series Recommendations, Feb. 1998.

[47] “Objective measurement of active speech level.” ITU-T Recommendation P.56, Mar. 1993.