Auditory Distortion Measures for Speech Coder Evaluation

by

Aloknath De

A thesis submitted to the Faculty of the Graduate
Studies and Research in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Department of Electrical Engineering
McGill University
Montréal, Canada
October, 1993

To my grandmother

Abstract

One of the important research problems in the area of speech coding is to determine the sound quality of coded speech signals. This quality can best be evaluated by a subjective assessment which is often difficult to administer and time-consuming. An objective measure which is consistent with subjective assessment could play a vital role in the evaluation as well as in the design of a low bit-rate speech coder. In this dissertation, we introduce two distortion measures for speech coder evaluation. Since the perceptual abilities of a human being determine the precision with which speech data must be processed, we consider the details of cochlear (inner ear) and other auditory processing. Using Lyon's auditory model, the time-domain speech signal is mapped onto a perceptual-domain (PD). Any speech utterance is communicated to the brain through a series of all-or-none electrical spikes (firings) and the PD representation provides information pertaining to the probability-of-firings in the neural channels. Our first measure, namely the cochlear discrimination information (CDI), evaluates the cross-entropy of the neural firings for the coded speech with respect to those for the original one. With this measure, we also compute the rate-distortion function determining the lowest bit-rate required for a specified amount of distortion. In the second measure, namely the cochlear hidden Markovian (CHM) measure, we attempt to capture the high-level processing in the brain with simple hidden Markov models (HMMs). We characterize the firing events by HMMs where the order of occurrence of PD observations and correlations among adjacent observations are modeled suitably. For computing the coder distortion, the PD observations of the coded speech are matched against the HMMs derived from the PD observations of the original speech. Experimental results show that these measures conform to subjective evaluation results in majority of the cases. Finally, the introduced measures are also applied in speech coder analysis, e.g., in the pitch frequency determination and the evaluation of noise weighting schemes.

Sommaire

L'un des problèmes de recherche importants dans le domaine du codage de la parole est de déterminer la qualité de son des signaux de parole codés. Cette qualité est évaluée à son meilleur par un jugement subjectif, ce qui est souvent difficile à organiser et assez long. Une mesure objective consistente avec l'évaluation subjective pourrait jouer un rôle vital dans la conception de codeurs de parole à bas taux de bits ainsi que dans le jugement quailitatif de la parole. Nous introduisons dans cette dissertation deux mesures de distortion pour l'évaluation de performance de codeurs de parole. Etant donné que la précision avec laquelle les données de parole devraient être traitées est déterminée par les capacités perceptuelles de l'être humain, nous considérons les détails du traitement de signaux par la cochlea (intérieur de l'oreille), ainsi que d'autres traitements par le système autitif. En utilisant le modèle auditif de Lyon, le parole dans le domaine temporel est transformée dans le domaine perceptuel (PD). Chaque phrase parlée est communiquée au cerveau à travers une série d'impulsions électriques sur une base de tout ou rien, et la représentation PD offre des informations pertinent à la probabilité d'envoi des impulsions dans les canaux neuronaux. Notre première mesure, plus exactement la discrimination de l'information par la cochlea (CDI), évalue l'entropie croisée des impulsions envoyées pour la parole codée avec ceux de la parole originale. Avec cette mesure, nous calculons aussi une fonction taux-distortion pour déterminer le plus bas taux de bits requis pour un niveau de distortion donné. Dans la seconde mesure de distortion, la Markovienne cachée de la cochlea (CHM), nous essayons de capturer le traitement de haut niveau dans le cerveau à travers de simples modèles de Markov cachés (HMM). Nous caractérisons les événements d'envoi d'impulsions par des HMM où l'ordre de lieu d'observations PD et la correlation entre observations adjacentes sont proprement représentés. Pour calculer la distortion du codeur, les observations PD de la parole codée sont comparées aux HMMs dérivés des observations PD de la parole d'origine. Les résultats expérimentaux démontrent que ces mesures sont conforme à l'évaluation subjective dans la majorité des cas. Finalement, les mesures introduites sont appliqués à l'analyse dans le codage de la parole, par exemple, pour la détermination de la fréquence fondamentale et l'évaluation de modèles de pondération bruités.

Acknowledgements

I express sincere gratitude to my supervisor Prof. P. Kabal for his technical guidance as well as financial assistance throughout this work. His magnetic personality, logical thinking and pleasant behavior have helped me overcome many difficult times in my doctoral study. My debt to him is simply immeasurable.

I am thankful to Prof. H. Leib who has contributed to my teaching and research abilities in various ways. I would also like to thank Prof. G. Zames for his healthy advices as one of my Ph.D. committee members. I am grateful to all the professors and staff members who have assisted me in making my professional career sound.

I owe to all my friends who have helped cherish my four years of stay in Montréal. I extend my heartfelt thanks to Mr. S. Valaee and Dr. A. K. Khandani with whom I have had many stimulating discussions. I also thank Mr. N. Maroun for the French translation of the thesis abstract. I have enjoyed the environment of McGill university as well as INRS-Télécommunications. The facilities provided by them have contributed greatly to the accomplishment of this work.

I acknowledge the Canadian Institute for Telecommunications Research for sponsoring this project. I am thankful to Dr. M. Slaney of Apple Computers Inc. for providing me the MacEar program. I would thank the CCITT Speech Quality Expert Group, some members of which gave me an opportunity to introduce this work to them at Orlando, USA. I would also like to thank the Canadian Acoustical Association for recognizing this research work with the '1993 Alexander Graham Bell Prize'.

My thanks are due to my dear parents whose loving care and proper advices have brought me to the point where I am today. I also express my thanks to my brother Arup and sister Ipsita for their affection. A special thank is for my wife Samapti who has always been very understanding and encouraging. I appreciate being associated with such a nice family.

Finally, I thank God for providing me with all the basic amenities and making the texture of my life beautiful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In a typical source coding problem, a continuous-time continuous-amplitude bandlimited signal is sampled in the time domain at or above the minimum sampling rate required. This time-discretized signal with amplitude having continuous probability density function has an infinite *entropy*. To transmit the output of such a source and recover it exactly, a communication channel of infinite *capacity* is required. In practice, every channel, due to perturbation by noise, has a finite capacity. Thus, it is not possible to transmit the output of a continuous source over any channel and recover it exactly [1]. Accepting the fact that there will inevitably be some distortion, a typical source coder minimizes it by removing deliberately some information which is deemed 'not very important' to the destination. The extent to which the information should be removed depends on the bit-rate of the coder; the lower the bit-rate, the more information is needed to be removed.

In speech communication, the ultimate recipient of information is a human being and hence his/her perceptual abilities govern the precision with which speech data must be processed and transmitted. Thus, to reduce the amount of distortion, the speech data can be modified by an intentional removal of some information in accordance with the limitations of the auditory system. Determining 'what is not very important' to the auditory system and 'how the auditory system assesses' the relative importance of information is the primary task involved in devising a distortion measure for speech coders.

The sound quality of a given speech coder can best be evaluated by listening to

it. However, an extensive subjective testing of speech coders is difficult to administer and also time-consuming. Often, it is found to give inconsistent result due to the inherent non-repeatability of human responses. Moreover, it does not provide much insight into the factors which may lead to an improvement of the speech coding system. It is obvious that an objective quality measure, if suitably defined, could play an important role in the evaluation as well as in the design of low bit-rate speech coders. One important advantage of an objective quality measure is that its repeated application at different time under different environment gives the same performance.

Defining an appropriate objective quality measure for coded/distorted speech has thus become one of the pressing tasks to maintain a 'good' speech quality with low bit-rate coding or to assess the perceptual quality of any speech coder. We provide a brief overview of speech coding techniques in Section 1.1. The utility of having a 'good' objective measure is discussed in Section 1.2 and the motivation for our research is explained in Section 1.3. We present an outline of the thesis in Section 1.4 and state our contribution to original knowledge in Section 1.5.

1.1 Brief Overview of Speech Coding Techniques

The primary objective in speech coding research is to determine strategies for generating synthesized signals with as high quality as possible and at the same time adhering to other constraints such as bit-rate and coding delay. Many coding techniques exist for rates starting from 64 kilobits per sec (kbps) all the way down to 2.4 kbps (or even lower). Speech coding algorithms vary from the high-rate/low-complexity waveform coders to the medium- to low-rate/high-complexity vocoders or hybrid coders [2]. Their variations are primarily in the following four aspects: (i) selection of information (features) to be encoded, (ii) representation of features by appropriate parameters for encoding, (iii) quantization technique adopted for parameter discretization and (iv) distortion measures applied for estimating quantization loss.

Waveform coders analyze, code and reconstruct speech on a sample-by-sample basis. Time-domain waveform coders exploit waveform redundancies, e.g., the periodicity and slowly varying intensity while the spectral-domain waveform coders take advantage of the nonuniform distribution of speech information across different fre-

quencies [3]. On the other hand, source coders or vocoders utilize speech-specific model. They generally identify certain aspects of the speech spectrum as being important to model and generate speech with good reproduction of these aspects [4]. In speech production, the source may be either periodic generating a voiced speech or noisy and aperiodic resulting in an unvoiced speech. The fundamental frequency of the vocal cord vibration, in the utterance of a voiced speech, is termed as the pitch frequency. The resonances, termed as the formants, occur due to the poles of the vocal tract frequency response while the spectral nulls (anti-resonances) occur due to its zeros. Currently, the code-excited linear prediction (CELP) algorithm is the most widely used speech coding method and a typical low or medium bit-rate such coder encodes the formant, pitch and residual information separately [5].

An all-pole linear prediction filter synthesizes formant information and the filter parameters are determined by autocorrelation or covariance method. These parameters may be encoded directly or may be expressed in other forms such as reflection coefficients, area ratio parameters, line spectral frequency parameters etc. and then coded. They differ from each other from the perspectives of computational efficiency, quantization sensitivity etc. [2]. A pitch prediction filter is generally characterized by the pitch gain and lag value parameters. Depending on the bits available, the number of pitch predictor taps and the codebook size for the pitch parameters are decided [6]. For sending information about the residual signal, a random excitation codebook or often a structured algebraic codebook is used [7] and an appropriate codebook entry is selected from it.

While designing a speech coder, once the parameters pertaining to different features are appropriately selected, they are quantized. The quantization may be scalar, vector (single- or multi-stage) and scalar-vector. For example, U.S. federal standard 4.8 kbps CELP coder [5] has formant filter with scalar quantization of line spectral frequency parameters. However, vector quantization provides a substantial bit reduction for the same speech quality at the expense of higher memory and computational complexity. Different variations and hybrid forms of scalar and vector quantizations are used for different coder rates [8]. Similarly, the pitch prediction parameters can also be quantized in various ways [9]. The residual signals are usually vector-quantized and stored in a stochastic codebook [5].

In an analysis-by-synthesis type speech coder, the selection of appropriate indices for different codebook entries requires minimizing a distortion criterion. It is

possible to define different types of distortion measures for different features. However, selecting a codeword with respect to a global distortion measure could yield better results as such a selection could even take care of interactions among the features. The overall perceptual quality of a given speech coder could be evaluated subjectively or by a properly defined objective measure. In this research, we have primarily concentrated on the formulation of distortion measures using an auditory model in the front-end. We have not attempted to use the measure in the coding process, but only in the evaluation of speech coder performances.

1.2 Utility of Objective Measure

In this section, we explain the utility of deriving an objective quality measure. Obtaining a suitable distortion measure could offer several advantages such as (a) its use in evaluating speech coder performances, (b) its application in a rate-distortion analysis which could indicate the lowest possible bit-rate required for a particular speech quality and (c) its use in the design procedure of speech coders.

1.2.1 Evaluation of Coder Performance

Speech coders of several standardized data rates are designed to 'match' to the capacities of different communication channels. These encoders vary from each other from the view point of the coder architecture, the type of features encoded, the number of bits allocated to the features and so on. A wide variety of encoding algorithms introduces a broad range of linear and nonlinear coder distortions. All of these distortions are not equally perceived by the auditory system. As a consequence, if we can devise a distortion measure incorporating the human perception procedure, then that can appropriately be used to evaluate the performances of different speech coders.

1.2.2 Rate-Distortion Analysis

The need for a strong mathematical foundation for the field of data compression has resulted in the development of rate-distortion theory. The performance achieved by various data compression systems can be compared with absolute bounds derived from

rate-distortion theory. With a particular source and a defined distortion measure, it is possible to draw a rate-distortion curve which determines the lowest possible rate for allowing a particular amount of coder distortion. However, if the distortion measure is not properly defined, this limit may not portray the real picture. Defining an appropriate distortion measure would facilitate the determination of the coder rate limit for attaining a particular speech quality.

1.2.3 Design of Speech Coders

A distortion measure can help the design procedure of speech coders in three ways:

(1) In an analysis-by-synthesis type speech coder, from a stochastic codebook, all innovation sequence entries (in the case of an 'optimal' coder) or selectively chosen entries (in the case of a 'suboptimal' coder) are used along with the formant and pitch synthesis filters to generate several coded speech signals. The index of that codebook entry is transmitted which results in the minimum distortion as measured by the defined fidelity criterion. The distortion measure can thus be instrumental in selecting an 'appropriate' codebook entry.

(2) With a limited number of bits available per second, a strategic allocation of bits to different feature parameters becomes very important. The bit allocation strategy adopted for an 8 kbps coder can neither be scaled down directly for a 4 kbps coder nor be scaled up for a 16 kbps coder. The relative importance of the information to be transmitted plays a significant role. In the design phase, the defined distortion measure can be used for improving the bit allocation policy of a particular speech coder, be it a waveform coder, an analysis-by-synthesis coder or a vocoder.

(3) While designing a speech coder, an appropriate distortion measure not only helps in making a sound bit allocation policy, but also in 'populating' (also called 'training') the codebook. In the training phase, determining the centroid for each class with the defined distortion measure results in the design of an 'optimum' (at least in the local sense) codebook. If the distortion measure properly reflects the perceptual importance of information, then a fixed size codebook designed in this way will also be filled up with the entries which contain perceptually important information.

1.3 Motivation for Our Research

Over the last two decades, several objective measures have been suggested in the literature (references are in Chapter 2). It is a well-known fact that the time-domain objective measures such as the signal-to-noise ratio and the segmental signal-to-noise ratio do not perform well in the assessment or in the design of a low or medium bit-rate speech coder. Spectral measures, e.g., the log likelihood ratio measure, the log area ratio measure, the log spectral distortion measure and the Itakura-Saito measure, exhibit a better performance. However, most of these measures are based on the parameters of linear prediction filter modeling the formant structure and thus do not adequately feature the perceptual phenomena. Only about 80% of the perceived degradation in speech quality can be explained by the distortions of the spectral peaks or speech formants [10]. Therefore, it is important for a 'good' quality measure to consider distortion not only in the formant information, but also in the pitch and the residual information.

A few psychoacoustically-motivated measures such as the information index and the Bark spectral distortion measure has also been studied. In the recent literature, several auditory models have been proposed and investigated (references are in Chapter 3). Some of these models emulate the psychoacoustic observations fairly well, at least at the level of auditory periphery. Thus, an application of one such auditory model in the formulation of a distortion measure could result in good performance. This may, to some extent, increase the complexity of computing the measure value. Nevertheless, we believe that the measure could be used widely in practice as the computational burden is eased with further progress in the signal processing technique and the VLSI technology. Keeping this view in mind, we have conducted research on the topic of *Auditory Distortion Measures for Speech Coder Evaluation.* We emphasize accuracy over computational considerations in the evaluation of speech coders.

1.4 Outline of the Thesis

The format of the dissertation is as follows. Chapter 2 reviews the existing time-domain, spectral-domain as well as perceptually-motivated distortion measures. Chapter 3 discusses psychoacoustic observations relevant to speech perception, describes

Lyon's auditory (cochlear) model and defines a perceptual-domain. Chapter 4 proposes a cochlear discrimination information measure which compares the set of perceptual-domain parameters for the original and the coded speech signals. With this measure, performance of several speech coders is evaluated objectively and also a rate-distortion-theoretic analysis is pursued. Chapter 5 puts forward another distortion measure methodology which uses hidden Markov models. This measure is computationally more intensive, but captures the basics of high-level processing in addition to the signal processing at the auditory periphery. Chapter 6 outlines some other applications of the measures, for example, in the pitch extraction or in the evaluation of perceptual weighting schemes usually incorporated in a speech coder. Chapter 7 concludes this dissertation with relevant remarks and future research directions.

1.5 Our Contribution

In this thesis, we consider an auditory model and suggest two distinct approaches for devising distortion measures for coded speech. The fundamental difference between our approaches and the existing perceptually-motivated measures is in addressing the issue of the 'cause' rather than that of the 'effect' involved in speech perception. In other words, instead of merely considering the important perceptual effects observed, we emulate the auditory system as it is and use it in the formulation of our distortion measure.

Our primary contribution is in the processing of neural information obtained at the output of Lyon's auditory model. As explained in the dissertation, in reality, a series of electrical spikes (firings) is transmitted from the auditory periphery to the brain through the neural pathways. Here, we treat the neural pathways to be communication channels with an input alphabet of size two, i.e., firing and non-firing. Our first distortion measure deals with the neural firing probabilities and evaluates the neural firing cross-entropy of the coded speech with respect to that of the original one. With this measure, we compute the rate-distortion function for speech coder determining the lowest bit-rate required for a given amount of distortion. Speech coders with rates ranging from 4.8 kbps to 32 kbps are studied from the viewpoint of their performance with respect to the rate-distortion limits. In the second measure, a two-state (one each for firing and non-firing events) fully-connected hidden Markov model (HMM) is associated with each of the neural channels and various model

parameters are derived with the pertinent neural firing information of the original signal. For computing the coder distortion, the neural firing observations from the coded speech are matched against the derived HMMs. We believe that the second measure is more powerful as it utilizes the contextual information present in the neural firing pattern. Experimental results show that these measures conform to subjective evaluation results in majority of the cases. The introduced measures are also applied in speech coder analysis, e.g., in the pitch frequency determination and the evaluation of noise weighting schemes usually incorporated in a low bit-rate coder.

Chapter 2

Distortion Measures for Speech Coding

2.1 Introduction

A distortion measure for speech quality is a measure which can be computed directly from an original speech waveform and its coded/distorted version; and which conforms to the results of a subjective measure of speech quality [11]. Regression analysis establishes a quantitative relationship between an objective quality measure and a subjective evaluation method. A *correlation coefficient* (ρ), defined as [12]

$$\rho = \frac{\sum_k (S_k - \overline{S})(O_k - \overline{O})}{\left[\sum_k (S_k - \overline{S})^2\right]^{1/2} \left[\sum_k (O_k - \overline{O})^2\right]^{1/2}}, \tag{2.1}$$

is often used as a figure-of-merit to measure the degree of correlation between a standard subjective measure and an objective measure. In (2.1), S_k and O_k, respectively, are the subjective and objective measure values for the k-th speech utterance in a particular database; and $\overline{S}$ and $\overline{O}$ are the corresponding average test scores, averaged over all the utterances of the database. One major disadvantage of applying the regression analysis technique is the necessity of knowing the form of the regression equation a priori. An alternative method [13] uses Bayesian estimation and a nonlinear relationship is automatically determined during the training.

Several subjective as well as objective measures have been proposed in the literature. For many such subjective and objective measure pairs, the degrees of correlation

have also been determined. In Section 2.2, some of the standard subjective evaluation procedures are outlined. We describe a major class of time-domain distortion measures in Section 2.3 and a few spectral distortion measures in Section 2.4. Some of the perceptually-motivated distortion measures are discussed in Section 2.5.

2.2 Subjective Quality Measures

Subjective quality measures can be classified into two primary categories [14]: *utilitarian* and *analytic*. The utilitarian quality measures are found to be reliable and reasonably efficient in the test administration. These measures are based on a unidimensional scale and the result is provided by a single number so that the coded speech qualities can be directly compared. On the other hand, the analytic measures typically use more than one dimension for assessing the speech quality and are directed towards characterizing the underlying psychological components that determine the perceived quality. With either of the classes, an extensive listener training procedure is needed to ensure the reliability of these tests under different test environments.

2.2.1 Utilitarian Tests

Subjective measures very often address the speech intelligibility and the articulation aspects separately. The intelligibility tests are scored by the percentage of correct understanding of the meaning conveyed by the transmitted speech while the articulation tests are evaluated by the percentage of correct recognition of the sounds, words or sentences. Fletcher and Steinberg [15] have constructed, for their articulation test, random lists of nonsense monosyllables (nullifying the associated semantic memory) in the form of consonant-vowel-consonant (C-V-C). Later, Fairbanks [16] has modified this test by specifying the trailing vowel-consonants to the listeners and asking them to choose the leading consonant based on his/her interpretation of the test speech. Many refined versions of such rhyme tests have subsequently been suggested where the listener responses are restricted in different manners (e.g., the choice being limited to a finite set of rhyming words).

These tests are found to be appropriate for speech coding systems that generate moderately to severely distorted speech [11]. However, for highly intelligible systems,

other perceivable attributes (e.g., pleasantness, naturalness) become important. In an isopreference evaluation procedure [17], test speech signals each having a different speech level and contaminated with different levels of additive noise are passed through the test transmission system. Test results are usually reported as 'isopreference contours' in the two-dimensional parameter space of speech level vs. noise level. Listeners usually judge the test speech in terms of a reference speech; hence, they are often compelled to consider a smaller perceptual descriptor space than that might be desired. The most widely used utilitarian type subjective evaluation method is the *mean opinion score* (MOS) [18] in which the listeners rate the speech distortion under test on a five-point absolute scale (Rate 5: imperceptible; Rate 4: just perceptible, but not annoying; Rate 3: perceptible and slightly annoying; Rate 2: annoying, but not objectionable; Rate 1: very annoying and objectionable). Since the listeners have freedom to interpret the scale 'ratings' in their own way, the MOS score provides an agglomerative measure value for different types of coder distortions.

2.2.2 Analytic Tests

An alternative subjective evaluation approach is to rate the test speech on a multidimensional scale. One such popular multidimensional measure is the *diagnostic acceptability measure* (DAM) [19]. The DAM evaluates a speech signal on sixteen separate scales (covering the signal quality, the background quality and the overall quality), all of which have a range from 0 to 100 points. In a multidimensional perceptual space, the distorted speech signals are represented as points so that the relationship between an individual preference and an acoustic distortion can be studied [20]. Signal degradations such as *fluttering* (amplitude-modulated speech), *thin* (high-pass speech), *rasping* (peak-clipped speech), *interrupted* (packetized speech with 'glitches'), *nasal*; background noise such as *hissing* (noise-masked speech), *buzzing* (tandemmed digital system), *babbling* (narrowband system with errors), *rumbling* (low-frequency noise-masked speech); and overall qualities such as *intelligibility, pleasantness, acceptability* are considered in the DAM test [11]. This measure attempts to minimize the errors involved in the measurement process as well as that associated with the human variability.

2.3 Time-Domain Objective Measures

The most popular class of the time-domain measures is the signal-to-noise ratio (SNR) with its varied forms (e.g., the segmental SNR, the granular segmental SNR).

2.3.1 Signal-to-Noise Ratio

The signal-to-noise ratio (SNR), for measuring the coded speech quality, is defined as

$$\text{SNR} = 10 \log_{10} \frac{\sum_n x^2[n]}{\sum_n (y[n] - x[n])^2} \text{ dB}, \tag{2.2}$$

where $x[n]$ and $y[n]$ are the n-th original and coded speech samples, respectively. Numerous studies [11] have exhibited that the SNR measure does not correlate well with subjective evaluation results. In practice, any phase distortion with a delay variation limited to a few milliseconds has such a small effect on the signal quality that it can be disregarded in the context of most synthetic speech quality [21]. However, the SNR measure degrades drastically with any time misalignment of $\{x[n]\}$ and $\{y[n]\}$. The correlation coefficient (with the MOS score) for the SNR measure has been found to be 0.24 correlated only across the waveform-coder distortions [11] where it is expected to perform relatively well.

2.3.2 Segmental SNR

A major drawback of the SNR measure is that it treats the entire speech utterance as a single vector thereby presuming an unrealistic idea of a single comparison made by the listener after listening to the entire utterance. A better measure, usually referred to as the segmental SNR (SNR_{seg}), is an average measure of the SNR values in dB. The averaging is done over the M speech 'segments' present in an utterance, each segment being of the order of 16 ms long (i.e., N =128 samples with 8,000 Hz sampling rate). Mathematically, this measure can be written as [22]

$$\text{SNR}_{\text{seg}} = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \left\{ \frac{\sum_{n=1}^{N} x^2[n + Nm]}{\sum_{n=1}^{N} (y[n + Nm] - x[n + Nm])^2} \right\} \text{ dB}. \tag{2.3}$$

The correlation coefficient associated with this measure has been determined to be 0.77 across a wide range of waveform coder distortions [11]. Though SNR_{seg} provides better accuracy than the SNR measure, it also can not be considered as a reliable measure for the speech quality. In segments where an original speech has almost no signal components, a little noise could give rise to a large negative SNR for that segment which in turn causes a considerable bias in the overall measure of SNR_{seg}. A threshold-adjusted or frequency-weighted SNR_{seg} measure could be used which alleviates this problem to a great extent [11]. Another variation of the SNR_{seg} measure is the granular SNR_{seg} which has been found to be appropriate only for the evaluation of delta modulation or differential waveform coders [23].

2.4 Spectral Objective Measures

Several spectral distortion measures have been proposed in the literature including the log likelihood ratio measure, the log area ratio measure, the line spectral frequency-based measure, the log spectral distortion measure, the cepstral distance, the Itakura-Saito measure and the coherence function. These distortion measures are generally computed using speech segments typically between 15 and 30 ms long. They are much more reliable than the SNR measure and are less sensitive to the occurrence of time misalignments between the original and the coded speech [11].

2.4.1 Log Likelihood Ratio

The log likelihood ratio (LLR) distance for a speech segment is defined based on the assumption that samples of a speech can be represented by a p-th order all-pole linear predictive coding (LPC) model of the form

$$x[n] = \sum_{m=1}^{p} a_m x[n-m] + G_x u[n], \tag{2.4}$$

where $x[n]$ is the n-th speech sample, a_m (for $m = 1, 2, \cdots, p$) are the coefficients of an all-pole filter $1/A_x(z)$ which models the resonances of the speech production mechanism, G_x is the gain of the filter and $u[n]$ is an appropriate excitation source

for the filter. The LLR measure then can be defined as [24]

$$\text{LLR} = \log\left[\frac{\boldsymbol{a}_x \boldsymbol{R}_y \boldsymbol{a}_x^T}{\boldsymbol{a}_y \boldsymbol{R}_y \boldsymbol{a}_y^T}\right], \tag{2.5}$$

where $\boldsymbol{a}_x$ is the LPC coefficient vector $[1, -a_1^x, -a_2^x, \cdots, -a_p^x]$ for the original speech $\{x[n]\}$ and $\boldsymbol{a}_y$ is the LPC coefficient vector $[1, -a_1^y, -a_2^y, \cdots, -a_p^y]$ for the coded speech $\{y[n]\}$. $\boldsymbol{R}_y$, the correlation matrix of $\{y[n]\}$, has elements as

$$r_y(i,j) = \sum_{n=1}^{N-|i-j|} y[n]y[n+|i-j|], \quad \text{for} \quad i,j = 0,1,\cdots,p, \tag{2.6}$$

where N is the number of samples used in the analysis. The denominator in (2.5) measures the prediction residual energy when $\{y[n]\}$ is filtered with its all-zero analysis filter $A_y(z)$, whereas the numerator measures the same when $\{y[n]\}$ is passed through the filter $A_x(z)$. A correlation coefficient of 0.59 is achieved with this measure [11].

2.4.2 Log Area Ratio Measure

The reflection coefficients $\{k_m\}$, another representation of the LPC coefficients $\{a_m\}$, are spectrally less-sensitive to quantization than their counterparts. However, the reflection coefficients can be sensitive to quantization errors when their magnitudes are near unity (i.e., they represent narrow-bandwidth poles). To reduce the sensitivity, a suitable nonlinear transformation expanding the region near $|k_m| = 1$ can be followed based on which a log area ratio (LAR) distortion measure is defined as [11]

$$\text{LAR} = \sum_{m=1}^{p}\left[\log\left(\frac{1-k_m}{1+k_m}\right) - \log\left(\frac{1-k'_m}{1+k'_m}\right)\right]^2, \tag{2.7}$$

where p is the number of predictor coefficients and k_m, k'_m (for $m = 1, 2, \cdots, p$) are the reflection coefficients corresponding to the original and the coded signals, respectively. A correlation coefficient of the order of 0.62 is attained with this measure [11].

2.4.3 Line Spectral Frequency-based Measure

The line spectral frequency (LSF) coefficients are derived by mapping the p-zeros of an all-zero analysis LPC filter $A(z)$ onto the unit circle through two orthogonal

polynomials $P(z)$ and $Q(z)$ of $(p+1)$-st order as [2]

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \tag{2.8}$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}). \tag{2.9}$$

The resulting polynomials, $P(z)$ and $Q(z)$, have their roots in conjugate pairs. A multiobjective functional measure is formulated by using the LSF transformation in determining the spectral peak locations and the spectral peak bandwidths for the original and the distorted speech frames. This measure compares six parameters which are (a) a shift in peak location, (b) a change in peak bandwidth, (c) a change in peak energy, (d) differences in inter-frame peak movement, (e) lost peaks and (f) distortion-induced extra peaks. This measure has exhibited a correlation coefficient of 0.78 [10].

2.4.4 Log Spectral Distortion Measure

The notions of one-step prediction error and spectral factorization are two important properties using which an L_p norm-based log spectral distortion (LSD) measure is defined between two log spectral densities as [25]

$$\text{LSD} = \left\{ \int_{-\pi}^{\pi} |V(\omega)|^p \frac{d\omega}{2\pi} \right\}^{1/p}, \tag{2.10}$$

where

$$V(\omega) = \log \left[\frac{G_x^2}{|A_x(e^{j\omega})|^2} \right] - \log \left[\frac{G_y^2}{|A_y(e^{j\omega})|^2} \right] \tag{2.11}$$

with G_x and G_y as the LPC gain coefficients; and $A_x(e^{j\omega})$ and $A_y(e^{j\omega})$ as the LPC model polynomials corresponding to the original and the coded speech signals, respectively. The most common choices in (2.10) for p are 1, 2 and ∞ giving rise to the mean absolute, the root mean square and the maximum deviation, respectively. A computational form of frequency-weighted log spectral distortion (FWLSD) measure is often given as

$$\text{FWLSD} = \left\{ \frac{\sum_{\nu=1}^{M} |X(\nu)|^\gamma |20 \log X(\nu)/Y(\nu)|^p}{\sum_{\nu=1}^{M} |X(\nu)|^\gamma} \right\}^{1/p}, \tag{2.12}$$

where M is an integer corresponding to M-point discrete Fourier transform (DFT), ν is the discrete frequency variable; and $X(\nu)$ and $Y(\nu)$ are the LPC spectra of $\{x[n]\}$ and $\{y[n]\}$, respectively. With $p = 2$ and $\gamma = 0.5$, a magnitude correlation coefficient of 0.60 is obtained [11]. Another version of this measure has recently been proposed in [26] where the kernel of the measure is not the original and the coded signal spectra, but the coded signal spectrum and the spectral representation of the nonlinear distortions incurred in the coding process.

2.4.5 Cepstral Distance

The basic problem with the LSD measures is the Fourier transform and logarithm computations involved in obtaining sufficient values of $V(\omega)$ in order to approximate the integral of (2.10) by summation. Computational efficiency and a high correlation with the root mean square LSD have thus made another measure, namely the cepstral distance (CD), popular [27]. The CD is a measure of the overall difference between an original and a corresponding coded speech cepstra. A cepstrum computed from the LPC coefficients, unlike that computed directly from the speech waveform, results in an estimate of the smoothed speech spectrum [28]. This can be written as

$$\log\left\{\frac{1}{A(z)}\right\} = \sum_{k=1}^{\infty} c[k]z^{-k}, \tag{2.13}$$

where $A(z)$ is the LPC analysis filter polynomial and $c[k]$ denotes the k-th cepstral coefficient. Accordingly, a CD measure is defined as

$$\mathrm{CD} = \frac{10}{\log 10}\sqrt{2\sum_{k=1}^{L}(c_x[k] - c_y[k])^2}, \tag{2.14}$$

where $c_x[k]$ and $c_y[k]$ are the k-th cepstral coefficients of the original and the distorted speech, and L is the number of the cepstral coefficients used. Although the sequence of the cepstral coefficients is infinite in (2.13), limiting it to three times the number of LPC parameters shows almost no deterioration in the result. A correlation coefficient of 0.80 has been obtained with this measure [11]. The quefrency-weighted CD [29], the liftering window-based CD [30] are some examples of weighted CD measures. A unifying framework for viewing different distortion measures in the cepstral domain has been laid out in [31].

2.4.6 Itakura-Saito Distortion Measure

With a maximum likelihood formulation of linear prediction, Itakura and Saito have defined a d_{IS} measure as [32]

$$d_{\text{IS}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[e^{V(\omega)} - V(\omega) - 1 \right] d\omega, \tag{2.15}$$

where $V(\omega)$ is as given in (2.11). The assumptions used in deriving the integral form of (2.15) are that the speech is generated by a Gaussian process, the result of uncorrelated noise passed through an all-pole LPC filter and that the analysis interval is much longer than the all-pole filter order. It has been shown in [25] that the d_{IS} measure is twice the asymptotic discrimination information under the above assumption. A frequency-weighted version of the Itakura-Saito measure has been found in [33] to give a better performance. The d_{cosh} measure, a symmetrical version of d_{IS}, is often defined as

$$d_{\text{cosh}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \{\cosh[V(\omega)] - 1\} \, d\omega. \tag{2.16}$$

It has been found in [27] that the d_{cosh} measure bounds the LSD measure from above, and in [25] that it becomes one half of the generalized Ornstein distance between two Gaussian processes. Computational costs for evaluating the d_{IS} and d_{cosh} measures are given in [34].

2.4.7 Coherence Function

In this method, the speech frames are first divided into four groups based on the four amplitude quartiles. The original and the coded signal power spectra as well as the cross-power spectrum are computed and averaged for all the frames in each quartile. The respective average spectra, denoted by $S_{xx}(f)$, $S_{yy}(f)$ and $S_{xy}(f)$, are used to compute the squared coherence function $\gamma^2(f)$ as [35]

$$\gamma^2(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f) S_{yy}(f)} \tag{2.17}$$

which can be interpreted as the correlation between the original and the coded signals at a frequency f. Next, the signal power $C(f)$ and the distortion power $D(f)$ are

estimated from $\gamma^2(f)$ and used to develop a modified signal-to-distortion ratio (SDR) for each quartile as

$$C(f) = \eta.\gamma^2(f)|S_{yy}(f)|^2, \tag{2.18}$$

$$D(f) = \eta.[1-\gamma^2(f)]|S_{yy}(f)|^2, \tag{2.19}$$

$$\text{SDR} = \frac{C(f).W_2(f)}{D(f).W_2(f)+W_1(f)}. \tag{2.20}$$

In (2.18) and (2.19), η is a scale factor. $W_1(\bullet)$ and $W_2(\bullet)$ in (2.20) are the weighting functions related to the hearing threshold and the handset receiver sensitivity. The regression-analyzed MOS value is calculated using a frequency-weighted quartile-weighted nonlinear function; the details are given in [35].

2.5 Perceptually-Motivated Objective Measures

Coder distortions can be perceived if the magnitude of the distortion is greater than the resolution of the human auditory system. The nature of the distortion is also important from the perception point of view. In the following, we discuss two perceptually-motivated distortion measures.

2.5.1 Information Index

An information index (II) measure which accounts for loss, noise and distortion in speech transmission over a telephone network has been proposed in [36]. The auditory system effect is roughly modeled by dividing the spectrum into sixteen critical bands and applying empirical frequency weights and hearing thresholds for each band. At first, a signal-to-distortion ratio for each critical band, denoted as $R(i)$, is computed by

$$R(i) = 10\log_{10}\frac{\sum\limits_{j\in b_i}|X(\omega_j)|^2}{|\sum\limits_{j\in b_i}|X(\omega_j)|^2-\sum\limits_{j\in b_i}|Y(\omega_j)|^2|}\ \text{dB}, \tag{2.21}$$

where j ranges over all the frequencies specified for the i-th band b_i. Here, $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of an original and a corresponding coded speech

frame. Assuming the bands to be independent, the II measure is computed as

$$\mathrm{II} = \sum_{i=1}^{16} W_2(i).\frac{3}{0.1 + 10^{-[\overline{R}(i)+W_1(i)/10]}}, \tag{2.22}$$

where $\overline{R}(i)$ is the average of $R(i)$ over all the frames and $W_1(i)$ and $W_2(i)$ are the appropriate weighting functions accounting for the hearing threshold and the perceptual importance of the i-th frequency band, respectively.

2.5.2 Bark Spectral Distortion

In [21], Schroeder et al. have described a method of calculating an objective measure for signal degradation based on the measurable properties of the auditory perception. Motivated by this work, a series of psychophysical experimental curves has been invoked in [37] to define a Bark spectral distortion (BSD) measure. At first, a nonlinear frequency transformation from Hertz f to Bark b is made via the relation [21]

$$f = 600 \sinh(b/6) \tag{2.23}$$

which transforms the original power spectral density function $X(f)$ to a critical band density function $Y(b)$. The function $Y(b)$ is 'smeared' by a prototype critical band filter $F(b)$ given as [38]

$$10 \log_{10} F(b) = 7 - 7.5(b - \alpha) - 17.5[0.196 + (b - \alpha)^2]^{1/2} \tag{2.24}$$

with $\alpha = 0.215$. The smearing is conceived of as a convolution operation between $F(b)$ and $Y(b)$ which yields a continuous spectrum $D(b)$. The fact that the ear is not equally sensitive to the amount of energy at different frequencies is exploited next. The well-known equal loudness level curves [39] have been used to translate the sound pressure levels (SPL) in dB to the loudness levels in *phons*. The increase in phons required to make the subjective loudness double depends on the loudness level and thus finally a *phon*-to-*sone* conversion is performed using [38]

$$S = \begin{cases} 2^{(P-40)/10} & \text{if } P \geq 40; \\ (P/40)^{2.642} & \text{if } P < 40 \end{cases} \tag{2.25}$$

to generate a Bark spectrum $S(i)$. The BSD measure is defined in [37] as the average of $\mathrm{BSD}^{(k)}$ with

$$\mathrm{BSD}^{(k)} = \sum_{i=1}^{N} [S_x^{(k)}(i) - S_y^{(k)}(i)]^2, \tag{2.26}$$

where N is the number of critical bands; and $S_x^{(k)}(i)$ and $S_y^{(k)}(i)$ are the Bark spectra in the i-th critical band for the k-th speech segment corresponding to the original and the coded speech, respectively.

The success of the BSD measure has demonstrated the advantage of considering important perceptual events while formulating a distortion measure. Recently, a software package, named PERCEVAL, is introduced in [40] which computes the probability of detection of the noise as a function of time for noise-corrupted audio and music signals.

2.6 Summary

In this chapter, we have reviewed some of the existing subjective and objective measures used in the speech coding area. The mean opinion score and the diagnostic acceptability measure are two of the widely used subjective measures. The most popular class of the time-domain measures is the SNR with its variants such as the segmental SNR, the granular segmental SNR etc. Among the spectral distortion measures, the log likelihood ratio measure, the log area ratio measure, the log spectral distortion measure, the cepstral distance and the Itakura-Saito distortion measure are quite well-known. Some of the existing objective measures have placed emphasis on the aspects which are perceptually important. Two such psychoacoustically-motivated measures are the information index and the Bark spectral distortion measure. The merit of considering important perceptual events has been demonstrated by the success of these measures.

Chapter 3

Auditory Representation of Speech

3.1 Introduction

The formulation of any distortion measure requires resolution of two important issues: (i) defining a suitable domain where the signal parameters should be compared and (ii) comparing them in a meaningful sense. This chapter is concerned with the first issue as relevant to speech signals. It has been observed that even the repeated utterances of a sentence by a speaker often differ considerably in the time-domain. In this regard, a spectral representation of speech has appeared to be a relatively steady one. However, we argue that neither the time-domain nor the frequency-domain, in isolation, is a good representation for speech signal. Since a human auditory system is the final information processor in speech communication, it would be meaningful to represent the speech signal in a *perceptual-domain* (PD). In this work, we use an auditory model for mapping the time-domain speech signal onto its corresponding PD representation.

The present chapter is organized as follows. Section 3.2 briefly studies the mechanism of the auditory system. Section 3.3 presents various well-established psychoacoustic observations pertinent to speech perception. Section 3.4 discusses four broad classes of analogous electrical model featuring primary auditory processing. In particular, we describe Lyon's auditory (cochlear) model which is used to define the

PD representation for the present work.

3.2 Mechanism of Auditory System

An ear consists of three sections—the outer ear, the middle ear and the inner ear. Speech pressure variations, directed towards the eardrum by the outer ear, are transformed into mechanical motion by the middle ear. Finally, the inner ear converts these mechanical vibrations into electrical firings (impulses) which are sensed by the hair cells and propagated to the brain following an ascending auditory pathway over nerve fibers [2, 41, 42]. In the following subsections, we concisely describe anatomy and functions of the prime components of the auditory system.

3.2.1 Outer and Middle Ear

The *pinna* which is the visible part of the outer ear channelizes sound waves into the ear canal (*meatus*) and finally hits the eardrum (*tympanic membrane*). This 2.7 cm long canal with about 0.7 cm diameter behaves as a quarter-wavelength resonator and amplifies energy between 3 kHz and 5 kHz by up to 12–15 dB. The middle ear which contains three tiny, dense bones (*malleus*, *incus* and *stapes*) transmit the sound wave vibrations to the *oval window* membrane of the inner ear. This way, it acts as an acoustic transformer matching the airborne-sound impedance of the outer ear to the fluid-borne sound impedance of the inner ear. The transformer action is due to the ratio of the area of the active parts of the eardrum to the area of the footplate of the stapes. The acoustic impedance of the inner ear fluid is about 4,000 times that of air and this impedance mismatch is such that, without the transformer effect of the ossicles, all but 0.1% of the pressure waves hitting the eardrum would be reflected back allowing very little energy to enter the inner ear. Additionally, the middle ear also helps in protecting the inner ear against very intense sounds.

3.2.2 Inner Ear (Cochlea)

The cochlea, a liquid-filled tube coiled in a snail-shaped spiral, converts mechanical vibrations at its oval window input into electrical excitation on its neural fiber outputs.

It has a cross-sectional area of about 4 mm^2 at its base near the stapes and tapers gradually to about 1 mm^2 at its apex. The interior of the cochlea is divided into three chambers—the *scala vestibuli*, the *scala media* and the *scala tympani*. Between the latter two chambers is the *basilar membrane* (BM) which increases from a width of 0.04 mm at its base to 0.50 mm at the apex. The stiffness of the BM varies smoothly over its length. It is stiff and thin at the basal end, but compliant and massive at the apical end (the ratio of stiffness between ends exceeds 100). Therefore, the cochlea near its base is most sensitive to high frequency sounds and as the wave travels down the cochlea, lower and lower frequencies are sensed. The prime feature of the cochlea is that energy in the acoustic wave is separated by frequency and each *place* in the cochlea responds best to one frequency, termed as its *characteristic frequency*. This way, it maps the spectral components of the signal onto the place domain and maintains a *tonotopic* organization.

3.2.3 Inner and Outer Hair Cells

On the top of the BM (within the *organ of Corti*), there are about 30,000 sensory *hair cells* arranged in several rows along the length of the cochlea. The endings of the auditory nerve terminate on these hair cells and each of them has about 40–140 hairs. The tips of the outer hair cells, placed in three or four rows, are embedded in the *tectorial membrane*. These cells usually do not send any information about the sound to the brain. Rather, they function as part of an active amplifier and signal-level controller. On the other hand, the single row of 3,500 inner hair cells that runs along the length of the BM is the primary source of the nerve pulses that travel to the *cochlear nucleus* and on up to the brain.

3.2.4 Neural Pathways

The chemical stimulation of the nerve endings attached to the hair cells produces an all-or-none electrical firings. The auditory firings pass via the cochlear nerve to the *ventral* and *dorsal* cochlear nuclei in the *medulla*. Subsequently, they traverse through the *superior olivary complex*, the *lateral leminscus*, the *inferior colliculus* and finally the *medial geniculate body* before entering the *brain cortex*. The stimuli received at the two ears may interact both at the medulla and mid-brain levels. The

exact neuro-electrical representation of sound stimuli at these various levels is not sufficiently understood.

3.3 Psychoacoustic Observations

Auditory system has been studied from different viewpoints by researchers in the field of psychoacoustics, physiology of hearing and speech processing [2, 42, 43]. We note here some of the psychoacoustic phenomena believed to be important in the perceptual event. This description, although supplementary to [44], is quite self-contained.

Observation 1 (*Ear Canal as an Organ Pipe*): The ear canal, about $l = 2.7$ cm long, is an air-filled cavity open at one end (at the pinna) and closed at the other (at the eardrum) [41]. To a rough approximation, the ear canal can be considered as a uniform pipe and it has normal modes of vibration which occur at frequencies where the pipe length is an odd multiple of a quarter wavelength. The first resonance therefore occurs at the frequency f_{res} given by

$$f_{\text{res}} = \frac{v_{\text{sound}}}{4l} \approx \frac{330 \text{ m/s}}{4 \times 0.027 \text{ m}} \approx 3,000 \text{ Hz} \tag{3.1}$$

which aids the ear's sensitivity in this frequency range.

Observation 2 (*Impedance Transformation in Middle Ear*): The lever action of the ossicles provides a force amplification (G) of about 1.3 [45]. Moreover, the vibrating area of the eardrum (A_{eardrum}) is approximately 55 mm^2, compared to the stapes area (A_{stapes}) of 3.2 mm^2. Therefore, the ratio (F) of pressure applied at the oval window to that applied at the eardrum is given by

$$F = \frac{GA_{\text{eardrum}}}{A_{\text{stapes}}} = 1.3 \times \frac{55}{3.2} \approx 22. \tag{3.2}$$

This impedance transformation (through pressure transformation) leads to an increase of about $20 \log_{10} 22$ dB $\approx$ 27 dB in sound pressure level (SPL) [*Note:* 0 dB SPL = 10^{-16} W/cm^2] within the middle ear [2]. When low-frequency sounds of more than 85–90 dB SPL reach the eardrum, the middle ear provides some automatic gain control effect via stapedial reflex [46].

Observation 3 (*Motion of BM in Cochlea*): The motion of the BM in cochlea is quite complicated; however, its total volume displacement at any instant of time

is equal to the volume displacement of the stapes or of the round window membrane [46]. The velocity of sound (v_{coch}) in cochlear fluid is 1,600 m/s and the length of the cochlea (L_{coch}) is around 35 mm [2]. The corresponding base-to-apex time-delay (τ_{coch}) of the sound is given by

$$\tau_{\text{coch}} = \frac{L_{\text{coch}}}{v_{\text{coch}}} = \frac{0.035 \text{ m}}{1,600 \text{ m/s}} \approx 20 \ \mu\text{s} \tag{3.3}$$

which indicates that there is essentially no phase delay in pressure along the BM. The mechanical properties (mass, stiffness, loss) of the cochlea change very slowly with place. As a consequence, no significant amount of wave energy is reflected back [41].

Observation 4 (*Resonances in BM*): The fluid current due to the motion of the BM tends to go through the point of least resistance where the BM compliance reactance annuls its mass reactance [47]. The BM appears to have a 'hole' in that point—to its left, the BM is very stiff (large capacitive reactance) and to its right, the BM is massive (large inductive reactance). Thus, each place along the BM resonates most strongly with a pressure wave of a characteristic frequency (CF) associated with it. The frequency response curves corresponding to different places, found by Nobel laureate von Békésy [45], were rather broad and later Mössbaur's gamma-ray-based experiment suggests much sharper frequency response curves [43]. It has also been observed that all the response curves have almost constant Q-factor, thereby implying a fixed ratio of center frequency to bandwidth for all the band-pass filters. Frequency resolution along the BM is best at low frequencies (apical end) whereas the time resolution is best at higher frequencies (basal end). This is primarily due to the fact that a hair cell attached to a high-CF location on the BM fires in response to a broader set of frequencies than does a low-CF hair cell [44].

Observation 5 (*Inner Hair Cells as Rectifiers*): Fine hairs, called *stereocilia*, protrude from the ends of the inner hair cells. They detect the shearing motion of the membranes and act as transducers converting this deflection to an ion current. When the cilia are bent one way, the hair cells stimulate the primary auditory neurons to fire. When the cilia are bent the other way, no pulses are generated. Thus, the inner hair cells act as half-wave rectifiers for the velocity of the motion of the fluid [41].

Observation 6 (*Outer Hair Cells as Coupled Gain-Controllers*): Studies on the cochlear echo and the oto-acoustic emission suggest that the BM behaves as an active system and the transfer characteristics of the BM system vary depending on the input signal level [48]. This is attributed to the fact that the outer hair cells interact with

the BM motion. Sounds with high SPLs are effectively diminished whereas sounds with low SPLs are enhanced by the 'superregenerative active' mechanisms of the outer hair cells [46].

An important aspect of hearing is the phenomenon of auditory masking in which the perception of low-energy sound is obscured by the presence of a high-energy sound [49, 50]. The outputs of the band-pass filters may be viewed as zero-mean 'carrier' signals which are 'amplitude-demodulated' by the half-wave detection nonlinearity. The phenomenon of auditory masking can thus be justified by the 'threshold effect' phenomenon [51] as observed in the envelope detection process of AM signals.

Effects of the outer hair cells can be emulated by automatic gain control (AGC) stages and some kind of inter-stage coupling of these AGCs can simulate the auditory masking feature. Any gain control effect (i.e., amplification or compression) is not instantaneous and the time required to adapt to any input signal is dependent on the signal level [44].

3.4 Perceptual-Domain Representation

We desire to deal with an accurate description of human perception as far as possible. But at the same time, since the computational speed of the model is also of importance, we prefer using a *functional* model of the auditory system for the PD representation of speech signal.

3.4.1 Auditory Models for Speech Representation

The interpretation of the cochlea as a spectrum analyzer goes back to Helmholtz [52] in the last century. The *timing* or *volley* theory states that low sound frequencies such as those corresponding to the fundamental frequency (F0) of speech, are perceived in terms of time-synchronous neural firings from the BM apex. On the other hand, the *place* theory suggests that, especially for higher frequencies such as those in the formants of speech, the spectral information is decoded via the BM locations of the neurons that fire most [53].

Current models for representing speech in the auditory periphery falls into

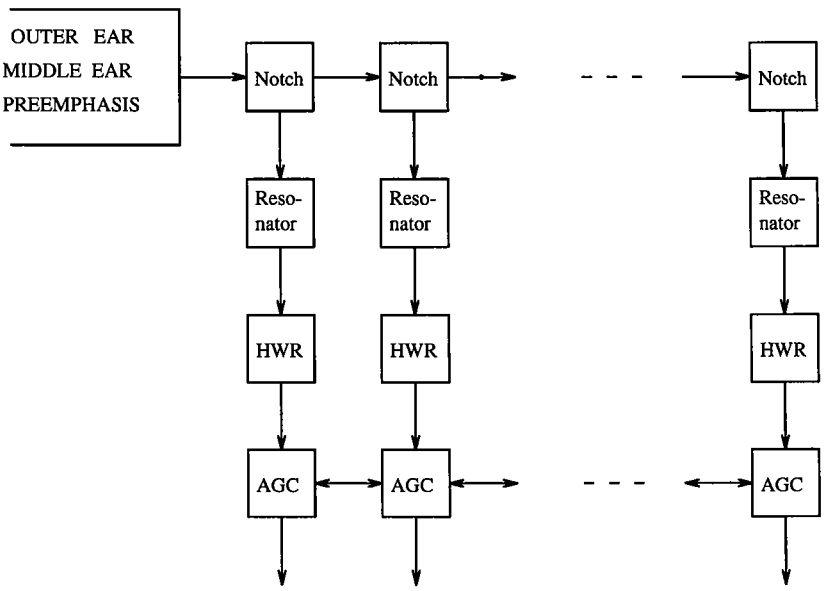


Figure 3.1: Block diagram of Lyon's cochlear model ('HWR' stands for the half-wave rectifier and 'AGC' stands for the automatic gain controller)

one of four broad classes [54]: rate/place, synchrony/place, synchrony/quasi-place and synchrony/place-independent. The rate/place representation [55], a well-defined average-rate-based spatial profile, functions well at low SPL. The synchrony/place representational form [56] is based on neural synchrony and requires the system to possess some knowledge of the tonotopic affiliation of each fiber with which to evaluate its temporal firing pattern. The synchrony/quasi-place model [57], in the form of *lateral inhibitory network*, considers simultaneous activity across adjacent channels. A proposition that a spectral representation based on synchrony need not be concerned with the tonotopic identity of the auditory nerve fibers gives rise to the synchrony/place-independent model [58] which works well only for high SPL.

Based on the psychoacoustic observations discussed in the previous section, we believe that a synchrony/quasi-place model [57, 59] is most appropriate for our work as it could operate satisfactorily for high, medium or even low signal levels. Consequently, we adopt one such synchrony/quasi-place model as suggested by Lyon [44] based on work described elsewhere such as [60].

3.4.2 Mapping Using Lyon's Cochlear Model

Time-domain speech representation is mapped onto a perceptual-domain where the time-place components become the fundamental bases of analysis. The conversion is achieved here using Lyon's cochlear model as described in [44, 61]. This model separates complex mixtures of sounds mainly by segregating different frequencies into different places, but also by preserving enough time resolution to separate the responses to different pitch pulses. Therefore, the voiced speech sounds that differ simultaneously in some formants as well as in pitch are separated into recognizably distinct patterns of activity at the output. By a detailed separation of sounds along the time and frequency dimensions, this model paves way for a robust speech analysis technique. Lyon's cochlear model, as shown in Fig. 3.1, integrates the prime features of the 'place' as well as the 'volley' theory. In the following, we describe the model in six steps.

Step 1 (*Outer-and-Middle Ear Filter*): The outer-and-middle ear effectively adds a slight high-pass response to the system. Assuming that the input speech signals are sampled at a frequency f_s of 8,000 Hz, a simple first-order high-pass discrete-time filter with a corner frequency of 300 Hz is designed to model roughly the effects of the outer and the middle ear. The frequency response of this filter $H_{\rm OM}(z)$, plotted in Fig. 3.2, is given by

$$H_{\rm OM}(z) = \frac{(1 - \exp\,[-2\pi\frac{300}{8000}]z)}{(1 - \exp\,[-2\pi\frac{300}{8000}]z)_{z=1}} = 4.76375(1 - 0.79008z). \tag{3.4}$$

This filter has unity gain at DC (i.e., at $z = 1$). For simplification, the AGC mechanism of the middle ear via stapedial reflex is not modeled here [62].

Step 2 (*Notch Filters and Resonators*): The cochlea is best described by a continuous differential equation [63]; however, it can be modeled by an ensemble of discrete stages in cascade. Lyon, in his proposed cochlear model, uses such discrete-place approximation. An implementation of the discrete-place stages involves combining a series of notch filters that model the traveling pressure waves with a series of resonators that model the conversion of pressure waves into BM motion [44, 61]. The notch filters operate at successively lower frequencies so that the net effect is to low-pass filter gradually the acoustic energy which are collected by the resonators corresponding to different places. We consider here sixty-four stages (covering up to 4,000 Hz) in cascade, each having a different frequency sensitivity representing the

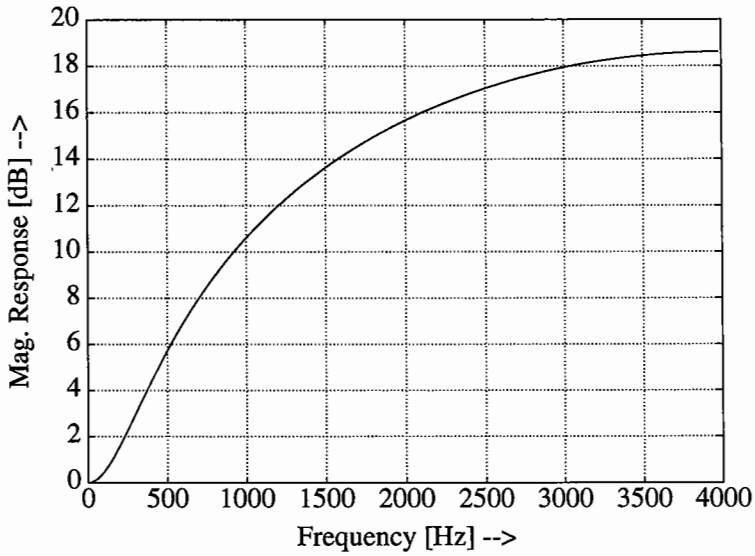


Figure 3.2: First-order outer-and-middle ear-filter

associated resonance and is characterized by the respective filter transfer function.

The notch filters and the resonators are approximated by biquadratic filter transfer functions. Though these stages are designed in discrete-time, Fig. 3.3 plots poles and zeros for some of the notch and the resonator filters, for the sake of clarity, in the s-plane. Each of the notch filters has a high-Q zero-pair near a low-Q pole pair whereas each of the resonators has a zero at DC with a high-Q pole pair located between the previous and the next notch filter zero-pairs. Several models of the cochlear mechanics include a micromechanical 'second filter' for a resonance in the organ of Corti that contributes a zero pair slightly below the BM resonance [64]. Presently, this not-so-well-accepted feature is left out. This can easily be incorporated in this model by putting another zero pair in the resonator section.

Step 3 (*Cascade Design of Stage Filters*): The combination of the notch filters and the resonators can be implemented in cascade/parallel form as shown in Fig. 3.1. However, to reduce the computations, the notch and the resonator filters of each stage can be integrated into a single ear-filter stage. The locations of the poles in the resonator filters are chosen to be at the same locations as the poles in the succeeding notch filter. This way, the zeros from each notch filter and the poles from a resonator and the next notch filter are integrated to yield a single ear-filter stage [61].

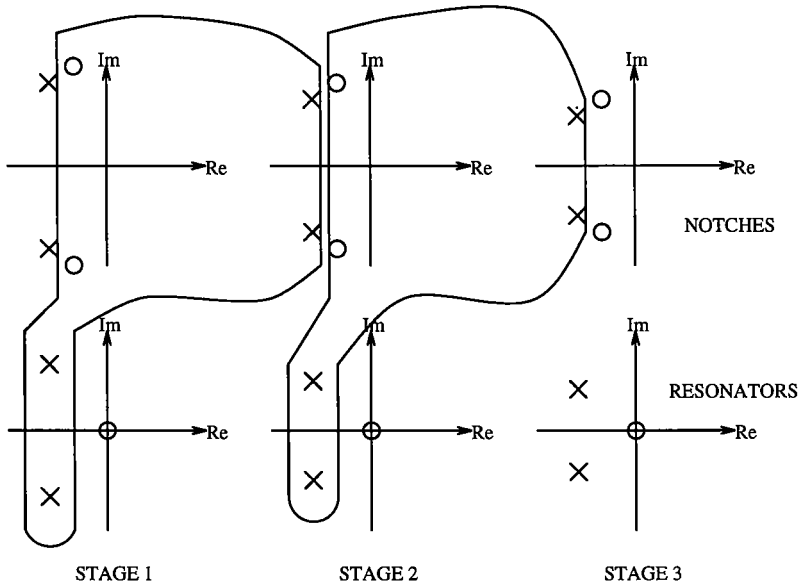


Figure 3.3: s-Domain pole-zero plots for typical stages (integrated notch and resonator filters)

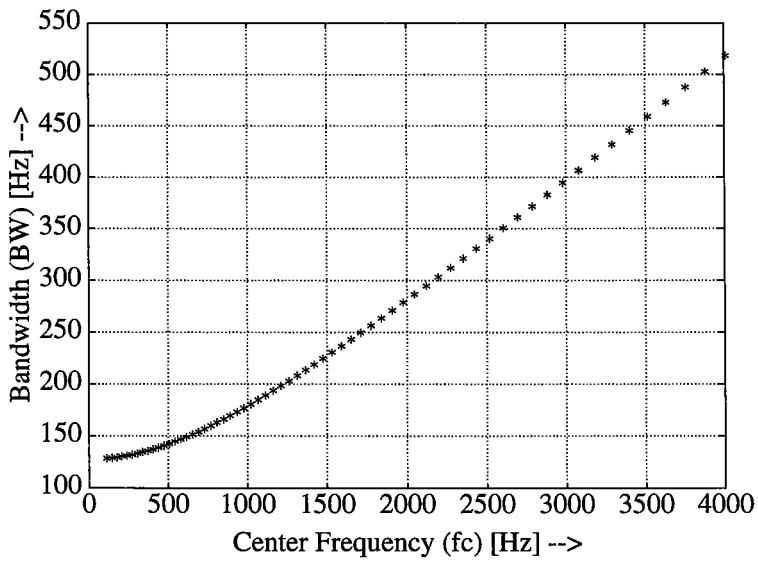


Figure 3.4: Bandwidths vs. center frequencies of sixty-four stages

The composite transfer function of each ear-filter stage is an asymmetric band-pass function. $W_{\text{ear}}(f_c)$, the 3-dB bandwidth of a band-pass filter with center frequency f_c, is defined as

$$W_{\text{ear}}(f_c) = \frac{\sqrt{f_c^2 + f_{\text{eb}}^2}}{Q_{\text{ear}}}, \tag{3.5}$$

where the ear-break frequency f_{eb} is 1,000 Hz and the constant Q-factor for all the band-pass filters Q_{ear} is 8. In conformance to psychoacoustical data, four successive ear-filter stages are overlapped within the 3-dB bandwidth of any one ear-filter and thus we have, S_{ear}, reciprocal of the number of overlapping ear-filter stages, as 0.25. Finally, the following parameters are obtained for any ear-filter stage corresponding to a particular characteristic frequency:

$$f_{cp} = f_c; \qquad Q_{cp} = \frac{f_{cp}}{W_{\text{ear}}(f_c)} \tag{3.6}$$

$$f_{cz} = f_c + W_{\text{ear}}(f_c).S_{\text{ear}}.Z_{\text{off}}; \quad Q_{cz} = h_{\text{ear}} \cdot \frac{f_{cz}}{W_{\text{ear}}(f_c)}, \tag{3.7}$$

where f_{cp} and f_{cz} are the center frequencies of the associated poles and zeros of a particular ear-filter stage having center frequency f_c. The center frequency of the associated zero is an extra stage higher than that of the pole. Thus, the Z_{off}, a factor that determines how far the zero is offset from the center frequency of the ear-filter stage, is chosen to be 1.5. Q_{cp} and Q_{cz} are the Q-factors for the corresponding poles and zeros and the parameter h_{ear}, which determines how much sharper the notch (zero) is than the resonator (pole), is selected to be 5.0.

The ear-filter stages are indexed from 1 (corresponding to the highest frequency) to 64 (corresponding to the lowest frequency) and the center frequency of each stage decreases by S_{ear} (here, 0.25) times the bandwidth of the previous stage. $W_{ear}(f_c)$ vs. f_c of all the sixty-four ear-filter stages are plotted in Fig. 3.4 where we observe that $\lim_{f_c \to 0} W_{\text{ear}}(f_c) \to \frac{f_{\text{eb}}}{Q_{\text{ear}}} = 125$.

Step 4 (*Other Adjustments in Stage Filters*): To implement the zeros at DC for every resonator, a differentiator is required for each stage. Since all the filtering used is linear, the differentiator (a term of the form $1 - z$) can be placed just once before the ear cascade. In addition, the differentiator is combined with a zero at the Nyquist rate $(1 + z)$ to compensate for the close spacing of the poles near $z = -1$ for high frequency. The frequency response for this combined filter is given as

$$H_{\text{comb}}(z) = 0.5(1 - z^2) \tag{3.8}$$

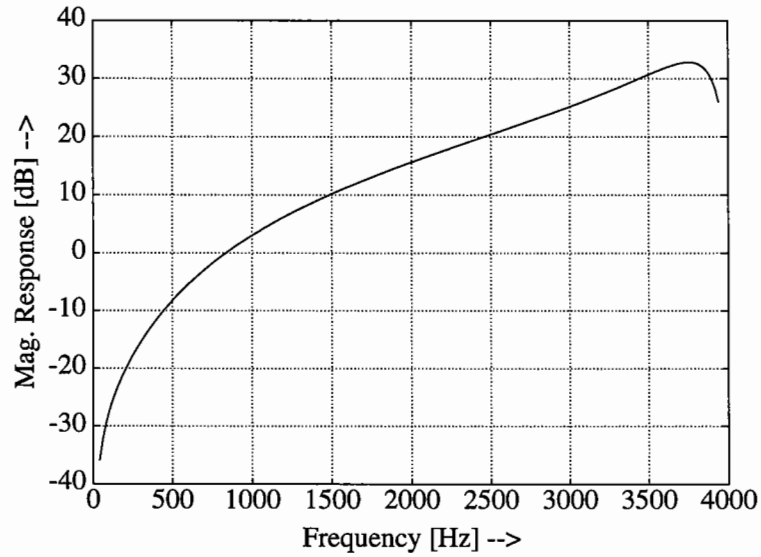


Figure 3.5: Initial stage filter

with unity gain at one-quarter of the sampling frequency.

In the cascade form, each of the ear-filter stages is implemented by a combination of two poles and two zeros. After the pole-zero integration, a pair of poles of the first stage is left aside. Thus, the ear-filter is redefined with an initial stage $H(z)$ which combines the effects of the outer-and-middle ear $H_{\mathrm{OM}}(z)$ and the differentiator-compensator $H_{\mathrm{comb}}(z)$ with the two poles of the first stage filter. The transfer function of this initial stage filter becomes

$$H(z) = \frac{(-0.77356 + 3.91442j)(1 - 0.79008z)(1 - z^2)}{0.67523 + 1.64342z + z^2} \tag{3.9}$$

and the corresponding magnitude frequency response plot is shown in Fig. 3.5.

The gain of an ideal differentiator is proportional to frequency. Preceding all stages of the ear-filter with a single differentiator causes the lower frequency stages to have a much lower output than the preceding stages. While within a single stage, it is desired to add a term that is proportional to frequency, the effect of differentiator at each stage is adjusted so that it has unity gain at the center frequency of the corresponding stage. Typical frequency responses for three ear-filter stages with center frequencies as 499 Hz, 1,013 Hz and 2,509 Hz are shown in Fig. 3.6.

Step 5 (*Half-wave Rectification*): The exact shape of the half-wave nonlinearity

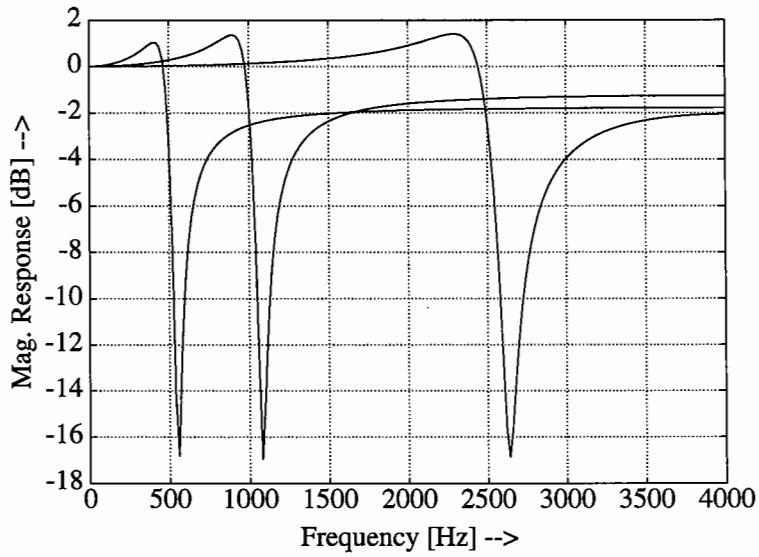


Figure 3.6: Magnitude responses for three typical ear-filter stages with f_c=499; 1,013 and 2,509 Hz

is not obvious; there are proposals for ideal as well as soft half-wave [65] rectification. In this work, an ideal half-wave rectifier is considered.

Step 6 (*Coupled Automatic Gain Controllers*): The effects of the BM and the hair cell nonlinearity are taken care of adequately by lumping them into a gain control mechanism. Other nonlinear effects, such as the cubic difference tones etc., are assumed to be relatively unimportant to normal hearing [41].

The most important adaptation mechanism in sensory systems is lateral inhibition by which the sensory neurons reduce their own gain as well as the gain of the others nearby. A logarithmic or simple non-coupled AGC mechanism does not adequately handle wide variations of energy across the frequency dimensions. Therefore, Lyon proposed a coupled AGC that adapts in the frequency domain [44]. One such coupled AGC, as described in [61], is shown in Fig. 3.7. Each stage is coupled directly only to its neighboring stages. However, in principle, any stage can affect all the other stages having an effect, perhaps, decaying exponentially with distance from it [66]. The gain offered to an input in an AGC stage varies between 0 and 1, and this gain factor is determined based on the previous states of the current, the left and the right stages as well as the previous output value.

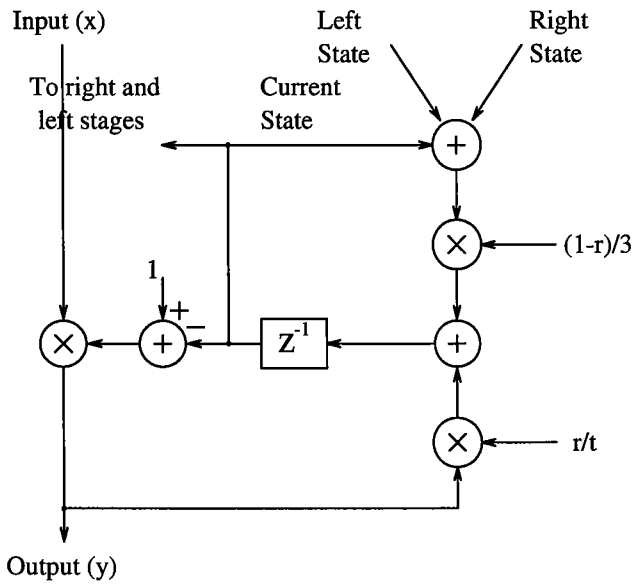


Figure 3.7: A typical automatic gain control (AGC) stage

The time constant of the coupled AGC is made dependent on the signal level. A cascade of four AGC blocks with different *time constants*, simulating the different adaptation times in the ear, are used [61]. A longer time constant implies that the AGC takes longer to respond to the input. Each AGC attenuates the incoming signal so that, under steady-state condition, it remains below the *target* value corresponding to that AGC. The target parameters (t) and the time constants (τ) of the four AGC blocks, respectively, are chosen as 0.0032, 0.0016, 0.0008 and 0.0004 units (on the same scale, the amplitude of a signal with +120 dB SPL is assumed to be unity) and 640 ms, 160 ms, 40 ms and 10 ms. The r parameters as indicated in Fig. 3.7 are related to the τ parameters as

$$r = \exp\left[-\frac{1}{\tau f_s}\right]. \tag{3.10}$$

For any one of the sixty-four stages, a typical steady-state response of the four cascaded AGC blocks is depicted in Fig. 3.8.

3.4.3 Auditory Representation

The outputs of the cochlear model vary over only about two orders of magnitude as the input signal varies over the entire range covering the threshold of hearing

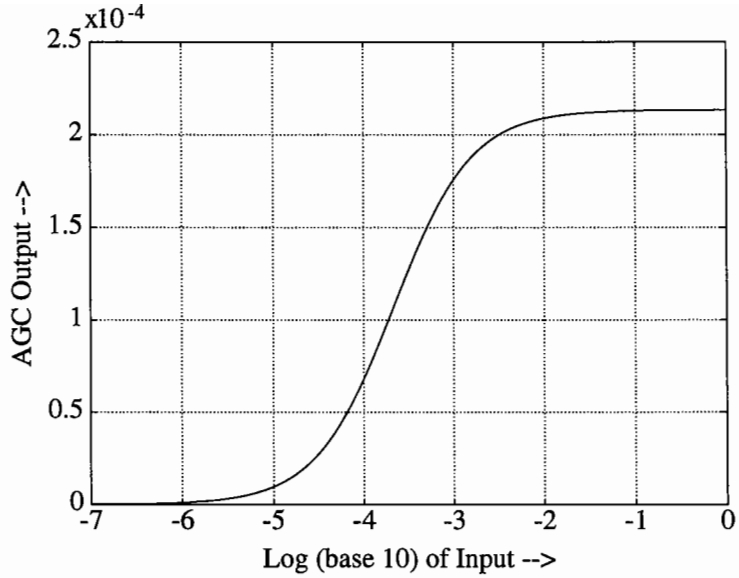


Figure 3.8: A typical steady-state response of four cascaded AGC blocks

to the threshold of pain. The neurons are attached to the hair cells at different places along the cochlear partition and they 'fire' (i.e., generate all-or-none electrical spikes) based on the gain-controlled signals as sensed by the corresponding hair cells. Essentially, these neural firing events are communicated from the auditory system to the brain through a large number of neural fibers. These neural pathways are termed hereafter as the 'neural channels' so as to keep conformity with the other communication channels. Although these neural fibers are spread densely along the BM, since we consider sixty-four discrete-place stages, we would visualize that all the neurons could be classified into sixty-four *characteristic* neural channels.

The normalized cochlear model output provides the probability-of-firing information in these sixty-four neural channels at each clock time. Here, the normalization is done with respect to the maximum possible output value (i.e., 0.000213 unit as shown in Fig. 3.8) of the four cascaded AGC blocks and the clock time is chosen to be same as the sampling time, i.e., 125 μs. Since we do not know the exact firing process, the neural activity patterns can be presented in a *cochleagram* matrix form which gives the probability-of-firings in all the neural channels for all the clock times. In our work, this auditory representation is referred to hereafter as the perceptual-domain (PD). We assert that, to devise a distortion measure for speech signals, the original and the coded/distorted signal should be compared in this perceptual (time-place)

domain, rather than just in the time or in the frequency domain. In the next two chapters, we propose two distinct approaches for comparing these PD parameters. These comparisons, in turn, provide measure values to assess the degree of overall coder distortions in a coded/distorted speech with reference to its original version.

3.5 Summary

This chapter of the dissertation has dealt with the issue of auditory representation for speech signal which is the first step in our formulation of distortion measures for coded speech. For comparing an original speech with its coded version, neither the time-domain nor the frequency-domain representation is sufficient. It is important to consider all the major perceptual events and represent the speech signal in a joint time-place domain. Towards this end, we have used Lyon's auditory model. This model has simulated the high-pass behavior of the outer-and-middle ear, the band-pass characteristics of the inner ear (cochlea), the half-wave nature of the inner hair cells and the automatic gain controlling feature of the outer hair cells. Temporal and spectral masking effects have also been emulated by inter-stage coupling. The final perceptual domain representation of the speech signal is in the form of firing probabilities in the neural channels at the clock times.

Chapter 4

Cochlear Discrimination Information (CDI) Measure

4.1 Introduction

In Chapter 3, we have addressed the issue of representing speech signal in a perceptual-domain (PD). This PD representation is a sequence of N-dimensional (in our work, $N = 64$) vectors at the clock times within a speech signal. Each of the N neural channels may be conceived as communication channels with an input alphabet of size two, i.e., firing and non-firing. Due to the lack of our knowledge about the exact neural conversion process, we compare the probability distributions for firing and non-firing, derived from an original and a coded signal, to quantify the degree of distortion. The discrimination information which has emerged as a powerful tool [67] for measuring the 'closeness' of two probability density or distribution functions is applied here for defining a *cochlear discrimination information* (CDI) measure [68, 69]. In the first part of this chapter, we formulate the CDI measure and study speech coder performances with it.

For any source-coder, a *source-destination* pair can be characterized by a probabilistic model of the source and a fidelity criterion measuring the degradation of the coded signal with reference to the original source. Based on the *rate-distortion theory*, a rate-distortion function $R(D)$ may be associated with any such source-destination pair. This function calculates the effective rate at which the source produces infor-

mation subject to the constraint that an average distortion of D is endured at the destination. A knowledge of the $R(D)$ is of considerable importance as it may prevent one from frivoling time as well as resources to achieve an impossible task. However, often, it becomes difficult to give an explicit closed-form or parametric solution to the $R(D)$, even for apparently simple sources and distortion measures. In such cases, a lower bound to the $R(D)$ or an algorithm to compute it proves to be helpful. The second part of this chapter provides a rate-distortion-theoretic analysis for speech coding based on the CDI distortion measure.

The remainder of the chapter is organized as follows. Section 4.2 puts forward the idea of CDI, a perceptual cross-entropy measure-based fidelity criterion for speech signals. Section 4.3 provides some experimental results with relevant remarks. Section 4.4 defines $R(D)$ mathematically, provides preliminary background and surveys pertinent literature. Section 4.5 addresses the $R(D)$ evaluation problem by characterizing a source-destination pair and computing an $R(D)$ function for speech coding directly using the Blahut algorithm. The performance of different speech coders is analyzed with respect to these limits.

4.2 Distortion Computation

The cochlear discrimination information (CDI) measure, in effect, determines the amount of new information (the increase in neural source entropy) associated with the coded signal when the neural source entropy associated with the original speech is known or vice versa.

Let P be a set of probability measures defined on a measure space $\mathcal{S}^{(J)}$ for a discrete information source with an alphabet of size J. The Rényi-Shannon entropy $H_\alpha(P)$ for such source with $P = \{p_1, p_2, \ldots, p_J\}$ is given as [70]

$$H_\alpha(P) = \begin{cases} -\sum_{j=1}^{J} p_j \log p_j, & \alpha = 1, \\ \\ \frac{1}{1-\alpha} \log(\sum_{j=1}^{J} p_j^\alpha), & \alpha \geq 0, \quad \alpha \neq 1. \end{cases} \tag{4.1}$$

It has been shown in [70, 71] that

1. $H_\alpha(P)$ is a continuous positive decreasing function of α and is also continuous in P.

2. $H_\alpha(P)$ is always non-negative and $H_\alpha(P) = 0$ if and only if all of the p_j's except one are equal to zero.

3. $H_\alpha(P)$ is strictly concave with respect to P for $0 < \alpha \leq 1$; i.e., $H_\alpha(\lambda P' + \overline{1-\lambda} P'') \geq \lambda H_\alpha(P') + (1-\lambda) H_\alpha(P'')$ $\forall P'$, P'' and all $\lambda \in (0,1)$.

4. Convexity or concavity of $H_\alpha(P)$ with respect to P depends on J for $\alpha > 1$.

Now, let us consider one neural channel for a specific clock time. Since there are only two events possible (i.e., firing and non-firing), the measure space can be written as

$$\mathcal{S}^{(2)} \triangleq \{P : P = (p_1, p_2); \quad p_1, p_2 \geq 0; \quad p_1 + p_2 = 1\}. \tag{4.2}$$

The Appendix A shows that with $P \in \mathcal{S}^{(2)}$, $H_\alpha(P)$ is strictly concave with respect to P not only for $0 < \alpha \leq 1$, but also for $1 < \alpha \leq 2$. Thus, here we consider α values in the range [0, 2) which ensures a global maximum of $H_\alpha(P)$ for $p_1 = p_2 = 1/2$.

In this work, time-domain speech representation $\mathcal{T}$ is mapped onto the PD $\mathcal{A}$ using Lyon's cochlear model $\mathcal{C}$. Mathematically, this mapping $\mathcal{B}$ can be expressed as $\mathcal{B} : \mathcal{T} \xrightarrow{\mathcal{C}} \mathcal{A}$. The PD representation $\mathcal{A}$ for an original speech signal can be written in a matrix form as

$$\mathcal{A} = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1N} \\ P_{21} & P_{22} & \cdots & P_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nN} \end{bmatrix} \tag{4.3}$$

with n clock times and N neural channels. An element P_{kl} of the matrix $\mathcal{A}$ implies that p_{1kl} and $p_{2kl} = 1 - p_{1kl}$ are the firing and the non-firing probabilities for the k-th neural channel at the l-th clock time corresponding to the original speech signal. Similarly, let q_{1kl} and $q_{2kl} = 1 - q_{1kl}$ be the firing and the non-firing probabilities for the coded/distorted speech. Accordingly, the directed divergence (a form of the

discrimination information measure) between P_{kl} and Q_{kl} can be written as [71]

$$D_\alpha(P_{kl};Q_{kl}) = \begin{cases} \sum_{j=1}^{2} p_{jkl} \log\left(\frac{p_{jkl}}{q_{jkl}}\right), & \alpha = 1, \\ \frac{1}{(\alpha-1)} \log\left(\sum_{j=1}^{2} \frac{p_{jkl}^{\alpha}}{q_{jkl}^{\alpha-1}}\right), & \alpha \geq 0, \quad \alpha \neq 1. \end{cases} \tag{4.4}$$

This measure is not a metric as it does not satisfy some of the conditions required for it to be a metric—(a) the *symmetry condition* [$D_\alpha(P_{kl};Q_{kl})$ is not the same as $D_\alpha(Q_{kl};P_{kl})$ when P_{kl} and Q_{kl} are different]; and (b) the *triangle inequality* [the sum of the measures $D_\alpha(P_{kl};Q_{kl})$ and $D_\alpha(Q_{kl};R_{kl})$ may be greater than, equal to or less than $D_\alpha(P_{kl};R_{kl})$ for any three probability distributions P_{kl}, Q_{kl} and R_{kl}]. However, the satisfaction of the *non-negativity condition* allows it to be considered as a fidelity criterion (even though it is not a metric). We define the directed divergence measure of order α for $0 < \alpha \leq 2$, the range in which $H_\alpha(P)$ has been shown to be concave with respect to $P \in \mathcal{S}^{(2)}$.

For simplicity, we assume that the neural firing events in different channels and at different clock times are independent. Thus, the neural sources corresponding to the N neural channels and the n clock times form a product source, i.e.,

$$\mathcal{S} = \times_{l\in\mathcal{L}} \times_{k\in\mathcal{K}} \mathcal{S}_{kl}^{(2)} \tag{4.5}$$

with $\times$ as the cartesian product of the probability spaces, $\mathcal{L} \equiv \{1,2,\cdots,n\}$ and $\mathcal{K} \equiv \{1,2,\cdots,N=64\}$. Under this assumption, the probability distribution of the product source is the product of the probability distributions of the individual sources [1] and the directed divergence values are additive, i.e.,

$$D_\alpha(P;Q) = \sum_{l\in\mathcal{L}} \sum_{k\in\mathcal{K}} D_\alpha(P_{kl};Q_{kl}). \tag{4.6}$$

The satisfaction of (4.6), along with the non-negativity of the directed divergence for $\alpha \geq 0$, are shown in the Appendix B.

One generalized form for the directed divergence measure is the f-divergence [72] based on which the distortion measure can be defined as

$$D_{\text{gen}}(P;Q) = \sum_{l\in\mathcal{L}} \sum_{k\in\mathcal{K}} \sum_{j=1}^{2} q_{jkl} f\left(\frac{p_{jkl}}{q_{jkl}}\right) \tag{4.7}$$

where $f(\bullet)$ is a convex function. This specializes to the directed divergence with $\alpha = 1$ (also known as the Kullback-Leibler divergence) if $f(x) = x \log x$; to the χ^2-divergence [72] if $f(x) = (x-1)^2$; to the K-directed divergence [73] if $f(x) = x \log\{2x/(1+x)\}$ and to the variational distance [74] if $f(x) = |x-1|$. It may be noted that there exist relationships among many of these measures (e.g., a lower bound for the Kullback-Leibler divergence in terms of the variational distance is given in [75]). In this work, we also use a 'symmetrized' divergence measure $S_\alpha(P;Q)$ defined as

$$S_\alpha(P;Q) = D_\alpha(P;Q) + D_\alpha(Q;P) \tag{4.8}$$

The divergence measures based on the entropies other than the Rényi-Shannon type can also be studied. One such common example is

$$C_{1.5}(P;Q) = \sum_{l\in\mathcal{L}} \sum_{k\in\mathcal{K}} \sum_{j=1}^{2} (\sqrt{p_{jkl}} - \sqrt{q_{jkl}})^2 \tag{4.9}$$

based on the Havrda-Charvat entropy $E_{1.5}(P)$ given as [74]

$$E_{1.5}(P) = 2(1 - \sum_{j=1}^{2} p_j^{1.5}). \tag{4.10}$$

In order to maintain the *boundedness* of the measure, in general, we impose a condition that the probability of firing or non-firing for the original and the coded signal can not be a complete certainty or uncertainty; and accordingly we associate a 1^- or a 0^+ probability, as appropriate.

4.3 Experimental Results

Twelve speech utterances, of 1–2 sec durations and spoken by male as well as female, were considered for the test. Digitized versions of these speech sentences (listed in the Appendix C) were stored in audio-files having SNR of 50 dB approximately. Each of these original utterances were passed through six different code-excited linear prediction (CELP)-type speech coders.

No database containing various types of coded/distorted speech with accompanying MOS ratings was available to us. Also, we did not attempt to develop MOS ratings as it implies substantial cost and considerable time. Obtaining such a subjective scale involves the great difficulty of repeatability and elimination of biases and

artifacts—especially without well-understood anchors. The quantization distortion unit (QDU), defined as the quantity of distortion subjectively equivalent to that of a single encoding of 64 kbps PCM, has often been used in practice as a distortion measure. Recent tests, however, indicate that the QDU may not be as stable and dependable as once it was thought to be [13]. Considering all these aspects, we decided to administer an informal subjective test against which the objective measure results were judged.

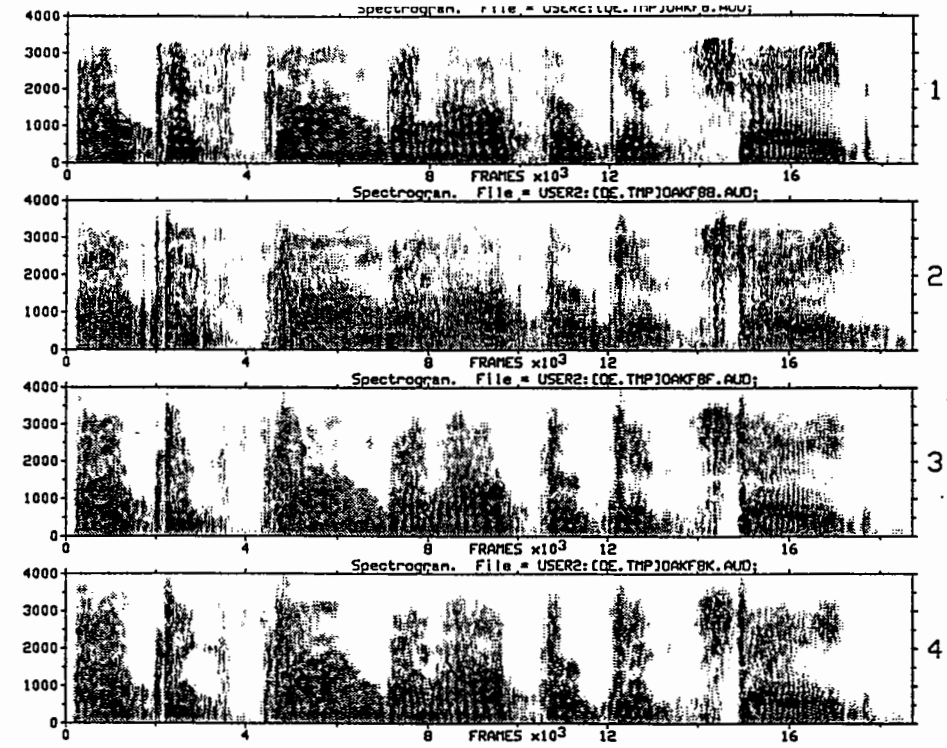
In this subjective test, twelve listeners ranked six different coded versions (two with 8 kbps coders C1, C2 and four with 4.8 kbps coders C3, C4, C5, C6) of all the twelve speech utterances. The overall perceptual quality of the coded signals was designated as the basis for the order of their preferences. Subsequently, we carried out an objective evaluation of these coded signals with reference to the original speech signal by considering eight variations of the proposed fidelity criterion. These measures were as follows.

1. The directed divergence with $\alpha = 1$ $[D_1(P;Q)]$,
2. The directed divergence with α=1.5 $[D_{1.5}(P;Q)]$,
3. The directed divergence with α=2 $[D_2(P;Q)]$,
4. The symmetrized divergence with α=1 $[S_1(P;Q)]$,
5. The variational distance $[V(P;Q)]$,
6. The χ^2-divergence $[\chi^2(P;Q)]$,
7. The K-directed divergence $[K(P;Q)]$ and
8. The Havrda-Charvat entropy-based $C_{1.5}$-divergence $[C_{1.5}(P;Q)]$.

A comparison of the informal listening test results and the objective measure values leads us to make the following remarks.

4.3.1 Performance of Objective Measures

In Fig. 4.1, the time-domain waveforms and the spectrograms of an original and three coded versions of a typical speech sentence, say, "*Oak is strong and also gives shade*"



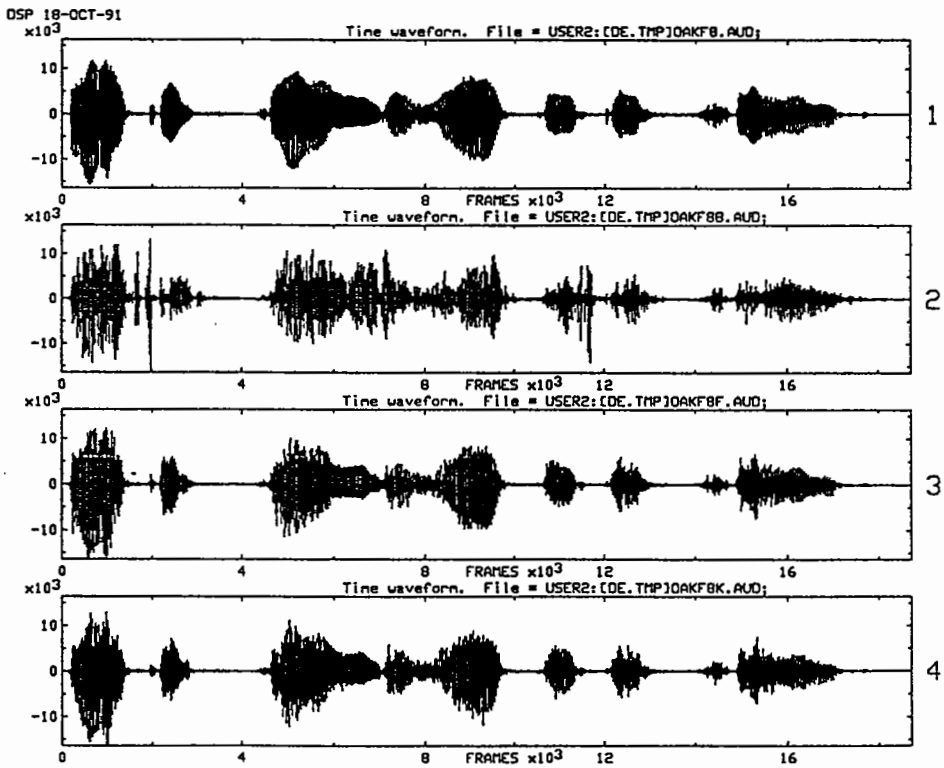


Figure 4.1: Time-domain waveforms and spectrograms of an original and three coded speech signals, "*Oak is strong and also gives shade.*"

(with 18,800 samples), are shown. Table 4.1 provides average distortion measure values per clock time (with a base-10 logarithm, wherever applicable) for the aforesaid speech utterance. We also tabulate the values of corresponding SNR_{seg} as well as SNR with and without scaling ('scaling' implies multiplication of all the coded speech samples by an appropriate factor so as to maximize the SNR value).

In Table 4.2, we provide subjective and objective measure values per clock time for each of the sentences. The subjective rankings (6 for the best and 1 for the worst) are averaged over the rankings made by the twelve listeners. These scores are average ordinal numbers and not the absolute quality scores. For each of the twelve utterances and six coded versions, the average ranking scores are mentioned in the first column (marked 'S'). As an example, if a coded signal is given a score of '6' by eight listeners, a score of '5' by three listeners and a score of '4' by one listener, the 'S' value becomes $(6 \times 8 + 5 \times 3 + 4 \times 1)/12 = 5.58$.

Measure Type	oakf8f	oakf8k	oakf8b	
Subjective Ranking	Best	Good	Poor	
$D_1(P;Q)$	2.721	2.756	4.273	
$D_{1.5}(P;Q)$	4.492	4.540	6.916	
$D_2(P;Q)$	6.751	6.812	10.165	
$S_1(P;Q)$	2.730	2.760	4.285	
$V(P;Q)$	8.777	8.845	11.454	
$\chi^2(P;Q)$	17.326	15.486	19.111	$\times$
$K(P;Q)$	0.795	0.806	0.909	
$C_{1.5}(P;Q)$	0.077	0.083	0.117	
SNR (w/o scaling [dB])	8.724	9.178	–2.597	$\times$
SNR (with scaling [dB])	8.979	9.334	0.009	$\times$
SNR_{seg} [dB]	6.815	7.080	–2.004	$\times$

Table 4.1: Different measure values for three coded signals (with three different 4.8 kbps speech coders) with reference to the original speech utterance F3 ('$\times$' indicates that the objective measures for 'oakf8f' and 'oakf8k' do not agree with the subjective rankings)

On the other side, we have computed the eight variations of the CDI mea-

sure values. However, here we tabulate only the $D_1(P;Q)$ measure values (in the second column marked 'D_1') as an example and make general remarks about the other measures. It is emphasized that the lower the amount of additional information (cross-entropy), better is the signal quality of the coded speech with reference to the original one. In Table 4.2, we observe that with the utterance M1, the C4, C5 coders and with the utterance F5, the C1, C2 coders were ranked same subjectively. Objective measures have shown slight preference towards C4 coder for M1 and towards C1 coder for F5. Besides that, for the utterance F4, the subjective and objective rankings were in contradiction for the coders C1, C2.

Sent.	C1		C2		C3		C4		C5		C6	
	S	D_1	S	D_1	S	D_1	S	D_1	S	D_1	S	D_1
M1	5.75	2.569	4.92	2.662	4.17	2.703	2.58	2.741	2.58	2.744	1.00	4.931
M2	5.50	2.630	5.17	2.651	4.25	2.678	2.75	2.702	2.25	2.793	1.08	4.817
M3	5.75	2.573	5.17	2.623	4.00	2.720	2.58	2.753	2.33	2.782	1.17	4.333
M4	5.00	2.672	5.67	2.654	4.25	2.716	2.50	2.752	2.58	2.747	1.00	4.776
M5	5.75	2.578	5.17	2.627	3.83	2.692	2.67	2.725	2.50	2.759	1.00	4.833
M6	5.58	2.621	5.25	2.666	3.83	2.696	2.75	2.719	2.42	2.760	1.17	4.669
F1	5.67	2.607	5.00	2.671	4.25	2.695	2.33	2.801	2.58	2.751	1.17	4.722
F2	5.67	2.612	5.00	2.678	3.91	2.737	2.67	2.766	2.50	2.774	1.25	4.285
F3	5.50	2.619	5.17	2.648	4.25	2.721	2.50	2.756	2.25	2.771	1.33	4.273
F4	5.41	2.661	5.25	2.649	4.17	2.700	2.75	2.729	2.17	2.793	1.25	4.562
F5	5.50	2.653	5.50	2.658	3.83	2.743	2.33	2.797	2.50	2.765	1.33	4.414
F6	5.67	2.602	4.83	2.674	4.08	2.694	3.08	2.701	2.17	2.791	1.17	4.379

Table 4.2: Subjective and objective measure values for coded signals with reference to the corresponding original speech utterances (M1–M6 (male) and F1–F6 (female) are speech utterances, C1–C6 are speech coders, 'S' denotes the average subjective ranking scores and 'D_1' gives the directed divergence measure values with $\alpha = 1$)

Over the test sentences, the human rankings were found to be almost consistent with the measures $D_1(P;Q)$, $D_{1.5}(P;Q)$, $D_2(P;Q)$ and $S_1(P;Q)$; and satisfactorily consistent with the measures $K(P;Q)$ and $C_{1.5}(P;Q)$. Furthermore, the $D_\alpha(P;Q)$ class of the measures has shown conformance to subjective evaluation results where

the SNR measure (with or without scaling) and also the SNR_{seg} measure have failed. However, the $V(P;Q)$ and the $\chi^2(P;Q)$ measures often disagreed with the subjective rankings, especially when two coded signals were very close in their perceptual quality.

4.3.2 Effect of Different Entropies

The $D_1(P;Q)$ and the $D_2(P;Q)$ measure profiles for one neural channel at a particular clock time are presented in Fig. 4.2 where the X-axis is the probability-of-firing for the original signal, the Y-axis is the probability-of-firing for the coded signal in the same channel and the Z-axis is the corresponding measure. It was noticed that the value of α in the $D_\alpha(P;Q)$ measure class has a consistent but small effect on its performance. For finer classification (i.e., classifying two coded signals almost equal in their perceptual quality), it has been found to be useful to apply an α value larger than one to increase the dynamic range of the measure values. It has also been observed that the measures based on the Rényi-Shannon entropy show better performance than that based on the Havrda-Charvat entropy.

4.3.3 Effect of Gain Changes

The $\chi^2(P;Q)$ and $V(P;Q)$ measure profiles with the same X, Y and Z axes as of Fig. 4.2 are shown in Fig. 4.3. In addition to the AGC nonlinearity, all the measure profiles (except the $V(P;Q)$) exhibit nonlinearity and the measure values are relatively very small in the neighborhood of the $X = Y$ region. This also makes them insensitive to small gain changes. We speculate that a linear profile of the $V(P;Q)$ measure is responsible for its poor performance. Due to its broad flatness around the $X = Y$ region, the $\chi^2(P;Q)$ measure shows less sensitivity to gain changes; however, this may be the reason for its unsatisfactory performance in the coder evaluation.

4.3.4 Effect of Sample Delays

The CDI measures, in general, were found to be relatively less sensitive (compared to the SNR measure) to a slight time misalignment of the coded signal with respect to the original one or vice versa. For example, let us consider the coded speech signals marked 'oakf8f' and 'oakf8k' of Fig. 4.1. Table 4.3 provides the SNR measure

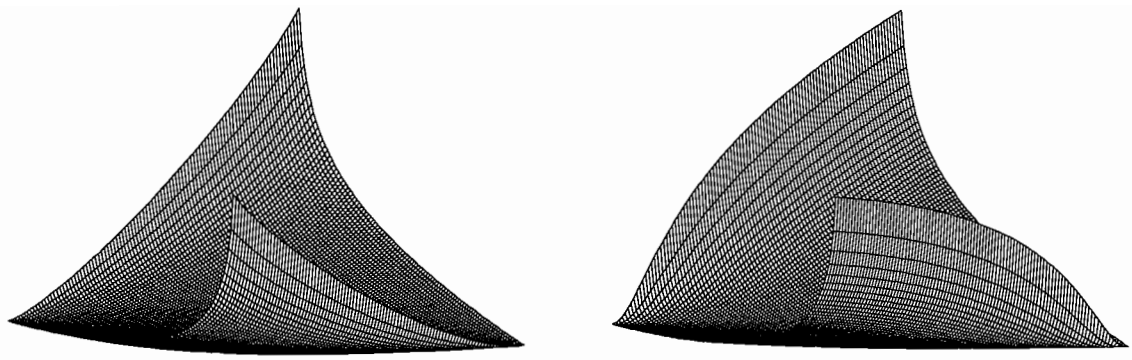


Figure 4.2: The discrimination measure profiles ($J = 2$)—(a) the directed divergence with $\alpha = 1$ and (b) the directed divergence with $\alpha = 2$.

(without scaling) values as well as the $D_1(P;Q)$ and the $D_2(P;Q)$ measure values with zero, one, two and three sample delays in the coded speech. These sample delays are with reference to the original signal and the misaligned sample places are filled in with zero values. In general, we observe that one sample delay does not cause much change in the CDI measure values, but two or three sample delays have considerable effect. With three sample delays, the measures show 'oakf8f' to be inferior to 'oakf8k' (which is aligned to the original signal) although subjectively the reverse is true.

Coded Speech	Measure	Sample Delays			
		Zero	One	Two	Three
oakf8f	SNR (w/o scaling [dB])	8.724	7.391	5.619	5.117
oakf8f	$D_1(P;Q)$	2.721	2.728	2.747	2.779
oakf8f	$D_2(P;Q)$	6.751	6.792	7.193	8.838
oakf8k	SNR (w/o scaling [dB])	9.178	7.503	6.108	7.027
oakf8k	$D_1(P;Q)$	2.756	2.762	2.791	3.128
oakf8k	$D_2(P;Q)$	6.812	6.855	7.124	8.950

Table 4.3: The directed divergence (with $\alpha = 1, 2$) measure values with zero, one, two and three sample delays for the coded signal 'oakf8f' and 'oakf8k' with reference to the original speech sentence

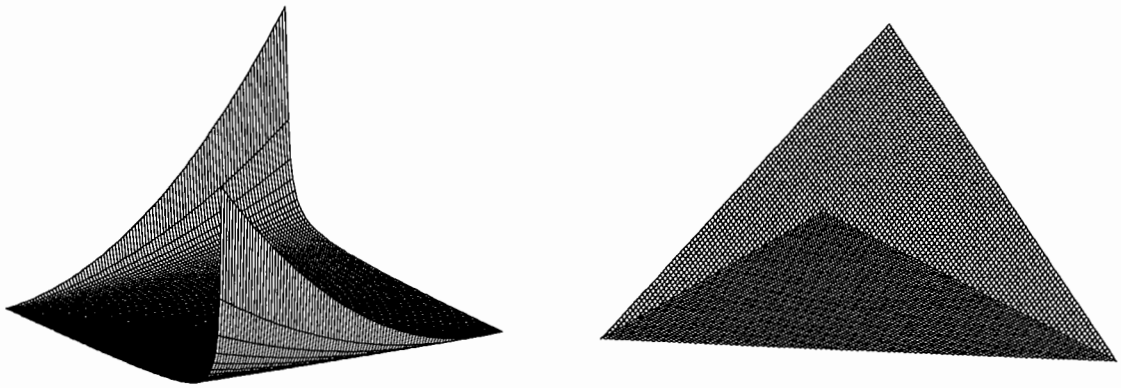


Figure 4.3: The discrimination measure profiles ($J = 2$)—(a) the χ^2 divergence and (b) the variational distance.

4.3.5 Speech Coder Identification

By considering the neural pathway to be a noisy channel, the subjective evaluation of the speech coders can be treated as a hypothesis testing problem. Csiszár and Longo [76] have shown that the probability-of-error of optimum hypothesis testers based on blocks of measurements decreases exponentially with the block length. Let us consider two coded speech of the same utterance and let γ^* be the smallest probability that 'C' is identified to be the samples of 'A' when it is actually the samples from 'B'. This probability is smallest over all the decision rules such that the probability of other type of error (i.e., 'C' chosen as samples of 'B' when it is actually from 'A') does not exceed β. Then, γ^*, for all β in (0,1) and with $\alpha = 1$, can be given as [76]

$$\gamma^* \sim \exp[-\sum_{l \in \mathcal{L}} \sum_{k \in \mathcal{K}} D(P_{kl}; Q_{kl})] \tag{4.11}$$

We conducted an experiment where the listeners were asked to listen to two coded speech sentences 'A' and 'B' and then a varying number of samples 'C' from one of them, not known to the listeners which one, were played. In such subjective evaluation testing, there is no precise way of determining γ^*. The γ^* could be estimated by carrying out the test with a large number of listeners and then considering their opinions (whether 'A' or 'B') about 'C'.

It would be of academic interest to investigate the validity of the relationship of (4.11). In our experiment, we only verified that to achieve a given probability of

Sentence	Sample Nos.	C3–C4	C4–C5	C5–C6
M1	3,000	5/12	4/12	7/12
	6,000	7/12	7/12	9/12
	9,000	11/12	10/12	12/12
	12,000	12/12	12/12	12/12
F3	3,000	6/12	4/12	8/12
	6,000	8/12	6/12	11/12
	9,000	11/12	9/12	12/12
	12,000	12/12	11/12	12/12

Table 4.4: Speech coder identification for two sentences M1 and F3 (the sample numbers played and the fraction of listeners who have correctly identified the coders are provided in the table)

decision error, it required more samples (i.e., longer durations) of 'C' to be played when 'A' and 'B' are of 'near equal' quality (as indicated by our measure) compared to that required when 'A' and 'B' are of 'substantially different' quality. Table 4.4 shows, for the same example sentence, the subjective identification of coders (i.e, the number of listeners out of twelve listeners correctly identified the coders) and the corresponding number of samples played. We have considered three coder pairs where C4–C5, C3–C4 and C5–C6 were ranked in the descending order from their perceptual quality 'closeness' point of view. For example, let us consider the utterance F3. In Table 4.4, we observe that by playing 6,000 samples, for C4–C5 coder pair, only one-half of the listeners could identify the coder correctly, the remaining listeners either identified wrongly or could not decide. On the other hand, with the same number of samples played, the correct coders were identified by two-third of the listeners for C3–C4 pair and by almost all the listeners for C5–C6 pair.

4.4 Rate-Distortion Analysis

Rate-distortion theory is a branch of information theory that establishes a mathematical foundation to a source encoding problem. For a particular source-destination pair, a rate-distortion function $R(D)$ could be computed which gives the lowest achievable

rate with an average distortion of D by the defined fidelity criterion. As D increases, $R(D)$ decreases monotonically and usually becomes zero at some finite value of distortion. In the following, we define the $R(D)$ analytically, discuss important results relevant to this work and review some of the pertinent literature.

4.4.1 Preliminary Background

We consider a time-discrete source $\{X_t, P\}$ that produces *i.i.d.* outputs described by a probability density function $p(x)$. The accuracy of reproduction of x by y is measured by a non-negative distortion measure $\rho(x, y)$. An average distortion

$$d(q) = \int\int p(x)q(y|x)\rho(x, y)\, dx\, dy \tag{4.12}$$

and an average mutual information

$$I(q) = \int\int p(x)q(y|x) \log\left\{\frac{q(y|x)}{q(y)}\right\} dx\, dy, \tag{4.13}$$

where

$$q(y) = \int p(x)q(y|x)\, dx \tag{4.14}$$

are assigned to every conditional probability density $q(y|x)$. Then, the rate distortion function $R(D)$ of $\{X_t, P\}$ with respect to the fidelity criterion is defined by

$$R(D) = \inf_{q \in Q_D} I(q), \tag{4.15}$$

where the set of all D-admissible conditional probability assignments is denoted by the symbol

$$Q_D = \{q(y|x) : d(q) = D\}. \tag{4.16}$$

$I(q)$ is a convex downward function of q which implies that any stationary point of $I(q)$ in Q_D must yield the absolute minimum, namely the $R(D)$. Since the above formulation is a convex programming problem, generalized Kuhn-Tucker conditions can be determined to identify the conditional probability distribution which attains the infimum in (4.15). The variational problem defining $R(D)$ can be solved using the method of Lagrange multipliers. An application of this method results in the following parametric expressions for D and R [77]:

$$D = \int\int \lambda(x)p(x)q(y)e^{s\rho(x,y)}\rho(x, y)\, dx\, dy \tag{4.17}$$

and

$$R = sD + \int p(x) \log \lambda(x)\, dx \tag{4.18}$$

where,

$$\lambda(x) = \left[\int q(y) e^{s\rho(x,y)}\, dx \right]^{-1}. \tag{4.19}$$

The slope of any $R(D)$ curve at the point (D_s, R_s) is represented by the parameter s which is generated parametrically from (4.17), (4.18) and (4.19). If Λ_s be the set of all non-negative functions $\lambda(x)$ satisfying

$$c(y) = \int \lambda(x) p(x) e^{s\rho(x,y)}\, dx \leq 1 \qquad \text{for all } y; \tag{4.20}$$

then,

$$R(D) = \sup_{s \leq 0, \lambda(x) \in \Lambda_s} \left[sD + \int p(x) \log \lambda(x)\, dx \right]. \tag{4.21}$$

For each $s \leq 0$, a necessary and sufficient condition for $\lambda(x)$ to attain the supremum in (4.21) is the existence of a probability density $q(y)$ that is related to $\lambda(x)$ by (4.19) and is such that $c(y) = 1$ in (4.20) for almost all y for which $q(y) > 0$.

4.4.2 Relevant Literature

The rate-distortion theory has been developed in the last two decades for discrete as well as continuous sources. For the evaluation of $R(D)$, two broad approaches are generally adopted.

One approach is to derive the Shannon lower bound $R_L(D)$ [77] and then to find conditions for the existence of a $D_c > 0$ *s.t.* $R(D) = R_L(D)$, for all $D \in [0, D_c]$. With difference-type distortion measures $R(D)$ functions have been calculated for Laplacian, Cauchy and Gaussian sources [77]. This idea is generalized and a parametric solution is provided for a weighted mean-square error distortion measure in [78]. For quotient-type distortion measure (i.e, a measure of the functional form $f(x/y)$) and a source with $p(x) = 0$ for $x < 0$, a logarithmic transformation of the source variables x and y yields $R(D)$ bounds from the results of the difference distortion measures [79]. With balanced distortion measures (i.e., with distortion matrix containing the same set of entries, perhaps permuted, in each column), the $R(D)$ functions are computed in [80] for discrete memoryless source and in [81] for

finite-alphabet sources with memory. However, it appears to be difficult with this approach to evaluate $R(D)$ with an arbitrary non-balanced fidelity criterion.

The second approach is to evaluate the $R(D)$ function directly. Some simple examples of the finite-alphabet source-destination pair for which the rate-distortion function that can be determined analytically are provided in [77]. Tan and Yao [82] have evaluated $R(D)$ for a Gaussian source and an absolute-magnitude difference criterion by making a suitable choice of the boundary set (i.e., the value of y for which the condition (4.20) is satisfied with equality). This method has also been applied to a large class of i.i.d. sources having probability densities with constrained tail decay [82]. An efficient algorithm for the direct evaluation of the $R(D)$ function for discrete as well as continuous sources has been suggested by Blahut in [83].

Historically, the application of the rate-distortion theory to the speech process has been hindered because of the lack of a widely accepted probabilstic model of the speech process as well as a meaningful distortion measure. The problem is further complicated by the mathematical difficulties in evaluating the rate-distortion function even if a reasonable source-destination pair is defined. A fairly large set of pdf models is suggested in the literature based on the first-order histograms of Nyquist samples of continuous speech waveforms. The gamma pdf based on the long-term statistics [3], the Laplacian pdf based on the medium-term statistics [84] and the Gaussian pdf based on the short-term statistics [85] are among the more popular ones. An evaluation of the first-order $R(D)$ functions based on these pdfs and difference distortion measures are available in [86]; and with Itakura-Saito distortion measure in [79].

4.5 Evaluation of Rate-Distortion Function

The objective of this section is to provide a rate-distortion-theoretic analysis for speech coders with the CDI measure. We formulate the problem by characterizing the source-destination pair precisely. Then, the $R(D)$ function is computed using the Blahut algorithm. Finally, the performances of different speech coders are studied with respect to these bounds.

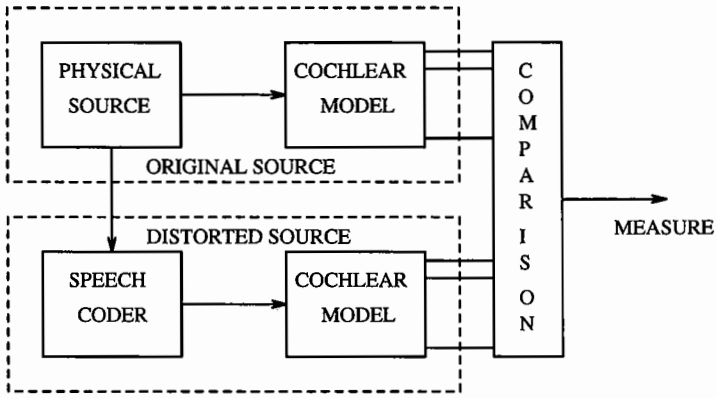


Figure 4.4: Source-destination pair characterization

4.5.1 Source-Destination Pair Characterization

The cochlear model is, in essence, a highly non-linear structure with the half-wave rectifiers, the AGC stages and the coupling among them simulating the auditory spectral and temporal masking phenomena. It may prove to be sufficiently difficult to express these signal processing operations, especially the coupling of the AGC stages, with the help of simple mathematical operators. Thus, we take a different outlook towards the source-destination pair model shown in Fig. 4.4. We merge the physical speech source with the cochlear model and consider this ensemble to be the source. Since there is as such no uniquely accepted pdf for the physical speech source, we are not in any further disadvantageous position by integrating the cochlear model with the speech source and determining the histogram of the cochlear model outputs. These outputs, being the probability-of-firing information, assume values in the range (0,1). The histogram for the firing-probability is determined by experimenting with twenty-four speech utterances (twelve male and twelve female voices) of 1–2 sec. durations. The firing-probability histogram for each of the sixty-four neural channels could be determined separately. For simplification purpose, we have assumed all the histograms to be identical and derived only one histogram based on the probability-of-firing information obtained from all the channels.

4.5.2 Calculation Based on Blahut's Algorithm

In [87], we have derived analytically a lower bound to the $R(D)$ with a single-letter cochlear variational distance measure. However, with the other distortion measure forms, it becomes difficult to give an analytical solution. Moreover, these are not exact solutions; they are merely lower bounds. Here, we use the Blahut algorithm for calculating the $R(D)$ functions exactly.

We treat the probability-of-firing information to be discrete-valued with symbols from one of the 255 uniformly spaced values between 0 and 1 (i.e., $1/256, 2/256, \cdots, 255/256$). Let the input alphabet (firing-probability corresponding to the original speech) u be reproduced in terms of an output alphabet (firing-probability corresponding to the coded speech) v. Then, the algorithmic steps could be written as follows.

Step 1 : An initial output probability distribution $\{Q_v\}$ is assumed, say, Q_v^0. The parameter set $\{A_{uv} = e^{s\rho_{uv}}\}$ is evaluated, where ρ_{uv} is the single-letter CDI measure between the input alphabet u and the output alphabet v.

Step 2 : The parameter s is chosen from the range of $-\infty$ to 0; and then Steps 3 and 4 are carried out with different values of s.

Step 3 : With the values of the input probability distribution P_u (obtained from the histogram of the cochlear model output) and the parameters A_{uv} the following parameters are calculated:

$$c_v = \sum_u P_u \frac{A_{uv}}{\sum_v A_{uv} Q_v}, \quad Q_v \leftarrow Q_v c_v, \tag{4.22}$$

$$L = \sum_v Q_v \log\ c_v, \quad U = \max_v \log\ c_v. \tag{4.23}$$

Step 4 : If $U - L \geq \epsilon$, then the previous step is repeated; otherwise, the program is terminated for this value of s by evaluating the following:

$$Q_{v|u} = \frac{A_{uv} Q_v}{\sum_v A_{uv} Q_v}, \tag{4.24}$$

$$D = \sum_u \sum_v P_u Q_{v|u} \rho_{uv}, \tag{4.25}$$

$$R(D) = sD - \sum_u P_u \log \left(\sum_v A_{uv} Q_v \right) - \sum_v Q_v \log c_v \tag{4.26}$$

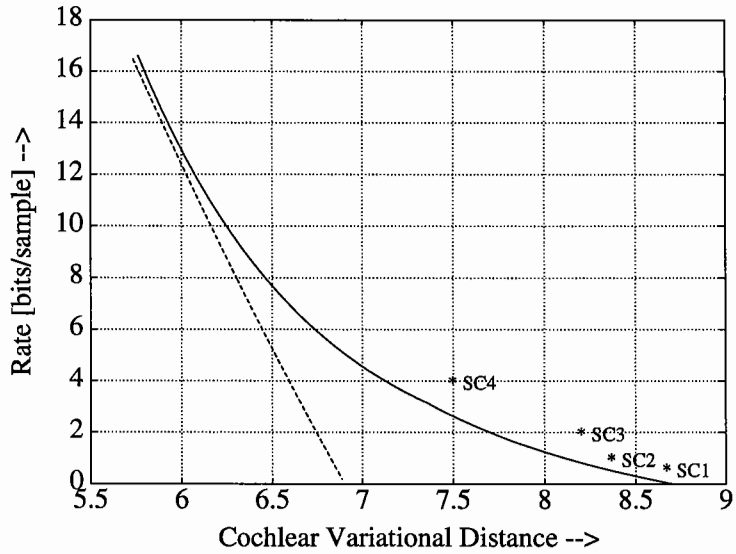


Figure 4.5: Speech coder rate in bits/sample vs. average cochlear variational distance measure (- - - line shows an analytically derived lower bound, —— line shows the exact rate-distortion curve using Blahut's algorithm and four '*' points [SC1–SC4] denote the performances of four speech coders)

Fig. 4.5 shows the $R(D)$ for the $V(P;Q)$ measure whereas Fig. 4.6 plots the $R(D)$ function for the $D_1(P;Q)$.

4.5.3 Measured Performances of Speech Coders

We have considered four state-of-the-art speech coders for the assessment of their average perceptual quality. These four coders (designated as SC1–SC4) were: CELP-based coder SC1 (4.8 kbps) [5], VSELP-based coder SC2 (8 kbps) [88], wideband CELP-based coder SC3 (16 kbps) [89] and ADPCM coder SC4 (32 kbps) [3]. For the first, second and the fourth coders with sampling rates of 8,000 Hz, sixty-four neural channels (covering up to 4,000 Hz band) were assigned as described in this chapter. On the other hand, for the wideband coder with sampling rate of 16,000 Hz, eighty-five neural channels (covering up to 8,000 Hz band) were assigned as described in Chapter 6. Although we considered only the CELP-type speech coders for comparing the CDI measure performance with subjective assessment, we do not foresee any difficulty in applying this measure to other types of speech coders. With this

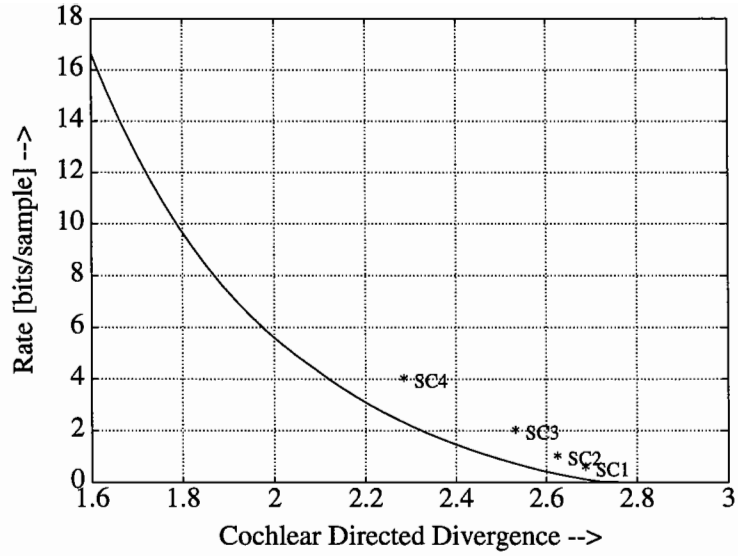


Figure 4.6: Speech coder rate in bits/sample vs. average cochlear directed divergence (with $\alpha = 1$) measure (—— line shows the rate-distortion curve using Blahut's algorithm and four '*' points [SC1–SC4] denote the performances of four speech coders)

understanding, we have included one ADPCM coder in this section to examine its quality with respect to the rate-distortion limit.

Twelve speech sentences of 1–2 sec durations were passed through each of the four coders to calculate the average distortion values over each sampling time. Fig. 4.5 and Fig. 4.6 plot the performances of the four speech coders (marked by '*') as evaluated by $V(P;Q)$ and $D_1(P;Q)$, respectively. Now, let us examine one of the figures, Fig. 4.6. We observe that the perceptual quality obtained (measured with the $D_1(P;Q)$) by SC1 coder is possible to achieve with much lower rate (as low as 1.5 kbps). Similarly, SC2, SC3 and SC4 coder performances are achievable with almost 3.8 kbps, 5.4 kbps and 20 kbps, respectively. From another perspective, we can say that a perceptual quality (a value of 2.575 units/sample) somewhere between those attained by SC2 and SC3 coders are attainable with a 4.8 kbps speech coder. A value of 2.485 units/sample which falls between the perceptual quality of SC3 and SC4 is theoretically achievable with an 8 kbps speech coder. Although the rate-distortion analysis does not provide with an answer to how to attain these limits, it gives an insight to what is possible and how close a specific speech coder is performing with respect to the $R(D)$ limits in terms of the perceptual quality.

4.6 Summary

In this chapter, the firing/non-firing probabilities of original and coded signals were compared in an information-theoretic sense to formulate the cochlear discrimination information measure. This fidelity criterion, in essence, evaluates the neural firing cross-entropy of the coded speech with respect to that of the original one. The performance of this objective measure was compared with subjective evaluation results. A low value in this measure has indicated superior quality of the corresponding speech coder. The last part of the chapter has dealt with the calculation of the rate-distortion functions for speech coding based on this distortion measure. For this purpose, we have applied the Blahut algorithm. Four speech coders with rates ranging from 4.8 kbps to 32 kbps were studied from the viewpoint of their performance (as assessed by the cochlear discrimination measure) with respect to the rate-distortion limits. Our study has shown that there is ample scope for the improvement of the coder architecture and the coding algorithm.

Chapter 5

Cochlear Hidden Markovian (CHM) Measure

5.1 Introduction

In Chapter 4, we have introduced a cochlear discrimination information measure which exploits the perceptual events at the auditory periphery. In this chapter, we attempt to capture the basics of high-level processing in the brain with simple hidden Markov models. We use these HMMs over the perceptual-domain speech representation and introduce a new measure [90, 91], namely the cochlear hidden Markovian (CHM) measure. Computing coder distortion with the CHM measure involves estimating the HMM parameters from the perceptual-domain observations of an original speech frame and calculating the likelihood (against the estimated HMM) of observing the PD representation corresponding to the coded version of the same speech frame. The proposed CDI measure compares the PD observations directly whereas the CHM measure is a parametric nonlinear model-based measure. Test results, model behavior, advantages/disadvantages of this method and also some other alternatives for measuring coder distortion are discussed.

The format of this chapter is as follows. Section 5.2 characterizes the hidden Markovian signal model. Section 5.3 provides some relevant background materials. Section 5.4 introduces a method to compute distortion for speech coders and also suggests briefly some other alternative approaches. Section 5.5 addresses the HMM

behavior and tabulates experimental results for speech coder evaluation.

5.2 Characterization of Hidden Markov Model

The cochlear model output is a sequence of K-dimensional vectors (in our work, K=64 corresponding to sixty-four neural channels) with one vector for each clock time t. The elements in each of the K-dimensional observation vectors represent information regarding the probability-of-firing. Based on this PD representation of a speech signal, what are transmitted through neural channels to the brain are series of all-or-none electrical spikes (firings). However, the exact conversion process of the PD representation to the firing/non-firing representation is not yet known. We attempt here to capture the underlying firing/non-firing event in each channel with discrete-time series analysis.

One such analysis technique involves using a hidden Markov model for modeling the observation sequence. The time-varying observation process is considered as a concatenation of many short-time segments of a fixed duration. However, it is expected that the properties of the process change neither synchronously with every analysis duration nor abruptly from each unit to the next one. The development of an efficient optimization technique [92] to estimate the model parameters so as to 'match' the observed signal patterns has culminated in the theory of HMM-based signal representation. The success of this hidden Markov modeling technique has been proven by its application in ecology (e.g., [93]), text analysis (e.g., [94]), coding theory (e.g., [95]) and speech recognition (e.g., [96]).

An HMM is a doubly embedded stochastic model with an underlying process that is not directly observable (it is hidden), but can be observed through another set of stochastic processes that produce the sequence of observations. In other words, the states of an HMM are hidden and the observation is a probabilistic function of the states. The order of occurrence of observations and the correlations among adjacent observations are suitably modeled by stochastic dependencies among the hidden states of an HMM. In the following, we characterize an HMM for our problem by selecting the model type, the number of hidden states and all the parameters associated with the model.

We consider K numbers of independent two-state ($N = 2$) fully-connected mod-

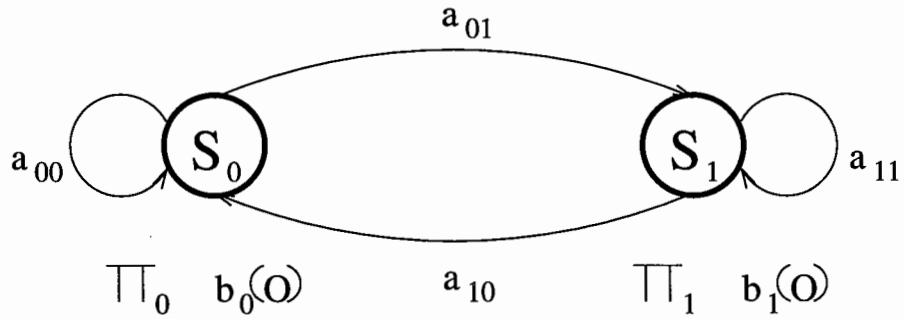


Figure 5.1: A two-state fully-connected hidden Markov model (S_0 and S_1 denote the non-firing and firing states, π_0 and π_1 are the initial state probabilities, a_{ij} gives the state transition probability from a state S_i to a state S_j, $b_0(O)$ and $b_1(O)$ are the observation probability density functions for the state S_0 and S_1 respectively)

els, as shown in Fig. 5.1, where either state is reachable from the other one. Although in many applications, the states do not have a physical meaning; here a state S_0 corresponds to a non-firing event whereas a state S_1 corresponds to a firing event. The initial state distribution (i.e., at t=1) is given as $\pi = \{\pi_i | i \in \mathcal{N}\}$ with

$$\pi_i = P[q_1 = S_i] \quad \text{for} \quad i \in \mathcal{N} \quad \text{and} \quad \sum_{i \in \mathcal{N}} \pi_i = 1, \tag{5.1}$$

where $\mathcal{N} \equiv \{0, 1\}$ and a state reached at any clock time t is denoted by q_t.

The HMM considered is of order one and hence the transition from one state to the next one occurs according to a transition probability distribution which depends only on the previous state. If we define an integer set $\mathcal{T} \equiv \{1, 2, \cdots, T-1\}$ then the state transition probability distribution $A = \{a_{ij} | i, j \in \mathcal{N}\}$ is given by

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i] \quad \text{for} \quad i, j \in \mathcal{N} \quad \text{and} \quad t \in \mathcal{T} \tag{5.2}$$

where every a_{ij} coefficient (i.e., $a_{00}, a_{01}, a_{10}, a_{11}$) is positive, and $\sum_{j \in \mathcal{N}} a_{ij} = 1$ for $i \in \mathcal{N}$.

Now, we consider any one of the neural channels for which the observation is represented by $\boldsymbol{O}$=$O_1 O_2 \cdots O_T$. To avoid significant degradation due to any quantization process, we treat the PD representation to be continuous-valued and accordingly

consider an HMM with continuous pdfs. However, the use of a continuous pdf requires some restrictions on its form so as to facilitate reestimation of the pdf parameters (e.g., mean, variance) in a consistent manner. The pdf for each of the two states is maintained fixed regardless of when and how the state is reached. The most general representation of the pdf, for which a reestimation procedure exists [92], is used here. Each state S_j is characterized by a continuous mixture pdf $b_j(x)$ of the form

$$b_j(x) = \sum_{m \in \mathcal{M}_L} c_{jm} b_{jm}(x) \quad \text{for} \quad j \in \mathcal{N}, \tag{5.3}$$

where $\mathcal{M}_L \equiv \{1, 2, \cdots, L\}$ with L as the number of components in the mixture and $b_{jm}(.)$ is any log-concave [92] or elliptically symmetric [97] density. The rationale behind choosing a mixture pdf and selecting the component pdf $b_{jm}(.)$ to be log-concave or elliptically symmetric is discussed later. In our present study, $b_{jm}(.)$ is assumed to be a beta density function and can be written as

$$b_{jm}(x) = \frac{\Gamma(d_{jm} + f_{jm} + 2)}{\Gamma(d_{jm} + 1)\Gamma(f_{jm} + 1)} x^{d_{jm}} (1 - x)^{f_{jm}} \quad \text{for} \quad d_{jm}, f_{jm} > 0, j \in \mathcal{N}, m \in \mathcal{M}_L, \tag{5.4}$$

where d_{jm} and f_{jm} are the parameters associated with the density function. The beta pdf of (5.4) is suitable as the observations are continuous-valued between 0 and 1. The Appendix D shows that the beta density function satisfies the log-conacavity condition.

The *observation probability density function* B is denoted as $B = \{b_j(x) | j \in \mathcal{N}\}$, where $b_j(x)\,dx$ is the probability of observing a value O_t in state S_j at clock time t. A coefficient c_{jm} is the m-th component mixture gain in state S_j and the set $\{c_{jm} | j \in \mathcal{N}, m \in \mathcal{M}_L\}$ satisfies the stochastic constraint

$$\sum_{m \in \mathcal{M}_L} c_{jm} = 1 \quad \text{for} \quad j \in \mathcal{N} \quad \text{with} \quad c_{jm} > 0 \quad \text{for} \quad j \in \mathcal{N} \text{ and } m \in \mathcal{M}_L \tag{5.5}$$

so that

$$\int_{-\infty}^{\infty} b_j(x)\,dx = 1, \quad j \in \mathcal{N}. \tag{5.6}$$

5.3 Preliminaries

In Section 5.2, an HMM has been defined by describing the complete parameter set of the model. The model is represented as $\boldsymbol{\lambda}=(\pi, A, B)$, where π is the state probability

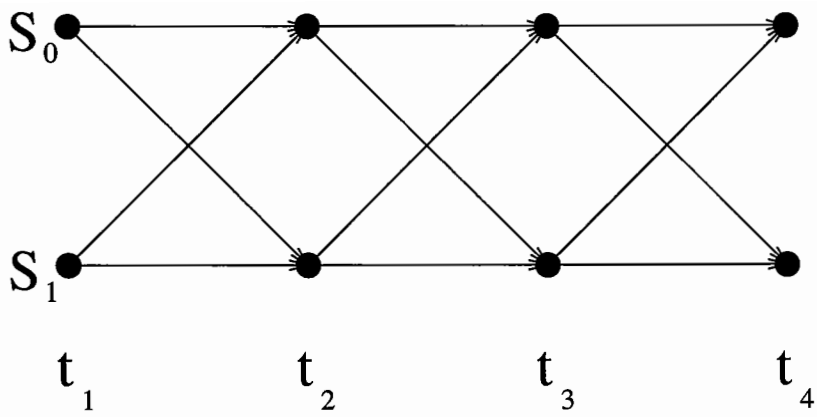


Figure 5.2: A two-state trellis diagram (S_0 and S_1 denote the non-firing and firing states)

vector, A is the state transition probability matrix and B is a set of two (N=2) continuous mixture pdfs, each with L mixtures. In this section, we provide some preliminaries required for computing the degree of distortion (similarity) of a coded speech with reference to its original version. A forward and a backward likelihood variables and an auxiliary function are defined below.

5.3.1 Forward and Backward Likelihood Variables

Let us extend the integer set $\mathcal{T}$ to $\mathcal{T}^+$ as $\mathcal{T}^+ \equiv \mathcal{T} + \{T\}$. Following Baum [92], a *forward likelihood variable* $\alpha_t(i)$ is then defined as

$$\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = S_i|\boldsymbol{\lambda}), \quad \text{for } i \in \mathcal{N} \text{ and } t \in \mathcal{T}^+ \tag{5.7}$$

which gives the probability of observing the *partial* sequence $O_1 O_2 \cdots O_t$ (until time t) and reaching the state S_i at clock time t *given* an HMM $\boldsymbol{\lambda}$. Likewise, a *backward likelihood variable* $\beta_t(j)$ is defined as

$$\beta_t(j) = P(O_{t+1} O_{t+2} \cdots O_T|q_t = S_j, \boldsymbol{\lambda}), \quad \text{for } j \in \mathcal{N} \text{ and } t \in \mathcal{T} \tag{5.8}$$

which gives the probability of observing the partial sequence $O_{t+1} O_{t+2} \cdots O_T$ (from $t+1$ to the end) *given* state S_j at time t and a model $\boldsymbol{\lambda}$.

The forward likelihood variable $\alpha_t(i)$ is initialized as the joint probability of

being in state S_i at $t = 1$ and an initial observation O_1, i.e.,

$$\alpha_1(i) = \pi_i b_i(O_1), \quad i \in \mathcal{N}. \tag{5.9}$$

With the help of the trellis diagram shown in Fig. 5.2, an iterative procedure is followed to compute the other forward likelihood variables from the initial one. Since $\alpha_t(i)$ is the probability of the joint event that $O_1 O_2 \cdots O_t$ are observed and the state S_i is reached at clock time t, the product $\alpha_t(i)a_{ij}$ becomes the probability of the joint event that $O_1 O_2 \cdots O_t$ are observed and the state S_j is reached at $t+1$ through the state S_i at t. Summation of this product over the possible two states S_i (for $i \in \mathcal{N}$) at time t yields the probability of reaching state S_j at $t+1$ with the corresponding partial observation sequence upto time t. Multiplication of the summed quantity by $b_j(O_{t+1})$, the probability of observing O_{t+1} at state S_j, results in the forward likelihood variable $\alpha_{t+1}(j)$ for time $t+1$. This evaluation procedure can be expressed by the following recurrence equation:

$$\alpha_{t+1}(j) = \left[\sum_{i \in \mathcal{N}} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \quad t \in \mathcal{T}, \quad j \in \mathcal{N}. \tag{5.10}$$

In a similar manner, let us now consider the backward variable $\beta_t(i)$. An initialization process arbitrarily defines

$$\beta_T(j) = 1, \qquad j \in \mathcal{N}. \tag{5.11}$$

Then, $\beta_t(i)$ is calculated recursively as follows:

$$\beta_t(i) = \sum_{j \in \mathcal{N}} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t \in \mathcal{T}, \quad i \in \mathcal{N}. \tag{5.12}$$

For a given model $\boldsymbol{\lambda}$, $\beta_t(i)$ is the probability of observing the particular partial sequence from time $t+1$ to the end when it is known that the state S_i is reached at time t. To compute this, it is evident from the trellis diagram of Fig. 5.2 that we need to consider both the states S_0 and S_1 at time $t+1$ accounting for the possible transitions from S_i to S_j, the observation O_{t+1} in state S_j and also the partial observation sequence $O_{t+2} O_{t+3} \cdots O_T$ (being in state S_j at time $t+1$).

5.3.2 Auxiliary Function

Following the concept of the Kullback-Leibler statistic, an auxiliary function $F(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ of two models $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$, for a given observation vector $\boldsymbol{O}$, can be defined [98] as

$$F(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}) \log P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}') \tag{5.13}$$

with $\boldsymbol{Q} = q_1 q_2 \cdots q_T$, $\boldsymbol{M} = m_1 m_2 \cdots m_T$, $q_k \in \mathcal{N}$ and $m_k \in \mathcal{M}_L$ for $k \in \mathcal{T}$. In the following, we show that if $F(\boldsymbol{\lambda}, \boldsymbol{\lambda}') \geq F(\boldsymbol{\lambda}, \boldsymbol{\lambda})$, then $P(\boldsymbol{O}|\boldsymbol{\lambda}') \geq P(\boldsymbol{O}|\boldsymbol{\lambda})$.

$$\begin{aligned}
P(\boldsymbol{O}|\boldsymbol{\lambda}) \log \frac{P(\boldsymbol{O}|\boldsymbol{\lambda}')}{P(\boldsymbol{O}|\boldsymbol{\lambda})} &= P(\boldsymbol{O}|\boldsymbol{\lambda}) \log \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} \frac{P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}')}{P(\boldsymbol{O}|\boldsymbol{\lambda})} \\
&= P(\boldsymbol{O}|\boldsymbol{\lambda}) \log \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} \frac{P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda})}{P(\boldsymbol{O}|\boldsymbol{\lambda})} \cdot \frac{P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}')}{P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda})} \\
&\geq P(\boldsymbol{O}|\boldsymbol{\lambda}) \cdot \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} \frac{P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda})}{P(\boldsymbol{O}|\boldsymbol{\lambda})} \cdot \log \frac{P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}')}{P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda})} \\
&= \left[F(\boldsymbol{\lambda}, \boldsymbol{\lambda}') - F(\boldsymbol{\lambda}, \boldsymbol{\lambda})\right] \geq 0
\end{aligned} \tag{5.14}$$

with strict inequality except when $P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}) = P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}')$. In the above, the fact that $\log x$ is strictly concave for $x > 0$ (since $d^2/dx^2(\log x) = -x^{-2} < 0$) has been used. The first inequality is the well-known Jensen's inequality whereas the second one is true by hypothesis. If the current model is defined as $\boldsymbol{\lambda}$=(π, A, B) and a reestimated model is $\boldsymbol{\lambda}'$=(π', A', B'); then either the initial model $\boldsymbol{\lambda}$ defines a critical point of the likelihood function (in that case $\boldsymbol{\lambda}' = \boldsymbol{\lambda}$), or the model $\boldsymbol{\lambda}'$ is better than the model $\boldsymbol{\lambda}$ in a sense that the observation sequence $\boldsymbol{O}$ is more likely to have been generated by $\boldsymbol{\lambda}'$. From (5.14), we observe that the maximization of $P(\boldsymbol{O}|\boldsymbol{\lambda})$ implies maximization of the auxiliary function; and hence a critical point of the auxiliary function gives an estimate about the HMM parameters.

5.4 Distortion Measure Methodology

An original speech segment and its coded version are passed through the cochlear model to obtain the PD representations. For each of these segments, the PD observations are sequences of 64-dimensional vectors corresponding to sixty-four neural

channels. A hidden Markov model is associated with each of the channels and the parameters are estimated from the PD observation sequence produced by the original speech segment. In a sense, all the sixty-four HMMs are 'trained' with the pertinent observation vectors corresponding to the original speech segment. Then, for the same speech segment, the observations from all the coded speech signals are 'matched' against the derived HMMs to compute the relative coder distortions. Now we describe the exact procedures for the model parameter estimation as well as the likelihood computation.

5.4.1 Parameter Estimation

There is no optimal way of estimating the model parameters from any finite-length observation sequence. Since the closed-form maximum likelihood is not possible, the HMM parameters are (re)estimated iteratively starting from an initial estimate. To solve this problem, Baum-Welch reestimation algorithm [99] is used here. An application of this algorithm is equivalent to solving a mathematical optimization problem for obtaining the maximum likelihood estimates of the HMM parameters. The scheme for estimating the HMM parameters is based on the maximization of the probability of the observation sequence given a model. This algorithm is quite powerful as it ensures a monotonic increase in the likelihood with the successive iterations of the algorithm [92].

Let us now consider the calculation of $P(\boldsymbol{O}|\boldsymbol{\lambda})$, the probability of the observation sequence $\boldsymbol{O}$ given the model $\boldsymbol{\lambda}$. Assuming the statistical independence of observations, for every given state sequence $\boldsymbol{Q}=q_1 q_2 \cdots q_T$, the probability of observing $\boldsymbol{O}$ can be written as $P(\boldsymbol{O}|\boldsymbol{Q},\boldsymbol{\lambda})$, where

$$P(\boldsymbol{O}|\boldsymbol{Q},\boldsymbol{\lambda}) = b_{q_1}(O_1) b_{q_2}(O_2) \cdots b_{q_T}(O_T). \tag{5.15}$$

The probability of the occurrence of such a state sequence $\boldsymbol{Q}$ is given as

$$P(\boldsymbol{Q}|\boldsymbol{\lambda}) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}. \tag{5.16}$$

Using (5.15) and (5.16), $P(\boldsymbol{O}|\boldsymbol{\lambda})$ can be computed as

$$P(\boldsymbol{O}|\boldsymbol{\lambda}) = \sum_{Q \in \mathcal{N}^T} P(\boldsymbol{O}|\boldsymbol{Q},\boldsymbol{\lambda}) P(\boldsymbol{Q}|\boldsymbol{\lambda}). \tag{5.17}$$

The global density function of (5.17) with the state density defined by (5.3) can be rewritten as

$$\begin{aligned} P(\boldsymbol{O}|\boldsymbol{\lambda}) &= \sum_{Q\in\mathcal{N}^T} \pi_{q_1} \prod_{t\in\mathcal{T}^+} \left[a_{q_t q_{t+1}} \left\{ \sum_{m\in\mathcal{M}_L} c_{q_t m} b_{q_t m}(O_t) \right\} \right] \\ &= \sum_{Q\in\mathcal{N}^T} \sum_{m_1\in\mathcal{M}_L} \sum_{m_2\in\mathcal{M}_L} \cdots \sum_{m_T\in\mathcal{M}_L} \left[\pi_{q_1} \prod_{t\in\mathcal{T}^+} a_{q_t q_{t+1}} c_{q_t m_t} b_{q_t m_t}(O_t) \right] \end{aligned} \quad (5.18)$$

assuming the parameter $a_{q_T q_{T+1}} = 1$. The direct computation of $P(\boldsymbol{O}|\boldsymbol{\lambda})$ as given by (5.18) involves enumerating every possible state sequence of length T. Instead, we exploit the trellis structure and use (5.10) and (5.12) for the forward and the backward likelihood parameters. In order to describe the procedure for an iterative update of the HMM parameters, we define a set of *transition likelihood variables* $\{\xi_t(i,j)|i,j\in\mathcal{N}, t\in\mathcal{T}\}$ as

$$\xi_t(i,j) = P(\boldsymbol{O}, q_t = S_i, q_{t+1} = S_j|\boldsymbol{\lambda}) \quad (5.19)$$

which gives the probability of observing the particular sequence $\boldsymbol{O}$, and being in the state S_i at time t and the state S_j at time $t+1$ given the model. From the trellis diagram of Fig. 5.2, it can be noted that $\xi_t(i,j)$ can be written as

$$\xi_t(i,j) = \sum_{m\in\mathcal{M}_L} \alpha_t(i) a_{ij} c_{jm} b_{jm}(O_{t+1}) \beta_{t+1}(j). \quad (5.20)$$

We note the following relationships among the three likelihood variables as defined in (5.10), (5.12) and (5.19):

1. A product of the forward and the backward likelihood variables for any clock time t is shown, using (5.3) and (5.12), equal to the sum of the transition likelihood variable $\xi_t(i,j)$ over the index j.

$$\begin{aligned} \alpha_t(i)\beta_t(i) &= \alpha_t(i) \left[\sum_{m\in\mathcal{M}_L} \sum_{j\in\mathcal{N}} a_{ij} c_{jm} b_{jm}(O_{t+1}) \beta_{t+1}(j) \right] \\ &= \sum_{j\in\mathcal{N}} \xi_t(i,j). \end{aligned} \quad (5.21)$$

2. Using (5.10) and (5.12), it is shown that a sum of the product of the forward and the backward likelihood variables, i.e., $\alpha_t(i).\beta_t(i)$ over i is independent of

the time index t.

$$\begin{aligned}\sum_{j\in\mathcal{N}} \alpha_{t+1}(j)\beta_{t+1}(j) &= \sum_{j\in\mathcal{N}} \left[\sum_{i\in\mathcal{N}}\sum_{m\in\mathcal{M}_L} \alpha_t(i)a_{ij}c_{jm}b_{jm}(O_{t+1})\right]\beta_{t+1}(j) \\ &= \sum_{i\in\mathcal{N}} \alpha_t(i)\left[\sum_{j\in\mathcal{N}}\sum_{m\in\mathcal{M}_L} a_{ij}c_{jm}b_{jm}(O_{t+1})\beta_{t+1}(j)\right] \\ &= \sum_{i\in\mathcal{N}} \alpha_t(i)\beta_t(i) \qquad \text{for } t\in\mathcal{T} \end{aligned} \tag{5.22}$$

3. Using (5.11) and applying (5.22) recurrently, $P(\boldsymbol{O}|\boldsymbol{\lambda})$ can be written as the sum of the *terminal forward likelihood variables* $\alpha_T(i)$ over i, i.e.,

$$P(\boldsymbol{O}|\boldsymbol{\lambda}) = \sum_{i\in\mathcal{N}} \alpha_T(i). \tag{5.23}$$

The logarithm of $P(\boldsymbol{O},\boldsymbol{Q},\boldsymbol{M}|\boldsymbol{\lambda}')$, the square bracketed term in (5.18), can be written as

$$\log P(\boldsymbol{O},\boldsymbol{Q},\boldsymbol{M}|\boldsymbol{\lambda}') = \log \pi'_{q_1} + \sum_{t\in\mathcal{T}^+} \log a'_{q_t q_{t+1}} + \sum_{t\in\mathcal{T}^+} \log c'_{q_t m_t} + \sum_{t\in\mathcal{T}^+} \log b'_{q_t m_t}(O_t). \tag{5.24}$$

It is seen that the HMM parameters π', A' and B' corresponding to the model $\boldsymbol{\lambda}'$ are segregated. Without any loss of generality, then the auxiliary function $F(\boldsymbol{\lambda},\boldsymbol{\lambda}')$ of (5.13) can also be written in a separated form as

$$\begin{aligned} F(\boldsymbol{\lambda},\boldsymbol{\lambda}') &= \sum_{Q\in\mathcal{N}^T}\sum_{M\in\mathcal{M}_L^T} P(\boldsymbol{O},\boldsymbol{Q},\boldsymbol{M}|\boldsymbol{\lambda})\left\{\log \pi'_{q_1} + \sum_{t\in\mathcal{T}^+} \log a'_{q_t q_{t+1}}\right. \\ &\quad \left. + \sum_{t\in\mathcal{T}^+} \log c'_{q_t m_t} + \sum_{t\in\mathcal{T}^+} \log b'_{q_t m_t}(O_t)\right\}. \end{aligned} \tag{5.25}$$

Since $F(\boldsymbol{\lambda},\boldsymbol{\lambda}')$ is considered as the basis for the maximum likelihood optimization procedure, separability of the individual auxiliary functions as given in Appendix E simplifies the (re)estimation procedure. Individual maximization of the first three summands subject to the constraints

$$\sum_{j\in\mathcal{N}} \pi_j = 1, \quad \pi_j \geq 0 \ \text{ for } j\in\mathcal{N}. \tag{5.26}$$

$$\sum_{j\in\mathcal{N}} a_{ij} = 1, \quad a_{ij} \geq 0 \ \text{ for } i,j\in\mathcal{N}. \tag{5.27}$$

$$\sum_{m \in \mathcal{M}_L} c_{im} = 1, \quad c_{im} \geq 0 \ \text{ for } i \in \mathcal{N}, \ \ m \in \mathcal{M}_L. \tag{5.28}$$

respectively, is well known. Each of the individual auxiliary functions has the same form $\sum_{j \in \mathcal{N}} u_j \log v_j$, which as a function of $\{v_j | j \in \mathcal{N}\}$ with the constraint $\sum_{j \in \mathcal{N}} v_j = 1$ and $v_j \geq 0$ for $j \in \mathcal{N}$ attains a global maximum at the single point $v_j = u_j / \sum_{i \in \mathcal{N}} u_i$ for $j \in \mathcal{N}$. The initial probability $\overline{\pi}$ can be reestimated as

$$\overline{\pi}_i = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i \in \mathcal{N}} \alpha_1(i)\beta_1(i)} = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i \in \mathcal{N}} \alpha_T(i)}, \quad \text{for } \ i \in \mathcal{N} \tag{5.29}$$

which is the expected frequency in state S_i at $t = 1$. Similarly, the reestimation formula for A results in a ratio of the expected number of transitions from state S_i to state S_j to the expected number of transitions out of state S_i, i.e.,

$$\overline{a}_{ij} = \frac{\sum_{t \in \mathcal{T}^+} \xi_t(i,j)}{\sum_{t \in \mathcal{T}^+} \sum_{j \in \mathcal{N}} \sum_{m \in \mathcal{M}_L} \xi_t^{(m)}(i,j)} = \frac{\sum_{t \in \mathcal{T}^+} \xi_t(i,j)}{\sum_{t \in \mathcal{T}^+} \alpha_t(i)\beta_t(i)}, \tag{5.30}$$

where $\xi_t^{(m)}(i,j)$ is the probability of being in state S_j at time $t+1$ and state S_i at time t with the m-th mixture component accounting for O_t, i.e.,

$$\xi_t^{(m)}(i,j) = \xi_t(i,j). \left[\frac{c_{im} b_{im}(O_t)}{\sum_{l \in \mathcal{M}_L} c_{il} b_{il}(O_t)} \right]. \tag{5.31}$$

with $b_{im}(O_t)$ as given by (5.3). $\overline{c}_{im}$ is the ratio of the expected number of transitions out of state S_i using the m-th mixture component to the expected number of total transitions out of state S_i. Thus, for $i \in \mathcal{N}$ and $m \in \mathcal{M}_L$, we get

$$\overline{c}_{im} = \frac{\sum_{t \in \mathcal{T}^+} \sum_{j \in \mathcal{N}} \xi_t^{(m)}(i,j)}{\sum_{t \in \mathcal{T}^+} \sum_{j \in \mathcal{N}} \sum_{m \in \mathcal{M}_L} \xi_t^{(m)}(i,j)} = \frac{\sum_{t \in \mathcal{T}^+} \sum_{j \in \mathcal{N}} \xi_t^{(m)}(i,j)}{\sum_{t \in \mathcal{T}^+} \alpha_t(i)\beta_t(i)}. \tag{5.32}$$

The parameters set $\{d_{im} | i \in \mathcal{N}, m \in \mathcal{M}_L\}$ and $\{f_{im} | i \in \mathcal{N}, m \in \mathcal{M}_L\}$ can be calculated from the following two equations.

$$\sum_{r=1}^{f_{im}+1} \frac{1}{(d_{im} + r)} = -\frac{\sum_{t \in \mathcal{T}^+} \sum_{j \in \mathcal{N}} \xi_t^{(m)}(i,j) \log(O_t)}{\sum_{t \in \mathcal{T}^+} \sum_{j \in \mathcal{N}} \xi_t^{(m)}(i,j)}, \tag{5.33}$$

$$\sum_{r=1}^{d_{im}+1} \frac{1}{(f_{im}+r)} = -\frac{\sum_{t\in\mathcal{T}+}\sum_{j\in\mathcal{N}} \xi_t^{(m)}(i,j)\log(1-O_t)}{\sum_{t\in\mathcal{T}+}\sum_{j\in\mathcal{N}} \xi_t^{(m)}(i,j)}, \tag{5.34}$$

where the parameters d_{im} and f_{im} are assumed, for reducing computations, to take up integer values.

5.4.2 Distortion Computation

We now discuss the CHM measure methodology. At first, we obtain the PD observation sequences from the original signal. For each of the sixty-four neural channels, we consider these PD observations for a frame of T consecutive clock times. An HMM is associated with each of such channels and the model parameters are determined starting from an initial estimate. Equations (5.29) to (5.34), derived based on the Baum-Welch algorithm, are used for estimating the model parameters. This technique iteratively chooses a 'better' model by maximizing $P(\boldsymbol{O}_n|\boldsymbol{\lambda}_n)$ where $\boldsymbol{O}_n$ is the n-th channel PD observation sequence for the original speech. After a reasonable number of iterations, the algorithm is terminated and the final model is denoted as $\boldsymbol{\lambda}_n^{(o)}$. Let the n-th channel PD observations for a corresponding coded speech be represented by $\boldsymbol{O}_n^{(c)}$. Using (5.23), we compute $P(\boldsymbol{O}_n^{(c)}|\boldsymbol{\lambda}_n^{(o)})$ for all the neural channels. This computation, in essence, evaluates the likelihood probability of the PD representation of the coded signal against the models derived from the PD representation of the original speech. We assume the neural channels to be independent and therefore the probability scores are multiplied. Upon taking logarithm, we obtain a similarity measure M_f for the frame as

$$M_f = \sum_{n=1}^{64} \log P(\boldsymbol{O}_n^{(c)}|\boldsymbol{\lambda}_n^{(o)}). \tag{5.35}$$

Finally, a cochlear hidden Markovian (CHM) distortion measure value is defined by taking average of M_f values over all the frames, negating it and also dividing it by 64 (i.e., CHM=$-\overline{M}_f/64$).

5.4.3 Alternative Approaches

Here, we suggest two other logical approaches for computing coder distortion although we have not carried out any tests with them.

State sequence approach

One alternative method is to determine the 'optimal' state sequences associated with the PD observation sequences of an original speech as well as its coded version. An optimality criterion chooses the state q_t that are individually most likely by maximizing the expected number of correct individual states. The individually most likely state q_t at time t is determined by computing

$$q_t = \underset{i \in \mathcal{N}}{\text{argmax}}[P(q_t = S_i | \boldsymbol{O}, \boldsymbol{\lambda})] \tag{5.36}$$

The bracketed term, i.e., the probability of being in state S_i at time t, given the observation sequence $\boldsymbol{O}$ and the model $\boldsymbol{\lambda}$, is written for the forward-backward technique in terms of the variables $\xi_t(i,j)$ as

$$P(q_t = S_i | \boldsymbol{O}, \boldsymbol{\lambda}) = \frac{\sum_{j \in \mathcal{N}} \xi_t(i,j)}{\sum_{i \in \mathcal{N}} \alpha_T(i)}. \tag{5.37}$$

The solution simply determines the most likely state at every instant without any regard to the probability of occurrence for sequence of states. A distortion measure could be defined based on calculating the Hamming distance between the estimated state sequences for the original and the coded speech signals. There is no unique way of selecting an 'optimality' criterion and the approach may even be modified to maximize the expected number of correct paths of pairs of states (q_t, q_{t+1}) or triples of states (q_t, q_{t+1}, q_{t+2}) etc.

Model distance approach

Another alternative is to estimate a model $\boldsymbol{\lambda}^{(c)}$ from the PD observations of the coded speech frame exactly the way we have estimated the model $\boldsymbol{\lambda}^{(o)}$ from the PD observation of the original speech frame. A model distance measure following the notion of discrimination information could be defined for comparing these pairs of HMMs [100]. One such measure form is

$$D(\boldsymbol{\lambda}^{(c)}, \boldsymbol{\lambda}^{(o)}) = \sum_{n=1}^{64} \log P(\boldsymbol{O}_n | \boldsymbol{\lambda}_n^{(c)}) - \sum_{n=1}^{64} \log P(\boldsymbol{O}_n | \boldsymbol{\lambda}_n^{(o)}). \tag{5.38}$$

This measure is non-symmetric and a symmetrized version could be used in practice.

5.5 Practical Considerations

A 'good' distortion measure should consider only the information relevant to perceptual events. However, the success of the measure also becomes heavily dependent on the accuracies of the implementation and the model description. Here, we discuss some practical aspects related to the evaluation of speech coders by the CHM measure.

5.5.1 Computational Issues

The forward probability calculation is, in effect, based upon the trellis structure shown in Fig. 5.2. Since there are only two possible states at each time in the trellis, all the possible state sequences will remerge into one of these two nodes, regardless of the length of the observation sequence. At any time t, computation of $\alpha_t(j)$ involves only two previous values of $\alpha_{t-1}(i)$ because each of the two grid points is reached from the same two grid points at the previous time slot. For computing each $\alpha_t(i)$ and $\beta_t(j)$, it requires on the order of N^2T calculations, rather than $2TN^T$ as required by the direct calculation.

Another important issue is that computing the likelihood variables involves multiplication of many terms having values smaller than 1. In a recursive procedure, each term of these variables starts to diminish towards zero exponentially and thus the number representation goes below the precision range of any machine. To circumvent this problem, the likelihood and other variables are multiplied by constants known as scaling coefficients [101]. The scaling procedure is not applied at every clock time, but once every few clock times.

5.5.2 Initial Estimates for HMM Parameters

Since a convergent reestimation procedure exists for the continuous mixture model considered here, it is theoretically possible to have arbitrary initial estimates for the HMM parameters obeying the stochastic constraints. The reestimation equations provide values for the HMM parameters corresponding to a local maximum of the likelihood function. The choice of 'good' initial estimates is thus important in making the convergence faster or ensuring the local maximum to be the global maximum of

the likelihood function. In fact, some of the parameters may be very sensitive to the initial estimates [102].

5.5.3 Training Data and Iterations

The PD observation sequence used for 'training' the models has a finite length and this causes problem in determining the HMM parameters via reestimation method. An insufficient number of occurrences of different model events does not truely portray the real scenario and therefore we have to have sufficiently long training data. On the other hand, we want the model parameters to be fixed for a specific period and then vary depending on the new PD observations. The Baum-Welch estimation algorithm also needs several iterations before the convergence occurs.

5.5.4 Mixture Processes

It is an usual practice to approximate a K-dimensional correlated random process by a mixture of few uncorrelated, K-dimensional random processes. The number of mixture components is heavily dependent on the degree of correlation. By assuming mixture uncorrelated processes, we effectively reduce the number of parameters to be estimated and thus help making the estimates more reliable. The trade-off is clearly between the increased error in the approximation process and the increased reliability in the estimation process.

5.6 Experimental Results

Before providing with the objective measure results, we describe the set-up procedure for some of the experimental parameters.

(i) We have 'trained' and 'matched' the HMMs with speech frames of 480 samples. For $N = 2$ and $T = 480$, only about 1920 computations were needed since the algorithm used was based on trellis structure.

(ii) The scaling procedure was used not at every instant, but after every ten clock times.

(iii) Although the length of the PD sequences over which the training and matching were done is 480, we overlapped each such frame with the previous frame by 50%. In other words, the observation window was shifted by 240 samples for dealing with each new model. This has allowed having sufficiently long training data and also has facilitated the models not to change the parameters drastically.

(iv) In our experiment, we have chosen models with three mixture components (i.e., $M = 3$). This has appeared to be a reasonable choice for making trade-off between the accuracy of modeling the histogram and the number of parameters to be estimated.

(v) Based on the psychoacoustic data, we have assumed the initial transition probability from a non-firing state to another non-firing state is 0.8 and that from a firing state to another firing state is 0.2. In accordance with this, the initial state probabilities were chosen to be 0.8 for non-firing state (S_0) and 0.2 for firing state (S_1).

(vi) The initial estimates for the beta pdf parameters $\{d_{im}\}$ and $\{f_{im}\}$ were chosen in such a fashion that the corresponding mean values were 0.25, 0.50 and 0.75 for $i \in \mathcal{N}$. The weighting factors $\{c_{im}\}$ were all assumed to be equal (i.e., 0.33) initially.

(vii) For any particular neural channel, the final estimate of HMM parameters obtained for a speech frame was considered as the initial estimate of the parameters for the subsequent frame.

(viii) While solving the simultaneous equations of (5.33) and (5.34), the $\{d_{im}\}$ and $\{f_{im}\}$ parameters were allowed to take up integral values between 1 and 40. Since the exact solution could not be found, we have determined the parameter values by choosing the best pair which minimizes the sum of the square errors. One more constraint imposed on the parameters was that the mean values (given by $d_{im}/(d_{im} + f_{im})$) for three different mixture components have been kept confined to three different regions—one between 0 and 1/3, the second between 1/3 and 2/3; and the third between 2/3 and 1. This also reduced the search for best solution by making some combinations of the parameter values to be invalid.

(ix) For model parameter estimations, we have made 30 iterations each time.

Sent.	C1		C2		C3		C4		C5		C6	
	S	H	S	H	S	H	S	H	S	H	S	H
M1	5.75	195	4.92	225	4.17	336	2.58	358	2.58	365	1.00	420
M2	5.50	250	5.17	231	4.25	280	2.75	310	2.25	390	1.08	414
F1	5.67	209	5.00	263	4.25	300	2.33	389	2.58	371	1.17	430
F2	5.67	220	5.00	276	3.91	347	2.67	378	2.50	312	1.25	398

Table 5.1: Subjective and objective measure values for six coded signals with reference to the corresponding original speech utterances ('S' gives the average subjective ranking scores and 'H' denotes the cochlear hidden Markov measure with single channel (CHM–SC))

Sent.	C1		C2		C3		C4		C5		C6	
	S	H	S	H	S	H	S	H	S	H	S	H
M1	5.75	146	4.92	188	4.17	256	2.58	314	2.58	320	1.00	408
M2	5.50	161	5.17	179	4.25	238	2.75	287	2.25	346	1.08	398
M3	5.75	157	5.17	183	4.00	261	2.58	310	2.33	304	1.17	401
M4	5.00	196	5.67	152	4.25	230	2.50	326	2.58	311	1.00	412
M5	5.75	138	5.17	170	3.83	277	2.67	301	2.50	335	1.00	421
M6	5.58	163	5.25	186	3.83	265	2.75	292	2.42	319	1.17	392
F1	5.67	154	5.00	182	4.25	244	2.33	326	2.58	307	1.17	416
F2	5.67	159	5.00	192	3.91	270	2.67	296	2.50	310	1.25	386
F3	5.50	170	5.17	177	4.25	221	2.50	319	2.25	352	1.33	381
F4	5.41	169	5.25	174	4.17	238	2.75	281	2.17	361	1.25	399
F5	5.50	162	5.50	155	3.83	272	2.33	330	2.50	304	1.33	373
F6	5.67	156	4.83	202	4.08	263	3.08	322	2.17	348	1.17	391

Table 5.2: Subjective and objective measure values for six coded signals with reference to the corresponding original speech utterances ('S' gives the average subjective ranking scores and 'H' denotes the cochlear hidden Markov measure with three channels (CHM–TC))

In this work, we have followed two strategies for computing coder distortions. Let us now consider determining the model parameters for the n-th neural channel. In the first strategy, while training the model, we have used only the n-th channel PD observation sequence corresponding to the original speech. We call this strategy as CHM-SC (with single channel). Table 5.1 shows subjective and objective measure values for six coded signals with reference to the original speech utterance. We tabulate here measure values for only four utterances. The CHM-SC measure was found to be not very satisfactory in ranking coded signals.

It has been our understanding that the training data length was not sufficient in the CHM-SC strategy to make a reliable estimate for the model parameters. Therefore, we formulated a new strategy where three adjacent channels—the $(n-1)$-th, n-th and $(n+1)$-th channel PD sequences were used in alternate manners for training. This strategy has been termed as the CHM-TC (with three channels). Table 5.2 provides subjective and CHM-TC measure values for all the twelve utterances given in the Appendix C.

Coded Speech	Measure	Sample Delays			
		Zero	One	Two	Three
oakf8f	SNR (w/o scaling [dB])	8.724	7.391	5.619	5.117
oakf8f	CHM-TC	221	227	224	229
oakf8k	SNR (w/o scaling [dB])	9.178	7.503	6.108	7.027
oakf8k	CHM-TC	319	321	326	323

Table 5.3: The SNR and the cochlear hidden Markovian—three channels (CHM-TC) measure values with zero, one, two and three sample delays for the coded signal 'oakf8f' and 'oakf8k' with reference to the original speech sentence

For the CHM distortion measure values, we have computed the logarithm (natural) of the likelihood probability scores, negated them and averaged over all the channels and all the speech frames. Tables 5.1 and 5.2 provide these measure values where a low value implies a better perceptual quality. Already in Chapter 4, we have noted that with the utterance M1, the C4, C5 coders and with the utterance F5, the C1, C2 coders were ranked same subjectively. The CHM-TC measure has found C4 coder for M1 and C2 coder for F5 to be slightly better than their counterparts. Other

than these tie cases, the subjective and objective measures were not in conformance for the C4, C5 with the utterance M3.

The success of the CHM measure is quite comparable with that of the CDI measure. However, the primary two advantages of the CHM measure are that (i) ample provisions (selecting better initial estimates, carrying out more iterations etc.) exist for its improvement and (ii) it attempts to take time correlations into account and is fairly robust against few sample delays. Unlike most of the other distortion measures, the CHM measure performs quite well without an explicit time-alignment. Table 5.3 provides the SNR measure (without scaling) values as well as the CHM measure values with zero, one, two and three sample delays in the coded speech. The misaligned sample places are filled in with very small (approximately zero) values. It is observed that the sample delays do not affect the measure values considerably.

5.7 Summary

In this chapter, we have introduced a cochlear hidden Markovian (CHM) measure for computing coder distortion. We have attempted to capture the basics of neural firing events with simple hidden Markovian models where the occurrence of perceptual-domain observations and correlation among adjacent observations are modeled appropriately. A two-state (one each for firing and non-firing events), fully-connected HMM has been associated with each of the neural channels.

For computing coder distortions, at first, all the HMMs are 'trained' (i.e., the HMM parameters are estimated) with the PD observation derived from the original signal. The Baum-Welch reestimation technique has been applied to derive the HMM parameters iteratively starting from an initial estimate. The PD observations obtained from the coded speech are 'matched' against these HMMs. A negated version of the log likelihood probability scores, averaged over all the speech speech frames and neural channels, has acted as the CHM distortion measure. This measure has shown promise by conforming appreciably with subjective evaluation results and also by exhibiting its robustness against coder delays.

Chapter 6

Applications in Coder Analysis

6.1 Introduction

The earlier chapters have dealt with an auditory representation of speech and two distinct approaches for computing distortions by comparing these perceptual-domain parameters of a coded signal vis-a-vis its original version. We have evaluated the performances of speech coders with these measures and also have computed the rate-distortion functions for speech coding with one of them. Although we have not attempted to use our measure formulation in a closed-loop fashion in any speech coder, it may very well be possible to use it for 'populating' a codebook in the training phase and/or for 'selecting' an appropriate codebook entry in the transmission phase. A typical speech coder has several components based on the features and its encoding parameters. For a low bit-rate speech coder, a proper bit allocation among these components plays a significant role in achieving a good perceptual quality for the coded speech. Thus, it would be helpful for the designer if there could be a way to assess the performances of these components in a separate manner.

State-of-the-art analysis-by-synthesis medium or low bit-rate speech coders comprise of a linear prediction filter to model the short-term spectrum, a pitch predictor to model the long-term periodicity and a stochastic codebook to represent the residual speech signal. In practice, some of these filter blocks and codebooks are often split into more than one components primarily to give different perceptual importance and also for computational reason. While transmitting, in the analysis phase,

different synthesized signals are compared with the original signal by a fidelity criterion. A mean-square distortion criterion has been found to be unsatisfactory as it does not even address perhaps the most important perceptual event, namely the auditory masking. To address this issue, in the recent literature, various noise weighting schemes have been coupled with the mean-square distortion criterion and/or other noise shaping filters have been suggested.

A detailed analysis on a component-by-component basis for different coders with the same bit-rate is beyond the scope of this thesis. Nonetheless, in this chapter, we will outline two related applications of the proposed measures. The first part describes an analysis procedure for determining the the pitch frequency by examining the output space of the cochlear model and applying the CDI measure. With the help of this analysis, it is possible to compare the pitch information of the original signal and that retained in a coded version. This, in turn, could provide a feedback to the designer regarding the deficiency of the pitch filter component. In the second part, we consider a wideband speech coder which uses three-way split vector quantization for the LPC parameters and fractional pitch lag value in the pitch predictor. We apply the CDI as well as the CHM measures for studying the performances of different noise weighting methods as incorporated in this coder. The coder was designed by K. Abboud [9] and the evaluation of the noise weighting schemes was carried out jointly by this author and Abboud.

The remainder of this chapter is formatted as follows. Section 6.2 briefly reviews some of the existing pitch frequency estimation algorithms. Section 6.3 suggests an algorithmic approach, using the CDI measure form, for the pitch frequency determination from the PD representation of a speech signal. Section 6.4 describes a 11.2 kbps code-excited linear prediction (CELP)-based wideband speech coder. Section 6.5 introduces some of the perceptual weighting schemes while Section 6.6 investigates their performances by the proposed objective measures. Thus, Sections 6.2 and 6.3 are related to the first application of pitch frequency estimation whereas Sections 6.4, 6.5 and 6.6 pertain to the second application of performance evaluation of perceptual weighting schemes implemented in a wideband coder.

6.2 Existing Pitch Estimation Algorithms

One simple time-domain approach for the pitch estimation is to low-pass filter all the energy from the speech signal except the fundamental harmonic and then detect the zero-crossing rate (ZCR). The ZCR measure is related to the pitch (F0) by

$$\text{F0} = \text{ZCR} * f_s/2, \tag{6.1}$$

where f_s is the sampling frequency in samples/sec. The prime difficulty of this method is in the determination of the cut-off frequency for the low-pass filter as it should be high enough to include the fundamental frequency from a high-pitched voice and low enough to exclude the first harmonic of a low-pitched voice. Moreover, the ZCR detects some time the first formant frequency rather than the F0 if the former has sufficiently high energy. A windowed autocorrelation function (ACF) is often calculated by taking the product of the speech sample $\{s[n]\}$ with its delayed version and passing it through a window filter $\{w[n]\}$ [103]. In the pitch determination, the ACF $R_n[k]$ given by

$$R_n[k] = \sum_{m=-\infty}^{\infty} s[m]w[n-m]s[m-k]w[n-m+k], \tag{6.2}$$

is evaluated for k ranging from the shortest possible period (e.g., 3 ms for a female voice) to the longest one (e.g., 20 ms for a male voice). Another alternative technique is to calculate the average magnitude difference function (AMDF) defined as

$$\text{AMDF}[k] = \sum_{m=-\infty}^{\infty} |s[m] - s[m-k]| \tag{6.3}$$

which shows minimum for a k value corresponding to the pitch period.

Frequency-domain techniques involve computing a windowed Fourier transform defined as

$$S_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} s[m]e^{-j\omega m}w[n-m]. \tag{6.4}$$

In a spectrogram, $S_n(e^{j\omega})$ is plotted with the sample-time n on the horizontal axis, the frequency ω on the vertical axis and the magnitude by darkness of the display. The pitch period can be detected by searching either the periodically-spaced vertical lines in a wideband spectrogram or the periodically-spaced horizontal lines in a narrowband spectrogram. For estimating the F0, a cepstral analysis technique has also been

employed where the complex cepstrum exhibits sharp pulses spaced at intervals typical of the pitch periods.

Unlike the conventional techniques, an auditory model-based pitch determination technique works for varying pitch effects, and is robust against a wide range of distortions [104]. Based on the *duplex theory* of the pitch perception, Lyon has published an 'auditory correlogram' [105]. Using this idea, Slaney et al. have recently proposed a perceptual pitch detector [104]. In this algorithm, a pre-processing step emphasizes the vertical structure in the correlogram, sums the value at each time-delay in the enhanced correlogram across all the frequencies and determines the pitch based on the location of the largest peak. Weintraub has used a cost-reduced correlogram version as a pitch tracker for his two voice sound separation experiments [106]. Seneff has used an auditory model and suggested a generalized synchrony detection (GSD) mechanism for detecting the pitch periodicities in the speech signal [107].

6.3 Pitch Frequency Estimation

Here we suggest an algorithm, using the CDI measure form, for estimating the pitch fundamental frequency.

A rectangular analysis window of 40 ms is chosen for a speech frame of 20 ms so that the successive windows overlap by 50% and at least two pitch periods are included in the analysis window. The output for each of the sixty-four neural channels is compared with itself delayed by τ samples (τ up to 20 ms =160 samples accounting for the lowest possible F0). With two sets $\mathcal{I} \equiv \{1, 2, \cdots, 160\}$ and $\mathcal{K} \equiv \{1, 2, \cdots, 64\}$, the comparison with the discrimination measure (e.g., with the $D_1(P;Q)$) takes the form

$$E_k(\tau) = \sum_{l \in \mathcal{I}} \sum_{j=1}^{2} p_{jkl} \log \left(\frac{p_{jkl}}{p_{jk(l+\tau)}} \right), \quad \text{for} \quad \tau \in \mathcal{I}, \quad k \in \mathcal{K}. \tag{6.5}$$

In this way, from a two-dimensional time-place representation $\mathcal{A}$, we derive a *cross-entropogram* $\mathcal{E} = \{E_k(\tau) | k \in \mathcal{K}, \tau \in \mathcal{I}\}$ which is also two-dimensional where the vertical direction corresponds to the channel $k \in \mathcal{K}$ and the horizontal direction corresponds to the sample delay $\tau \in \mathcal{I}$. To enhance the vertical structure of $\mathcal{E}$, a convolutional operator $O = [-1 \ +2 \ -1]$ is used to give $\mathcal{G} = \{G_k(\tau) | k \in \mathcal{K}, \tau \in \mathcal{I}\}$

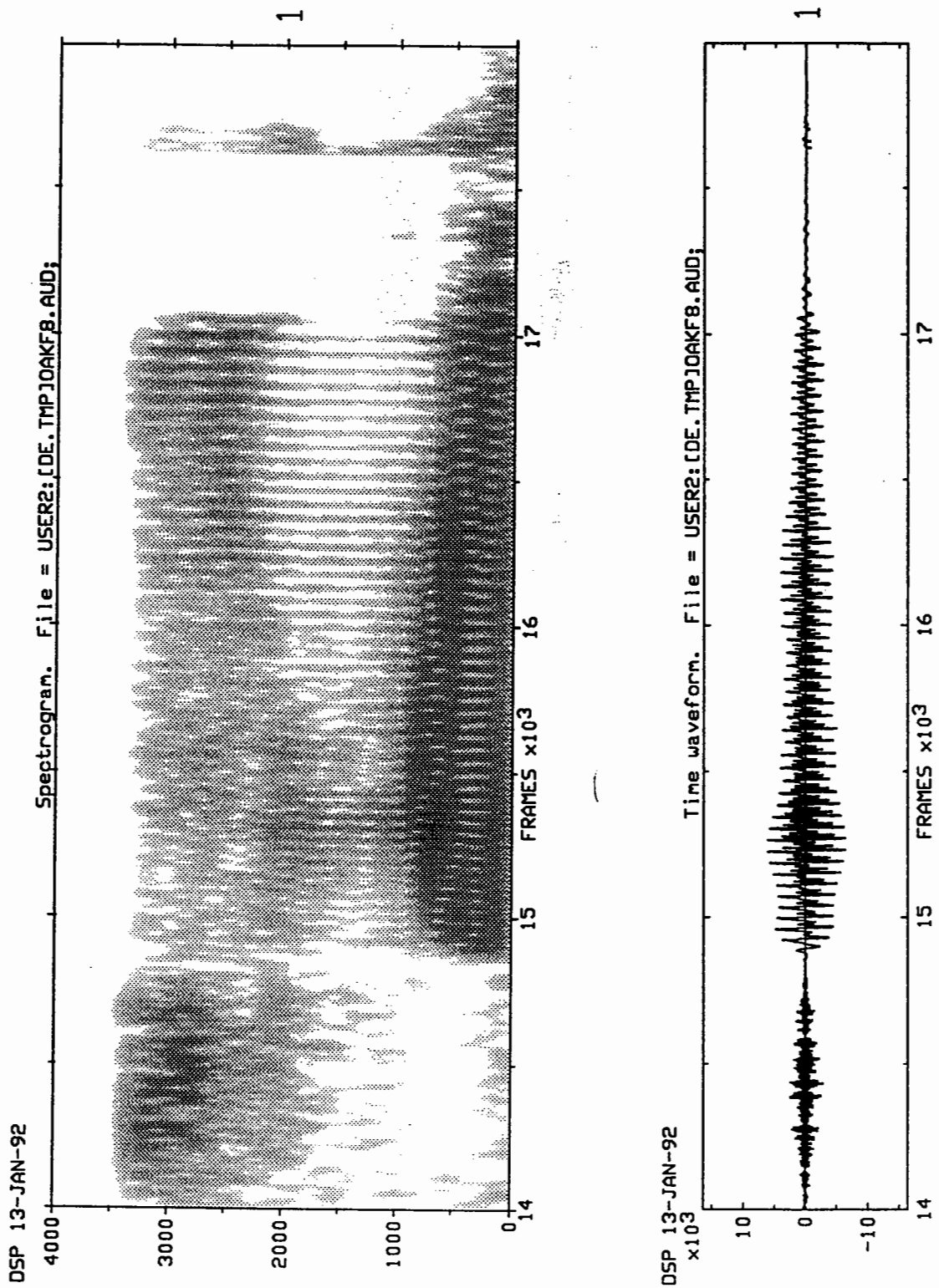


Figure 6.1: Time-domain waveform and spectrogram plot of the vowel /a/ in the word 'shade' (female voice)

with

$$G_k(\tau) = -E_k(\tau - 1) + 2E_k(\tau) - E_k(\tau + 1), \quad \text{for } \tau \in \mathcal{I}, \quad k \in \mathcal{K}, \tag{6.6}$$

where $E_k(0) = E_k(161) = 0$, $\forall k \in \mathcal{K}$. An exponential weighting set $\{w[k]|k \in \mathcal{K}\}$ of the form

$$w[k] = e^{-a(N-k)/N}, \quad \text{for } k \in \mathcal{K} \tag{6.7}$$

is defined where the number of channels N is sixty-four and the decay factor a is chosen to be 6. The exponentially-weighted discrimination measure values for all the sample delays, summed over all the sixty-four channels, can then be written as

$$G(\tau) = \sum_{k \in \mathcal{K}} w[k] G_k(\tau), \quad \text{for } \tau \in \mathcal{I}. \tag{6.8}$$

The evidences from the higher harmonics are combined this way to make the pitch estimate robust. At the same time, the contributions from the formant frequencies are minimized by giving exponentially decaying weights to the higher frequency channels. In this flattened one-dimensional cross-entropogram $\mathcal{M} = \{G(\tau)|\tau \in \mathcal{I}\}$, the measure $G(\tau)$ shows the first significant dip at a τ value corresponding to the pitch period. Thus, an average F0 for a frame is calculated as

$$\text{F0} = f_s * \left[\min_{\tau} \{\tau \in \mathcal{I} : G(\tau) < H\}\right]^{-1} \tag{6.9}$$

with $f_s = 8,000$ Hz and an appropriate threshold H.

Fig. 6.1 shows the time-domain waveform and spectrogram plot for the word 'shade' (female voice). We execute our pitch estimation algorithm to determine the pitch period for one frame (160 samples starting from the sample number 15,000) of /a/ in that word. In the one-dimensional cross-entropogram plot (using the directed divergence measure with $\alpha = 1$) of Fig. 6.2, the first dip in $G(\tau)$ is observed at τ =40 samples (equivalently, 5 ms) with an H value, say, -30 units. The perceptual pitch period is thus calculated to be 200 Hz. It is expected that this scheme, due to merging of information from all the channels, could estimate the pitch correctly even when the 'fundamental frequency' component is filtered out from the original signal. However, the selection of H may become more stringent. A post-processor may be used for the pitch estimates of successive frames to correct any serious error, e.g., pitch doubling or halving.

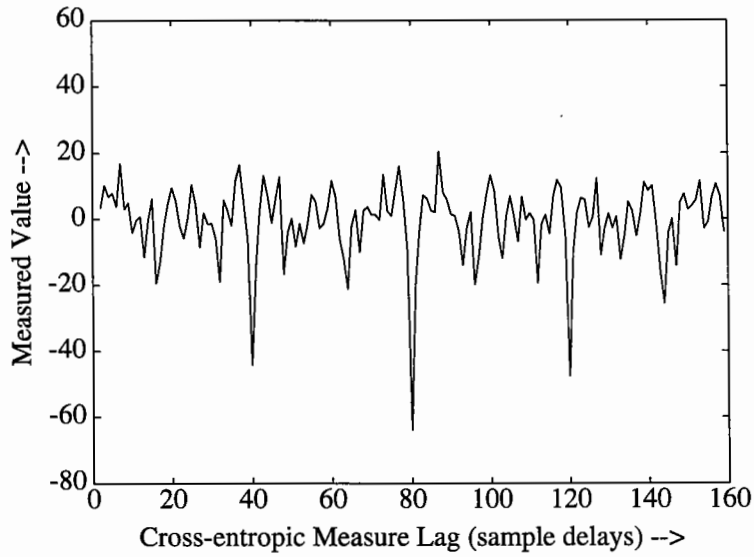


Figure 6.2: One-dimensional cross-entropogram (directed divergence with α=1) for one particular frame (160 samples starting from the sample number 15,000) of /a/ in the word 'shade'

6.4 Wideband Coder Architecture

In the second application, our intention is to examine the performances of some of the noise shaping schemes generally incorporated in a medium or low bit-rate speech coder. Here, we consider a 11.2 kbps CELP-based wideband speech coder. A wideband (50–7,000 Hz) speech shows superiority in the perceived quality over a narrowband (200–3,400 Hz) speech. This stems from the fact the added low frequencies increase the naturalness of a voice while the added high frequencies make the speech sound more intelligible, especially for fricative sounds. Obviously, more bits are required to code the additional information which leads us to have a trade-off between preserving acceptable speech quality of the reconstructed signal and maintaining a relatively low operating bit-rate. However, here our objective is not efficient coding; but assessing the effectiveness of the perceptual weighting schemes.

The CELP analysis-by-synthesis speech coders treat the input speech samples on a frame-by-frame basis. Linear prediction operations are used to exploit the fact that the speech exhibits a high degree of intersample correlations—correlation observed between adjacent samples (near-sample redundancy) and also, for voiced speech, cor-

relation between samples separated by the pitch period (far-sample redundancy). Generally, a CELP coding technique consists of three basic functions— (i) short-term prediction (in the analysis phase) to determine the LPC (or equivalent) coefficients, (ii) pitch search (in the synthesis phase) to calculate the pitch lag (i.e., the pitch period in samples) and the pitch coefficient (i.e., the corresponding gain parameter) values and (iii) codebook search (in the synthesis phase) to determine the index of an excitation waveform and the associated scale factor.

Accordingly, two codebook indices and two quantized gain values are determined along with the formant predictor coefficients. In fact, the CELP coder does not directly need an analysis stage; ideally, the formant synthesis filter could also be optimized for each candidate excitation waveform. However, the formulation of an optimal (in a mean-square sense) formant synthesis filter leads to a highly nonlinear set of equations which is not amenable to solution. Thus, a formant synthesis filter is generally implemented as the inverse of a formant filter determined in the analysis step. These parameters are selected in a systematic way for matching a synthesized speech to the original one with minimum error as defined by a distortion measure. All of them are updated at regular intervals and transmitted over a communication channel in order to reconstruct the speech signal in the decoder side. A wide variety of wideband CELP-based algorithms has been proposed in the literature (e.g., [89]). In this work, we use a 11.2 kbps wideband speech coder as designed by Abboud [9] and described below briefly.

6.4.1 LSF-based Short-term Prediction

For this wideband coder, a 16-th order LPC filter is chosen for the short-term prediction. The LPC parameters are determined by an autocorrelation method in which each frame of speech samples is multiplied by a Hamming window before getting filtered by the inverse formant filter. This method involves minimizing the residual signal (of the filtered version) energy and requires solving a set of Yule-Walker equations. The LPC parameters obtained are not well-suited for direct transmission because an error in any one parameter can cause the filter to become unstable and their wide dynamic range may make an efficient quantization practically impossible. Thus, they are transformed into a 'better-behaved' set of parameters such that the synthesis filter characteristics vary smoothly as a function of those parameters.

The line spectral frequency (LSF) parameters represent the phase angle of an ordered set of poles on the unit circle. They are used quite often because they simplify the quantization procedure, ensure synthesis filter stability (if LSFs are ordered) and are closely related to the formant frequencies. An use of the LSF parameters has an added advantage of localized spectral sensitivity, i.e., that an error in one LSF only affects the synthesized spectrum near that frequency. An efficient technique for the computation of the LSFs is followed. The polynomial roots are determined by applying a Chebyshev transformation so as to map the upper semi-circle in the z-plane to the [–1,+1] range and searching for sign changes in this interval [108].

These LSF parameters are quantized before transmission. In scalar quantization, each LSF coefficients are quantized individually while in the vector quantization (VQ), all the LSF coefficients are quantized together. For the same performance, the first one demands a high number of bits for quantization while the second one suffers from high complexity in terms of the amount of training data needed, the memory required and the number of computations involved. Here, a three-way split VQ technique for the LSF parameters is adopted. For each frame of 15.625 ms, 30 bits are distributed among the 16 LSFs. They are divided into three subgroups—13 bits (for the first 8 LSFs), 9 bits (for the middle 4 LSFs) and 8 bits (for the last 4 LSFs). A training data of LSF vectors is used to construct three different optimal (at least in the local sense) codebook sets using the Linde-Buzo-Gray (LBG) algorithm [109].

During the transmission phase, an unquantized LSF vector is compared with the codebook entries for the LSF vectors. The algorithm chooses that codebook vector which minimizes a weighted-Euclidean LSF distance measure where the weighting factor considers the frequency sensitivity and also the LSF positions. A nested search technique is followed with priority given to the first LSF subgroup where most of the perceptual information prevails. The optimal first vector is combined with the second LSF codebook entries to yield the second LSF vector; and finally, the optimal first and second vectors are combined with the third LSF codebook entries to generate the overall LSF vector. With a transmission of 30 bits per frame and an update rate of 64 Hz, the operating rate for the short-term predictor is 1,920 bits/sec.

6.4.2 Long-term Prediction with Fractional Delays

The long-term linear predictor parameters are the pitch coefficient and the pitch lag. The pitch coefficient is a scaling factor related to the degree of waveform periodicity. It is zero for a signal without a periodic structure and is approximately one for steady-state voiced speech. The pitch lag tends to vary smoothly in the voiced segments with only occasional departure from the smooth trajectory [110]. However, in the unvoiced segments, the pitch lag tends to jump around. To avoid the problem of locking onto the correct pitch during the transition from silence to voiced speech, a good pitch delay resolution should be maintained at all times during the analysis and synthesis stages of the CELP coder.

The pitch coefficient values may be positive or negative. In general, the negative values tend to occur in the speech regions with low energy and the large positive values tend to occur in the transition regions (silence to speech). A restriction on the pitch period to be integer multiples of the sampling interval results in the partial destruction of the harmonic structure, especially in the high frequency regions. To increase the temporal resolution, non-integer (fractional) pitch lag values can be used. For this purpose, a multitap or a pseudo-multitap pitch prediction filter [111] can also be used. However, in this coder, a single-tap pitch predictor with fractional lag values is adopted. This is implemented by the use of interpolation and polyphase filters [112].

The pitch lag in wideband speech ranges from 40 to 320 samples with some delays occuring more often than others. With the use of 38,400 pitch subframes of 3.125 ms (five subframes in a formant frame of 15.625 ms) each, a pitch delay distribution is generated. Based on this, a nonuniform distribution of non-integer delays is set up; the highest resolution is given to the pitch lags in the range of 71 to 100 while the lowest resolution is given to the end of the lag range. For each subframe, the pitch gain is represented by 4 bits whereas the pitch lag by 10 bits. Thus, with an update rate of 320 Hz, the pitch parameters altogether require 4,480 bits/sec.

6.4.3 Residual Signal Codebook

The residual signal codebook is filled up with codevectors containing sparse ternary elements. These excitations are generated by center-clipping and using a zero-mean unit-variance Gaussian sequence. The center-clipping threshold is set to ± 1.2 in order

to satisfy a specific percent (75%) of sparsity inside the vector. All values between +1.2 and −1.2 in the Gaussian distribution are set to 0, values greater than +1.2 are set to +1, and values less than −1.2 are set to −1. The number of codewords is set to 1,024 requiring 10 bits for each subframe of 3.125 ms. A 4 bit differential quantizer with a leaky predictor is used to code the differences in successive subframe magnitudes and an extra bit to code the sign. Thus, 4,800 bits/sec are required for residual signal representation with an update rate of 320 Hz.

6.5 Perceptual Noise Weighting

At a low bit-rate, the mean-squared error criterion between the original and the reconstructed speech has been found to provide an unsatisfactory result. This indicates a requirement to shape the noise based on the auditory masking phenomenon in speech perception. In noise spectral shaping, the noise components at certain frequencies can only be diminished at the price of increased noise components at some other frequencies. At low bit-rates where the average noise level is quite high, it is difficult, if not impossible, to maintain noise below the masking threshold at all frequencies. The noise components in spectral valleys may exceed the threshold; nonetheless, they can be attenuated substantially by a postfilter used at the last stage of the decoding process. On the other hand, the postfiltering operation introduces distortion in the speech signal to some degree [113]. In this coder, the filter parameters and the codebook entries are selected by minimizing a noise-weighted mean square error and/or using noise shaping filters. This section discusses three perceptually-weighted filtering schemes—(i) a simple noise weighting filter, (ii) a codebook shaping filter and (iii) an enhanced noise weighting filter. Although no postfilter is used here, it could be used with any of the above schemes.

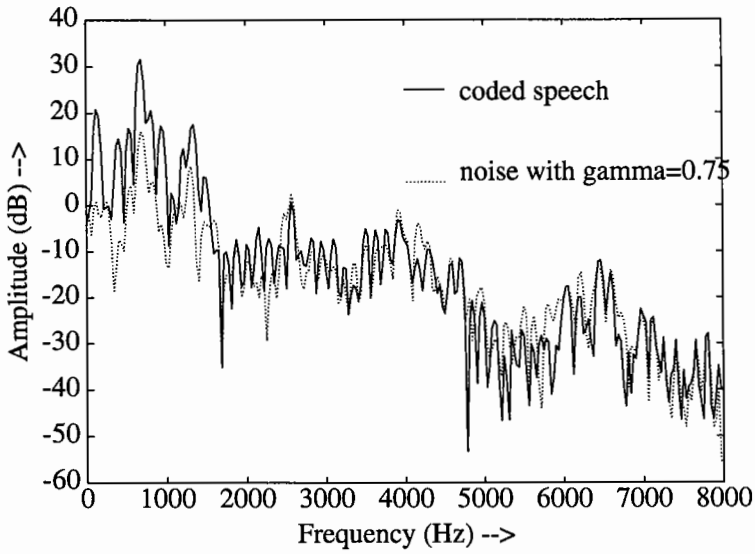


Figure 6.3: Noise weighting with $\gamma = 0.75$.

6.5.1 Simple Noise Weighting

This auditory masking scheme is accomplished by minimizing a weighted mean-square error with a noise shaping filter $W(z)$ defined as

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}}, \tag{6.10}$$

where $A(z)$ is the formant analysis filter with the predictor order p. The value of γ $(0 < \gamma < 1)$ is determined by the degree desired to de-emphasize the formant regions in the error spectrum. Decreasing the value of γ moves the poles of the filter $1/A(z/\gamma)$ inward and therefore increases the bandwidth of the poles of $W(z)$. A good value of γ is 0.75 [9]. Fig. 6.3 shows the effect of using such a noise weighting filter in the reconstruction process of the input speech. The resulting noise level is no longer flat, but has the spectral shape of $W^{-1}(z)$ and therefore is boosted in the formant peaks and attenuated in the formant valleys.

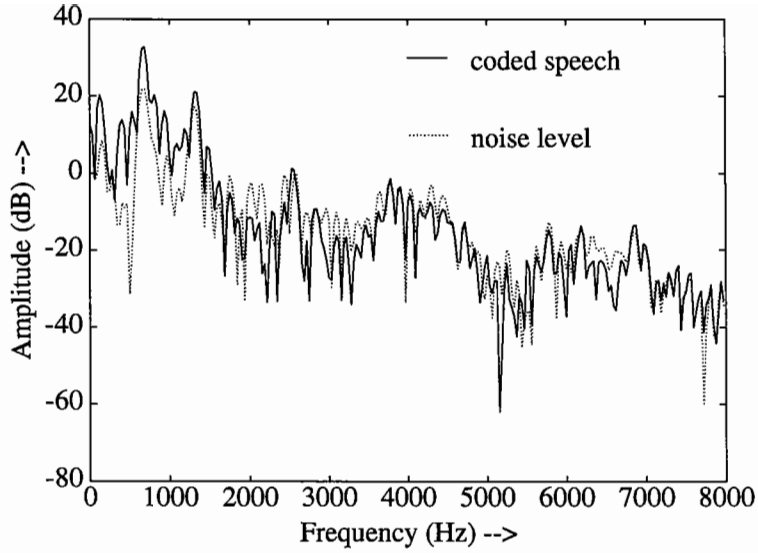


Figure 6.4: Noise level using codebook shaping filter.

6.5.2 Codebook Shaping Filter

In the second approach, an excitation codebook is cascaded with a shaping filter $F(z)$ to form a modified codebook structure. The filter $F(z)$, defined as [114]

$$F(z) = (1 - \mu z^{-1}) \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \tag{6.11}$$

is dynamically changed to control the statistical properties of the codebook in time and in frequency. The parameters γ_1 and γ_2 are de-emphasizing constants while μ is a parameter compensating for the spectral tilt. A differencer uses $\mu = 1$, but an optimum preemphasis filter which maximizes the output spectral flatness measure has $\mu = \frac{r(1)}{r(0)}$, where $r(n)$ represents the autocorrelation sequence for the excitation signal before shaping. For unvoiced sounds, this fraction is relatively small and the effect of the preemphasis filter becomes negligible. On the other hand, for voiced sounds where $r(1)$ is very close to $r(0)$, the preemphasis filter acts almost as a differencer.

6.5.3 Enhanced Noise Weighting

The prime disadvantage of a simple noise weighting filter $W(z)$ is inadequate balancing of low and high frequency components due to interdependency of the tilt and

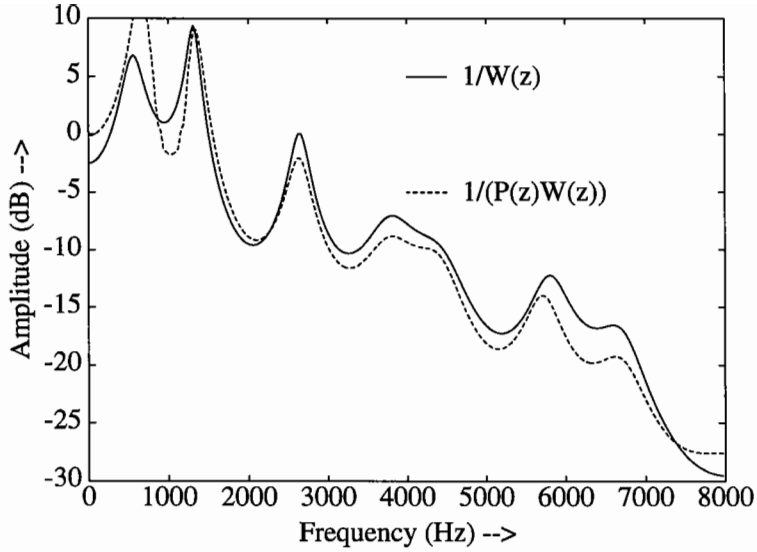


Figure 6.5: Performance of the weighting filter with $N = 2$ and $\delta = 0.7$.

formant parameters. An accurate modeling of one requires a sacrifice in modeling the other. An enhanced noise weighting technique [115] introduces a decoupling factor that results in an independent control of the tilt with respect to the formants. In general, the corresponding weighting filter is implemented as

$$W'_N(z) = W(z)P_N(z) = \frac{A(z)}{A(z/\gamma)} \frac{1}{1 + \sum_{k=1}^{N} p_k \delta^k z^{-k}}, \tag{6.12}$$

where the coefficients p_k are determined by an LPC analysis on the first $(N + 1)$ correlation coefficients of the inverse filter $A(z)$ and the parameter δ controls the spectral tilt.

6.6 Performance Evaluation

We kept the wideband coder architecture in tact and applied different noise weighting schemes. In this second application, the primary contribution of this author, was in the performance evaluation of these schemes by the introduced objective measures [116]. The following seven configurations were used for the evaluation purpose.

1. no weighting,

2. simple noise weighting with $W(z)$,

3. simple noise weighting with $W(z)$ and also codebook shaping with $F(z)$,

4. enhanced noise weighting with $W_3'(z)$,

5. enhanced noise weighting with $W_2'(z)$,

6. enhanced noise weighting with $W_3'(z)$ and also codebook shaping with $F(z)$ and

7. enhanced noise weighting with $W_2'(z)$ and codebook shaping with $F(z)$.

For the filter $W(z)$, the parameter γ was chosen to be 0.75. The filter $F(z)$ used parameters $\gamma_1 = 0.80$, $\gamma_2 = 0.95$ and an optimum μ. The parameter δ for $W_3'(z)$ was taken to be 0.5 whereas that for $W_2'(z)$ was considered to be 0.7.

Config.	SNR_{seg} (dB)	CDI	CHM-TC
1	16.06	2.731	362
2	12.89	2.684	284
3	11.00	2.649	247
4	11.90	2.543	183
5	11.59	2.518	166
6	9.69	2.606	212
7	9.95	2.623	225

Table 6.1: Distortion measures for different noise weighting configurations (the segmental SNR values (SNR_{seg}), the cochlear discrimination information measure values with $\alpha = 1$ (CDI) and the cochlear hidden Markovian measure with three channels (CHM-TC) are tabulated)

For this part of the work with wideband speech, the cochlear model was extended to have eighty-five neural channel outputs. The transfer functions corresponding to all the filters are provided in [61]. It is worth mentioning that although sixty-four neural channels were needed to cover from 0 to 4,000 Hz, only an additional twenty-one channels were sufficient for covering the 4,000 to 8,000 Hz band. This is attributed to the fact that the center frequencies of the stages corresponding to these neural channels are logarithmically and not linearly placed.

By evaluating the coded signal quality with the CDI and CHM measures, the effectiveness of the perceptual weighting schemes described in the previous section were assessed. The results with the SNR_{seg}, the CDI (with $\alpha = 1$) and the CHM measures are presented in Table 6.1. These experimental results indicate the following points.

(a) As has been discussed earlier, the SNR_{seg} measure does not perform well as an objective criterion. Ironically, we observe here that the SNR_{seg} with the first configuration (no noise weighting filter) is highest whereas the perceptual quality measured subjectively as well as by the CDI and CHM measures is lowest among the others.

(b) The use of a codebook shaping filter $F(z)$ in combination with a simple noise weighting filter $W(z)$ yields better results in the perceptual domain than the configuration (1) with no shaping filter. This is due to appropriate shaping of the excitation codevectors.

(c) The two-pole weighting filter $W_2'(z)$ exhibits better performance than the three-pole weighting filter $W_3'(z)$. The three-pole filter effectively boosts the high frequencies near the half sampling rate due to the presence of a real pole. However, this is achieved at the cost of a broadband increase in the level of distortion at lower frequencies. By getting rid of the real pole, the two-pole filter has been able to attain low level of distortion at lower frequencies while maintaining an acceptable level of high frequency noise.

(d) Among the configurations tested, the enhanced noise weighting scheme with two-pole filter (i.e., the choice (5)) appears to be the best one. A pausible explanation for such a good performance of this configuration is that the weighting filter controls the tilt as well as the formant parameters efficiently.

(e) We note that an integration of codebook shaping filter along with the enhanced noise weighting scheme (two or three poles), in fact, degrades the perceptual quality of the synthesized speech compared to the one generated by using only the enhanced noise weighting technique.

(f) In Table 6.1, we observe that the CDI and CHM measures are consistent with each other in ordering the coded signals.

(g) For this part of the work, we did not conduct an informal listening test with all the twelve listeners; instead, only two of us (i.e., myself and Abboud) made a cursory subjective assessment. Our assessments agreed with the objective measure results. Despite the agreement made by the concerned objective measures about the relative superiority of the configuration (6) over (7), subjectively we assessed them to be of equal quality.

6.7 Summary

In this chapter, we have sketched applications of the proposed distortion measures in the analysis of speech coder components. Using the CDI measure form, the output space of the cochlear model was examined to estimate the pitch frequency. We carried out, although not presented here, some preliminary work for formant estimation [87] similar to the work presented in [107]. In another application, several noise weighting schemes were used in a wideband speech coder. As a coordinated work with K. Abboud, the perceptual impacts of these techniques were studied with the CDI and CHM measures. We believe that this type of analysis could help the designer to study any particular section of the speech coder, adopt a new strategy and/or redistribute the available bits in a more efficient manner.

Chapter 7

Concluding Remarks

In this dissertation, we have proposed two auditory distortion measures and investigated their performance in speech coder analysis and evaluation. Section 7.1 summarizes the key points of our work while Section 7.2 provides a future research direction related to our work.

7.1 Summary of Our Work

In Chapter 1, we have given a brief overview of speech coding techniques. The coding algorithms vary in the selection of features, in the parametric representation of features, in the quantization of parameters and in the computation of distortion. We have explained the importance of deriving an objective quality measure for speech coding. A 'good' measure could be used in the evaluation of speech coder performance, in the computation of rate-distortion function, in the analysis of speech coder components and also in the design of speech coder. In this dissertation, our purpose has been to introduce and investigate auditory distortion measures for coded speech.

In Chapter 2, we have reviewed some of the subjective and objective quality measures used in the speech coding area. Among the existing subjective measures, the MOS and DAM scores are more popular than the others. The time-domain objective measures such as the SNR and the segmental SNR are used widely for their simplicity even though they do not correlate well with subjective measures. We have studied numerous parametric distortion measures (e.g., the log likelihood ratio, the

cepstral distance) based on all-pole speech synthesis models. We have also discussed two recently proposed psychoacoustically-motivated objective measures, namely the information index and the Bark spectral distance.

In this work, we have introduced two types of perceptual distortion measures for the purpose of speech coder evaluation. Towards this end, we have represented the speech signal onto a perceptual-domain using an auditory model. In Chapter 3, we have described the mechanism of auditory system and also analyzed some of the important psychoacoustic observations related to the speech perception. Among various functional auditory models, we have chosen Lyon's cochlear model. The outer-and-middle ear filter is modeled by a simple high-pass filter. The band-pass characteristics of the basilar membrane in the inner ear (cochlea) are simulated by sixty-four combinations of second-order notch filters and resonators. The activities of the inner hair cells are mimicked by half-wave rectification process while those of the outer hair cells are imitated by the automatic gain control stages. Unlike many other models, Lyon's auditory model considers the temporal as well as spectral masking effects. The final representation of the cochlear model output is the probability-of-firing information in the neural channels at the clock times.

In Chapter 4, we have introduced and studied a distortion measure, namely the cochlear discrimination information (CDI) measure, which compares the neural-firing information corresponding to an original speech and its coded version in a cross-entropic sense. An insufficient knowledge about the exact neural firing processes has prompted us to use the probabilistic information of firing/non-firing in the comparison. We have investigated several variants of the CDI measure based on different types of entropy, the associated parameters and also the cross-entropic measure form. The effects of gain changes and sample delays etc. have also been studied. The directed divergence measure form based on the Rényi-Shannon entropy has shown very good performance by conforming strongly with informal subjective test in terms of ranking coded speech from six different coders. Subsequently, a rate-distortion analysis for speech coder has also been carried out with this measure. We have evaluated the rate-distortion function directly using the Blahut algorithm and also determined performances of four speech coders. We have observed that there is ample scope for improving the coded speech quality at a specified bit-rate.

We have suggested another approach towards formulating a perceptual distortion measure in Chapter 5. This method has used hidden Markovian model in an

effort to capture the basic firing/non-firing process operative in the brain. We have considered two-state fully-connected model of order one for each neural channel; the two states of the model are corresponding to the firing and non-firing events. These models have been assumed to be stationary over a fixed duration (in our work, 480 sample times). The model parameters have been determined based on the PD observations corresponding to the original signal. The Baum-Welch optimization technique has been applied for the parameter estimation. Finally, the PD representations of the coded speech have been passed through the respective models so as to calculate the corresponding likelihood probabilities. The logarithms of these probability scores have been added and negated to give the cochlear hidden Markovian (CHM) distortion measure. This measure has shown promise by agreeing with subjective evaluation results to a large extent and also by demonstrating its robustness against sample delays.

Chapter 6 has outlined some of the possible applications of these measures in the analysis of speech coder components. The present-day analysis-by-synthesis medium or low bit-rate coders use several filters and codebooks. Keeping all but one component intact and having various configurations for the specific component under test, several coded versions could be synthesized for a speech utterance. As a first application, an algorithm for pitch frequency estimation has been suggested. This algorithm has involved examining the output space of the cochlear model with the CDI measure form and integrating information across channels. As a second application, different noise weighting schemes have been included in a wideband speech coder and their effect on performance has been evaluated by the CDI and CHM measures. An enhanced noise weighting scheme which controls the tilt as well as the formant parameters efficiently shows the best performance among the configurations.

While converting the time-domain speech signal into its corresponding PD representation by an auditory model, the resonating nature of the cochlea, the perceptual nonlinearity as well as the temporal and spectral masking effects have been considered. An inclusion of the spectral masking feature has allowed the probability-of-firing information in a particular neural channel at a specific clock time to depend not only on the strength of the gain-controlled signal of that channel but also on those of the other channels. Similarly, the same probability-of-firing information depends not only on the strength of the gain-controlled signal at that clock time but also on those at the other times. Thus, the PD representation for speech signal has exploited reasonably

the interdependencies at the auditory periphery level.

In the CDI measure, we have compared element-by-element of the cochleagram matrices (whose elements are the probability-of-firing information) for the original and the coded speech signals. However, this measure has been found to be not very robust against the coder delays. Thus, estimating and removing time-delay between the original and the coded speech are, in some sense, necessary first steps in applying the CDI measure. The CHM measure which has considered the temporal ordering in the firing pattern has shown a greater robustness against the coder delays. An explicit removal of the coder delays is not a necessity when the delays are confined to just a few samples. We believe that, even if the original and the coded speech signals are properly aligned, the CHM measure methodology is more powerful in the sense that it utilizes the contextual information present in the neural firing patterns.

7.2 Future Research Directions

In this section, we provide a future research direction by outlining some of the issues involved to improve this work.

7.2.1 Improvement of Model Structure

Lyon's auditory model which we have used in our work is, no doubt, a simplification of the complex behavior of the cochlea. The main simplification is in separating the interacting behaviors of the basilar membrane and the organ of Corti into non-interacting models—simple time-invariant filtering followed by a detection nonlinearity and an automatic gain control mechanism. A further refinement (e.g., [117]) of the model structure may improve the performance of the distortion measure. Some of the aspects for refinement are—fine-tuning the model parameters (e.g., Q_{ear}, f_{eb}), dynamically adjusting the Q_{ear} value, making the model structure to be two- or three-dimensional, incorporating the binaural feature etc. Many of these aspects may be important for other reasons such as localization of sound source etc. and thus may not contribute significantly in the distortion measure for coded speech. If we want to create a more biology-like condition by having a large number of neurons, it may become necessary to use a massively-parallel computer architecture based on 'connec-

tionist' model or a neural network architecture based on Kohonen's self-organizing feature map [118].

7.2.2 Reduction of Computational Complexity

It takes approximately seventy times (run on a SUN-SLC workstation) the real time system to provide the perceptual-domain representation for a speech signal. However, most of the signal processing tasks (except the coupling stages) may be performed in parallel to make the operation real-time. Advances in VLSI and signal processing technology have resulted in the fabrication of an application-specific integrated circuit (ASIC) for cochlea [119]. An application of such an ASIC would make the processing very fast compared to the software simulation.

7.2.3 Administration of Formal Subjective Test

An objective measure is considered to be useful if its result comports with the result of a formal subjective test (generally, the MOS). A regression analysis is usually performed to determine an analytic relationship between the objective measures and the MOS scores. Since different coded signals with accompanying MOS scores were not available in our academic environment, we had to rely on the results of informal test with twelve listeners. As a consequence of this, we have not carried out any regression analysis because finding a relationship between our objective measure and any such informal listening test result would only be misleading. With a limited time-duration for doctoral work, we had to make a choice between the two—(i) confining to the CDI measure approach and pursuing a more rigorous testing, or (ii) along with the CDI measure, addressing the issues of temporal ordering in the firing pattern and robustness of the measure against coder delays. The second option appealed to us. Although our experimental results show enough promise, correlation with a formal subjective test result is needed to validate our approach.

In a speech coder standardization process, the perceptual qualities of several coders are evlauated by subjective testing. Often, the coders are assessed under different test conditions. For example, the Telecommunications Industries Association (TIA) is currently setting up a 6.5 kbps speech coder standard (half-rate North American standard) for mobile communication purposes [120]. From a large pool of

candidate coders, they have selected nine of them. All these coders are being tested subjectively under fourteen different conditions (e.g., channel conditions, background noise, tandemming). Such a testing procedure involves a great deal of money and also consumes a large amount of time. We believe that our measures could, at least, be used in bringing down the number of candidate coders to a few for final subjective assessment. This would substantially reduce the amount of time and money involved for testing.

7.2.4 Derivation of Firing Pattern

For continuous speech, its perception depends not only on the acoustic cues, but also on the semantic cues (the meaning of preceding and following words and the subject matter), the syntactic cues (grammatical rules) and the circumstantial cues (speaker identity, listening environment etc.). It is quite likely that the processing of speech does not occur in a hierarchical way from one level to the next and that there are extensive links between levels [43]. However, the speech coders typically do not produce distortions that are specifically related to the semantic, syntactic or circumstantial cues. Therefore, it is reasonable to hypothesize that the proposed measures are, by and large, sufficient from this perspective.

The CDI measure compares the probability-of-firing information whereas the CHM measure compares implicitly the neural firing patterns for the original and the coded signals. With further progress in psychoacoustic research, it may be possible to derive the actual neural firing patterns from the cochlear model outputs by a suitable trigger mechanism [121] and compare them explicitly for the original and coded signals. Since all the information related to the speech perception are conveyed to the brain only as a sequence of neural firings through neural fibers, in future, an explicit comparison of these patterns may become an effective way for devising a distortion measure.

7.2.5 Application of Measures in Speech Coding

The present day state-of-the-art low bit-rate speech coders generally use, in the closed-loop analysis, a mean square error criterion with some form of perceptual weighting filter. For an use of the introduced measures in a speech coding process, the cochlear

model transformation has to be expressed in a more analytically tractable form. Also, because of the temporal masking effects, speech coding with such measures would imply additional coding delays. Motivated by the in-synchrony characteristics of the timing information in the auditory nerve firing patterns, in [122], the in-synchrony-bands spectrum has been used in an analysis/synthesis system. In [123], wavelet functions have been used to incorporate the multiresolution signal nature at the cochlea and represent the speech signal onto a joint time-frequency domain. Recent efforts (e.g., [124, 125]) have been made to propose empirical but perceptually advantageous time-frequency frameworks for speech processing. Further research is necessary to express cochlear functions, preserving all the major perceptual events, in a compact mathematical form and apply it in the speech coding process.

7.3 Epilog

In this dissertation, our primary contributions are—(i) reviewing the existing subjective and objective distortion measures, (ii) studying the auditory system and various cochlear models, (iii) applying Lyon's cochlear model for auditory representation of speech, (iv) devising a cochlear discrimination information measure and evaluating speech coder performance with it, (v) pursuing a rate-distortion analysis with this measure for speech coding, (vi) formulating a cochlear hidden Markovian measure and assessing speech coder quality with it, (vii) suggesting an algorithm for pitch frequency estimation from the cochlear model outputs, (viii) comparing different perceptual weighting strategies adopted in low bit-rate speech coders and (ix) providing a future research direction in the context of our work.

Determining a 'good' distortion measure for speech coding is an extremely difficult problem due to its very basic nature. At the same time, finding such a measure would surely have a significant impact on the speech coding and coder evaluation procedures. Our objective has never been to give a 'final' answer for this complex problem, rather we have tried to take an incremental step towards the solution. With the progress of time, we expect an improvement of the cochlear model structure, a determination of an analytically tractable expression for it and also a reduction in the computational complexity. However, the basic framework of comparing the neural firing information for original and coded signal could still be maintained.

Since the work of von Békésy, the auditory system has been studied from different perspectives. Some of these research findings are well-accepted in the literature. On the other side, since the pioneering work of Shannon, the field of information theory has grown substantially. Through the proposed work of distortion measures for coded speech, we have made an endeavor to use a physiological model for auditory processing and apply information-processing techniques from information theory.

Appendix A

It is known [70, 71] that $H_\alpha(P)$ is strictly concave with respect to P for $0 < \alpha \leq 1$, but its convexity or concavity depends on J for $\alpha > 1$. In this appendix, we show that for $J = 2$ (i.e., with $P = \{p_1, p_2\}$), (a) $H_\alpha(P)$ is strictly concave with respect to P for $0 < \alpha \leq 2$ and (b) for every $\alpha > 2$, $H_\alpha(P)$ is neither convex nor concave with respect to P.

It is shown in [71] that the $H_\alpha(P)$ is concave for $0 < \alpha \leq 1$. So, to prove (a), we have to show that for $J = 2$, the concavity is also satisfied for $1 < \alpha \leq 2$. We demonstrate this by showing that the second derivative of $H_\alpha(p, (1-p))$ with respect to P is negative in the range $1 < \alpha \leq 2$.

$$H_\alpha(P) = \frac{1}{(1-\alpha)} \log(p_1^\alpha + p_2^\alpha), \quad \text{where } p_2 = 1 - p_1, \quad p_1, p_2 \geq 0. \tag{A.1}$$

$$\frac{d^2 H_\alpha(P)}{dp_1^2} = \frac{\alpha}{(1-\alpha)} \cdot \frac{(\alpha-1)(p_1^\alpha + p_2^\alpha)(p_1^{\alpha-2} + p_2^{\alpha-2}) - \alpha(p_1^{\alpha-1} - p_2^{\alpha-1})^2}{(p_1^\alpha + p_2^\alpha)^2}$$

$$= \frac{\alpha}{(1-\alpha)} \cdot \frac{(p_1^{\alpha-2} p_2^{\alpha-2}).\{\alpha - (p_1^\alpha + p_2^\alpha)(p_1^{2-\alpha} + p_2^{2-\alpha})\}}{(p_1^\alpha + p_2^\alpha)^{-2}} \tag{A.2}$$

It is noted that for $\alpha > 1$,

$$(p_1^\alpha + p_2^\alpha) < (p_1 + p_2)^\alpha = 1. \tag{A.3}$$

Furthermore, $p_1 = p_2 = 1/2$ maximizes the expression $(p_1^{2-\alpha} + p_2^{2-\alpha})$ for $1 < \alpha \leq 2$. We note that $\alpha > (\frac{1}{2})^{1-\alpha}$ for $1 < \alpha \leq 2$. Additionally, we observe that the denominator factor $(1-\alpha)$ of (A.2) is negative for $\alpha > 1$. Thus, $H_\alpha(P)$ is proved to be concave in the range $1 < \alpha \leq 2$.

Now, we investigate the concavity for $J = 2$ and $\alpha > 2$. With sufficiently small $\delta > 0$ and $p_1 = \delta$ or $p_2 = \delta$, we obtain

$$\frac{d^2 H_\alpha(P)}{dp_1^2} > 0, \tag{A.4}$$

On the other hand, with $p_1 = p_2 = 1/2$, we have

$$\frac{d^2 H_\alpha(P)}{dp_1^2} = -4\alpha < 0. \tag{A.5}$$

From (A.4) and (A.5), we observe that for $J = 2$ and $\alpha > 2$, $H_\alpha(P)$ is neither convex nor concave.

Appendix B

In this appendix, we show that (a) the directed divergence measure based on the Rényi-Shannon entropy is non-negative and (b) it is additive for random measurements that are independent under both probability distributions. For showing the part (a), the inequality $\log x \geq 1 - (1/x)$ is used.

(a) Non-negativity of the measure:

$$D_1(P_{kl}; Q_{kl}) = \sum_{j=1}^{2} p_{jkl} \log\left(\frac{p_{jkl}}{q_{jkl}}\right) \geq \sum_{j=1}^{2} p_{jkl}\left(1 - \frac{q_{jkl}}{p_{jkl}}\right) = \sum_{j=1}^{2} p_{jkl} - \sum_{j=1}^{2} q_{jkl} = 0. \quad \text{(B.1)}$$

$$D_\alpha(P_{kl}; Q_{kl}) = \frac{1}{\alpha - 1} \log\left(\sum_{j=1}^{2} \frac{p_{jkl}^{\alpha}}{q_{jkl}^{\alpha-1}}\right) \geq \frac{1}{\alpha - 1}\left[1 - \frac{1}{\sum_{j=1}^{2} \frac{p_{jkl}^{\alpha}}{q_{jkl}^{\alpha-1}}}\right], \quad \alpha \neq 1. \quad \text{(B.2)}$$

To show that $D_\alpha(P_{kl}; Q_{kl}) \geq 0$, we need to show that

$$Y_\alpha(P_{kl}; Q_{kl}) \equiv \sum_{j=1}^{2} \frac{p_{jkl}^{\alpha}}{q_{jkl}^{\alpha-1}} \begin{cases} \leq 1 & \text{for } 0 \leq \alpha < 1 \\ \geq 1 & \text{for } \alpha > 1. \end{cases} \quad \text{(B.3)}$$

We note that $p_{1kl} = q_{1kl}$ and $p_{2kl} = q_{2kl}$ maximizes $Y_\alpha(P_{kl}; Q_{kl})$ for $0 \leq \alpha < 1$ and minimizes it for $\alpha > 1$. Thus, the $Y_\alpha(P_{kl}; Q_{kl})$ conditions of (B.3) are met and hence the non-negativity of the divergence measure (the Rényi-Shannon type) is also satisfied. The measure becomes equal to zero if and only if the distributions P_{kl} and Q_{kl} become the same.

(b) Addivitivity of the measure: With $w \in \mathcal{L} \times \mathcal{K}$ and $m = nN$, we obtain

$$\begin{aligned} D_1(P; Q) &= \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \cdots \sum_{j_m=1}^{2} \left(\prod_{w=1}^{m} p_{j_w}\right) \left[\sum_{w=1}^{m} \log \frac{p_{j_w}}{q_{j_w}}\right] \\ &= \left\{ \sum_{j_1=1}^{2} p_{j_1} \log\left(\frac{p_{j_1}}{q_{j_1}}\right) \sum_{j_2=1}^{2} p_{j_2} \cdots \sum_{j_m=1}^{2} p_{j_m} \right. \\ &+ \sum_{j_1=1}^{2} p_{j_1} \sum_{j_2=1}^{2} p_{j_2} \log\left(\frac{p_{j_2}}{q_{j_2}}\right) \cdots \sum_{j_m=1}^{2} p_{j_m} \\ &+ \left. \cdots + \sum_{j_1=1}^{2} p_{j_1} \sum_{j_2=1}^{2} p_{j_2} \cdots \sum_{j_m=1}^{2} p_{j_m} \log\left(\frac{p_{j_m}}{q_{j_m}}\right) \right\} \end{aligned}$$

$$= \sum_{w=1}^{m} \left\{ \sum_{j_w=1}^{2} p_{j_w} \log \left(\frac{p_{j_w}}{q_{j_w}} \right) \right\}$$

$$= \sum_{l \in \mathcal{L}} \sum_{k \in \mathcal{K}} D_1(P_{kl}; Q_{kl}) \tag{B.4}$$

Similarly, for $\alpha \neq 1,\ \alpha \geq 0$, we get

$$\begin{aligned} D_\alpha(P;Q) &= \frac{1}{(\alpha-1)} \log \left\{ \sum_{j_1=1}^{2} \cdots \sum_{j_m=1}^{2} \left(\frac{\prod_{w=1}^{m} p_{j_w}^{\alpha}}{\prod_{w=1}^{m} q_{j_w}^{\alpha-1}} \right) \right\} \\ &= \frac{1}{(\alpha-1)} \log \left\{ \left(\sum_{j_1=1}^{2} \frac{p_{j_1}^{\alpha}}{q_{j_1}^{\alpha-1}} \right) \left(\sum_{j_2=1}^{2} \frac{p_{j_2}^{\alpha}}{q_{j_2}^{\alpha-1}} \right) \cdots \left(\sum_{j_m=1}^{2} \frac{p_{j_m}^{\alpha}}{q_{j_m}^{\alpha-1}} \right) \right\} \\ &= \sum_{w=1}^{m} \left\{ \frac{1}{(\alpha-1)} \log \left(\sum_{j_w=1}^{2} \frac{p_{j_w}^{\alpha}}{q_{j_w}^{\alpha-1}} \right) \right\} \\ &= \sum_{l \in \mathcal{L}} \sum_{k \in \mathcal{K}} D_\alpha(P_{kl}; Q_{kl}) \end{aligned} \tag{B.5}$$

Appendix C

The reference audio files were obtained by digitally filtering the speech and sampling it at a rate of 8,000 Hz. The digital filter (255 tap FIR) applied was designed to be unity between 0 and 3,200 Hz. For the purpose of speech coder evaluation, the following test sentences [male (M) and female (F) voices] were used.

1. Add the sum to the product of these three (M1, F1).
2. Cats and dogs each hate the other (M2, F2).
3. Oak is strong and also gives shade (M3, F3).
4. Open the crate but don't break the glass (M4, F4).
5. The pipe began to rust while new (M5, F5).
6. Thieves who rob friends deserve jail (M6, F6).

Appendix D

A beta density function is given as

$$b(x) = \frac{\Gamma(d+f+2)}{\Gamma(d+1)\Gamma(f+1)} x^d (1-x)^f. \tag{D.1}$$

In this appendix, we prove that this function satisfies the log-concavity condition, i.e., the logarithm of the function is concave. Taking logarithm of (D.1), we get,

$$\begin{aligned} \phi(x) &= \log\Gamma(d+f+2) - \log\Gamma(d+1) - \log\Gamma(f+1) \\ &\quad + d\log x + f\log(1-x). \end{aligned} \tag{D.2}$$

To show the log-concavity nature of (D.1), we need to show that $\phi(x)$ is concave w.r.t. x. Defining $\overline{\lambda} \equiv (1-\lambda)$, we write

$$\begin{aligned} & \phi(\lambda x^{'} + \overline{\lambda} x^{''}) - \lambda\phi(x^{'}) - \overline{\lambda}\phi(x^{''}) \\ = \; & d\lambda\log(\lambda x^{'} + \overline{\lambda} x^{''}) + f\lambda\log(1-\lambda x^{'} - \overline{\lambda} x^{''}) + d\overline{\lambda}\log(\lambda x^{'} + \overline{\lambda} x^{''}) \\ & + f\overline{\lambda}\log(1-\lambda x^{'} - \overline{\lambda} x^{''}) - d\lambda\log x^{'} - f\lambda\log(1-x^{'}) \\ & - d\overline{\lambda}\log x^{''} - f\overline{\lambda}\log(1-x^{''}) \\ = \; & d\lambda\log\left(\frac{\lambda x^{'} + \overline{\lambda} x^{''}}{x^{'}}\right) + f\lambda\log\left(\frac{1-\lambda x^{'} - \overline{\lambda} x^{''}}{1-x^{'}}\right) \\ & + d\overline{\lambda}\log\left(\frac{\lambda x^{'} + \overline{\lambda} x^{''}}{x^{''}}\right) + f\overline{\lambda}\log\left(\frac{1-\lambda x^{'} - \overline{\lambda} x^{''}}{1-x^{''}}\right) \\ \geq \; & d\lambda\left(1 - \frac{x^{'}}{\lambda x^{'} + \overline{\lambda} x^{''}}\right) + f\lambda\left(1 - \frac{1-x^{'}}{1-\lambda x^{'} - \overline{\lambda} x^{''}}\right) \\ & + d\overline{\lambda}\left(1 - \frac{x^{''}}{\lambda x^{'} + \overline{\lambda} x^{''}}\right) + f\overline{\lambda}\left(1 - \frac{1-x^{''}}{1-\lambda x^{'} - \overline{\lambda} x^{''}}\right) \\ = \; & d - d\left(\frac{\lambda x^{'} + \overline{\lambda} x^{''}}{\lambda x^{'} + \overline{\lambda} x^{''}}\right) + f - f\left(\frac{\lambda(1-x^{'}) + \overline{\lambda}(1-x^{''})}{1-\lambda x^{'} + \overline{\lambda} x^{''}}\right) \\ = \; & 0. \end{aligned} \tag{D.3}$$

Since it has been shown that $\phi(\lambda x^{'} + \overline{\lambda} x^{''}) \geq \lambda\phi(x^{'}) - \overline{\lambda}\phi(x^{''})$, the beta pdf of (D.1) is proven to be log-concave.

Appendix E

In our work, an auxiliary function $F(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ is considered as the basis for the maximum likelihood optimization procedure. The Baum-Welch (re)estimation procedure is used for determining different model parameters. Separability of the individual auxiliary functions has made this procedure elegant and reduced the complexity. Here, we write the expressions for individual auxiliary functions. We can rewrite (5.25) as

$$\begin{aligned} F(\boldsymbol{\lambda}, \boldsymbol{\lambda}') &= F_{\pi}(\boldsymbol{\lambda}, \pi') + \sum_{i \in \mathcal{N}} F_{a_i}(\boldsymbol{\lambda}, \{a'_{ij}\}_{j \in \mathcal{N}}) \\ &\quad + \sum_{i \in \mathcal{N}} \sum_{m \in \mathcal{M}_L} F_b(\boldsymbol{\lambda}, b'_{im}) + \sum_{i \in \mathcal{N}} F_{c_i}(\boldsymbol{\lambda}, \{c'_{im}\}_{m \in \mathcal{M}_L}), \end{aligned} \tag{E.1}$$

where

$$\begin{aligned} F_{\pi}(\boldsymbol{\lambda}, \pi') &= \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}) \log \pi'_{q_1} \\ &= \sum_{i \in \mathcal{N}} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, q_1 = S_i, \boldsymbol{M}|\boldsymbol{\lambda}) \log \pi'_i, \end{aligned} \tag{E.2}$$

$$\begin{aligned} F_{a_i}(\boldsymbol{\lambda}, \{a'_{ij}\}_{j \in \mathcal{N}}) &= \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}) \sum_{t \in \mathcal{T}^+} \log a'_{q_t q_{t+1}} \delta(q_t - S_i) \\ &= \sum_{j \in \mathcal{N}} \sum_{t \in \mathcal{T}^+} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, q_t = S_i, q_{t+1} = S_j, \boldsymbol{M}|\boldsymbol{\lambda}) \log a'_{ij}, \end{aligned} \tag{E.3}$$

$$\begin{aligned} F_b(\boldsymbol{\lambda}, b'_{im}) &= \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}). \sum_{t \in \mathcal{T}^+} \log b'_{q_t m_t}(O_t) \delta(q_t - S_i) \delta(m_t - m) \\ &= \sum_{t \in \mathcal{T}^+} P(\boldsymbol{O}, q_t = S_i, m_t = m|\boldsymbol{\lambda}) \log b'_{im}(O_t) \end{aligned} \tag{E.4}$$

and

$$\begin{aligned} F_{c_i}(\boldsymbol{\lambda}, \{c'_{im}\}_{m \in \mathcal{M}_L}) &= \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}|\boldsymbol{\lambda}) \sum_{t \in \mathcal{T}^+} \log c'_{q_t m_t} \delta(q_t - S_i) \\ &= \sum_{m \in \mathcal{M}_L} \sum_{t \in \mathcal{T}^+} P(\boldsymbol{O}, q_t = S_i, m_t = m|\boldsymbol{\lambda}) \log c'_{im}, \end{aligned} \tag{E.5}$$

where δ in the above expressions is the Kronecker delta function.

References

[1] R. E. Blahut, *Principles and Practice of Information Theory.* Addison-Wesley, 1987.

[2] D. O'Shaughnessy, *Speech Communication.* Academic Press, 1987.

[3] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video.* Prentice Hall, 1984.

[4] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals.* Prentice Hall, 1979.

[5] J. P. Campbell Jr., T. E. Tremain, and V. C. Welch, "The federal standard 1016 4800 bps CELP voice coder," *Digital Signal Processing*, vol. 1, pp. 145–155, July 1991.

[6] W. B. Kleijn, *Analysis-by-Synthesis Speech Coding Based on Relaxed Waveform-Matching Constraints.* PhD thesis, Delft University of Technology, Dec. 1991.

[7] J.-P. Adoul, P. Mabilleau, M. Delprat, and S. Morissette, "Fast CELP coding based on algebraic codes," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 49.4.1–49.4.4, 1987.

[8] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Kluwer Academic Pub., 1992.

[9] K. Abboud, "Wideband CELP speech coding," Master's thesis, McGill University, Feb. 1993.

[10] H. J. Coetzee and T. P. Barnwell III, "An LSP based speech quality measure," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 596–599, 1989.

[11] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality.* Prentice Hall, 1988.

[12] T. P. Barnwell III, "Correlation analysis of subjective and objective measures for speech quality," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 706–709, 1980.

[13] R. F. Kubichek, "Standards and technology issues in objective voice quality assessment," *Dig. Signal Process.*, vol. 1, pp. 38–44, Jan. 1991.

[14] M. H. L. Hecker and C. E. Williams, "Choice of reference conditions for speech preference tests," *J. Acoust. Soc. Am.*, vol. 40, pp. 946–952, Nov. 1966.

[15] H. Fletcher and J. C. Steinberg, "Articulation testing methods," *Bell Syst. Tech. J.*, pp. 806–854, 1929.

[16] G. Fairbanks, "Test of phonemic differentiation: The rhyme test," *J. Acoust. Soc. Am.*, pp. 596–600, July 1958.

[17] W. A. Munson and J. E. Karlin, "Isopreference method for evaluating speech-transmission circuits," *J. Acoust. Soc. Am.*, vol. 31, pp. 762–774, June 1962.

[18] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Aud. and Electroacoust.*, pp. 227–246, Sept. 1969.

[19] W. D. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 204–207, 1977.

[20] S. R. Quackenbush, "Objective measures of speech quality," Tech. Rep. DSPL-85-4, Georgia Institute of Technology, 1985.

[21] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Am.*, vol. 66, pp. 1647–1652, Dec. 1979.

[22] P. Mermelstein, "Evaluation of a segmental SNR measure as an indicator of the quality of ADPCM coded speech," *J. Acoust. Soc. Am.*, vol. 66, pp. 1664–1667, Dec. 1979.

[23] B. J. McDermott, C. Scagliola, and D. J. Goodman, "Perceptual and objective evaluation of speech processed by adaptive differential PCM," *Bell Syst. Tech. J.*, pp. 1597–1619, May 1978.

[24] R. E. Crochiere, J. E. Tribolet, and L. R. Rabiner, "An interpretation of the log likelihood ratio as a measure of waveform coder performance," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-28, pp. 318–323, June 1980.

[25] R. M. Gray, A. Buzo, A. H. Gray Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-28, pp. 367–376, Aug. 1980.

[26] U. Halka and U. Heute, "A new approach to objective quality-measures based on attribute-matching," *Speech Commun.*, vol. 11, pp. 15–30, Mar. 1992.

[27] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-24, pp. 380–391, Oct. 1976.

[28] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 242–248, Feb. 1988.

[29] K. K. Paliwal, "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition," *Speech Commun.*, vol. 1, pp. 151–154, May 1982.

[30] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-35, pp. 1414–1422, Oct. 1987.

[31] Y.-T. Lee, "Information-theoretic distortion measures for speech recognition," *IEEE Trans. Signal Process.*, vol. 39, pp. 330–335, Feb. 1991.

[32] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. 6th Int. Cong. Acoust., Japan*, pp. C 17–C 20, 1968.

[33] P. L. Chu and D. G. Messerschmitt, "A frequency weighted Itakura-Saito spectral distance measure," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-30, pp. 545–560, Aug. 1982.

[34] B. A. Carlson and M. A. Clements, "A computationally compact divergence measure for speech processing," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 13, pp. 1255–1260, Dec. 1991.

[35] Bell Northern Research, "Evaluation of nonlinear distortion via the coherence function," *Contribution to CCITT, COM-XII-no. 60-E*, Apr. 1982.

[36] J. Lalou, "The information index: An objective measure of speech transmission performance," *Ann. Telecommun.*, vol. 45, pp. 47–65, Jan. 1990.

[37] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 819–829, June 1992.

[38] R. Bladon, "Modeling the judgement of vowel quality differences," *J. Acoust. Soc. Am.*, vol. 69, pp. 1414–1422, May 1981.

[39] D. Robinson and R. Dadson, "A redetermination of the equal-loudness relations for pure tones," *Brit. J. Appl. Physics*, vol. 7, pp. 166–181, 1956.

[40] B. Paillard, J. Soumagne, P. Mabilleau, and S. Morissette, "PERCEVAL: Perceptual evaluation of the quality of audio signals," *J. Audio Eng. Soc.*, vol. 40, pp. 21–31, Jan.–Feb. 1992.

[41] J. Flanagan, *Speech Analysis, Synthesis and Perception.* Springer-Verlag, 1972.

[42] D. Green, *An Introduction to Hearing.* Erlbaum, 1976.

[43] B. C. J. Moore, *Introduction to the Psychology of Hearing.* Academic Press, 1989.

[44] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 1282–1285, 1982.

[45] G. von Békésy, *Experiments in Hearing.* McGraw-Hill, 1960.

[46] J. B. Allen, "Cochlear modeling," *IEEE ASSP Mag.*, pp. 3–29, Jan. 1985.

[47] J. J. Zwislocki, "Five decades of research on cochlear mechanics," *J. Acoust. Soc. Am.*, vol. 67, pp. 1679–1685, 1980.

[48] D. T. Kemp, "Towards a model for the origin of cochlear echoes," *Hearing Res.*, vol. 2, pp. 533–548, 1980.

[49] M. J. Penner, "Forward masking with equal-energy maskers," *J. Acoust. Soc. Am.*, vol. 66, pp. 1719–1724, Dec. 1979.

[50] E. D. Young and P. E. Barta, "Rate responses of auditory nerve fibers to tones in noise near masked threshold," *J. Acoust. Soc. Am.*, vol. 79, pp. 426–442, Feb. 1986.

[51] A. B. Carlson, *Communication Systems.* McGraw Hill, 1986.

[52] H. V. Helmholtz, *On the Sensations of Tone.* Dover Pub., 1954.

[53] C. D. Geisler, "Representation of speech sounds in the auditory nerve," *J. of Phonetics*, vol. 16, pp. 19–35, Jan. 1988.

[54] S. Greenberg, "The ear as a speech analyzer," *J. of Phonetics*, vol. 16, pp. 139–149, Jan. 1988.

[55] M. B. Sachs, C. C. Blackburn, and E. D. Young, "Rate-place and temporal-place representations of vowels in the auditory nerve and anteroventral cochlear nucleus," *J. of Phonetics*, vol. 16, pp. 37–53, Jan. 1988.

[56] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. of Phonetics*, vol. 16, pp. 55–76, Jan. 1988.

[57] S. A. Shamma, "Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," *J. Acoust. Soc. Am.*, vol. 78, pp. 1622–1632, Nov. 1985.

[58] O. Ghitza, "Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment," *J. of Phonetics*, vol. 16, pp. 109–123, Jan. 1988.

[59] L. Deng, C. D. Geisler, and S. Greenberg, "A composite model of the auditory periphery for the processing of speech," *J. of Phonetics*, vol. 16, pp. 93–108, Jan. 1988.

[60] G. Zweig, R. Lipes, and J. R. Pierce, "The cochlear compromise," *J. Acoust. Soc. Am.*, vol. 59, pp. 975–982, 1976.

[61] M. Slaney, "Lyon's cochlear model," Tech. Rep. 13, Apple Computer Inc., 1988.

[62] J. O. Pickles, *An Introduction to the Physiology of Hearing.* Academic Press, 1982.

[63] L. Deng, "Processing of acoustic signals in a cochlear model incorporating laterally coupled suppressive elements," *Neural Networks*, vol. 5, pp. 19–34, 1992.

[64] T. Hall, "Cochlear models: Evidence in support of mechanical nonlinearity and second filters (a review)," *Hearing Res.*, vol. 2, pp. 455–464, 1980.

[65] M. R. Schroeder and J. L. Hall, "Model for mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Am.*, vol. 55, pp. 1055–1060, 1974.

[66] R. F. Lyon and L. Dyer, "Experiments with a computational model of the cochlea," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 37.6.1–37.6.4, 1986.

[67] S. Kullback, *Information Theory and Statistics.* John Wiley & Sons, 1959.

[68] A. De and P. Kabal, "Cochlear discrimination: An auditory information-theoretic distortion measure for speech coders," in *Proc. 16 th Biennial Symp. on Commun., Kingston, Canada*, pp. 419–423, May 1992.

[69] A. De and P. Kabal, "Auditory distortion measure for coded speech—discrimination information approach," *Speech Commun. (being revised for publication)*, 1993.

[70] J. Aczél and Z. Daróczy, *On Measures of Information and Their Characterizations.* Academic Press, 1975.

[71] A. Rényi, *Probability Theory.* North-Holland, 1970.

[72] J. Aczél, "Some recent results on characterizations of measures of information related to coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 592–595, Sept. 1978.

[73] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inform. Theory*, vol. 37, pp. 145–151, Jan. 1991.

[74] C. R. Rao and T. K. Nayak, "Cross entropy, dissimilarity measures, and characterizations of quadratic entropy," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 589–593, Sept. 1985.

[75] G. T. Toussaint, "Sharper lower bounds for discrimination information in terms of variation," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 99–100, Jan. 1975.

[76] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 405–417, July 1974.

[77] T. B. Berger, *Rate Distortion Theory.* Prentice Hall, 1971.

[78] D. J. Sakrison, "The rate distortion function of a Gaussian process with a weighted square error criterion," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 506–508, May 1968.

[79] A. Buzo, F. Kuhlmann, and C. Rivera, "Rate-distortion bounds for quotient-based distortions with applications to Itakura-Saito distortion measures," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 141–147, Mar. 1986.

[80] J. T. Pinkston, "An application of rate-distortion theory to a converse to the coding theorem," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 66–71, Jan. 1969.

[81] R. M. Gray, "Rate distortion functions for finite-state finite-alphabet Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 127–134, Mar. 1971.

[82] H. H. Tan and K. Yao, "Evaluation of rate-distortion functions for a class of independent identically distributed sources under an absolute-magnitude criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 59–64, Jan. 1975.

[83] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.

[84] P. Noll, "Adaptive quantizing in speech coding systems," in *Zurich Seminar Dig. Commun., Zurich, Switzerland*, Mar. 1974.

[85] D. H. Richards, "Statistical properties of speech signals," *Proc. IEEE*, vol. 52, pp. 941–949, 1964.

[86] H. Abut and N. Erdöl, "Bounds on $R_1(D)$ functions for speech probability models," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 225–228, Mar. 1979.

[87] A. De and P. Kabal, "Rate distortion function for speech coding based on perceptual distortion measure," in *Proc. of IEEE Globecom'92*, pp. 452–456, Dec. 1992.

[88] B. S. Atal, V. Cuperman, and A. Gersho, *Advances in Speech Coding.* Kluwer Academic Pub., 1991.

[89] G. Roy and P. Kabal, "Wideband CELP speech coding at 16 kbits/sec," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 17–20, 1991.

[90] A. De and P. Kabal, "Hidden Markov model-based auditory distortion measure for speech coder evaluation," in *Abstracts of Canadian Inst. Telecommun. Res. Conf., Montréal, Canada (to appear)*, Aug. 1993.

[91] A. De and P. Kabal, "Auditory distortion measure for coded speech—hidden Markovian approach," *Speech Commun. (being prepared for submission)*, 1993.

[92] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.*, vol. 37, pp. 1554–1563, 1966.

[93] L. E. Baum and J. A. Egon, "An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology," *Bull. Amer. Meteorol. Soc.*, vol. 73, pp. 360–363, 1967.

[94] R. L. Cave and L. P. Neuwirth, "Hidden Markov models for english," *Hidden Markov models for Speech (J. Ferguson [ed.])*, vol. IDA-CRD, pp. 16–56, 1980.

[95] R. W. Chang and J. C. Hancock, "On receiver structures for channels having memory," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 463–468, Oct. 1966.

[96] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–536, Apr. 1976.

[97] L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 729–734, Sept. 1984.

[98] B.-H. Juang, "Maximum-likelihood estimation for mixture multivariate stochatic observations of Markov chains," *Bell Syst. Tech. J.*, pp. 1235–1249, July-Aug. 1985.

[99] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[100] B.-H. Juang, "On the hidden Markov model and dynamic time warping for speech recognition," *Bell Syst. Tech. J.*, pp. 1213–1243, Sept. 1984.

[101] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, pp. 1035–1074, Apr. 1983.

[102] L. R. Rabiner, B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Some properties of continuous hidden Markov model representations," *Bell Syst. Tech. J.*, pp. 1251–1269, July-Aug. 1985.

[103] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices.* Springer-Verlag, 1983.

[104] M. Slaney and R. F. Lyon, "Visualizing sound with auditory correlograms," *in submission for J. Acoust. Soc. Am.*, 1991.

[105] R. F. Lyon, "Computational models of neural auditory processing," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 36.1.1–36.1.4, 1984.

[106] M. Weintraub, "A computational model for separating two simultaneous talkers," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, 1986.

[107] S. Seneff, "Pitch and spectral estimation of speech based on auditory synchrony model," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 36.2.1–36.2.4, 1984.

[108] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. ASSP-34, pp. 1419–1426, 1986.

[109] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.

[110] J. L. Moncet and P. Kabal, "Codeword selection for CELP coders," Tech. Rep. 87-35, INRS-Telecommunications, 1987.

[111] Q. Yasheng and P. Kabal, "Pseudo-three-tap pitch prediction filters," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. II.523–II.526, 1993.

[112] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing.* Prentice Hall, 1983.

[113] N. S. Jayant and V. Ramamoorthy, "Adaptive postfiltering of 16 kb/s-adpcm speech," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 16.4.1–16.4.4, 1986.

[114] C. Laflamme, J.-P. Adoul, R. Salami, S. Morissette, and P. Mabilleau, "16 kbps wideband speech coding technique based on algebraic CELP," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 13–16, 1991.

[115] E. Ordentlich and Y. Shoham, "Low-delay code-excited linear-predictive coding of wideband speech at 32 kbps," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. 9–12, 1991.

[116] A. De, "Auditory distortion measures for coded speech quality evaluation," in *Proc. Canadian Acoust. Assoc. Annual Symp., Toronto, Canada (to appear)*, Oct. 1993.

[117] J. M. Kates, "A time-domain digital cochlear model," *IEEE Trans. Signal Process.*, vol. 39, pp. 2573–2592, Dec. 1991.

[118] T. Kohonen, *Self-Organization and Associative Memory.* Springer Verlag, 1988.

[119] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. 36, pp. 1119–1134, July 1988.

[120] "Half-rate speech codec test plan V 6.0," Tech. Rep. TR 45.35, Telecommun. Industries Assoc., 1993.

[121] R. D. Patterson, "Auditory/connectionist techniques for speech," Tech. Rep. 2, ESPRIT Basic Research Action 3207, 1991.

[122] O. Ghitza, "Auditory nerve representation criteria for speech analysis/synthesis," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-35, pp. 736–740, June 1987.

[123] X. Yang, K. Wang, and S. A. Shamma, "Auditory representation of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38 (II), pp. 824–839, Mar. 1992.

[124] Y. Shoham, "High quality speech coding at 2.4 to 4.0 kbps based on time-frequency interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. II.167–II.170, 1993.

[125] D. Sen, D. H. Irving, and W. H. Holmes, "Use of an auditory model to improve speech coders," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pp. II.411–II.414, 1993.