Perceptual Postfiltering for Low Bit Rate Speech Coders

Wei Chen

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

November 2007

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Engineering.

Abstract

Adaptive postfiltering has become a common part of speech coding standards based on the Linear Prediction Analysis-by-Synthesis algorithm to decrease audible coding noise. However, a conventional adaptive postfilter is based on empirical assumptions of masking phenomena, which sometimes makes it hard to balance between noise reduction and speech distortion.

This thesis introduces a novel perceptual postfiltering system for low bit rate speech coders. The proposed postfilter works at the decoder, as is the case for the conventional adaptive postfilter. Specific human auditory properties are considered in the postfilter design to improve speech quality. A Gaussian Mixture Model based Minimum Mean Squared Error estimation of the perceptual postfilter is performed with the received information at the decoder. Perceptual postfiltering is then applied to the reconstructed speech to improve speech quality. Test results show that the proposed system gives better perceptual speech quality over conventional adaptive postfiltering.

Sommaire

Le post-filtrage adaptatif est devenu monnaie courante pour le codage de la parole basé sur l'algorithme de prédiction linéaire par analyse/synthèse afin de diminuer le bruit de codage audible. Toutefois, un post-filtre adaptatif conventionnel utilise des phénomènes de masquage basés sur des hypothèses empiriques, ce qui rend parfois difficile le compromis entre la réduction du bruit et la distorsion de la parole.

Cette thèse propose un nouveau système de post-filtrage perceptuel pour les codeurs à faible débit binaire. Le post-filtre proposé fonctionne au niveau du décodeur, comme dans le cas du post-filtre adaptatif conventionnel. Des propriétés spécifiques du système auditif humain sont considérées dans la conception du post-filtre afin d'améliorer la qualité de la parole. Un modèle de mixture gaussienne basé sur l'estimation de l'erreur quadratique moyenne minimale est considéré selon l'information reçue au décodeur. Un post-filtrage perceptuel est ensuite appliqué à la parole reconstruite pour en améliorer la qualité. Des résultats expérimentaux démontrent que le système proposé donne une meilleure qualité perceptuelle de la parole par rapport au post-filtrage adaptatif conventionnel.

Acknowledgments

I would like to thank my supervisor, Prof. Peter Kabal, for his valuable advice and academic guidance during my study at McGill University. His deep knowledge of the research field and his great patience made this work possible.

I would like to express my gratitude to all my fellow Telecommunications and Signal Processing Laboratory graduate students for their companionship, discussions and technical assistance. I am much obliged to François Duplessis-Beaulieu for the French abstract.

I also appreciate the support given by my friends here in Montreal. Their friendship and encouragement mean a lot to me and make me much stronger.

Special thanks are going to my family for their endless love and support. I am deeply indebted to their understanding and patience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech is the most natural form of human communications. Speech coding algorithms have made the communication and the storage of voice data effective and efficient. Due to increasing demand for speech communication, speech coding technology has received increased interest from research, standardization, and business communities. Speech coding algorithms have been employed in many applications including personal wireless communication systems, multimedia and Internet communication systems.

In speech coding, researchers have studied ways of efficiently representing acoustic speech waveforms in the digital domain. The ultimate goal in the design of speech codecs is to achieve the best possible quality at low bit rates, with constraints on complexity and delay. Both statistical redundancy removal and perceptual irrelevancy removal are considered. First, speech is produced by people as part of a physical process (air flow and moving muscles), and the corresponding signal has certain properties (such as correlations) which can be exploited to do more efficient processing. Also, speech is listened to by people and we can take into account the properties of human hearing system. Our ears are quite good but they are not perfect. In speech and audio coding, to achieve good coding efficiency, processing methods usually remove the perceptually irrelevant information and enhance the perceptually sensitive information. Currently, the psychoacoustic properties of human hearing are considered in both speech coding and audio coding to reduce the information to be transmitted while maintaining good fidelity.

Adaptive postfiltering has been commonly applied in low bit rate *linear prediction analysis-by-synthesis* (LPAS) speech coders. Lower bit rates are usually associated with poorer speech quality. Audible noise becomes more noticeable in the reconstructed speech at lower bit rates.

Postfiltering is used to reduce this noise by exploiting psychoacoustic properties while not significantly degrading speech.

1.1 Overview of Speech Coding

One possible way of lowering the bit rate in speech coding is to choose a low sampling frequency. By far the two most popular choices of speech sampling frequency are 8 and 16 kHz. Codecs using 8 kHz sampling frequency are referred to as narrowband codecs and those using 16 kHz sampling frequency are called wideband codecs.

Most speech coding systems were designed to support telecommunication applications, with the frequency contents limited to between 300 and 3400 Hz with 8 kHz sampling frequency. This kind of speech is usually called narrowband telephone speech or narrowband speech. We will only consider narrowband speech in our work.

Fig. 1.1 represents the encoder/decoder structure of a *speech coder*. A digital speech signal with 16 bits/sample (i.e. 16 bits $\times$ 8 kHz $=$ 128 kbps) is the input to the speech coder. The encoder attempts to reduce the bit rate. The output of encoder represents the encoded information about the speech and should have substantially lower bit rate than that of the input. The decoder takes the encoded bit-stream as its input to generate decoded speech signal, which is a discrete-time signal having the same rate as the signal to the encoder. Different design approaches of the encoder/decoder pair provide differing speech quality and bit rate, as well as implementation complexity.

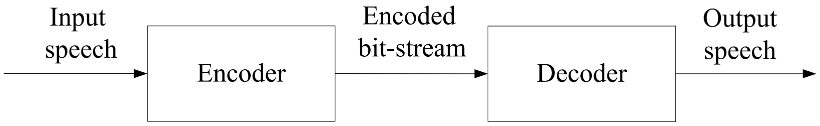


Fig. 1.1 Block diagram of a speech coder

All speech coders are designed to reduce the reference bit-rate of 128 kbps towards lower values. Depending on the bit-rate of the encoded bit-stream, it is common to classify the speech coders according to Table 1.1. Different techniques lead to different bit-rates.

According to coding techniques, modern speech coders are classified into following three types [1, 2]:

Table 1.1 Classification of speech coders by bit-rate [1]

Category	Bit-Rate Range
High bit-rate	>15 kbps
Medium bit-rate	5 to 15 kbps
Low bit-rate	2 to 5 kbps
Very low bit-rate	<2 kbps

Waveform Coders As the name implies, the goal of waveform coding is to reproduce the original waveform as accurately as possible. It is sample-by-sample coding and often not speech-specific. Waveform coding can deal with non-speech signals without difficulty. However, the cost of this fidelity is a relatively high bit rate. These coders work best at a bit rate of 32 kbps and higher. Example coders of this class include *pulse code modulation* (PCM), *adaptive differential PCM* (ADPCM) and subband coders.

Parametric Coders Speech is assumed to be generated from a *model*, which is controlled by some *parameters*. During encoding, parameters of the model are estimated from the input speech signal frame-by-frame, and are transmitted after being coded. This type of coders makes no attempt to preserve the original shape of the waveform. Perceptual quality of the decoded speech is directly related to the accuracy and sophistication of the underlying model. Speech quality tends to be synthetic and variable between speakers, although intelligible. This coder is signal specific, having poor performance for nonspeech signals. The most successful model is based on *linear prediction* (LP). This type of coder works well at low bit rates. Examples of this type are *liner prediction coding* (LPC) and *mixed excitation linear prediction* (MELP) [3, 4].

Hybrid Coders This type combines features from both waveform coders and parametric coders to provide good-quality, efficient speech coding. Like a parametric coder, it relies on a speech model. During encoding, parameters of the model are estimated. Additional parameters of the model are optimized in such a way that the decoded speech is as close as possible to the original waveform, with the closeness often measured by a perceptually weighted error signal. As above in waveform coders, an attempt is made to match the original signal with the decoded signal in the time domain. By an *analysis-by-synthesis* technique, good quality coding is achieved at rates between about 4 kbps and 16 kbps. *Coded-excited linear prediction* (CELP) and its variants are the most outstanding representatives [5–7].

For speech coding to be useful in public telecommunication applications, it has to be standardized (i.e. it must conform to the same algorithm and bit format) to ensure universal interoperability. Speech coding standards are established by various standards organizations [8], for example, International Telecommunications Union, Telecommunications Standardization Sector (ITU-T, formally CCITT), Telecommunications Industry Association (TIA), Research and Development Centre for Radio Systems (RCR) in Japan, European Telecommunications Standards Institute (ETSI), and other government agencies.

Since CELP can achieve relatively high coding quality at the bit-rate range from 4 to 16 kbps, CELP-based coders have been deployed in a wide range of recent standardizations including ITU-T Recommendation G.723.1 at rate of 6.3/5.3 kbps and G.729 at rate of 8 kbps.

1.1.1 Speech Production Model

Human speech, which is represented by speech waveforms, is generated by a voluntary movement of anatomical structures. A source-tract modelling is widely used in speech coding. The model is inspired by observations of the basic properties of speech signals and represents an attempt to mimic the human speech production mechanism. The human vocal tract is an acoustic tube, which has one end at the glottis and the other end at the lips. The vocal tract changes shape continuously with time, creating an acoustic filter with a time-varying frequency response. As air from the lungs travels through the tract, the frequency spectrum is shaped by the frequency selectivity of the tract. By the action of the glottis constricting the air-flow from the lung periodically or not, the source signal is nearly-periodic or noise-like. The resonance frequencies of the vocal tract tube are called *formant frequencies* or simply *formants*, which depend on the shape and dimensions of the vocal tract.

The source-tract model leads to a representation that consists of a description of an excitation (source) signal that is periodic or aperiodic and a time-varying linear filter that has a transfer function representing the vocal tract. The property of the excitation gives *voiced* or *unvoiced* speech. In the time domain, voiced sound is characterized by strong periodicity present in the signal, with the fundamental frequency referred to as the *pitch frequency*, or simply *pitch*. For adult males, pitch ranges from 50 to 250 Hz, while for adult females the range usually falls somewhere in the interval of 120 to 500 Hz [1]. Unvoiced sounds, on the other hand, are essentially random in nature. The energy distribution of the speech signal in the frequency domain is controlled by the time-varying filter in this model. Linear prediction analysis is the most successful technique to find the coefficients of the time-varying linear filter.

For most speech coders, the signal is processed on a frame-by-frame basis, where a frame consists of a finite number of samples. The length of the frame is selected in such a way that the statistics of the signal remain almost constant within the interval.

1.1.2 Speech Perception

A human auditory model is a mathematical model, which describes the behaviour of the human auditory system. Human auditory models have been widely applied in audio coding to get near transparent coding quality while saving bits. Also in speech coding, with the knowledge of how sound is perceived, resources in the coding system can be allocated in the most efficient manner.

Our human ears are the ultimate receiver for sound. The pinna is the surface surrounding the canal in which sound is funnelled [1]. Sound waves are guided by the canal toward the eardrum—a membrane that acts as an acoustic-to-mechanic transducer. The sound waves are then translated into mechanical vibrations that are passed to the cochlea through a series of bones known as the ossicles. Presence of the ossicles improves sound propagation by reducing the amount of reflection and is accomplished by the principle of impedance matching.

The cochlea is a rigid snail-shaped organ filled with fluid. Mechanical oscillations impinging on the ossicle cause an internal membrane, known as the *basilar membrane*, to vibrate at various frequencies. Each point along the basilar membrane has a characteristic frequency to which it vibrates maximally. A simple modelling technique is to use a bank of filters to describe this behaviour. Displacement of the basilar membrane at different places is sensed by the inner hair cells and causes neural activities that are transmitted to the brain through the auditory nerve [9].

Along the basilar membrane, different points are affected differently depending on the frequencies of the incoming sound waves. Hair cells located at different positions along the membrane are excited by sounds of different frequencies. The neurons, which contact the hair cells and transmit the excitation to higher auditory centres, maintain the frequency specificity. This arrangement makes the human auditory system behave very much like a frequency analyzer. The auditory system characterization is simpler if done in the frequency domain. The frequency resolution[1] is greatest at low frequencies.

[1] The use of spectral components for the extraction of the biological meaning from communication sounds is built upon the ability of the auditory systems to resolve frequency components of the sounds. The frequency resolution or frequency selectivity of the human ear is its ability in detecting differences in pitch.

The *absolute threshold of hearing* of a sound is the minimum detectable level of that sound in the absence of any other external sounds [1]. It characterizes the amount of energy needed in a pure tone such that it can be detected by a listener in a noise-free environment. The absolute threshold of hearing is frequency dependent. The ear's sensitivity is best for frequencies in the range of 1 to 4 kHz, while thresholds increase rapidly at very high and very low frequencies. It is commonly accepted that below 20 Hz and above 20 kHz, the auditory system does not respond.

The absolute threshold of hearing can be applied in speech coding. Any signal with an intensity below the absolute threshold need not be considered, since it does not have any impact on the final quality of the coder.[2] More resources should be allocated for the representation of the signal within the most sensitive frequency range, since distortion in this range is more perceptible.

Masking is a phenomenon in sensory perception. It is about a sound being inaudible because of the presence of a stronger sound, and has received significant attention from researchers in the field of psychoacoustics [12]. The stronger signal is called *masker*, while the masked signal is referred as *maskee*. A *masking threshold* corresponds to the increased threshold of audibility, resulting from a masker. The amount of masking is influenced by various factors including signal level, frequency and duration. In general, masking capability increases with the relative intensity of the masker. Masking theory is mainly used in audio coding and objective measures of perceived audio quality [13]. Masking can also be exploited for speech coding developments. For example, by analyzing the spectral contents of a signal, it is possible to locate the frequency regions that are most susceptible to distortion. Perceptual weighting filtering and adaptive postfiltering have been widely used in speech coding, which are motivated from masking theory. More will be discussed in Chapter 2.

1.2 Motivation and Objective of Our Research

Speech coding is a balancing game between quality, bit rate, delay and complexity [14]. The quality is a function of the bit rate. In order to meet the strong need to have a common means for communication, many speech coding standards have been created. These standards have been widely used in speech communications.

[2] In the case of sound intensity, 0 dB sound pressure level (SPL) is chosen to be the average absolute threshold of humans for a 1 kHz sinusoid [10]. The SPLs of all frequencies that have the same loudness as 0 dB SPL 1 kHz sound form the absolute threshold of hearing. When designing signal processing algorithms, it is often not possible to know beforehand the playback levels of signals. Therefore, a common assumption about the playback level is the lowest possible signal power of a 1 kHz sound corresponds to approximately 0 dB SPL [11].

A speech coder derives a set of parameters at the encoder to control a speech production model at the receiver. The goal of speech coding is either to maximize the perceived quality at a particular bit rate, or to minimize the bit rate for a particular perceptual quality. At low bit rates, it is hard to get a good speech quality, so a trade-off is often found to satisfy the necessity of a given application. With the development of sophisticated signal processing algorithms and technologies, a lot of research has been done to reduce the number of parameters representing speech signal at the encoder, while maintaining the coded speech quality. For a given bit rate, the speech quality can be improved to some extent by employing more complex encoding algorithms. Although changes in speech encoding are usually used to improve the speech quality, there should be other means which can improve the coded speech without changing the speech coder structure.

In current LPAS speech coders, the properties of speech production, as well as the human hearing system properties, are exploited. At low bit rates, there is still audible noise in the coded speech. Postfiltering is a tool to reduce the coding noise in the decoded speech based on the local characteristics of the speech spectrum at the decoder. It acts as an add-on component, which makes it widely used in different types of speech coders with the same general structure. Adaptive postfiltering algorithm [15] achieves significant noise reduction without introducing significant distortion in speech. Since its initial introduction, its variations have been successfully used in many speech coding standards, such as ITU-T G.729 and G.723.1. However, the conventional postfilter [15] still has problems comparing with the ideal postfilter[3], which the conventional postfilter is built on.

The conventional postfiltering is empirically designed by general masking phenomena considerations. Although perceptual models have been successfully implemented in audio coding by exploiting the masking property, specific perceptual models have not been applied in postfiltering for speech coders. This motivates us to study an improved postfilter exploiting more precisely perceptual properties. It is possible to design a postprocessor, which improves the coded speech quality, with the same available information at the decoder as the conventional postfilter. We believe that a better postprocessing paradigm exploiting the encoding information and the characteristics of human hearing system can give an improved quality to the reconstructed speech.

The goal of our research is to develop a postfilter which incorporates the knowledge of perceptual properties. Comparing with the conventional postfilter [15], our postfilter uses some specific perceptual properties to improve the coded speech quality. By using knowledge of the human

[3] An ideal postfilter should not alter the formant information and should attenuate null information in the speech spectrum in order to achieve reduction and produce better speech quality [16].

auditory properties, it is expected that perceptual quality of the processed speech may sound better than the conventional adaptive postfilter. The perceptual postfilter only uses the information available at the decoder. The structure of the coding system should not be changed.

1.3 Thesis Contribution

In this thesis, we design a novel perceptual postfilter for low bit rate speech coders. The proposed postfilter is based on the characteristics of the human hearing system. The proposed postfilter is an add-on part and embedded in the decoder. The encoder is not modified and no extra side information is sent to the decoder.

The originality of the proposed perceptual postfilter is a combination of the following two features.

Perceptual Postfiltering

- A perceptual postfilter is derived from clean speech and its coded version based on the properties of psychoacoustic models. It operates on a frame-by-frame basis. The postfilter is then applied to the decoded speech to improve the speech quality. However, in practice, we do not have the information about the perceptual postfilter gains at the decoder, if they are not sent as side information by the encoder.

Optimal MMSE Estimator Based on GMM

- Without additional side information received, we estimate the perceptual postfilter with an optimal *Minimum Mean Squared Error* (MMSE) estimator with the available information at the decoder. We use the available information as an “input” vector, and the perceptual postfilter gains as a “target” vector. A feature vector is constructed with “input” and “target” vectors.

- In order to find a MMSE estimate of the “target” vector, *a priori* information of the *joint probability density function* (joint pdf) of the feature vector is required. A *Gaussian mixture model* (GMM) is used to model the joint density.

1.4 Thesis Organization

This thesis consists of 6 chapters. Chapter 2 presents a brief review of adaptive postfiltering. Starting with the fundamentals of LPAS speech coder, Chapter 2 discusses the algorithms of adaptive postfiltering. It explains how the masking concept is used in both encoder and decoder to achieve better quality in low bit rate speech coding. Some methods of speech quality measure are also provided.

Chapter 3 introduces three popular psychoacoustic models. One masking model has applications in speech enhancement and coding noise control of speech and audio. The other two models are used for objective perceptual quality measurement.

In Chapter 4, we describe our new perceptual postfilter. The idea of utilizing auditory properties for speech quality improvement is developed. The derivation of a novel perceptual postfilter is presented.

Chapter 5 details the implementation of the proposed algorithm. A detailed description of the system implementation is provided. ITU-T G.723.1 speech coder at rate of 5.3 kbps is examined. The simulation and the comparison with the conventional adaptive postfilter is presented.

Finally, Chapter 6 concludes our work and presents future work directions.

Chapter 2

Adaptive Postfiltering

In medium and low rate speech coding systems, most coders are based on an underlying model of the human speech production mechanism. The properties of human auditory system have also been considered. However, the perceptual properties are only implemented intuitively in speech coding. Most speech coders operating below 8 kbps compromise quality. The degradation in speech quality in low rate speech coders is ascribed to not only the coding method itself, but poor approximation of the properties of the human hearing mechanism.

Adaptive postfiltering is proposed to perceptually suppress audible coding noise, which is inevitable at low encoding rates. In speech perception, the formants of speech are perceptually much more important than spectral valley regions. Conventional adaptive postfiltering algorithm uses a strategy of sacrificing valley regions and preserving the formants.

This chapter begins with a description of the popular linear prediction analysis-by-synthesis (LPAS) speech coding. Perceptually motivated approaches in LPAS coders for lessening audible coding noise—noise shaping and adaptive postfiltering—are discussed. We will describe the conventional adaptive postfilter and some of its variants.

2.1 Linear Prediction Analysis-by-Synthesis Speech Coding

LPAS speech coding utilizes short-term and long-term linear prediction models for speech synthesis, and incorporates an excitation codebook which is searched during encoding to locate the best excitation sequence. It is among one of the most influential ideas in speech coding. Many standardized coders are based on LPAS principles.

2.1.1 Analysis-by-Synthesis Principle

In parametric coders and hybrid coders, a combination of parameters is used to represent the speech signal. A straightforward method to quantize each parameter is to compare its value to stored values in a quantization table, and to select the nearest quantized value. The corresponding index of this value is then stored or transmitted, and used to retrieve the quantized parameter value for synthesis later. This quantization method is called *open-loop* quantization. *Analysis-by-synthesis* is also known as *closed-loop* quantization [2]. It selects the quantized parameter value to synthesize a signal which gives the most accurate reconstruction of the original speech signal. The analysis-by-synthesis procedure is most effective when it is performed simultaneously for a number of parameters. The principle of an analysis-by-synthesis coder is illustrated in Fig. 2.1. In the encoder, a decoding structure identical to that used at the decoder is incorporated. For each of a large number of quantized parameter configurations, an error criterion comparing the original and reconstructed signals is computed. This criterion is usually a frequency weighted *mean squared error* (MSE) computed on the difference signal between the original and the reconstructed signals. Based on this criterion, the best configuration of the quantized coder parameters is selected and its index or indices are transmitted to the receiver. At the receiver, the decoder uses the same decoding structure as in the encoder to reconstruct the original speech signal.

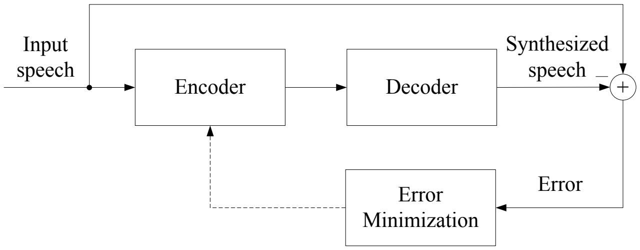


Fig. 2.1 Diagram of a simplified analysis-by-synthesis coder

Properties of speech signals constantly change with time. Speech signals are usually processed on a frame-by-frame basis. A frame consists of a certain number of samples. Within this interval, speech properties remain roughly constant. Typically, the frame length is selected between 10 and 30 ms, or 80 to 240 samples for narrowband speech sampled at 8 kHz.

2.1.2 Linear Prediction (LP)

LP [17] is based on the high correlation of consecutive speech samples: a speech signal sample could be approximately predicted by a linear combination of its past values. This is called *short-time* spectral analysis. The short-term correlations can be effectively removed from speech signals with a linear analysis filter $A(z)$

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

where a_i, $i = 1, 2, \ldots, p$, are the estimates of the *linear prediction coefficients*. The coefficients of this filter are typically updated frame-by-frame. It also gives an all-pole LP synthesis filter (often called the *LPC filter*), $1/A(z)$. The spectrum of this LPC filter (called the *LPC spectrum*) is a short-term estimate of the speech spectral envelope. The all-pole filter uses an order p between 8 and 16. A prediction order of 10 is in general enough to capture the spectral envelope [2].

The all-pole modelling is usually derived from the autocorrelation sequence of a segment of speech. The speech signal $s(n)$ is usually multiplied by a window function $w(n)$ with length N, within which speech is assumed quasi-stationary. The windowed speech segment $s_w(n)$ is

$$s_w(n) = s(n)w(n), \qquad 0 \le n \le N-1. \tag{2.2}$$

A window such as a Hamming or Hanning window is often used. The value of $s_w(n)$ is approximated by a linear combination of past values. Let $\hat{s}_w(n)$ denote the prediction

$$\hat{s}_w(n) = \sum_{i=1}^{p} a_i s_w(n-i). \tag{2.3}$$

The difference signal $e(n)$ is

$$e(n) = s_w(n) - \hat{s}_w(n) = s_w(n) - \sum_{i=1}^{p} a_i s_w(n-i). \tag{2.4}$$

The goal of LP is to minimize the total MSE of this segment

$$J = \sum_{n=0}^{N-1} e^2(n) = \sum_{n=0}^{N-1} \big(s_w(n) - \sum_{i=1}^{p} a_i s_w(n-i)\big)^2. \tag{2.5}$$

By minimizing the difference between the speech samples and the estimated signal samples, the linear prediction is formulated. The LP coefficients (LPCs) can be derived from [17]

$$\sum_{i=1}^{p} a_i R(k-i) = R(k), \qquad 1 \leq k \leq p, \tag{2.6}$$

where $R(k)$ is the autocorrelation function of the $s_w(n)$ with $R(k) = \sum_{n=0}^{N-1-k} s_w(n)s_w(n+k)$.

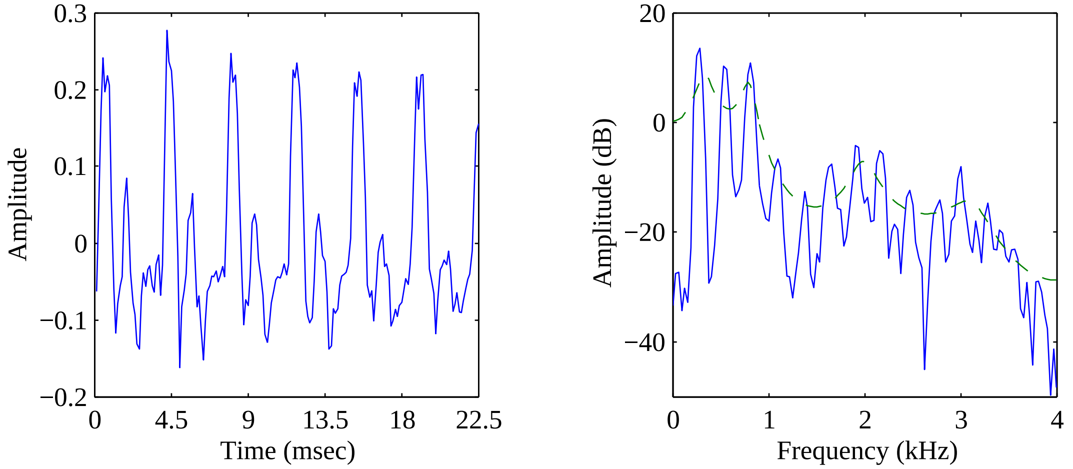
The *linear spectral frequencies* (LSFs) [18] are a popular parametric representation of the LPC filter as an alternative to LPCs. The LSFs form the roots of symmetric and antisymmetric polynomials constructed from LPCs. There is one-to-one correspondence between LPCs and LSFs [18]. Due to many desirable properties (for instance, guaranteed stability of the resultant synthesis filter after quantization), the LSFs have received widespread acceptance in speech coding applications.

The short-term synthesis filter models the short-term correlation (spectral envelope) in speech. For a segment of a speech signal, its LPC spectrum models the frequency response of the vocal tract while the fine structure in the Fourier spectrum is a manifestation of the source excitation or driving function. The predictor filter tracks the time-varying characteristics of the vocal tract. The effect of prediction in coding is the reduction of signal variance (the prediction error signal or residual has a smaller variance than that of the original signal) and whitening of the signal spectrum (the error signal is largely uncorrelated since most the signal redundancy is represented by the predictor coefficients). Fig. 2.2 shows a frame of voiced female speech with 180 samples in the time domain and in the frequency domain. Fig. 2.2.a gives its time-domain waveform. In Fig. 2.2.b, the corresponding spectrum is given. The LPC spectrum is also shown by the dashed line. The peaks in the spectral envelope are called formants, and the low parts between adjacent peaks are called valleys.

Fig. 2.3 shows the relationship between the LP analysis and synthesis filters. If the prediction error signal is the input to the LP synthesis filter, i.e. $u(n) = e(n)$, the original speech $s(n)$ is precisely recovered from the synthesized speech $\hat{s}(n)$.

Another type of LP used in speech coding is long-term LP. A long-term predictor targets correlation between samples one pitch period apart. It is also called *pitch predictor*.[1] A commonly

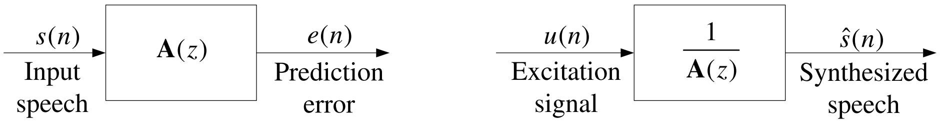
[1] For easy description, only one-tap pitch predictor is presented. However, fractional delay pitch predictor and multiple-tap pitch predictor are often applied in practical speech coders. Both of them are realized with multiple taps and can achieve a higher prediction gain than one-tap pitch predictor of Eq. (2.7).



(a) Time domain speech segment

(b) Frequency domain speech segment and its envelope (dashed line)

Fig. 2.2 A segment of voiced speech in time domain and its spectrum



(a) LP Analysis Filter

(b) LP Synthesis Filter

Fig. 2.3 Diagram of LP filters

used pitch prediction filter with input $e_s(n)$ and output $e(n)$ is

$$H(z) = 1 - g_l z^{-T}, \tag{2.7}$$

where T is the pitch period and g_l is the long-term gain. The procedure to determine g_l and T is referred to as long-term LP analysis. A long-term predictor predicts the current signal sample from a past sample that is a one or more pitch periods apart. Let $\hat{e}_s(n)$ denote the prediction of $e_s(n)$ by a long-term predictor

$$\hat{e}_s(n) = g_l e_s(n - T). \tag{2.8}$$

Within a given time interval of interest, parameters g_l and T are found by minimizing the sum of the squared error

$$J = \sum_n \big(e_s(n) - \hat{e}_s(n)\big)^2 = \sum_n \big(e_s(n) - g_l e_s(n-T)\big)^2. \tag{2.9}$$

By differentiating Eq. (2.9) with respect to g_l and equating to zero to get the optimal long-term gain, we have

$$g_l = \frac{\sum_n e_s(n)e_s(n-T)}{\sum_n e_s^2(n-T)}. \tag{2.10}$$

Substituting Eq. (2.10) back into Eq. (2.9) leads to

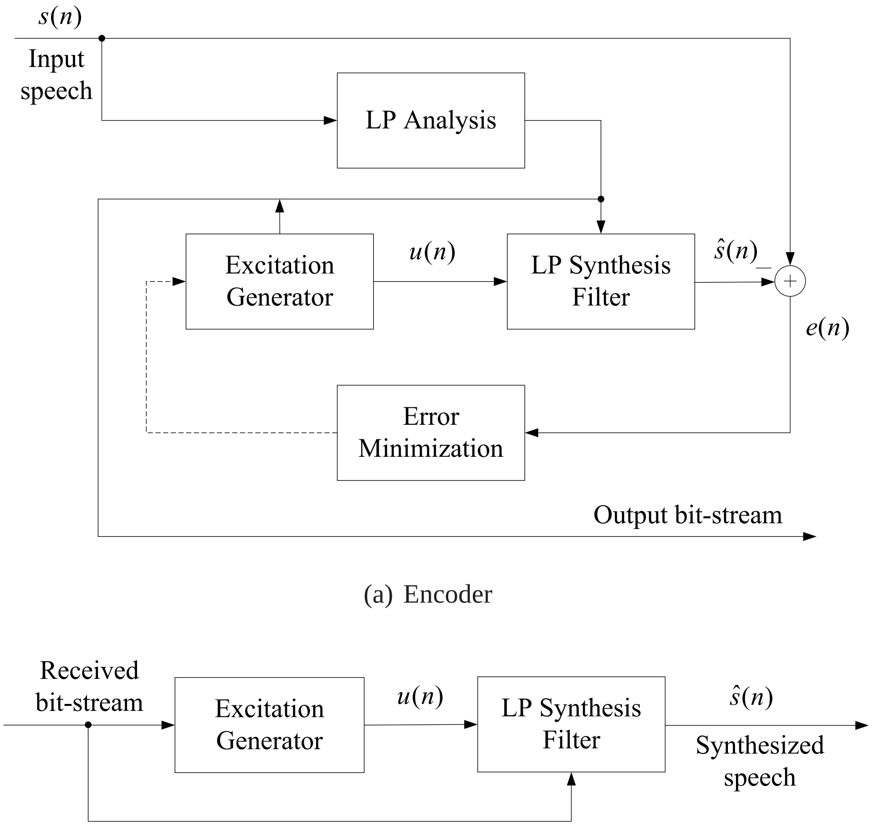
$$J = \sum_n e_s^2(n) - \frac{\big(\sum_n e_s(n)e_s(n-T)\big)^2}{\sum_n e_s^2(n-T)}. \tag{2.11}$$

An exhaustive search procedure can be applied to Eq. (2.11) to find the optimal T within a possible pitch period range $[T_{\min}, T_{\max}]$.

A pitch estimation, which is expressed as an integer multiple of the sampling interval, contains a time quantization error. This error may lead to audible distortion. Also, for periodic signals, the current period is not only similar to the previous one but also to periods that occurred multiple periods ago. Eq. (2.11) of integer pitch period estimation may cause the phenomenon of pitch multiplication and produce a multiple of the pitch period. Fractional pitch period is introduced as a means to increase temporal resolution by allowing the pitch period to have a fractional part plus the integer part [19]. Its introduction reduces both the reverberant distortion related to pitch multiplication, as well as the roughness of speakers with short pitch period.

2.1.3 Linear Prediction Analysis-by-Synthesis (LPAS) Speech Coder

The excitation signal for the LP synthesis filter in a LPAS speech coder is generated by passing each candidate excitation signal through the LP synthesis filter and comparing the synthesized speech with the original speech. In modern speech coders, the excitation is generated from a codebook or codebooks. The combination of the parameters from the codebook or codebooks that gives the least MSE is chosen. A common LPAS coder is the *Coded-Excited Linear Prediction* (CELP) coder. The excitation from CELP is composed of two components: an *adaptive codebook* contribution and a *fixed codebook* contribution. The adaptive codebook contribution models the periodicity of the excitation signal which occurs for voiced speech. It approximates



(a) Encoder

(b) Decoder

Fig. 2.4 Diagram of a generic LPAS speech coder

the excitation in the current subframe by a scaled segment of previously constructed excitation. The adaptive codebook plays the role of the *pitch-predictor synthesis filter*. The fixed codebook is used to model the part of the excitation which the adaptive codebook does not adequately model. It generates a noise-like sequence which is superimposed on the adaptive prediction to form a candidate excitation signal. Several successful models have been widely used, such as the multi-pulse model [5], the regular-pulse model [20] and the algebraic model [5, 6]. The algebraic model is the most widely used and the corresponding LP speech coder is called an *algebraic-CELP* (ACELP) coder. Fig. 2.4 shows a generic LPAS speech coder.

Generally, the LPCs are estimated from the windowed original speech signal once per frame, and then converted to LSFs and quantized. The excitation is determined and quantized over blocks which are shorter in duration than the frame, and which are referred to as subframes.

Almost all recent speech coding standards belong to the class of LPAS coders. This class includes ITU Recommendations G.723.1 [5], G.728 [21] and G.729 [6] and all the current digital cellular standards, such as EVRC [22] and SMV [7].

2.2 Distortions from LPAS Coders

LPAS speech coders can not give a satisfactory quality at bit rates below 8 kbps [1]. It suffers from a degradation described as “roughness”. In voiced speech, this distortion is more noticeable for female speech than for male speech. This can be partly explained by Skoglund and Kleijn [23]. They studied the pitch-dependent temporal behaviour of masking. Their results show that the auditory system sensitivity to low-frequency noise is strongest in the valleys between the harmonics in the spectral domain for high-pitched sounds, while the sensitivity to high-frequency noise is strongest in the valleys between the pulse peaks in the time domain for low-pitched sounds. This gives a suggestion for speech coding. For female speakers, it is important to maintain the harmonic structure of the short-term Fourier magnitude spectrum at low frequencies but that low accuracy suffices for the Fourier phase spectrum of the pitch cycle. For male speakers, more bits should be allocated to the Fourier phase spectrum of the pitch cycle, but a degradation in the harmonic structure is not audible.

In CELP speech coders, the MMSE criterion is used in the time domain for coding, which means many bits are essentially spent on the description of the phase of the pitch-cycle waveform for voiced speech. This makes the male speakers sound relatively good. However, the reconstruction accuracy of the harmonic structure of the short-term magnitude spectrum is relatively low in CELP coders. This is a result of inadequate performance by the long-term predictor.

Kroon and Altal pointed out that two major facts cause the CELP coder distortion in [24]. One fact is the shortcomings of the coding concept itself, and the other is the quantization of the side information of LP coefficients and excitation parameters. The coder itself can not reproduce high frequencies well and the rapid changes in the speech signal are not adequately tracked. Limited size of the codebook and the block-adaptation of the coder parameters may be part of the reason.

In LPAS coders, the quantization errors often lead to a deemphasis of the formant structure of the speech signal. This is shown in Fig. 2.5. After the quantization of LP coefficients of a voiced frame by ITU-T G.723.1 at rate of 5.3 kbps, the formants become lower and a bit wider.

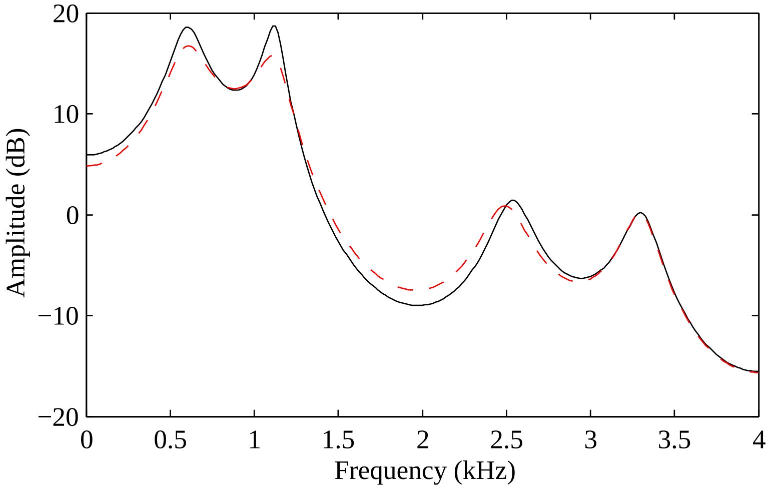


Fig. 2.5 LP Spectra and its quantized version (dashed line)

2.3 Masking

All waveform coders, which use properties of human hearing to keep the perceptual distortion low, rely on auditory masking. Masking is the property that one signal, the *masker*, can render another signal, the *maskee*, inaudible [2]. In the case of speech and audio coding the masker is the input signal and the maskee is the error signal or coding/quantization noise, as $s(n)$ and $e(n)$ shown in Fig. 2.4.a.

Masking phenomena are common in sensory perception. Masking reflects limited frequency and temporal resolutions of human hearing system. Generally, there are two masking effects: *simultaneous masking* and *temporal masking*. Simultaneous masking occurs when two or more stimuli in different frequencies are presented at the same time. It is the most significant masking property, since it produces the largest amount of masking. Temporal masking occurs when the masker and maskee have a temporal offset with respect to each other. The masker and the maskee are presented close in time, but not simultaneously. When the maskee is presented prior to the masker onset, it is called *backward masking* (see [25]), while *forward masking* happens when the maskee is present after the masker is turned off. Backward masking is considered far less important. Forward masking is the more prominent form of temporal masking.

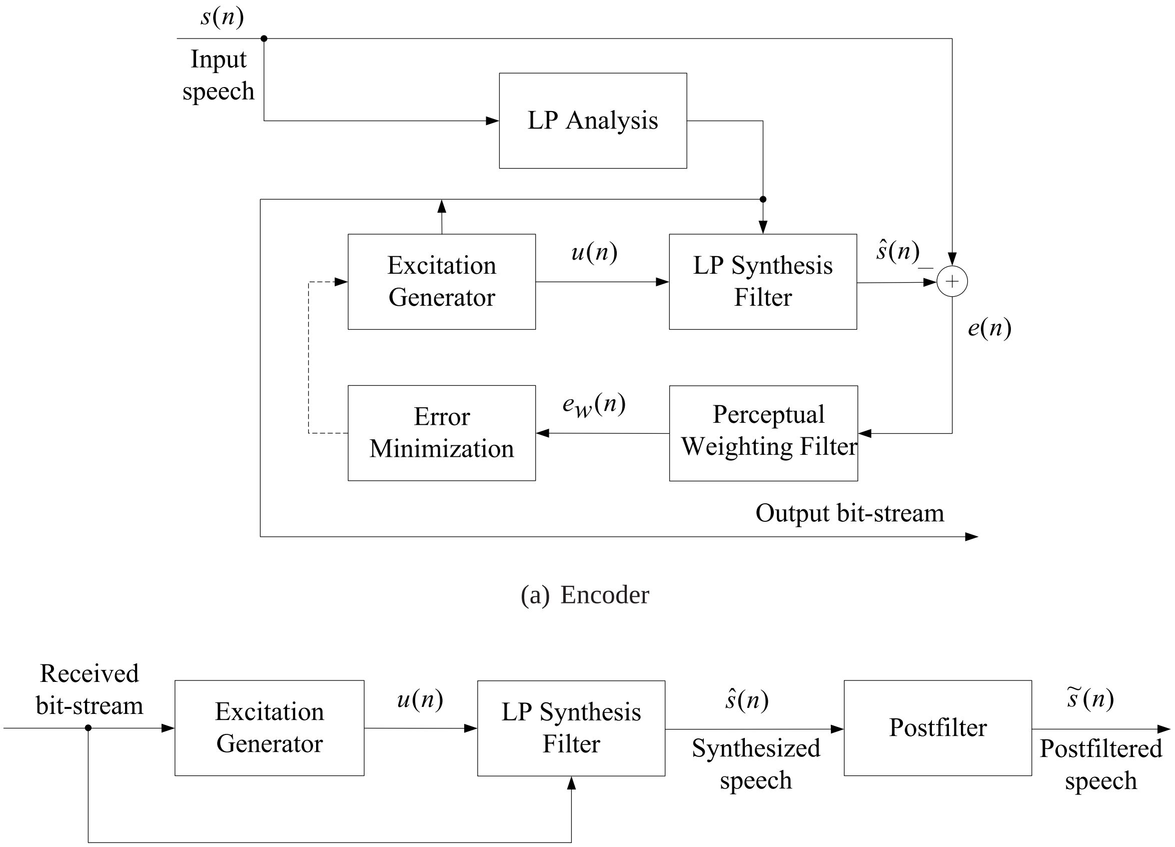
As we discussed in Chapter 1, an isolated stimulus is audible if it has a sufficiently high level and a frequency content that falls within the audible range. This is measured by the *absolute threshold of hearing*. In a masking condition, for the stimulus to be audible in the presence of a masker, its level has to be higher than the so-called *masking threshold*. Masking threshold is the combination effect of both simultaneous masking and temporal masking as well as simultaneous maskers. An optimal coding scenario is that all coding noise lies below the masking threshold. However, only the masking concept and the empirical masking properties are adopted in speech coding standards and the “true” masking threshold is never computed.

2.4 Perceptual Properties Applied in LPAS Speech Coders

Speech coding is related to human perception, and therefore a degree of fuzziness exists, in the sense that no absolute right or wrong can be established for certain situations. Therefore, solutions are often presented and justified on an empirical basis.

Lowering the bit rate of a codec by employing powerful coding techniques will result in higher distortion, but, by exploiting knowledge about the human auditory system, techniques that mask the distortion can achieve high perceptual quality at lower bit rates. Two perceptually-based approaches are widely use in LPAS speech coders: *noise spectral shaping* and *adaptive postfiltering*. At low encoding rates, it is impossible to push all of the coding noise under the masking threshold in both formant and valley regions. Noise spectral shaping is used to make the coding noise to follow the speech LP spectrum. A perceptual weighting filter is applied in a speech encoder to shape the coding noise. This is based on the assumption that the original speech has most of its energy in the spectral formant regions, and more noise is masked in these regions. Noise spectral shaping alone is not sufficient to make the coding noise inaudible at low coding rates. Lowering noise components at certain frequencies can only be achieved at the price of increased noise components at other frequencies [15]. It is very difficult to force noise below the masking threshold at all frequencies at a low bit rate. While the coding noise spectral shaping follows the speech spectrum, most of the perceived coding noise comes from spectral valleys, including the valleys between pitch harmonic peaks. However, an adaptive postfilter [15] is introduced to attenuate these noise components at the speech decoder output. A useful postfilter may attenuate the frequency components between pitch harmonics as well as the components between formants, while the spectral envelope peaks corresponding to the formants have roughly the same height as before the postfilter. However, noise spectral shaping in coding only affects coding noise, while adaptive postfiltering in decoding has to modify both speech and noise.

Fig. 2.6 illustrates the LPAS speech coder that incorporates the perceptually-motivated approaches: *perceptual weighting filter* at the encoder and *adaptive postfilter* at the decoder.



(a) Encoder

(b) Decoder

Fig. 2.6 Diagram of a LPAS speech coder with perceptual approaches

2.4.1 Noise Shaping

Auditory masking theory motivates the use of noise shaping in speech encoding. An unweighted MMSE criterion for speech signal does not ensure perceptually low distortion. It is important to consider the relationship between the spectrum of the quantization noise and the spectrum of the speech signal to achieve perceptually low distortion. Since most of the noise in the formant regions could be partially or totally masked by the speech, a large portion of perceived noise comes from spectral valleys.

Atal and Schroeder [26] proposed noise spectral shaping first in 1979. The basic idea is to shape the spectrum of the coding noise so that it follows the speech spectrum to some extent. Due to the masking effect of human auditory system, the spectrally shaped coding noise is less audible to human ears. In modern LPAS speech coders, a *perceptual weighting filter* is widely used. It has the form of

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \tag{2.12}$$

where $0 < \gamma_2 < \gamma_1 < 1$, and $A(z)$ is the LP analysis filter that is defined in Eq. (2.1). The perceptual weighting filter shapes the coding noise to follow the trend of the spectral peaks and valleys of the speech spectrum, which gives a no-longer white coding noise and makes the coding noise less audible to human ears. In ITU-T G.723.1 [5], the perceptual weighting filter has $\gamma_1 = 0.9$ and $\gamma_2 = 0.5$. Fig. 2.7 shows the frequency response of the perceptual weighting filter with the LP spectrum from Fig. 2.2.

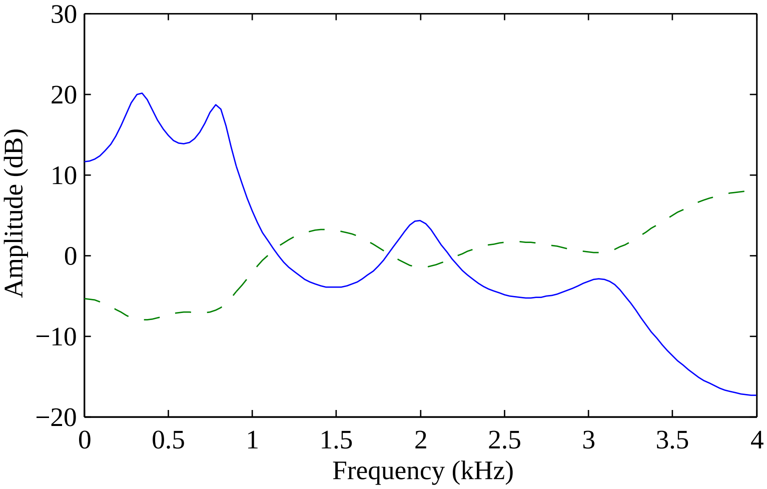


Fig. 2.7 Perceptual weighting filter response (dashed line) corresponding to the LP spectrum (solid line)

While properly tuned, the perceptual weighting filter allows more noise in the formant regions, but below the masking threshold, and decrease the amount of quantization noise in the spectral valleys. The noise components in some valley regions may exceed the masking threshold. This audible noise affects the perceptual quality of the speech.

2.4.2 Adaptive Postfiltering

Postfiltering is used as a postprocessing technique at the decoder to enhance the reconstructed speech. According to Chen [15], Smith and Allen first proposed a postfilter in 1981 for enhancing the output of an adaptive delta modulation. The usage became popular until 1984 when

Ramamoorthy and Jayant proposed a new postfiltering technique in [27] to move the poles and zeros of the synthesis filter radially toward the origin by suitably chosen factors. It was further developed in [28, 29]. Adaptive postfiltering was first combined with noise spectral shaping in a speech coder in 1986 by Yatsuzuka, Iizuka and Yamazaki [30]. They were also the first to propose explicitly an additional long-term postfilter section based on the pitch periodicity in speech. However, the postfilters mentioned above had a muffling (low-pass) effect of the speech sound. Chen proposed a postfilter which significantly reduced the low-pass effect in [31] in 1987. This postfilter was elaborated in [15]. Since 1987, the use of the postfilter proposed by Chen in CELP-like coders has become very popular. It has become a common part of speech coding standards based on LPAS .

An adaptive postfilter is preferred due to the variations of the local characteristics of speech spectrum. It is usually used on a frame-by-frame basis, and thus is based on the local characteristics of the speech spectrum. Adaptive postfiltering is based on empirical results for low bit rate coders [15]: a) the masking threshold follows to some extent the spectral peaks and valleys of the speech spectrum; b) the noise shaping by a perceptual weighting filter at the encoder makes the coding noise fall below the masking threshold around the spectral peaks but appear above the masking threshold in the spectral valleys. While attenuating audible noise in some valley regions, the speech components in these regions will also be attenuated. Fortunately, the intensity of the spectral valleys can be altered as large as 10 dB without any audible effect [32]. Therefore, by doing so, the postfilter can achieve a substantial noise reduction with only minimal perceptual distortion of the speech itself. Unlike weighting at the encoder (where the clean speech signal is available) which shapes the coding noise only, postfiltering at the decoder affects both the speech and the coding noise. It is a compromise between speech distortion and noise reduction. Other than the conventional postfilter by Chen, other postfiltering algorithms were also proposed by researchers in [16, 33, 34].

2.5 Adaptive Postfiltering

We will discuss conventional postfiltering [15] in detail in this section. Some variations [16, 34] will also be introduced.

2.5.1 Conventional Postfilter

According to speech perception, the formants of speech are perceptually much more important than spectral valley regions. At low coding rates, even though a perceptual weighting filter is applied, it is impossible to push all the noise components below the masking threshold. The noise components in some of the valley regions may exceed the threshold, which makes most of the perceived coding noise coming from spectral valleys, including the valleys between pitch harmonic peaks. As mentioned in Section 2.2, in LPAS speech coding, quantization errors often lead to a deemphasis of the formant structure and a decreased periodicity. To have more flexibility in the shape of the postfilter, the adaptive postfilter proposed in [15] contained elaborate short-term and long-term postfilter sections which achieved significant noise reduction by emphasizing the formant structure and increasing the periodicity, respectively. A general postfilter transfer function is given by

$$H(z) = GH_l(z)H_s(z), \tag{2.13}$$

where $H_l(z)$ represents a long-term postfilter, $H_s(z)$ represents a short-term postfilter and G is an overall gain factor. The long-term postfilter emphasizes pitch harmonics and attenuates the spectral valleys between pitch harmonics. It is also called a *pitch postfilter*. On the other hand, the short-term postfilter emphasizes speech formants and attenuates the spectral valleys between formant. It is also called a *formant postfilter*.

The general postfilter depends on both the short-term and long-term correlations in the speech signal. This information usually is transmitted to the decoder in most LPAS coders. However, the postfilters can derive this information from the decoded speech signal. In some implementations it was found that, even when the parameters are transmitted, it is better to recompute them, to take into account the interaction effects with the excitation signal [2]. Moreover, in many implementations the postfilter is integrated with the decoder synthesis filter and does not just operate on the reconstructed output signal. For example, long-term postfiltering is usually done on the excitation signal, so that the LP synthesis filter can smooth out discontinuities, which appear at frame boundaries where the long-term postfilter is updated. This is the case in a number of standard speech coders [5–7].

Short-Term Postfilter

The frequency response of an ideal short-term postfilter should follow peaks and valleys of the spectral envelope of speech without giving an overall spectral tilt. Since the LP synthesis filter spectrum closely follows the spectral envelope of the input speech, it is natural to derive the short-term postfilter from the LPC predictor. Conventionally, a short-term postfilter is given by [15]

$$H_s(z) = \frac{A(z/\lambda_1)}{A(z/\lambda_2)}(1 - \mu z^{-1}), \tag{2.14}$$

where $0 < \lambda_1 < \lambda_2 < 1$. The optimal values of λ_1 and λ_2 depend on the bit rate and the type of the speech coder used. They generally need to be determined empirically based on subjective listening tests. The difference between λ_1 and λ_2 introduces a low pass spectral tilt in the spectrum, which makes the voiced speech muffled. The first-order filter with a transfer function $(1 - \mu z^{-1})$ is used to reduce the lowpass effect. It is referred to as the tilt-compensation filter. It is usually made to be adaptive to better track the spectral tilt of $A(z/\lambda_1)/A(z/\lambda_2)$. For example, in ITU-T G.723.1 [5], the short-term postfilter is given by the following equations:

$$H_s(z) = \frac{A(z/\lambda_1)}{A(z/\lambda_2)}(1 - 0.25k_1 z^{-1}), \tag{2.15a}$$

$$k_1 = \frac{3}{4}k_{1\mathrm{old}} + \frac{1}{4}k, \tag{2.15b}$$

where $\lambda_1 = 0.65$, $\lambda_2 = 0.75$, k is the first reflection coefficient and $k_{1\mathrm{old}}$ is the value of k_1 from the previous subframe. $k = R[1]/R[0]$ is estimated from a subframe of the synthesized speech. $R[0]$ and R[1] are the autocorrelation values of the corresponding subframe. In Eq. (2.15a), the tilt factor is made adaptive as a function of the overall spectral slope of the input signal.

Long-Term Postfilter

The function of a long-term postfilter is to attenuate frequency components between pitch harmonic peaks. Also, no overall spectral tilt should be introduced. Such a long-term postfilter is typically derived from the pitch predictor. Since zeros in a transfer function can provide more flexibility and more control of the frequency response, the long-term postfilter with both poles and zeros can be represented by the following function

$$H_l(z) = G_l \frac{1+\alpha_1 z^{-T}}{1-\alpha_2 z^{-T}}, \tag{2.16}$$

where G_l is an adaptive scaling factor, T is the pitch period and $0 < \alpha_1, \alpha_2 < 1$. The coefficients, G_l, α_1 and α_2, are determined by the degree of periodicity in speech. In ITU-T G.723.1 [5], the long-term postfilter is of the form

$$H_l(z) = G_l(1+\alpha_1 z^{-T}), \tag{2.17}$$

where G_l is an overall gain which is chosen to make the energy of the output signal equal to the energy of the input signal, and α_1, T are derived from the decoded excitation signal. G_l is the square root of the ratio between the energies of the input signal and the postfiltered signal. T can only be positive [5, 35].

From Eq. (2.16), the long-term postfilter has its own scaling factor G_l, but the short-term postfilter does not have a similar scaling factor. In general, the power gain of the short-term postfilter would be high for those speech frames where the prediction gain of the LPC predictor is high, and vice-versa. The gain factor G in Eq. (2.13) is needed to ensure that the energy of the postfiltered signal is the same as that of the input signal before postfiltering. To avoid possible discontinuities, the scaling factor is lowpass filtered. For example, in [5], the gain is updated on a sample by sample basis using

$$g(n) = ag(n-1) + (1-a)g_s, \tag{2.18}$$

where g_s is the square root of the ratio between the energies of the input signal and the short-term postfiltered signal and $a = 15/16$. Each sample of the short-term postfiltered output signal is multiplied with the corresponding value of $g(n)$.

Fig. 2.8 shows an example of the postfilter frequency response of Eq. (2.13) for a segment of voiced speech.

The conventional postfiltering technique has been implemented successfully. It has been widely used in modern speech coders such as ITU-T Recommendation for multimedia communication G.723.1 [5] and G.729 [6].

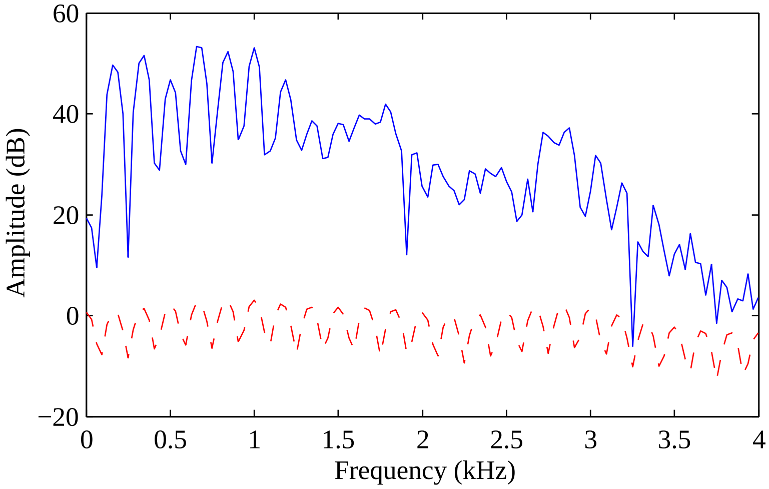


Fig. 2.8 An example of speech spectrum and the corresponding overall postfilter frequency response (dashed line)

2.5.2 Adaptive Formant Postfilter Proposed by Mustapha and Yeldener [16]

Conventional postfiltering uses the same constants, λ_1 and λ_2, for all of the formants and causes the formants to be weighted in the same way. However, it is difficult to adapt these coefficients from one frame to another and still produce a postfilter without spectral tilt. Conventional time-domain postfiltering produces varying spectral tilt from one frame to another affecting speech quality. The parameters of the high-pass tilt compensation filter are difficult to control well. The purpose of the tilt-compensation filter in Eq. (2.14) is to compensate the tilt of the first part of Eq. (2.14) so as to reduce the lowpass effect. The coefficient μ is made adaptively proportional to the first reflection coefficient k. For highly correlated voiced frames, $k = R[1]/R[0] \approx 1$. For proper λ_1 and λ_2, the resulting postfilter tends to have less spectral tilt, but preserves the peaks and valleys. For unvoiced frames, however, the magnitude of k tends to decrease, and k might change from positive to negative. This is due to the fact that correlation among adjacent samples is weakened. Also, the spectra of unvoiced frames tend to develop a high-pass tilt. Therefore, it is better to either diminish the amount of tilt compensation or even change to lowpass filtering in order to cancel the high-pass tilt [1]. However, this is not an easy task.

The performance of conventional postfiltering is not optimal without adjusting the postfilter parameters λ_1 and λ_2 (see [36]). Mustapha and Yeldener [16, 37] developed a new time-domain postfiltering technique which eliminates the problem of spectral tilt in speech spectrum and can be applied to various speech coders. This postfilter uses the pole information in the LPC spectrum and finds the relation between poles and formants. The formants, nulls and their bandwidths are first found to get a desired postfilter response. A modified least squares approach based on the modified Yule-Walker (MYW) method is used to give a postfilter with better speech quality than the conventional technique.

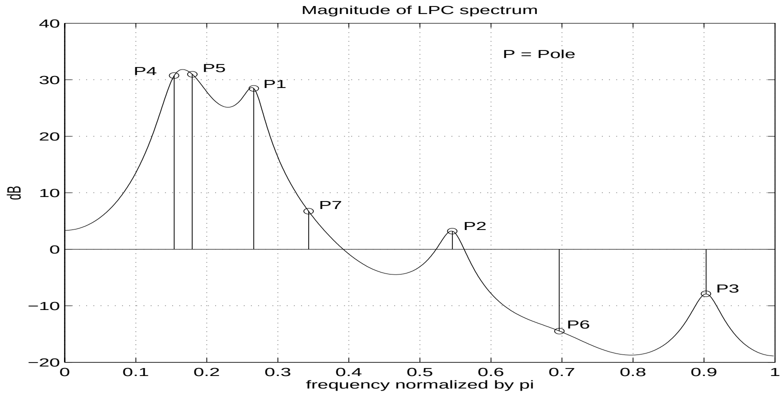


Fig. 2.9 LP frequency spectrum for the modified Yule-Walker method [16]

The new postfilter is based on the MMSE approach as

$$E = \sum_{k=0}^{L-1} \big(d(n) - h(n)\big)^2, \tag{2.19}$$

where $d(n)$ and $h(n)$ are the impulse responses of the desired and estimated postfilters, respectively. The transfer function of the estimated postfilter based on MYW filter is

$$H(z) = \frac{B(z)}{A(z)} = \frac{b_0 + \displaystyle\sum_{k=1}^{N} b_k z^{-k}}{1 + \displaystyle\sum_{k=1}^{M} a_k z^{-k}}. \tag{2.20}$$

In the desired postfilter, the aim is to preserve the formant information. Therefore, the postfilter has a unity gain in the formant regions of spectrum. Outside of the formant regions, the aim is to have some controllable attenuation factor, τ, that controls the depth of the postfiltering. In [16], τ is set to 0.6. However, τ is adaptable from one frame to another depending on how much postfiltering is needed and the type of the speech coder used. For a LPC spectrum as Fig. 2.9, the frequency response of the desired postfilter is shown in Fig. 2.10.

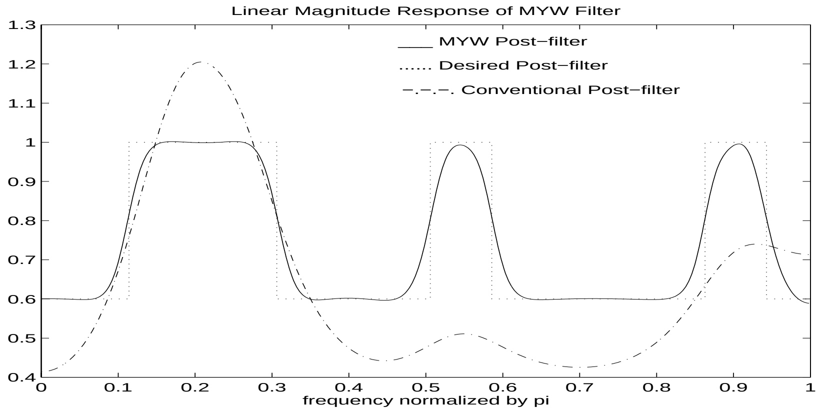


Fig. 2.10 Frequency response of postfilters (the modified Yule-Walker method [16])

The denominator coefficients of the filter $A(z)$ are computed from the autocorrelation method for LP. The autocorrelation coefficients are derived from the power spectrum of the desired postfilter by inverse Fourier transformation. The numerator coefficients of the filter $B(z)$ are computed as in [38]. This postfilter has a flat frequency response that overcomes the spectral tilt and other problems present in conventional postfilter mentioned earlier. Fig. 2.10 also shows the frequency response of the estimated postfilter and the conventional short-term postfilter with corresponding LPC spectrum in Fig 2.9. It is clear that the formant peaks are flat in the frequency response of the new MYW postfilter, while those of the conventional one are not. The new and conventional postfilter LPC spectra are shown in Fig 2.11. It is also clear that the new postfilter has no spectral tilt at all comparing with the original LPC spectrum, while the conventional one has a spectral tilt. The new postfilter has the desired property of preserving the formant peaks and attenuating the nulls. Furthermore, the attenuation of nulls is more controllable in the new postfilter than the conventional one.

2.5.3 Adaptive Pitch Postfilter Proposed by Kleijn [34]

In order to emphasis the coded speech spectrum to improve the quality, Kleijn [34] gave an enhancement algorithm based on constrained optimization to enhance speech fine structure. It can be considered as a long-term postfilter. The spectral fine structure offers particular large potential for enhancement because of the large dynamic range of the harmonic structure of voiced speech.

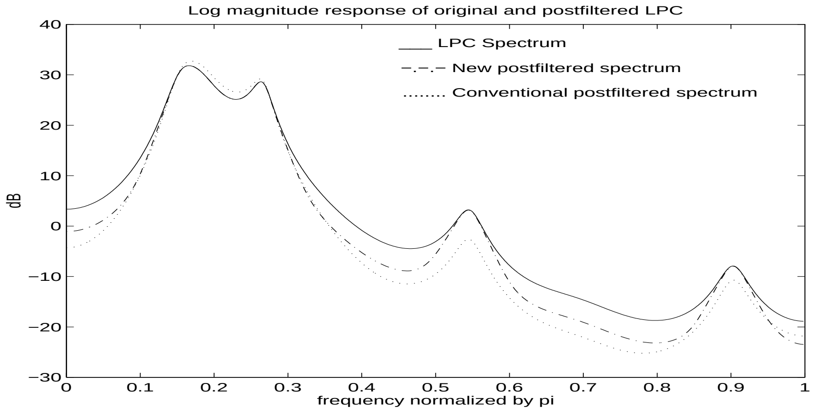


Fig. 2.11 Postfiltered LPC spectra (for the modified Yule-Walker method [16])

Conventional adaptive postfilters often give a spectral emphasis that is too strong or too weak within different segments of a signal. [34] also points out that the time synchronization between the spectral envelope and the spectral fine structure is generally incorrect in current fine-structure postfilters, because the inherent delay is neglected.

The criterion is to increase the periodicity of the speech signal on a block-by-block basis. Two constraints are applied. One is to ensure the preservation of the signal power, and the other is a modification constraint to ensure that the power of the difference signal between the enhanced and unenhanced signal is less than a fraction of the power of the unenhanced signal. This method can increase the periodicity of voiced speech segment, while unvoiced speech segments are perceptually unaffected due to the modification constraint.

Let $\mathbf{s}_j$ be a discrete speech segment of K subsequent speech samples, with time label j. $\mathbf{s}_{j,m}$ denotes a sample sequence of K samples, and each sample in this sequence is m pitch cycles removed from the corresponding sample of the sequence $\mathbf{s}_j = \mathbf{s}_{j,0}$. $\mathbf{s}_{j,m}$ and $\mathbf{s}_{j,m+1}$ can overlap. Let $\hat{\mathbf{s}}_j$ be the enhanced segment corresponding to $\mathbf{s}_j$. The measure of periodicity of the enhanced signal is given as

$$\eta_{\mathcal{J}} = \sum_{j \in \mathcal{J}} \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \langle \hat{\mathbf{s}}_j, \hat{\mathbf{s}}_{j,m} \rangle, \tag{2.21}$$

with constraints:

$$\|\hat{\mathbf{s}}_j\| = \|\mathbf{s}_j\|, \tag{2.22a}$$

$$\|\mathbf{s}_j - \hat{\mathbf{s}}_j\|^2 \leq \beta \|\mathbf{s}_j\|^2, \tag{2.22b}$$

where α_m describes a discrete window function, $\mathcal{I}$ is a set of integers that describes the support of this window (e.g., $\mathcal{I} = \{-3, -2, \cdots, 3\}$, $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, $\| \cdot \|$ denotes the Euclidean norm ($\langle \mathrm{s}, \mathrm{s} \rangle = \|\mathrm{s}\|^2$), and $\mathcal{J}$ is a set of consecutive-block indices. The window $\{\alpha_m\}_{m \in \mathcal{I}}$ should be defined based on perception and $\beta \in [0, 1]$. The criterion in Eq. (2.21) can be maximized by iteratively maximizing the criteria

$$\eta_j = \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \langle \hat{\mathbf{s}}_j, \mathbf{s}_{j,m} \rangle. \tag{2.23}$$

In order to simplify the procedure, one iteration for each η_j defined in Eq. (2.23) is maximized based on the original $\mathrm{s}_{j,m}$. The constraints are applied to the individual optimizations.

This algorithm has been implemented in iLBC [39] for Internet coding of 8 kHz sampled speech by Global IP Sound (GIPS). The sequence length $K = 80$, and $\{\alpha_m\}_{m \in \mathcal{I}}$ is set to a Hanning window with seven-sample support. The parameter β is set to 0.05, corresponding to a signal to modification power ratio of about 13 dB. The constraints contribute to inherent robustness by preventing large changes to the signal.

2.6 Speech Quality Assessment

Speech quality assessment is of primary concern in speech coding and speech enhancement. There are many dimensions in quality perception, and *intelligibility* and *naturalness* are the most important. In digital communications speech quality is classified into four general categories [40]:

- *commentary or broadband* quality refers to wide-bandwidth (typically 50–7000 Hz, but 20–20,000 Hz for compact disk) high-quality speech that can generally be achieved at rates, at least 32–64 kbps

- *network or toll or wireline* quality describes speech as heard over the switched telephone network (approximately the 300–3400 Hz bandwidth range, with a signal-to-noise ratio of more than 30 dB and with less than 2–3% harmonic distortion). It can be achieved at rate between 8 kbps and 32 kbps.

- *communications* quality implies somewhat degraded speech quality which is natural and highly intelligible. Communications speech can be achieved at rates above 4 kbps.

- *synthetic* speech is usually intelligible but can be unnatural and associated with a loss of speaker recognizability.

To establish a fair means of comparing speech coding or enhancement algorithms, a variety of quality assessment techniques have been formulated. Generally speaking, tests fall into two classes: *subjective quality measures* and *objective quality measures*. Subjective measures are based on comparisons of original and the processed speech by a listener or group of listeners, who subjectively rank the quality of speech along a predetermined scale. Objective quality measures are based on a mathematical comparison of the original and the processed speech signals. Most objective quality measures quantify quality with a numerical distance measure or a model of how the auditory system interprets quality.

2.6.1 Subjective Quality Measures

In subjective testing, the individual ratings are gathered and averaged to yield the final score. The test is normally done for a wide variety of conditions so as to obtain a general performance appreciation for a particular coder. There are three commonly used procedures to perform subjective testing [1]

- *Absolute Category Rating* (ACR): The listeners are required to make a single rating for each speech passage. Five choices are given in Table 2.1. The average of all votes is known as the *mean opinion score* (MOS). The MOS is a widely used measure to quantify coded speech quality. It usually involves 12–24 listeners.

Table 2.1 MOS Five-Point Scale [40]

Rating	Speech Quality	Level of Distoriton
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Bad	Very annoying and objectionable

- *Degradation Category Rating* (DCR): In this test, the listeners are presented with the original signal as a reference before they listen to the synthetic signal, and are asked to compare the two and give a rating according to the amount of degradation perceived. The five choices are given in Table 2.2. The average of all votes is known as the degradation mean opinion score (DMOS).

Table 2.2 DMOS Five-Point Scale [1]

Rating	Level of Degradation
5	Not perceived
4	Perceived but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

- *Comparison Category Rating* (CCR): In the DCR test, the final score might be biased because of the order by with the speech materials are presented. A better approach is to present two samples and ask the listeners to compare and rate the second with respect to the first. The order of the original speech and the processed speech can be made arbitrary or random. The choices are given in Table 2.3.

Table 2.3 CCR Scale [1]

Rating	Level of Comparison
3	Much better
2	Better
1	Slightly better
0	About the same
-1	Slightly worse
-2	Worse
-3	Much worse

Pair comparison [8], sometimes called A-B test, is also commonly used for informal speech quality tests. In the pair comparison test, each test utterance is compared with various other utterances, and the fraction that the test utterance is judged to be better than the other utterances is calculated as the preference score.

Since the goal for coding and enhancement is to produce speech that is perceived by the auditory system to be natural and free of degradation, it is understandable that subjective quality measures are the preferable means for quality assessment. However, it is clear that subjective tests are expensive to implement and highly time consuming. Therefore, it is desirable to build objective evaluation methods producing evaluation results which correspond well with the subjective evaluation results. Current research efforts are being directed toward perceptually-based objective measures.

2.6.2 Objective Quality Measures

Objective speech quality measures are reliable, repeatable, easy to implement and in some cases have been shown to be good predictors of subjective quality.

In the time domain, some forms of signal-to-noise ratio are the major types of objective measures.

- *Signal-to-Noise Ratio* (SNR) The SNR is the most widely used measure for analog and waveform coding systems. Given the original speech $x(n)$ and the processed version $y(n)$, the SNR is defined by

$$\mathrm{SNR} = 10\,\log_{10}\left(\frac{\sum\limits_{n} x^2(n)}{\sum\limits_{n}\big(x(n)-y(n)\big)^2}\right) \tag{2.24}$$

  with the range of the time index n covering the measurement interval.

- *Segmental Signal-to-Noise Ratio* (SEGSNR) The SNR is a long-term measure for the accuracy of speech reconstruction. It tends to ignore temporal noise, which could affect the perceived quality significantly. SEGSNR is a frame-based measure. It is an average of SNR values obtained from isolated frames, with the frame being a block of samples (typically 15–25 ms). The definition of SEGSNR is

$$\mathrm{SEGSNR} = \frac{1}{N}\sum_{m=1}^{N}\mathrm{SNR}_m, \tag{2.25}$$

  where SNR_m is the SNR value of the mth frame. SEGSNR compensates for the underemphasis of the weak-signal performance in conventional SNR measure.

The SNR and SEGSNR are only meaningful for waveform reconstruction. They are extremely sensitive to waveform misalignments and phase distortion, which are not always perceptually relevant [1]. However, most low bit rate coders do not preserve the original speech waveform. For low bit rate coding, LPC spectrum preservation is essential to perceived quality. Some distortion measures in spectral domain, such as *Itakura Measure, Log Spectral Distortion Measure* and *Weighted Euclidean Distance Measure* [41], have been proposed for low bit rate coders. These objective measures are based on the comparison of LP spectra of the original speech and the processed speech.

In speech processing, the root-mean-square (RMS) log spectral measure is used to determine the error or difference between two spectral models on a log magnitude versus frequency scale [41]. A similar measure is the *spectral distortion* (SD) [1], which has become the standard measure for evaluating the performance of spectrum coding. SD is defined as

$$\mathrm{SD}^2 = \frac{1}{2\pi}\int_0^{2\pi}\left(10\,\log_{10}\frac{S(e^{j\omega})}{\tilde{S}(e^{j\omega})}\right)^2 d\omega, \tag{2.26}$$

where $S(e^{j\omega})$ and $\tilde{S}(e^{j\omega})$ are the *power spectral densities* (PSDs) of the original and estimated synthesis autoregressive signals with LP coefficients $a_i, \tilde{a}_i$, $i = 1, 2, \cdots, p$, and input noise variances $g, \tilde{g}$, respectively, for a current frame. Thus

$$S(e^{j\omega}) = \frac{g}{|A(e^{j\omega})|^2}, \tag{2.27}$$

$$\tilde{S}(e^{j\omega}) = \frac{\tilde{g}}{|\tilde{A}(e^{j\omega})|^2}, \tag{2.28}$$

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i}, \quad \tilde{A}(z) = 1 + \sum_{i=1}^{p} \tilde{a}_i z^{-i},$$

where p is the prediction order.

When the integral in Eq. (2.26) is approximated in practice by a N_F-point *fast Fourier transform* (FFT), the relation in Eq. (2.26) may be rewritten for full-band spectral distortion as

$$\mathrm{SD}^2 = \frac{1}{N_F/2+1}\sum_{k=0}^{N_F/2}\left(10\,\log_{10} S(e^{j\omega_k}) - 10\,\log_{10}\tilde{S}(e^{j\omega_k})\right)^2. \tag{2.29}$$

In practice, it is often written as

$$\text{SD}^2 = \frac{1}{n_1 - n_0} \sum_{k=n_0}^{n_1} \left(10 \, \log_{10} S(e^{j\omega_k}) - 10 \, \log_{10} \tilde{S}(e^{j\omega_k}) \right)^2, \tag{2.30}$$

where $0 \leq n_0 \leq n_1$. Typically, if the sampling frequency is f_s, for $f_s = 8$ kHz, $n_0 = 4$, $n_1 = 100$, and $N_F = 256$, so that only the spectrum values between 125 Hz and approximately 3.1 kHz are taken into account for the computation of SD. Thus only the most perceptually sensitive part of the spectrum is considered.

The average SD has been used extensively to measure the performance of LP coefficient quantizers. It is highly desirable to reduce the average SD as well as the number of outlier frames. However, SD does not account for the frequency-domain or time-domain masking effects of the human auditory system. Therefore, it might not totally correlate with subjective evaluation results.

The objective measures mentioned above are only related to a numeric distance, while the perceptual properties of the human ear are ignored. Ideally, the outcomes of the objective tests should be highly correlated with the subjective test scores. Since the 1980s, the ITU has been investigating many proposals for objective quality measurements based on psychoacoustic sound perception modelling. Objective quality measurements should give a MOS value. The most well-known one is the ITU-T Recommendation P.861 [42], the *perceptual speech quality measure* (PSQM) algorithm, which is correlated well with the subjective quality of coded speech. In 2001, ITU-T finalized another refined method through recommendation P.862 [43] to replace P.861, and make it suitable for real systems which include filtering and variable delay, as well as distortions due to channel errors. ITU-T P.862 uses the *Perceptual Evaluation of Speech Quality* (PESQ) algorithm for cognitive perceptual model. For wide band audio codecs, ITU-R recommended the PEAQ algorithm implemented in recommendation ITU-R BS.1387 [13].

Chapter 3

Psychoacoustic Models

Nowadays, audio coding applications often use a psychoacoustic model. Our human ear is a rather complex system. To model the human auditory system, masking models are usually applied [13, 44, 45]. A masking model delivers a masking threshold along with the amount of the allowable distortion in the frequency domain. Signal energy lying below the masking threshold is inaudible. In audio processing, masking is used for bit allocation and audio enhancement.

Another field of extensive interest of psychoacoustic models is objective measurements of perceived speech and audio quality. In speech and audio coding, the quality can be determined either objectively or subjectively. Subjective tests are difficult to reproduce. It is also expensive, and time consuming. Therefore, objective quality measurement methods are in great demand. Objective methods map the signals for comparison onto an internal representation which is as close as possible to the subjective quality domain. Various perceptual models have been proposed with different levels of accuracy and complexity. ITU has proposed some recommendations for speech and audio codecs, such as ITU-T Recommendation P.862 (PESQ) for speech, and ITU-R Recommendation BS.1387 (PEAQ) for audio.

This chapter will present three auditory models. One model, Johnston's model, is related to audio coding. The other two models, PAQM and PESQ, are related to objective perceptual measure of audio quality.

3.1 Critical Bands

The frequency resolution of our human ear is represented by *critical bands*, which have nonlinear mapping to the frequency value (Hz) of the stimulus. The ear integrates signal energy within a critical band, which makes it difficult to separate signals within one critical bandwidth for a human observer. The overall energy of the masker affects perception. Critical bands correspond to approximately 1.5 mm spacings along the basilar membrane and are scaled by *Bark*. One Bark spans the width of a critical band. A Bark bandwidth is smaller at low frequencies (in Hz) and larger at high ones. Schroeder et al. [46] proposed an expression to relate the frequency and critical band rate

$$z = 7\,\text{arc}\sinh(f/650). \tag{3.1}$$

It is almost linear below 500 Hz and exponential above 1 kHz. There are also other expressions for Hz to Bark transformation, such as *equivalent rectangular bandwidth* (ERB) [12].

3.2 Johnston's Masking Model

A masking threshold is derived by weighting an excitation pattern, which is obtained by frequency and time domain spreading. The excitation pattern predicts the physical activity of hair cells along the basilar membrane in the ear. Johnston [44] proposed a masking model to shape the quantization noise to below the masking threshold in a transform coder. It operates on 64 ms frames of 15 kHz audio signals which are sampled at 32 kHz. The square root of a Hanning window is used for the 1/16th overlapped section of each frame. This model calculates the short-term spectral masking threshold to determine the noise-shaping function for the coder.

Given the N_F-point discrete Fourier transform (DFT) coefficients of the windowed signal frame $x_w(n)$ are $X(k)$, the short-term power spectrum is $X_p(k) = |X(k)|^2, k = 0, 1, \cdots, N_F/2$. First, the *critical band analysis* is done by calculating the energy presented in each critical band.

$$X_b(i) = \sum_{k=b_{li}}^{b_{hi}} X_p(k), \tag{3.2}$$

where b_{li} and b_{hi} are the lower and upper boundaries of critical band i, respectively, and $X_b(i)$ is the energy in critical band i, where $i = 1$ to i_{max}, and i_{max} is dependent on the sampling rate. For this model there are 26 critical bands in the 15 kHz bandwidth, i.e., $i_{max} = 26$.

Each critical band energy is spread across all the critical bands to estimate the masking effects. The *spreading function* $S(i)$, which Johnston used in [44], is proposed in [46]. The proposed spreading function is the same for each critical band masker without dependency on frequency or intensity. $S_{\mathrm{dB}}(i,j)$ has the expression of

$$\begin{aligned} S_{\mathrm{dB}}(i,j) &= 10\ \log_{10}\big(S(i-j)\big) \\ &= 15.81 + 7.51\big((i-j)+0.474\big) - 17.5\big(1+((i-j)+0.474)^2\big)^{1/2}\ \mathrm{dB}, \end{aligned} \tag{3.3}$$

where i is the bark frequency of the masked signal, and j is the bark frequency of the masking signal. An excitation pattern spectrum, $X_e(i)$, is obtained by convolving the bark spectrum with the spreading function

$$X_e(i) = S(i) * X_b(i). \tag{3.4}$$

A *noise masking threshold* is derived by subtracting an offset (in decibels) from the excitation pattern spectrum. Depending on the nature of the masking signal, the offset is different for tonal maskers and noise maskers. The *Spectral Flatness Measure* (SFM) is used to characterize the tonality of the signal. The SFM is defined as a ratio of the *geometric mean* (GM) to the *arithmetic mean* (AM) of the power spectrum $X_p(k)$

$$\mathrm{SFM}_{\mathrm{dB}} = 10\ \log_{10}\frac{GM}{AM}, \tag{3.5}$$

where $GM \triangleq (\prod_{k=0}^{N_F/2} X_p(k))^{-(N_F/2+1)}$ and $AM \triangleq \sum_{k=0}^{N_F/2} X_p(k)/(N_F/2+1)$. A tonality coefficient α is generated from the SFM

$$\alpha = \min\big(\frac{\mathrm{SFM}_{\mathrm{dB}}}{\mathrm{SFM}_{\mathrm{dBmax}}}, 1\big), \tag{3.6}$$

where $\mathrm{SFM}_{\mathrm{dBmax}} = -60$ dB is used to represent an entirely tonelike signal with the tonality coefficient $\alpha = 1$. An entirely noiselike signal has $\alpha = 0$. For an entirely tonelike signal, the noise threshold is estimated to be $14.5 + i$ dB below the spreading spectrum $X_e(i)$, while an entirely noiselike signal has a uniform offset of 5.5 dB across the Bark spectrum. With this tonality coefficient α, the offset in decibels for each critical band is set as

$$O(i) = \alpha(14.5 + i) + (1-\alpha)5.5. \tag{3.7}$$

The *spread threshold* $T(i)$ is obtained by subtracting the offset in decibels from the excitation pattern spectrum

$$T(i) = 10^{(\log_{10} X_e(i)) - O(i)/10}. \tag{3.8}$$

To get the noise masking threshold in the frequency domain, the threshold $T(i)$ should be deconvolved. This procedure is very unstable because of the shape of the spreading function. Johnston proposed a renormalization of the threshold instead of deconvolution. Since the spreading function increases the energy estimates in each band, the renormalization multiplies each $T(i)$ by the inverse of the energy gain per band, assuming each band has unity energy. This compensates for the energy increase from spreading convolution of other critical bands. After renormalization, the threshold is denoted by $T'(i)$.

At last, the final threshold is derived by comparing $T'(i)$ with the absolute threshold of hearing $T_q(i)$. The maximum value between $T'(i)$ and $T_q(i)$ is chosen within each band to give the final masking threshold.

3.3 Psychoacoustic Models for Objective Quality Measurement

Loudness is a fundamental element of sound perception. It belongs to the category of intensity sensations. The intensity of sound, denoted by I, is defined as the amount of sound energy, P, flowing across a unit area surface in a second [47]. Conventionally, a sound is measured in *sound pressure level* (SPL),

$$L = 10\ \log_{10} |I/I_0| = 20\ \log_{10} |p/p_0|, \tag{3.9}$$

where p is the sound pressure, and I_0 and p_0 are the standard references corresponding to the hearing threshold value at 1 kHz. L is an objective measurement of sound, which indicates the relative intensity of a sound with respect to the hearing threshold at 1 kHz. Human sensation is not flat. Even with the same SPL, tones at different frequencies would sound different. The loudness level of a sound is the sound pressure level of a 1 kHz tone of a plane and frontal incident wave that is as loud as the measured sound. Its unit is *phon*. The subjective measurement of loudness is called *sone*. It is measured by how much louder a sound is heard relative to a standard reference. This standard reference of one sone is a tone of 1 kHz at an intensity of 40 dB SPL.

Classical objective measures, for example SNR, determine the quality of a speech/audio codec under test on the basis of differences in the physical signal characteristics for a certain set of test signals. These methods do not use the characteristics of the human auditory system. Moreover, classical objective measurements are not meaningful when applied to modern speech/audio codecs which exploit signal redundancy and the masking properties of the auditory system. Human auditory models have been developed to study the correlation between the aspects of objective measurements and the subjectively perceived quality.

For objective quality measurement, a psychoacoustic model involves two steps. First, it maps the input and output signals of an audio device, such as an audio coder, onto internal representations. Then the quality of the device is calculated based on the difference of the internal representations. The psychoacoustic model does not need the difference signal, nor does it explicitly calculate a masking threshold. However, they can also derive a masking pattern as an intermediate. Here, we are only interested in perceptual models (representations) in the first step, which we will discuss below.

3.3.1 PAQM Model

Beerends and Stemerdink [45] introduced a psychoacoustic model to measure the objective quality of audio devices. A model of the human auditory system is used to calculate the internal representation of the input and output signals of an audio device. The transform from the physical domain to the psychophysical (internal) domain is performed by way of two operations: time-frequency spreading and level compression.

The processed signal is transformed in the frequency domain using overlapping frames, each consisting of N samples. Let $x(m,n)$, $0 \leq m \leq N-1$, be the n-th frame of a windowed discrete-time signal. The short-term DFT coefficients of $x(m,n)$ are represented by $X(k,n)$. The power spectrum is $X_p(k,n) = |X(k,n)|^2$, $0 \leq k \leq N-1$. Analysis is performed in discrete frequency bands. These bands are analogous to *critical bands* (CB), although each CB is now divided into narrower frequency regions. These frequency regions have the same bandwidth, dz, in the Bark domain. Assume there are B such frequency bands. Similar to Eq. (3.2), the total energy per frequency band, $X_b(i,n)$, is calculated from the power spectrum of the signal

$$X_b(i,n) = \sum_{k=b_{li}}^{b_{hi}} X_p(k,n), \quad 0 \leq i \leq B-1, \tag{3.10}$$

where b_{li} and b_{hi} are the lower and upper bounds of the frequency band i, respectively. The outer to inner ear transformation is performed with this perceptual domain spectrum

$$X_a(i,n) = a_0(i)\, X_b(i,n), \quad 0 \leq i \leq B-1, \tag{3.11}$$

where $a_0(i)$ is an outer-to-inner ear transformation function. This pitch representation $X_a(i,n)$

is then combined with that of a previous frame to perform time-domain masking operation

$$X_t(i,n) = X_a(i,n) + T_f(i,n-1)X_a(i,n-1) = \sum_{j=n-1}^{n} T_f(i,j)X_a(i,j), \quad 0 \le i \le B-1, \tag{3.12}$$

where $T_f(i,n) = 1$ and $T_f(i,n-1)$ is an exponential function given by

$$T_f(i) = \exp(-d/\tau(i)), \tag{3.13}$$

where d is the time distance between adjacent short-time frames, and $\tau(i)$ is derived from psychoacoustic time-domain masking experiments.

The *time-domain smeared pitch representation* $X_t(i,n)$ is then convolved with a level dependent basilar membrane spreading function $S(i, L(i))$ to get the excitation intensity, $X_e(i)$. The spreading function from the i-th frequency band to the j-th frequency band is a two sided exponential with slopes as

$$\begin{aligned} S_l(i,L(i)) &= S_l \;=\; 31 \text{ dB/Bark} && j \le i, \\ S_u\big(i,L(i)\big) &= -22 - \min(230/f_c(i), 10) + 0.2L(i) && \\ &= S_0(i) + 0.2L(i) \text{ dB/Bark} && j > i, \end{aligned} \tag{3.14}$$

where $L(i)$ is the level in dB SPL of the i-th frequency band with $L(i) = 10\log_{10} X_t(i,n)$ and $f_c(i)$ is the centre frequency value of the i-th frequency band in Hz. A parametric nonlinear form is used to model the nonlinear additivity of maskers

$$X_e(i) = \left\{ \sum_{j=i}^{B-1} \left[10^{-S_l(j-i)dz/10} X_t(j)\right]^{\alpha/2} + \sum_{j=0}^{i-1} \left[10^{S_0(j)(i-j)dz/10} \big(X_t(j)\big)^{1+0.2(i-j)dz}\right]^{\alpha/2} \right\}^{2/\alpha}, \tag{3.15}$$

where the parameter α is optimized to produce maximum correlation of the excitation value $X_e(i)$ with subjective tests. Experiments have shown that simultaneous stimuli result in an excitation value which is considerably higher than the sum of the contributions. Then the value of α is set to be less than 2. In [45], the optimal setting of α is 0.8.

At last, from this quantity a compressed loudness function is calculated according to the expression in [47],

$$X_l(i) = c\left[\frac{E_0(i)}{s}\right]^{\gamma} \left\{ \left[1 - s + s\frac{X_e(i)}{E_0(i)}\right]^{\gamma} - 1 \right\}, \tag{3.16}$$

where c and s are experimentally derived parameters, γ is a parameter that is also optimized for maximum correlation with the subjective test, and $E_0(i)$ is the absolute threshold of hearing multiplied by the outer-to-inner ear transformation. The function $X_l(i)$ corresponds to the psychoacoustic representation of the short-time frame power spectrum $X_p(k)$.

3.3.2 PEAQ Model

The Perceptual Evaluation Audio Quality (PEAQ) is used to rate the quality of an audio coder. It is described in ITU-R Recommendation BS.1387 [13]. The psychoacoustic model used by PEAQ estimates the masking threshold and loudness among other intermediate model variables. The PEAQ model consists of two versions: basic version and advanced version. The basic version only uses an FFT-based perceptual model. This section describes those steps involved in the computation of the masking threshold and loudness in the basic version.

The PEAQ model operates with $f_s = 48$ kHz sampled input segments of 0.042 second ($N_F = 2048$ samples) with 50% overlap. A short-term FFT is computed following the multiplication with a Hann window. Assuming the maximum level to be 92 dB SPL, the resulting frequency domain coefficients are scaled by a factor to get the transformed input signal $X(k)$ for $0 \leq k \leq N_F - 1$.

The combined filtering effect of the outer and middle ear is expressed as

$$A_{dB}(f) = -0.6 \cdot 3.64(f/1000)^{-0.8} + 6.5 \cdot e^{-0.6(f/1000-3.3)^2} - 10^{-3}(f/1000)^{3.6}. \tag{3.17}$$

The outer and middle ear weighted FFT outputs are

$$X_w(k) = |X(k)| \cdot 10^{A_{dB}(f(k))/20}, \tag{3.18}$$

where $f(k) = kf_s/N_F$.

The weighted spectrum $|X_w(k)|^2$ are grouped into quarter-bark bands in order to transform into the perceptual domain. Each perceptual band is characterized by a lower frequency, $f_l(i)$, a centre frequency, $f_c(i)$, and an upper frequency, $f_u(i)$. In the case that a frequency bin is across two bands, the energy contributed to each band is obtained by multiplying the frequency bin energy by the percentage of the frequency bin lying within the frequency group. For the ith frequency band, the contribution from the energy in DFT bin k is [48]

$$\mu(i,k) = \frac{\max\big[0,\, \min(f_u(i), \frac{2k+1}{2}\frac{f_s}{N_F}) - \max(f_l(i), \frac{2k-1}{2}\frac{f_s}{N_F})\big]}{\frac{f_s}{N_F}} \tag{3.19}$$

The resulting energies of the frequency groupings are denoted by $P_e(i)$ with

$$P_e(i) = \sum_{k=k_l(i)}^{k_u(i)} \mu(i,k)|X_w(k)|^2, \tag{3.20}$$

where $\mu(i,k)$ is non-zero over the interval $k_l(i) \le k \le k_u(i)$. The *pitch patterns* $P_p(i)$ are obtained by adding the frequency dependent internal noise of the inner ear, P_{Thres}, to $P_e(i)$

$$P_p(i) = P_e(i) + P_{\text{Thres}}, \tag{3.21}$$

where the internal noise is $P_{\text{Thres}} = 10^{0.4 \cdot 3.64 \cdot (f_c(i)/1000)^{-0.8}}$.

The pitch patterns $P_p(i)$ are smeared out over frequency using a level dependent spreading function. The spreading function from the i-th band to the j-th band is a two sided exponential with slopes as

$$\begin{aligned} S_l\big(i, L(i)\big) &= S_l \;=\; 27 \text{ dB/Bark} && j \le i, \\ S_u\big(i, L(i)\big) &= -24 - 230/f_c(i) + 0.2L(i) \text{ dB/Bark} && j > i, \end{aligned} \tag{3.22}$$

where $L(i)$ represents the signal power (in dB SPL) in the i-th perceptual band with $L(i) = 10\,\log_{10}\big(P_p(i)\big)$. To model the nonlinear additivity of maskers, a power law is used. The *unsmeared excitation pattern* in band i, $E_2(i)$, is the normalized sum of the spread energy contributions from all bands

$$E_2(i) = \frac{1}{\text{Norm}_{SP}(i)} \Big(\sum_{j=0}^{B-1} E_{\text{line}}(j,i)^{0.4} \Big)^{\frac{1}{0.4}}, \tag{3.23}$$

where $E_{\text{line}}(j,i)$ represents the energy spread of the j-th band to the i-th band, $\text{Norm}_{SP}(i)$ is the sum of the spread energies of all bands with unit energy, and B is the total number of frequency groups. $E_{\text{line}}(j,i)$ is defined by

$$E_{\text{line}}(j,i) = \begin{cases} \dfrac{10^{\frac{L(j)}{10}} 10^{\frac{-0.25(j-i)S_l(j,L(j))}{10}}}{\displaystyle\sum_{l=0}^{j-1} 10^{\frac{-0.25(j-l)S_l(j,L(j))}{10}} + \sum_{l=j}^{B-1} 10^{\frac{0.25(l-j)S_u(j,L(j))}{10}}} & \text{if } i \leq j, \\ \dfrac{10^{\frac{L(j)}{10}} 10^{\frac{0.25(i-j)S_u(j,L(j))}{10}}}{\displaystyle\sum_{l=0}^{j-1} 10^{\frac{-0.25(j-l)S_l(j,L(j))}{10}} + \sum_{l=j}^{B-1} 10^{\frac{0.25(l-j)S_u(j,L(j))}{10}}} & \text{if } i > j. \end{cases} \tag{3.24}$$

$\text{Norm}_{SP}(i)$ is calculated according to

$$\text{Norm}_{SP}(i) = \left(\sum_{j=0}^{B-1} \tilde{E}_{\text{line}}(j,i)^{0.4} \right)^{\frac{1}{0.4}} \tag{3.25}$$

with

$$\tilde{E}_{\text{line}}(j,i) = \begin{cases} \dfrac{10^{\frac{-0.25(j-i)S_l(j,0)}{10}}}{\displaystyle\sum_{l=0}^{j-1} 10^{\frac{-0.25(j-l)S_l(j,0)}{10}} + \sum_{l=j}^{B-1} 10^{\frac{0.25(l-j)S_u(j,0)}{10}}} & \text{if } i \leq j, \\ \dfrac{10^{\frac{0.25(i-j)S_u(j,0)}{10}}}{\displaystyle\sum_{l=0}^{j-1} 10^{\frac{-0.25(j-l)S_l(j,0)}{10}} + \sum_{l=j}^{B-1} 10^{\frac{0.25(l-j)S_u(j,0)}{10}}} & \text{if } i > j. \end{cases} \tag{3.26}$$

Forward masking is modelled by smearing out the energies in each frequency group over time by a first order low pass filter. The time constants of the filters are frequency dependent and are calculated by

$$\tau(i) = \tau_{\text{min}} + \frac{100}{f_c(i)} \cdot (\tau_{100} - \tau_{\text{min}}), \tag{3.27}$$

where $\tau_{100} = 0.030$s, $\tau_{\text{min}} = 0.008$s, and $f_c(i)$ is the centre frequency value of the i-th band in Hz. The final *excitation patterns*, $E(i,n)$, of the current segment are calculated by

$$E_f(i,n) = a(i) \cdot E_f(i,n-1) + \big(1 - a(i)\big) \cdot E_2(i,n), \tag{3.28}$$

$$E(i,n) = \max\big(E_f(i,n), E_2(i,n)\big), \tag{3.29}$$

where n is the frame index, and $a(i) = \exp\Big(-1/\big(f_{ss} \cdot \tau(i)\big)\Big)$ and f_{ss} is the frame rate given by

$$f_{ss} = \frac{f_s}{N_F/2}. \tag{3.30}$$

A *masking patterns*, $M(i,n)$ is determined by applying a weighting function, $m(i)$, to the excitation patterns, $E(i,n)$.

$$M(i,n) = \frac{E(i,n)}{10^{m(i)/10}}, \tag{3.31}$$

where

$$m(i) = \begin{cases} 3 & 0.25i \leq 12, \\ (0.25)^2 i & 0.25i > 12. \end{cases} \tag{3.32}$$

This masking threshold is defined in the perceptual domain. If the masking threshold in the frequency domain is necessary, the effects of the internal noise, middle and outer ear need to be removed from $M(i,n)$ for each frequency group.

The loudness patterns of the signal is derived from the excitation patterns $E(i,n)$ with the same expression as Eq. (3.15) in [47].

Chapter 4

Design of a Perceptual Postfilter Based on GMM Estimation

From Section 2.4, it is clear that human perceptual modelling in speech coding is very empirical. Noise shaping alone is not enough to make the encoding noise below the masking threshold at low bit rates. Adaptive postfiltering has been shown to improve the decoded speech quality efficiently. Conventional postfiltering uses the available decoded information, and is empirically designed according to human perception. However, only a few improvements (for instance, [16, 34]) have been made to adaptive postfiltering despite the development in our understanding of the human auditory system.

Also, due to the complexity of speech encoding, the coding noise is correlated to the speech signal to some extent. This makes conventional speech enhancement methods designed for reducing background acoustic noise inappropriate to deal with coding noise, because most speech enhancement algorithms are based on the assumption that the speech signal and the noise signal are independent and the noise signal is stationary. Furthermore, in speech encoding, the coded speech always has an average energy smaller than the original speech signal. This is different from the scenario of speech enhancement where additive noise is assumed [49].

Speech quality can be enhanced if we match speech coders (which are based on voice production models) to the human ear with a good auditory model [50]. For those speech coders which have been implemented in practice, it is preferable to improve the speech quality by an embedding part, instead of changing the coding structure. In this chapter, we will introduce a novel perceptual postfilter to improve the quality of the decoded speech without change in the encoder.

This postfilter exploits properties of psychoacoustic models, and can be applied directly to the frequency domain of the decoded signal to suppress perceptible noise. Section 4.2 discusses the perceptual postfiltering idea. In Section 4.3, we present our perceptual postfilter algorithm by a *Minimum Mean Squared Error* (MMSE) estimator based on *Gaussian mixture model* (GMM).

4.1 Postprocessing Model

Our postprocessing model at a receiver is shown in Fig. 4.1. The postprocessor has the same purpose as the conventional postfilter or its variations to improve the perceptual quality of the reconstructed speech. However, it exploits properties of psychoacoustics.

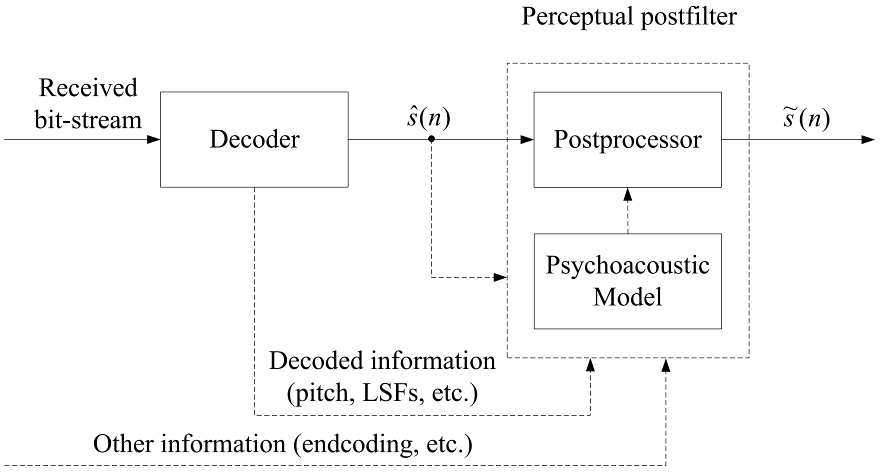


Fig. 4.1 The proposed perceptual postfiltering model

After decoding a received stream, the decoder gives a decoded speech, $\hat{s}(n)$. The postprocessor modifies the decoded speech to produce a enhanced speech, $\tilde{s}(n)$, with improved quality. The modification is done with the knowledge of the decoded information, the decoded speech, the psychoacoustic model, and other available information (for instance, how the encoder works). A perceptually-based postfiltering algorithm performs the modification using internal psychoacoustic properties, which is described in the following section.

Our postprocessing model is carried out in the frequency domain. Frequency domain techniques have the advantage of modifying different parts of the frequency spectrum independently. Also, the perceptual properties are well modelled in the frequency domain. Since speech is perceived by the hearing system, frequency domain approaches are the proper choice to incorporate the perceptual concepts in our system. We then can enhance the speech with frequency-by-frequency gain modification.

Clearly, our postfiltering idea is a superset of conventional postfiltering. A conventional postfilter is controlled by its parameters. For example, the short-term postfilter in Eq. (2.14) is mainly determined by LPCs, while λ_1, λ_2 and μ are used to tune the postfilter shape to some limited extent. Our postfilter gains in each critical band can be set more freely according to its theoretical basis.

4.1.1 Proposed System

Our proposed complete system is shown in Fig. 4.2. The top diagram gives the generation of a training data set and the GMM training. A low bit rate speech codec encodes the corresponding information of the excitation signal and LSFs for each frame of speech.

A feature vector for each processing block is formed for GMM training. Training vectors are generated from processing blocks. For each processing block, a *decoding feature vector* derivable from the coded information is obtained. A vector of perceptual postfilter gains is derived from each processing block. By passing all speech for training through the speech encoder, a data set composed of feature vectors for GMM training are generated.

Our proposed perceptual postfilter works at the receiver end, as shown in the bottom part of Fig. 4.2. For each speech frame, a coded stream is sent from the encoder and the decoder decodes this received stream to generate coded information about the speech. A decoding feature vector is derived with the same process as generating training vectors. A MMSE estimate of the postfilter gains given the decoding feature vector for each processing block is obtained. The postfiltering is performed on windowed blocks of the decoded speech. A modified decoded speech is then obtained.

It is clear that the key issues in our system are the perceptual postfilter, GMM training and MMSE estimation of the perceptual postfilter. We will discuss each of them in this chapter.

4.2 Perceptual Postfilter

In psychoacoustic modelling, a neural excitation called loudness is assumed to directly affect perceived strength. A loudness distribution is predicted from the excitation intensity by a nonlinear

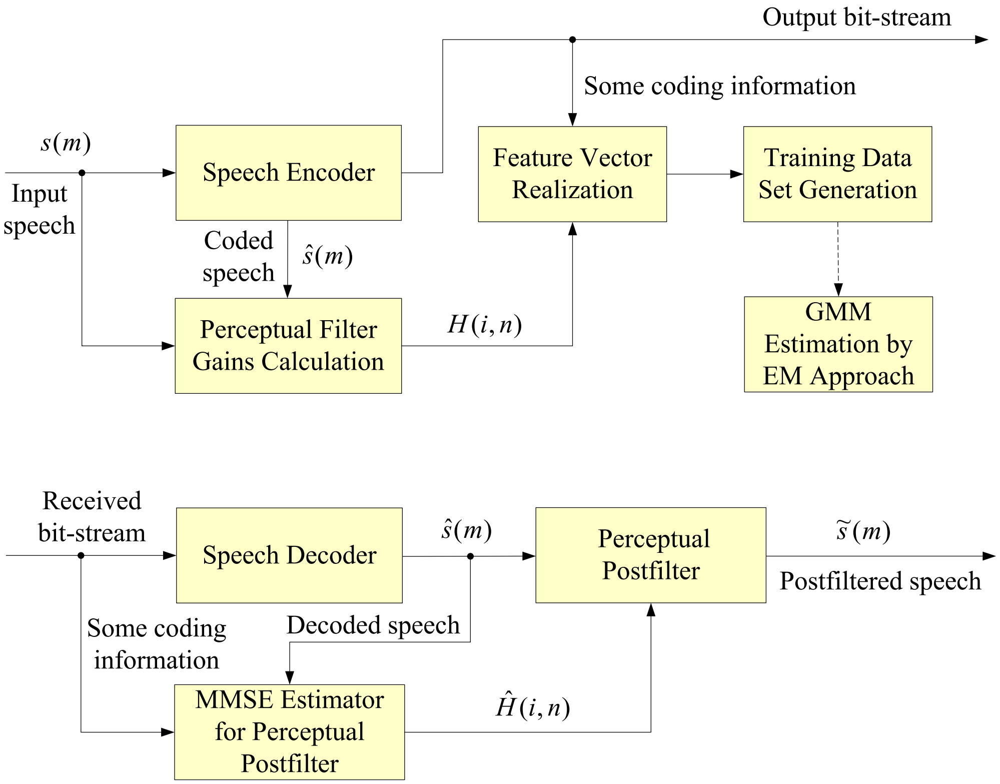


Fig. 4.2 System Diagrams. Top: GMM training at the encoder; Bottom: Perceptual postfiltering by MMSE estimation at the decoder.

transformation. Masking has been widely used in audio coding. Recent research also considers loudness in audio coding [49].

If we use masking in postfiltering, it is more complicated since both the masker (the original speech) and the coding noise are unknown at the decoder. However, both loudness and masking are directly connected to the excitation pattern with operations independent of the signal level. The excitation patterns of a sound represent the activity or excitation evoked by that sound as a function of characteristic frequency along the basilar membrane. A global masking curve is calculated by applying frequency dependent offsets (in dB SPL) to the excitation patterns. The transform from the excitation pattern to the specific loudness pattern is given by a warping function [47]. The excitation pattern model implies that human hearing can detect distortion, if, in any critical band, there is more than 1 dB distortion in the excitation pattern [47].

Similar to Wiener filtering in speech enhancement [51], the estimation of a psychoacoustic representation of the clean signal can be derived from a modification of such a representation of the coded signal. A perceptual filter is designed to reduce the audible coding noise by equalizing the excitation psychoacoustic representation of the original signal and the coded signal. If the coding stream which is sent by the encoder is received without error, the same perceptual postfilter is applicable to enhance the decoded speech at the receiver.

4.2.1 Perceptual Filter Proposed by Lam and Stewart [52]

Lam and Stewart [52] designed a generalized perceptual audio filter in low rate audio coding. The perceptual filter is based on a human auditory perception model which attempts to model the psychoacoustic behaviour of the ear. It tries to perceptually suppress coding noise in the subjective domain, i.e. the loudness representation of the coded signal after filtering is set to the same level of the original signal. The psychoacoustic model used is the loudness patterns described in Section 3.3.1. The generalized linear perceptual filter is finally realized by restoring the excitation pattern in the perceptual (critical band) domain of the reconstructed signal.

Conditions for Noise Suppression

Let us denote the n-th frame of the original signal as $s(m,n)$, and the coded signal as $\hat{s}(m,n)$. m is a time counter inside a frame. Let the short-time power spectra of windowed frames of the original signal and the coded signal be $S_p(k,n)$ and $\hat{S}_p(k,n)$, respectively. Also, let their psychoacoustic loudness representations of Eq. (3.16) in Section 3.3.1 be $S_l(i,n)$ and $\hat{S}_l(i,n)$, representatively. According to [45], the difference between these two representations, $S_l(i,n)$ and $\hat{S}_l(i,n)$ is a measure for the coding noise in the perceptual domain. This difference will be audible by a listener. Therefore, in terms of enhancement, it is proposed to modify the power spectrum of the coded signal so that the resulting psychoacoustic representation corresponds to that of the original signal. Let the power spectrum of the modified signal be $\tilde{S}_p(k,n)$ and its corresponding psychoacoustic representation be $\tilde{S}_l(k,n)$. A linear filter $H(i,n)$ is proposed to modify the coded signal. The gain of this filter is assumed to be constant within the same critical band i so that the enhanced signal is given by

$$\tilde{S}_p(k,n) = H(i,n)\,\hat{S}_p(k,n), \quad b_{li} \leq k \leq b_{hi}, \quad 0 \leq i \leq B-1. \tag{4.1}$$

Then we have the equation of the critical band intensities after critical band energy grouping with Eq. (3.10)

$$\tilde{S}_b(i,n) = H(i,n)\,\hat{S}_b(i,n), \quad 0 \le i \le B-1. \tag{4.2}$$

A suitable condition for psychoacoustic signal enhancement is setting the psychoacoustic representation of the modified signal $\tilde{S}_l(i,n)$ to that of the original signal $S_l(i,n)$, which is given by the following expression

$$\tilde{S}_l(i,n) = S_l(i,n), \quad 0 \le i \le B-1. \tag{4.3}$$

From the expression of the psychoacoustic loudness representation in Eq. (3.16), it can be clearly concluded that the condition for psychoacoustic signal enhancement in Eq. (4.3) is equivalent to

$$\tilde{S}_e(i,n) = S_e(i,n), \quad 0 \le i \le B-1, \tag{4.4}$$

where $\tilde{S}_e(i,n)$ is the excitation intensity of the modified signal and $S_e(i,n)$ is the excitation intensity of the original signal. The excitation patterns of the modified signal are then restored to those of the original signal [53].

Generalized Perceptual Filter

The perceptual filter gains are derived from Eq. (4.4) with the psychoacoustic model in PAQM [45], which is described in Section 3.3.1. Lam and Stewart ignored level-dependent effects on the spreading function and proposed a generalized perceptual filter for low bit rate audio coding. In the derivation of the generalized perceptual filter [52] (see Appendix A), the spreading function was assumed to be level-independent. This assumption is essential to an analytical expression of the generalized perceptual filter. The generalized perceptual filter gains are sent to the audio decoder as side information for perceptual suppression of quantization noise in the decoded signal.

4.2.2 Proposed Perceptual Postfilter

The above perceptual filter exploits the properties of a psychoacoustic model, and can be directly applied to the frequency domain of the coded signal to suppress the perceptible noise. It gives us a new outlook for adaptive postfiltering with a specific psychoacoustic model.

Actually, the psychoacoustic model used by Lam and Stewart [52] is an invertible auditory model after their level-independent approximation of the frequency domain spreading function and the omission of temporal masking. With such an assumption of an invertible auditory model, we then have the excitation level of a speech frame $s(m,n)$

$$S_e^{\alpha/2}(i) = \sum_{v=0}^{B-1} S_b^{\alpha/2}(v)C_{i,v}, \quad 0 \le i \le B-1, \tag{4.5}$$

where $S_b(v)$ is the v-th critical band intensity defined by Eq. (3.2) and $C_{i,v}$ is the Bark domain spreading value only related to i, v (see Eq. (A.2))

$$C_{i,v} = \begin{cases} \left[10^{-\frac{S_l(v-i)dz}{10}}\right]^{\alpha/2} & \text{for } i \le v, \\ \left[10^{\frac{S_0(v)(i-v)dz}{10}}\right]^{\alpha/2} & \text{for } i > v, \end{cases} \tag{4.6}$$

with S_l and S_0 defined in Eq. (3.14) in Section 3.3.1.

From Eq. (4.2) and Eq. (4.5) we have

$$\begin{aligned} \tilde{S}_e^{\alpha/2}(i) &= \sum_{v=0}^{B-1} \tilde{S}_b^{\alpha/2}(v)C_{i,v} \\ &= \sum_{v=0}^{B-1} H^{\alpha/2}(v)\hat{S}_b^{\alpha/2}(v)C_{i,v}. \end{aligned} \tag{4.7}$$

Combining Eq. (4.4), Eq. (4.5), and Eq. (4.7), we get

$$\sum_{v=0}^{B-1} H^{\alpha/2}(v)\hat{S}_b^{\alpha/2}(v)C_{i,v} = \sum_{v=0}^{B-1} S_b^{\alpha/2}(v)C_{i,v}. \tag{4.8}$$

This is equivalent to

$$\mathbf{A} \begin{bmatrix} \hat{S}_b^{\alpha/2}(0) & 0 & \cdots & 0 \\ 0 & \hat{S}_b^{\alpha/2}(1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{S}_b^{\alpha/2}(B-1) \end{bmatrix} \mathbf{h}^{\alpha/2} = \mathbf{A} \begin{bmatrix} S_b^{\alpha/2}(0) \\ S_b^{\alpha/2}(1) \\ \vdots \\ S_b^{\alpha/2}(B-1) \end{bmatrix}, \tag{4.9}$$

where $\mathbf{A}$ is a $B \times B$ matrix with elements $a_{i,j} = C_{i-1,j-1}$, $i, j = 1, 2, \cdots, B$, and $\mathbf{h} = [H(0), H(1), \cdots, H(B-1)]^t$ is the perceptual filter vector.

With the optimal value of α set to be 0.8 in PAQM [45], we can compute with MATLAB that the determinant of $\mathbf{A}$ is 0.8914 and the 2-norm condition number of $\mathbf{A}$ is 1.4443. That means $\mathbf{A}$ is a full rank matrix. Then we can obtain the perceptual filter $\mathbf{h}$ from Eq. (4.9)

$$\mathbf{h} = \begin{bmatrix} S_b(0)/\hat{S}_b(0) \\ S_b(1)/\hat{S}_b(1) \\ \vdots \\ S_b(B-1)/\hat{S}_b(B-1) \end{bmatrix}. \tag{4.10}$$

We can see that the conversion from the critical band intensities to the excitation intensities is unnecessary, although Eq. (4.4) is the enhancement condition. The gains of the perceptual filter in Eq. (4.10) are just the ratio of the critical band intensities of the original signal to those of the coded signal.

Actually, since all the psychoacoustic representations are originated from the critical band intensities, the modified signal will have all the same psychoacoustic representations as those of the original signal if the critical band intensities of the coded signal are set to the same level as those of the original signal. Therefore, we get a general perceptual postfilter which equalizes the energy in perceptual domain

$$\tilde{S}_b(i) = S_b(i), \quad 0 \le i \le B-1, \tag{4.11}$$

where $S_b(i)$ and $\tilde{S}_b(i)$ are the critical band intensity of the original signal and the perceptually filtered signal. With the grouping method of Eq. (3.2) as in Johnston's model, the energy in each critical band of $s(m,n)$ is summed up to give the critical band spectrum $S_b(i)$

$$S_b(i) = \sum_{k=bl_i}^{bh_i} S_p(k), \quad 0 \le i \le B-1. \tag{4.12}$$

where bl_i and bh_i are the lower and upper bounds of the critical band i, respectively.

Applying the grouping to Eq. (4.1) and combining with Eq. (4.11), our new perceptual postfilter has the expression

$$H(i) = S_b(i)/\hat{S}_b(i), \quad 0 \le i \le B-1, \tag{4.13}$$

where $\hat{S}_b(i)$ is the critical band spectrum of the coded signal. Eq. (4.13) is the same as Eq. (4.10). These are the postfilter gains given the knowledge of the original and coded critical band spectra. Tests with speech signals show that this postfilter filters the coded speech and produces a modified speech signal which is indistinguishable from the original speech with the human hear. In the next section, we propose a method to estimate these gains at the decoder.

4.3 Perceptual Postfilter with MMSE Estimation Based on GMM

The perceptual filter from [52], which is discussed in Section 4.2.1 for low bit-rate audio coding, motivates us to build a similar postfilter for *linear prediction analysis-by-synthesis* (LPAS) speech coders in Section 4.2.2. The perceptual filter can be derived from each processing frame and applied to the decoded speech to improve the speech quality. However, from the discussion in Section 4.2, direct information about the perceptual filter is unavailable to the receiver unless it is sent as side information while we need a postfilter which works as an add-on part at the receiver without requiring additional bits. A novel postfiltering method combining perceptual properties and statistical estimation together has been introduced by the present author in [54]. A *Minimum Mean Squared Error* (MMSE) estimation of the perceptual postfilter based on *Gaussian mixture model* (GMM) was proposed. The postfilter gains are estimated from a MMSE estimator given a feature vector which is from the information at the decoder. We call this feature vector the *decoding feature vector*. This operation works on a frame-by-frame basis at the receiver. The output of the MMSE estimator is determined by the estimator parameters and the decoding feature vector. The parameters of the MMSE estimator are from a trained GMM. That means they are available at the receiver and do not need to be transmitted as side information by the encoder. For a LPAS speech coder, the decoded speech and coded information (which is sent to the receiver by a coding stream) are available at the encoder. Therefore, we can generate the training data at the encoder. A GMM is used to model the joint pdf of the training vector. Each training vector in the training data set is obtained from encoding of a speech frame. A training vector is composed of perceptual filter gains and a decoding feature vector. The *Expectation-Maximization* (EM) algorithm is commonly used for the training of a GMM.

In [54], we only considered the static features for perceptual postfilter estimation. Incorporating the locally sequential speech property as well as the individual frame, we also study the model with joint static and frame-differential feature components.

4.3.1 GMM Estimation by the EM Algorithm

The GMM is popularly used to approximate a *probability density function* (pdf) of a random vector with relatively small number of parameters. Its ability to represent some general speech spectral shapes by the Gaussian components makes it popular in speech recognition and speaker identification [55], as well as in neural information processing [56]. A GMM is also used for vector quantization of LSFs [57, 58]. The underlying pdf of vectors in a database can be modelled by a *Gaussian mixture* (GM) pdf and the parameters of the model can be estimated. Qian and Kabal [59, 60] used the GMM for bandwidth extension by estimating the missing high band information from the low band LSFs.

A GM pdf for a d-dimensional random feature vector $\mathbf{x}$ is a mixture of M joint Gaussian densities $\{\omega_1, \cdots, \omega_M\}$

$$p_{\mathbf{x}}(\mathbf{x}) = \sum_{i=1}^{M} P(\omega_i)\, p_{\mathbf{x}}(\mathbf{x}|\omega_i), \tag{4.14}$$

where $p_{\mathbf{x}}(\mathbf{x}|\omega_i)$ is the i-th Gaussian component, and $P(\omega_i)$ is *a priori* probability.

For notation convenience, let $\alpha_i = P(\omega_i)$, and $\mathcal{N}(\mathbf{x}|\boldsymbol{\theta}_i) = p_{\mathbf{x}}(\mathbf{x}|\omega_i)$, we have

$$p_{\mathbf{x}|\boldsymbol{\Theta}}(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{i=1}^{M} \alpha_i\, \mathcal{N}(\mathbf{x}|\boldsymbol{\theta}_i), \tag{4.15}$$

$$\boldsymbol{\Theta} = \{\alpha_1, \cdots, \alpha_M, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}, \tag{4.16}$$

where α_i is a nonnegative constant and $\sum_{i=1}^{M} \alpha_i = 1$. $\mathcal{N}(\mathbf{x}|\boldsymbol{\theta}_i)$ is an individual Gaussian density parameterized by $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\theta}_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\bigl(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\bigr). \tag{4.17}$$

Therefore, a GM pdf is defined by the mean vectors, the covariance matrices and the mixture weights for the Gaussian components, i.e. $\boldsymbol{\Theta}$ in Eq. (4.16). With the EM algorithm, we can train a GMM to approximate the pdf of certain features in speech.

The parameter set $\boldsymbol{\Theta}$ can be estimated by the *maximum likelihood* (ML) method. The EM algorithm is a widely used approach for ML estimation in cases where a closed-form analytical expression for the optimal parameters is difficult to derive. EM is an iterative algorithm where a monotonic increase in the log-likelihood, L, is guaranteed [57], i.e. $L(\boldsymbol{\Theta}^{(k+1)}) \geq L(\boldsymbol{\Theta}^{(k)})$, in each iteration over a given database. $\mathbf{\Theta}^{(k)}$ is the value of the parameter set $\mathbf{\Theta}$ at iteration k.

One key issue for applications of mixture modelling is the number of parameters in $\mathbf{\Theta}$. The larger the number of parameters, the greater is the possibility to describe the fine structure of the underlying data distribution. On the other hand, with a high degree of freedom in the modelling, there is a risk for overfit. A rich set of parameters may lead to undue complexity. Thus, the selection of the number of parameters must be a compromise [57]. Both full and diagonal covariance matrices are widely used in the GM density. With a GMM of M Gaussian densities for a d-dimensional random feature vector, the number of parameters to be estimated during training is $M(d+\frac{d(d+1)}{2}+1)$ for full covariance Gaussians, and $M(2d+1)$ for diagonal covariance Gaussians. A GMM with diagonal covariance Gaussian components is usually preferred, because of fewer parameters and its potentiality of modelling the underlying pdf just as well if enough mixtures are used.

Assuming we have a data set $\mathbf{X} = \{\mathbf{x}_n\}, n = 1, \cdots, N$ of N observations of the feature vector $\mathbf{x}$, the log-likelihood function can be expressed as

$$\begin{aligned} L(\mathbf{\Theta}) &= \ln \prod_{n=1}^{N} p_{\mathbf{x}|\mathbf{\Theta}}(\mathbf{x}_n|\mathbf{\Theta}) \\ &= \sum_{n=1}^{N} \ln\, p_{\mathbf{x}|\mathbf{\Theta}}(\mathbf{x}_n|\mathbf{\Theta}) \\ &= \sum_{n=1}^{N} \ln \sum_{i=1}^{M} \alpha_i \mathcal{N}(\mathbf{x}_n|\boldsymbol{\theta}_i). \end{aligned}$$

It is not easy to express optimal parameters in a closed form since the function contains a logarithm of a sum. Given an initial set of M Gaussian component pdfs $\mathcal{N}(\mathbf{x}_n|\boldsymbol{\theta}_i^{(k)})$ and M mixture weights $\alpha_i^{(k)}$, $i = 1, \cdots, M$, $k = 0$, a GMM with a parameter set $\mathbf{\Theta}$ is trained by the EM approach iteratively:

1. **E-Step:** Compute the likelihoods $\mathcal{N}(\mathbf{x}_n|\boldsymbol{\theta}_i^{(k)})$ and determine the posterior probabilities $\nu_i^{(k)}(n) = p(\omega_i|\mathbf{x}_n, \mathbf{\Theta}^{(k)})$ of each mixture component for each training data point $\mathbf{x}_n$ as

$$\nu_i^{(k)}(n) = \frac{\alpha_i^{(k)} \mathcal{N}\big(\mathbf{x}_n|\boldsymbol{\theta}_i^{(k)}\big)}{\sum\limits_{j=1}^{M} \alpha_j^{(k)} \mathcal{N}\big(\mathbf{x}_n|\boldsymbol{\theta}_i^{(k)}\big)}. \tag{4.18}$$

2. **M-Step:** Re-estimate component pdfs and weights, based on data, likelihoods and posterior probabilities [57]

$$\alpha_i^{(k+1)} = \frac{1}{N}\sum_{n=1}^{N} \nu_i^{(k)}(n), \tag{4.19a}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \frac{\displaystyle\sum_{n=1}^{N} \nu_i^{(k)}(n)\,\mathbf{x}_n}{\displaystyle\sum_{n=1}^{N} \nu_i^{(k)}(n)}, \tag{4.19b}$$

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\displaystyle\sum_{n=1}^{N} \nu_i^{(k)}(n)\big(\mathbf{x}_n - \boldsymbol{\mu}_i^{(k+1)}\big)\big(\mathbf{x}_n - \boldsymbol{\mu}_i^{(k+1)}\big)^T}{\displaystyle\sum_{n=1}^{N} \nu_i^{(k)}(n)}. \tag{4.19c}$$

3. Repeat steps 1 and 2 with $k = k+1$ until $L(\boldsymbol{\Theta})$ of Eq. (4.18) of the entire data set does not change appreciably, or a limit on the number of iterations is reached.

When we assume $\boldsymbol{\Sigma}_i$ be diagonal, i.e. $\boldsymbol{\Sigma}_i = \text{diag}\{\lambda_{i,1}, \cdots, \lambda_{i,d}\}$, the update equation for the diagonal elements $\lambda_{i,j}$ corresponding to Eq. (4.19c) becomes

$$\begin{aligned}\lambda_{i,j} &= \frac{\displaystyle\sum_{n=1}^{N} \nu_i^{(k)}(n)\big(x_{n,j} - \mu_{i,j}^{(k+1)}\big)^2}{\displaystyle\sum_{n=1}^{N} \nu_i^{(k)}(n)} \\ &= \frac{\sum_{n=1}^{N} \nu_i^{(k)}(n)\,x_{n,j}^2}{\sum_{n=1}^{N} \nu_i^{(k)}(n)} - (\mu_{i,j}^{(k+1)})^2,\end{aligned} \tag{4.20}$$

where $x_{n,j}$ and $\mu_{i,j}^{(k+1)}$ are the j-th vector component of $\mathbf{x}_n$ and $\boldsymbol{\mu}_i{}^{(k+1)}$, respectively.

4.3.2 Prior Model

Modern speech coders take advantage of the short-term and long-term correlations of speech. Speech-signal segments are often characterized in terms of the properties of their power spectra.

A relationship exists between the autocorrelation and power-spectral domains: the fine structure of the power spectrum corresponds to the long-term autocorrelation of the time-domain signal, and the power-spectral envelope corresponds to the short-term autocorrelation [2]. We choose a d_1-dimensional postfilter gains, $\mathbf{h}$, and the d_2-dimensional speech properties, $\mathbf{y}$, as a d-dimensional feature vector for each frame with $d = d_1 + d_2$. $\mathbf{y}$ is also the *decoding feature vector* and composed of a subvector of the short-term property, $\mathbf{b}$, and a subvector of the long-term property, $\mathbf{n}$. Then the d-dimensional feature vector is denoted by $\mathbf{s}$

$$\mathbf{s} = [\mathbf{h}; \mathbf{y}], \quad \text{with} \quad \mathbf{y} = [\mathbf{b}; \mathbf{n}]. \tag{4.21}$$

Dynamic features are the local (weighted) time difference of static features. While considering the dynamic features in our system, we use the simplest dynamic property of the frame-differential ("delta") features $\mathbf{s}$, defined by

$$\Delta\mathbf{s}_n \equiv \mathbf{s}_n - \mathbf{s}_{n-1}, \tag{4.22}$$

where n is the frame index.

With static features only, the pdf of $p(\mathbf{s}_n)$ is

$$p(\mathbf{s}_n) = \sum_{i=1}^{M} \alpha_i \mathcal{N}(\mathbf{s}_n | \boldsymbol{\mu}_i^{\mathbf{s}}, \boldsymbol{\Sigma}_i^{\mathbf{s}}). \tag{4.23}$$

The joint probability distribution function for both the static, $\mathbf{s}_n$, and delta, $\Delta\mathbf{s}_n$, features is modelled by GMMs assuming the static and dynamic features are uncorrelated with each other. The pdf $p(\mathbf{s}_n, \Delta\mathbf{s}_n)$ is given by

$$p(\mathbf{s}_n, \Delta\mathbf{s}_n) = \sum_{i=1}^{M} \alpha_i \mathcal{N}(\mathbf{s}_n | \boldsymbol{\mu}_i^{\mathbf{s}}, \boldsymbol{\Sigma}_i^{\mathbf{s}}) \mathcal{N}(\Delta\mathbf{s}_n | \boldsymbol{\mu}_i^{\Delta\mathbf{s}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{s}}). \tag{4.24}$$

In Eq. (4.23) and Eq. (4.24), the Gaussian density parameters in the i-th Gaussian densities

$\mathcal{N}(\mathbf{s}_n|\boldsymbol{\mu}_i^{\mathbf{s}}, \boldsymbol{\Sigma}_i^{\mathbf{s}})$ and $\mathcal{N}(\Delta\mathbf{s}_n|\boldsymbol{\mu}_i^{\Delta\mathbf{s}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{s}})$ can be written in block matrices as follows

$$\boldsymbol{\mu}_i^{\mathbf{s}} = \begin{bmatrix} \boldsymbol{\mu}_i^{\mathbf{h}} \\ \boldsymbol{\mu}_i^{\mathbf{y}} \end{bmatrix}, \tag{4.25a}$$

$$\boldsymbol{\Sigma}_i^{\mathbf{s}} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{\mathbf{hh}} & \boldsymbol{\Sigma}_i^{\mathbf{hy}} \\ \boldsymbol{\Sigma}_i^{\mathbf{yh}} & \boldsymbol{\Sigma}_i^{\mathbf{yy}} \end{bmatrix}, \tag{4.25b}$$

$$\boldsymbol{\mu}_i^{\Delta\mathbf{s}} = \begin{bmatrix} \boldsymbol{\mu}_i^{\Delta\mathbf{h}} \\ \boldsymbol{\mu}_i^{\Delta\mathbf{y}} \end{bmatrix}, \tag{4.25c}$$

$$\boldsymbol{\Sigma}_i^{\Delta\mathbf{s}} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}\Delta\mathbf{h}} & \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}\Delta\mathbf{y}} \\ \boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{h}} & \boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}} \end{bmatrix}. \tag{4.25d}$$

In Eq. (4.25a) and Eq. (4.25c), the vector blocks with superscripts $\mathbf{h}$ and $\Delta\mathbf{h}$ are of length d_1, which is the dimension of the postfilter gain vector $\mathbf{h}$, and the vector blocks with the superscripts $\mathbf{y}$ and $\Delta\mathbf{y}$ are of length d_2, which is the dimension of the decoding feature vector $\mathbf{y}$. In Eq. (4.25b) and Eq. (4.25d), the matrix blocks with superscripts $\mathbf{hh}$ and $\Delta\mathbf{h}\Delta\mathbf{h}$ are $d_1 \times d_1$ matrices, those with superscripts $\mathbf{hy}$ and $\Delta\mathbf{h}\Delta\mathbf{y}$ are $d_1 \times d_2$ matrices, those with superscripts $\mathbf{yh}$ and $\Delta\mathbf{y}\Delta\mathbf{h}$ are $d_2 \times d_1$ matrices, and those with superscripts $\mathbf{yy}$ and $\Delta\mathbf{y}\Delta\mathbf{y}$ are $d_2 \times d_2$ matrices.

The new delta features make $\mathbf{s}$ no longer independent of its previous frame, and so $\mathbf{s}$ captures the trajectory information of speech by part of the prior information. The new dynamic parameters $\boldsymbol{\mu}_i^{\Delta\mathbf{s}}$ and $\boldsymbol{\Sigma}_i^{\Delta\mathbf{s}}$ provides additional information which can not be inferred from the static parameters $\boldsymbol{\mu}_i^{\mathbf{s}}$ and $\boldsymbol{\Sigma}_i^{\mathbf{s}}$. The dynamic features partly captures the strong, locally defined trajectory property of speech, while the static features captures only the global, loosely specified temporal information of speech [61]. It is speculated that this scheme improve the overall quality of coded speech.

4.3.3 MMSE Estimator

A MMSE estimator $\hat{\mathbf{h}}$ of $\mathbf{h}$ given the observation vector $\mathbf{y}$ is a conditional expectation

$$\hat{\mathbf{h}} = E\{\mathbf{h}|\mathbf{y}\}. \tag{4.26}$$

The conditional pdf of $\mathbf{h}$ given $\mathbf{y}$ is computed from the joint pdf of $\mathbf{s}$. Section 4.3.2 gives a GMM to approximate the joint pdf. The GMM is trained with a training set of speech signals by the EM algorithm (which is described in Section 4.3.1) with the encoder beforehand. Assuming the same environmental condition at the encoder as that of training, the trained GMM parameters are used at the receiver for postfiltering without additional information from the encoder.

Estimation of the Perceptual Postfilter with Static Features

While using the MMSE estimator Eq. (4.26) with a GMM pdf, we need the conditional pdf of the “target” postfilter gain vector $\mathbf{h}$ given the “input” vector $\mathbf{y}$. The conditional pdf and any marginal pdf of jointly Gaussian random variables are still Gaussian densities [62]. Assuming the Gaussian density $\mathcal{N}(\mathbf{s}_n|\boldsymbol{\theta}_i) = \mathcal{N}(\mathbf{s}_n|\boldsymbol{\mu}_i^{\mathbf{s}}, \boldsymbol{\Sigma}_i^{\mathbf{s}})$, the i-th GM component in Eq. (4.23), is a joint density of the variate $\mathbf{s}$ defined in Eq. (4.21), it can be factored into a conditional Gaussian pdf $\mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_i^{\mathbf{h}|\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}})$ of $\mathbf{h}$, given $\mathbf{y}$, with mean vector $\boldsymbol{\mu}_i^{\mathbf{h}|\mathbf{y}}$ and covariance matrix $\boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}}$, and a marginal Gaussian pdf $\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_i^{\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{yy}})$ of $\mathbf{y}$ with mean vector $\boldsymbol{\mu}_i^{\mathbf{y}}$ and covariance matrix $\boldsymbol{\Sigma}_i^{\mathbf{yy}}$

$$\mathcal{N}(\mathbf{s}_n|\boldsymbol{\mu}_i^{\mathbf{s}}, \boldsymbol{\Sigma}_i^{\mathbf{s}}) = \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_i^{\mathbf{h}|\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}})\,\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_i^{\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{yy}}), \tag{4.27}$$

where

$$\boldsymbol{\mu}_i^{\mathbf{h}|\mathbf{y}} = \boldsymbol{\mu}_i^{\mathbf{h}} + \boldsymbol{\Sigma}_i^{\mathbf{hy}}(\boldsymbol{\Sigma}_i^{\mathbf{yy}})^{(-1)}(\mathbf{y} - \boldsymbol{\mu}_i^{\mathbf{y}}), \tag{4.28a}$$

$$\boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}} = \boldsymbol{\Sigma}_i^{\mathbf{hh}} - \boldsymbol{\Sigma}_i^{\mathbf{hy}}(\boldsymbol{\Sigma}_i^{\mathbf{yy}})^{(-1)}\boldsymbol{\Sigma}_i^{\mathbf{yh}}, \tag{4.28b}$$

and $\boldsymbol{\mu}_i^{\mathbf{h}}, \boldsymbol{\mu}_i^{\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{yy}}, \boldsymbol{\Sigma}_i^{\mathbf{hy}}, \boldsymbol{\Sigma}_i^{\mathbf{yh}}$ and $\boldsymbol{\Sigma}_i^{\mathbf{hh}}$ are defined in Eq. (4.25).

For the GMM-modelled joint density of $\mathbf{h}$ and $\mathbf{y}$ defined by Eq. (4.23), the marginal joint density function of $\mathbf{y}$ is

$$p(\mathbf{y}) = \sum_{k=1}^{M} \alpha_k\,\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_k^{\mathbf{y}}, \boldsymbol{\Sigma}_k^{\mathbf{yy}}). \tag{4.29}$$

Therefore, the conditional pdf of $\mathbf{h}$ given $\mathbf{y}$ is expressed in terms of a GMM as

$$\begin{aligned} p(\mathbf{h}|\mathbf{y}) = \frac{p(\mathbf{y},\mathbf{h})}{p(\mathbf{y})} &= \frac{\sum_{i=1}^{M} \alpha_i \mathcal{N}(\mathbf{s}_n|\boldsymbol{\mu}_i^{\mathbf{s}}, \boldsymbol{\Sigma}_i^{\mathbf{s}})}{\sum_{k=1}^{M} \alpha_k \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_k^{\mathbf{y}}, \boldsymbol{\Sigma}_k^{\mathbf{yy}})} \\ &= \frac{\sum_{i=1}^{M} \alpha_i \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_i^{\mathbf{h}|\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}}) \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_i^{\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{yy}})}{\sum_{k=1}^{M} \alpha_k \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_k^{\mathbf{y}}, \boldsymbol{\Sigma}_k^{\mathbf{yy}})} \\ &= \sum_{i=1}^{M} \beta_i(\mathbf{y}) \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_i^{\mathbf{h}|\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}}), \end{aligned} \tag{4.30}$$

where [62]

$$\beta_i(\mathbf{y}) = \frac{\alpha_i \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_i^{\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{yy}})}{\sum_{k=1}^{M} \alpha_k \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_k^{\mathbf{y}}, \boldsymbol{\Sigma}_k^{\mathbf{yy}})}. \tag{4.31}$$

The MMSE estimate of $\mathbf{h}$ is derived with Eq. (4.26) (4.30) and (4.28a)

$$\hat{\mathbf{h}} = \sum_{i=1}^{M} \beta_i(\mathbf{y}) \boldsymbol{\mu}_i^{\mathbf{h}|\mathbf{y}}. \tag{4.32}$$

When diagonal covariance matrices are used for the GM densities, the MMSE estimator is reduced to

$$\hat{\mathbf{h}} = \sum_{i=1}^{M} \beta_i(\mathbf{y}) \boldsymbol{\mu}_i^{\mathbf{h}}. \tag{4.33}$$

Estimation of the Perceptual Postfilter with Static and Dynamic Features

Now with both static and dynamic features to derive a MMSE estimator, we use the pdf of Eq. (4.24). Given the estimated postfilter gain feature in the immediately past frame, $\hat{\mathbf{h}}_{n-1}$, and the corresponding realization of $\Delta\mathbf{y}_n$, the conditional MMSE estimator of the current frame becomes

$$\hat{\mathbf{h}}_{n|n-1} \equiv E\{\mathbf{h}_n|\mathbf{y}_n, \hat{\mathbf{h}}_{n-1}, \Delta\mathbf{y}_n\}. \tag{4.34}$$

As the factorization of $\mathcal{N}(\mathbf{s}_n|\boldsymbol{\mu}_i^{\mathbf{s}}, \boldsymbol{\Sigma}_i^{\mathbf{s}})$ in Eq. (4.27), $\mathcal{N}(\Delta\mathbf{s}_n|\boldsymbol{\mu}_i^{\Delta\mathbf{s}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{s}})$ in Eq. (4.24) can be factored into a conditional Gaussian pdf $\mathcal{N}(\Delta\mathbf{h}|\boldsymbol{\mu}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}})$ of $\Delta\mathbf{h}$, given $\Delta\mathbf{y}$, with mean vector $\boldsymbol{\mu}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}}$ and covariance matrix $\boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}}$, and a marginal Gaussian pdf $\mathcal{N}(\Delta\mathbf{y}|\boldsymbol{\mu}_i^{\Delta\mathbf{y}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}})$ of $\Delta\mathbf{y}$ with mean vector $\boldsymbol{\mu}_i^{\Delta\mathbf{y}}$ and covariance matrix $\boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}}$

$$\mathcal{N}(\Delta\mathbf{s}_n|\boldsymbol{\mu}_i^{\Delta\mathbf{s}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{s}}) = \mathcal{N}(\Delta\mathbf{h}|\boldsymbol{\mu}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}})\,\mathcal{N}(\Delta\mathbf{y}|\boldsymbol{\mu}_i^{\Delta\mathbf{y}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}}), \tag{4.35}$$

where

$$\boldsymbol{\mu}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}} = \boldsymbol{\mu}_i^{\Delta\mathbf{h}} + \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}\Delta\mathbf{y}}(\boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}})^{(-1)}(\Delta\mathbf{y} - \boldsymbol{\mu}_i^{\Delta\mathbf{y}}), \tag{4.36a}$$

$$\boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}} = \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}\Delta\mathbf{h}} - \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}\Delta\mathbf{y}}(\boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}})^{(-1)}\boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{h}}, \tag{4.36b}$$

and $\boldsymbol{\mu}_i^{\Delta\mathbf{h}}$, $\boldsymbol{\mu}_i^{\Delta\mathbf{y}}$, $\boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}}$, $\boldsymbol{\Sigma}_i^{\Delta\mathbf{h}\Delta\mathbf{y}}$, $\boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{h}}$ and $\boldsymbol{\Sigma}_i^{\Delta\mathbf{h}\Delta\mathbf{h}}$ are defined in Eq. (4.25).

Similar to the derivation of Eq. (4.32), the conditional pdf of $\mathbf{h}$ given $\mathbf{y}$ and $\hat{\mathbf{h}}_{n-1}$ is

$$\begin{aligned} p(\mathbf{h}_n|\mathbf{y}_n, \hat{\mathbf{h}}_{n-1}, \Delta\mathbf{y}_n) &= \frac{p(\mathbf{h}_n, \mathbf{y}_n, \Delta\mathbf{y}_n|\hat{\mathbf{h}}_{n-1})}{p(\mathbf{y}_n, \Delta\mathbf{y}_n|\hat{\mathbf{h}}_{n-1})} \\ &\approx \frac{p(\mathbf{s}_n, \Delta\mathbf{y}_n|\mathbf{h}_{n-1})}{p(\mathbf{y}_n, \Delta\mathbf{y}_n)}, \end{aligned} \tag{4.37}$$

where the approximation simplifies the estimator dramatically to avoid dynamic programming, and $p(\mathbf{s}_n, \Delta\mathbf{y}_n|\mathbf{h}_{n-1})$ has the form [61]

$$\begin{aligned} p(\mathbf{s}_n, \Delta\mathbf{y}_n|\mathbf{h}_{n-1}) &= \sum_{i=1}^{M} \alpha_i \mathcal{N}(\mathbf{h}_n|\boldsymbol{\mu}_i^{\mathbf{h}|\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}})\,\mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_i^{\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{y}\mathbf{y}}) \\ &\quad \mathcal{N}(\Delta\mathbf{y}_n|\boldsymbol{\mu}_i^{\Delta\mathbf{y}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}})\,\mathcal{N}(\mathbf{h}_n - \mathbf{h}_{n-1}|\boldsymbol{\mu}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}}) \\ &= \sum_{i=1}^{M} \alpha_i \mathcal{N}(\mathbf{h}_n|\boldsymbol{\mu}_i', \boldsymbol{\Sigma}_i')\,\mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_i^{\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{y}\mathbf{y}})\,\mathcal{N}(\Delta\mathbf{y}_n|\boldsymbol{\mu}_i^{\Delta\mathbf{y}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}}). \end{aligned} \tag{4.38}$$

$\mathcal{N}(\mathbf{h}_n|\boldsymbol{\mu}_i', \boldsymbol{\Sigma}_i')$ is a GM density of $\mathbf{h}_n$ with mean vector

$$\boldsymbol{\mu}_i' = (\boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}} + \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}})^{-1}\boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}}\boldsymbol{\mu}_i^{\mathbf{h}|\mathbf{y}} + (\boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}} + \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}})^{-1}\boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}}(\mathbf{h}_{n-1} + \boldsymbol{\mu}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}}) \tag{4.39}$$

and covariance matrix

$$\boldsymbol{\Sigma}_i' = (\boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}} + \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}})^{-1}\boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}}\boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}}. \tag{4.40}$$

Combining Eq. (4.37) and (4.38) together, we get the conditional pdf

$$p(\mathbf{h}_n|\mathbf{y}_n, \hat{\mathbf{h}}_{n-1}, \Delta\mathbf{y}_n) \approx \frac{\sum_{i=1}^{M} \alpha_i \mathcal{N}(\mathbf{h}_n|\boldsymbol{\mu}'_i, \boldsymbol{\Sigma}'_i)\,\mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_i^{\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{yy}})\,\mathcal{N}(\Delta\mathbf{y}_n|\boldsymbol{\mu}_i^{\Delta\mathbf{y}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}})}{\sum_{k=1}^{M} \alpha_i \mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_k^{\mathbf{y}}, \boldsymbol{\Sigma}_k^{\mathbf{yy}})\,\mathcal{N}(\Delta\mathbf{y}_n|\boldsymbol{\mu}_k^{\Delta\mathbf{y}}, \boldsymbol{\Sigma}_k^{\Delta\mathbf{y}\Delta\mathbf{y}})} \tag{4.41a}$$

$$= \sum_{i=1}^{M} \beta_i(\mathbf{y}_n, \Delta\mathbf{y}_n)\,\mathcal{N}(\mathbf{h}_n|\boldsymbol{\mu}'_i, \boldsymbol{\Sigma}'_i), \tag{4.41b}$$

where

$$\beta_i(\mathbf{y}_n, \Delta\mathbf{y}_n) = \frac{\alpha_i \mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_i^{\mathbf{y}}, \boldsymbol{\Sigma}_i^{\mathbf{yy}})\,\mathcal{N}(\Delta\mathbf{y}_n|\boldsymbol{\mu}_i^{\Delta\mathbf{y}}, \boldsymbol{\Sigma}_i^{\Delta\mathbf{y}\Delta\mathbf{y}})}{\sum_{k=1}^{M} \alpha_i \mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_k^{\mathbf{y}}, \boldsymbol{\Sigma}_k^{\mathbf{yy}})\,\mathcal{N}(\Delta\mathbf{y}_n|\boldsymbol{\mu}_k^{\Delta\mathbf{y}}, \boldsymbol{\Sigma}_k^{\Delta\mathbf{y}\Delta\mathbf{y}})}. \tag{4.42}$$

The final MMSE estimation of $\mathbf{h}$ is obtained by substituting Eq. (4.41) into Eq. (4.34)

$$\begin{aligned} \hat{\mathbf{h}}_{n|n-1} &\approx \sum_{i=1}^{M} \beta_i(\mathbf{y}, \Delta\mathbf{y}_n)\,\boldsymbol{\mu}'_i \\ &\approx \sum_{i=1}^{M} \beta_i(\mathbf{y}, \Delta\mathbf{y}_n)\,[\boldsymbol{\Psi}_\mathbf{1}(i)\boldsymbol{\mu}_i^{\mathbf{h}|\mathbf{y}} + \boldsymbol{\Psi}_\mathbf{2}(i)(\hat{\mathbf{h}}_{n-1} + \boldsymbol{\mu}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}})], \end{aligned} \tag{4.43}$$

where

$$\boldsymbol{\Psi}_\mathbf{1}(i) = (\boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}} + \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}})^{-1}\boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}} \tag{4.44a}$$

$$\boldsymbol{\Psi}_\mathbf{2}(i) = (\boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}} + \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}|\Delta\mathbf{y}})^{-1}\boldsymbol{\Sigma}_i^{\mathbf{h}|\mathbf{y}} \tag{4.44b}$$

with

$$\boldsymbol{\Psi}_\mathbf{1}(i) + \boldsymbol{\Psi}_\mathbf{2}(i) = \mathbf{I} \qquad \forall i.$$

With diagonal covariance matrices for the GM densities, Eq. (4.43) is reduced to

$$\hat{\mathbf{h}}_{n|n-1} \approx \sum_{i=1}^{M} \beta_i(\mathbf{y}, \Delta\mathbf{y}_n)\,[\boldsymbol{\Psi}_\mathbf{1}(i)\boldsymbol{\mu}_i^{\mathbf{h}} + \boldsymbol{\Psi}_\mathbf{2}(i)(\hat{\mathbf{h}}_{n-1} + \boldsymbol{\mu}_i^{\Delta\mathbf{h}})] \tag{4.45}$$

and

$$\Psi_1(i) = (\boldsymbol{\Sigma}_i^{\mathbf{hh}} + \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}\Delta\mathbf{h}})^{-1}\boldsymbol{\Sigma}_i^{\Delta\mathbf{h}\Delta\mathbf{h}} \tag{4.46a}$$

$$\Psi_2(i) = (\boldsymbol{\Sigma}_i^{\mathbf{hh}} + \boldsymbol{\Sigma}_i^{\Delta\mathbf{h}\Delta\mathbf{h}})^{-1}\boldsymbol{\Sigma}_i^{\mathbf{hh}}. \tag{4.46b}$$

Given the trained GMM parameters in Eq. (4.23) or Eq. (4.24), the postfilter gains can be easily estimated by the MMSE estimator of Eq. (4.32) or Eq. (4.46) and applied to the decoded speech at the receiver.

Chapter 5

Experimental Results

In this chapter, we present the integration of the perceptual postfiltering method into a LPAS speech coder and the experimental results. The ITU-T Recommendation G.723.1 speech codec [5] at rate of 5.3 kbps is chosen for the simulation. In Section 5.1, details of algorithm implementation of the G.723.1 standard are described. Section 5.2 presents the probabilistic dependency between "input" and "output" features of the MMSE estimator by information measure. The experimental results are presented in Section 5.3.

5.1 Algorithm Implementation

We incorporate the perceptual postfilter based on GMM, which was introduced in Section 4.3, into a low bit rate speech codec to improve the decoded speech quality. In the experiment, all speech is sampled at 8 kHz with 16-bit PCM resolution. Two sets of clean speech signals recorded under the same condition are used as the test material. One set is for training, and the other one is for evaluation.

The experiment involves three steps:

1. Generation of the training data set
2. GMM training
3. Implementation of the perceptual postfilter with the trained GMM

The proposed perceptual postfilter has been designed to reduce the perceived level of noise in low rate speech coders. We did the simulation of the proposed perceptual postfilter with the

G.723.1 speech codec [5] at rate of 5.3 kbps. The G.723.1 speech codec operates on frames of 240 samples. Each frame is divided into four subframes of 60 samples each. For each subframe, 10th order LP analysis is used on a Hamming windowed block of 180 samples centered on the subframe. The LP coefficients for the last subframe are converted to LSFs and quantized. For each subframe, linear interpolation is performed between the quantized LSFs of the current frame and the quantized LSFs of the previous frame to derive the quantized LSFs for the current subframe. The excitation signal is coded with a pitch period and *algebraic-code-excitation* for each subframe.

A local speech database was used. The database was composed of speech of 23 speakers (12 females and 11 males). In considering the limited size of the speech database, we choose a section of the database with 10 female and 9 male speech for GMM training, and the rest of the database with 2 females and 2 males was used for performance evaluation.

The G.723.1 speech coder encoded the corresponding information about excitation and LSFs. For each frame, only the information about the first and the third subframes are used in training. A feature vector was constructed from each processing block of 180 samples (3 coding subframes) centered on the first or the third subframe of each frame. There was an overlapping of 60 samples for adjacent processing blocks. First, a decoding feature vector $\mathbf{y}$ of dimension 12 was derived from the coded information of the current subframe. The 10 quantized LSFs were used as a sub-feature vector representing the speech short-term spectral property, $\mathbf{b}$, while pitch and its corresponding *long-term prediction* (LTP) gain represented the long-term spectral property, $\mathbf{n}$. The LSFs and pitch were obtained directly from the coded information of the center subframe. The LTP gain was calculated from the coded speech with the coded pitch and corresponding subframe. Then a sine-squared window was applied to the first 60 and the last 60 samples of the processing blocks of the original and decoded speech as shown in Fig. 5.1. A 512-point FFT was used on each windowed block. A 17 bark-scale perceptual postfilter gain vector $\mathbf{h}$ was derived from Eq. (4.13). Therefore, a realization of $\mathbf{s}$ with dimension 29 in Eq. (4.21) was obtained with the realizations of $\mathbf{y}$ and $\mathbf{h}$ for each processing block. We actually used the dB value of the perceptual postfilter gains for $\mathbf{h}$. By passing the training speech set through the encoder, a training set consisting of 338,916 vectors was generated.

Diagonal covariance matrices were used for both the GMM of static features only (Eq. (4.23)) and the GMM of static and dynamic features (Eq. (4.24)). The GMM of static features was trained with the training set. For training the GMM of static and dynamic features, the dynamic features were created with Eq. (4.22). The first static feature in each sentence was set as the initial feature for that sentence. Consequently, the static and dynamic features in the rest of training sentences formed the training data set for the GMM of static and dynamic features. The EM algorithm is generally satisfactory to train a GMM when the number of parameters to be estimated is small with respect to the number of training observations. Usually, the size of the training set should be at least 50 to 100 times of the number of estimated parameters [63]. For the GMM of static features with dimension $17 + 12 = 29$ and the GMM of static and dynamic features with dimension $17+12+12 = 41$, the mixture number should be less than 100 for better modelling.

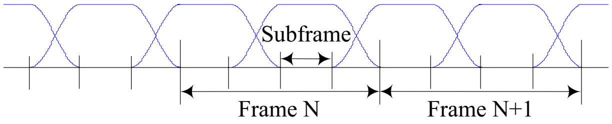


Fig. 5.1 Windowing for the training and perceptual postfiltering.

By incorporating the GMM parameters into Eq. (4.33) or (4.45), the perceptual postfilter was estimated with decoding feature vectors from the received coding information. For every two subframes of the decoded speech, the postfiltering was performed on windowed blocks of 180 samples, with 60 sample overlaps. The decoded LSFs, pitch and corresponding calculated LTP gain of the center subframe of the processing block were used as the decoding feature vector to derive the perceptual postfilter with Eq. (4.33) or (4.45). The same window in Fig. 5.1 was used and the windowed processing block was transformed into the frequency domain with the same length FFT[1] in training data set generation. For those frequency components within a bark band i at the n-th frame, the same postfilter gain $\hat{H}(i,n)$ was applied to their Fourier magnitudes. The modified Fourier magnitudes were then transformed back to the time domain with IFFT combined with the phase of the decoded speech frame. The overlap-add method [64] was used to combine the processed blocks into the final modified signal. Speech in the evaluation set were passed through the G.723.1 speech encoder at rate of 5.3 kbps, and then decoded and postfiltered with the bottom system in Fig. 4.2.

The generation of the training data set has been executed in the C language. The training of the GMM parameters has been done in Matlab. We have implemented the proposed perceptual postfilter in C.

[1] For linear FIR filtering in the frequency domain, the size of the FFT and IFFT must be at least $N_F = L+M-1$ to avoid the aliasing that results in the time domain [64]. Here, L is the size of the processing block and M is the size of the estimated perceptual postfilter.

5.2 Probabilistic Dependency between “Input” and “Output” Features

It would provide us with a better understanding about the possibilities to successfully estimate the perceptual postfilter gains with the MMSE estimator based on GMM, if we check the dependency between the “input” features and the “output” features.

For the case of estimation with static features only, we need to find how large the remaining uncertainty of the perceptual postfilter gains is given the decoding features. This was done by determining the ratio between the mutual information $I(\mathbf{h},\mathbf{y})$ of $\mathbf{h}$ and $\mathbf{y}$ and the entropy $H(\mathbf{h})$ of $\mathbf{h}$ [65]. The joint density function of $\mathbf{h}$ and $\mathbf{y}$ was modelled by a GMM $p(\mathbf{h},\mathbf{y})$. We can easily obtain the marginal densities $p(\mathbf{h})$ and $p(\mathbf{y})$ with $p(\mathbf{h},\mathbf{y})$ (See Section 4.3.3). The GMM training set was generated from Section 5.1. The estimate of the mutual information, $\widehat{I}(\mathbf{h},\mathbf{y})$, and the estimate of the differential entropy, $\widehat{h}(\mathbf{h})$, were obtained from [65]

$$\widehat{I}(\mathbf{h},\mathbf{y}) = \frac{1}{N}\sum_{n=1}^{N}\Big(\log_2\big(\frac{p(\mathbf{h}_n,\mathbf{y}_n)}{p(\mathbf{h}_n)p(\mathbf{y}_n)}\big)\Big), \tag{5.1}$$

$$\widehat{h}(\mathbf{h}) = -\frac{1}{N}\sum_{n=1}^{N}\log_2(p(\mathbf{h}_n)), \tag{5.2}$$

where the sample vector sets $\{\mathbf{h}_n\}$, $\{\mathbf{y}_n\}$ and $\{\mathbf{h}_n,\mathbf{y}_n\}$ were generated from the GMM and each contained N vectors. We used $N = 10^6$.

The relationship between the entropy $H(\mathbf{h})$ and the differential entropy can be express as [65]

$$H(\mathbf{h}) \approx h(\mathbf{h}) - \log_2(\Delta^{d_1}), \tag{5.3}$$

where d_1 is the dimension of the vector $\mathbf{h}$ and Δ is the quantization step. Since we applied the dB value of the perceptual postfilter gains for $\mathbf{h}$, we selected $\Delta = 1$ according to [65]. Then the entropy is equal to the differential entropy of $\mathbf{h}$.

For the estimation with both static and dynamic features, the estimate of the mutual information, $\widehat{I}(\mathbf{h},\Delta\mathbf{h},\mathbf{y},\Delta\mathbf{y})$, and the estimate of the entropy, $\widehat{H}(\mathbf{h})$, were obtained similarly. Due to the limited size the the training set, we tested mixture components less than 100. Table 5.1 and Table 5.2 present the mutual information from GMM pdf with M=8, 16, 32, 64, and 80 mixture components, the entropies, and the ratios between the mutual information and the entropies.

From Table 5.1 and Table 5.2, we can see that the mutual information is only a small fraction of the “target” entropy, while the dependency increases slightly with more mixtures. The results

Table 5.1 Information Results for Static Features.

Gaussian Mixtures	$\widehat{I}(\mathbf{h},\mathbf{y})$	$\widehat{H}(\mathbf{h})$	$\widehat{I}(\mathbf{h},\mathbf{y})/\widehat{H}(\mathbf{h})(\%)$
8	3.85	66.49	5.80
16	3.77	66.28	5.69
32	3.99	65.84	6.06
64	4.26	65.43	6.51
80	4.38	65.42	6.70

Table 5.2 Information Results for Static+Dynamic Features.

Gaussian Mixtures	$\widehat{I}(\mathbf{h},\Delta\mathbf{h},\mathbf{y},\Delta\mathbf{y})$	$\widehat{H}(\mathbf{h})$	$\widehat{I}(\mathbf{h},\Delta\mathbf{h},\mathbf{y},\Delta\mathbf{y})/\widehat{H}(\mathbf{h})(\%)$
8	4.49	66.50	6.76
16	4.59	66.31	6.93
32	5.70	65.90	8.66
64	5.46	65.67	8.31
80	6.18	65.55	9.44

show GMMs with both static and dynamic features have less uncertainty than those only with static features. However, mutual information is a statistical tool. Although its value is not high, the “input” and the “target” may still be perceptually bounded well.

5.3 Results and Discussion

Comparison of Postfilter Gains and Spectrograms

A GMM with M=80 was used for perceptual postfiltering. The ideal perceptual postfilter was described in Eq. (4.10). We compared the G.723.1 standard formant adaptive postfilter with our perceptual postfilter based on GMM in postfilter gains and spectragrams.

A female speech was used for comparison. Its waveform is given in Fig. 5.2. Also, Fig. 5.3 shows the postfilter gains of 30 sequential frames. In Fig. 5.2, the corresponding time interval of those frames is specified between the two lines.

Comparing the postfilter gains in Fig. 5.3, we see that the MMSE estimations generally follow the ideal postfilter trend to some extent, while the G.723.1 standard formant postfilter has very little effect on the speech. This is because the G.723.1 standard formant postfilter is mainly determined by LPCs which change slowly frame by frame, and λ_1, λ_2 and μ constrain its dynamic range to be very limited comparing with the ideal postfilter.

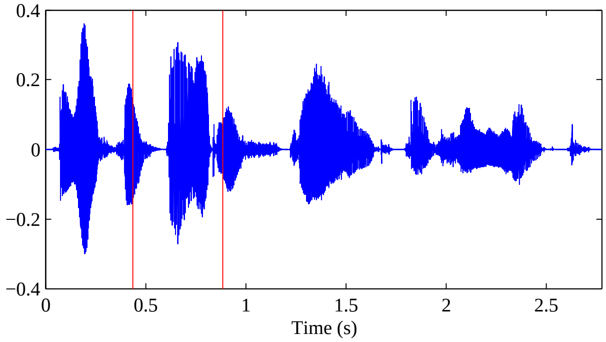


Fig. 5.2 A female speech waveform.

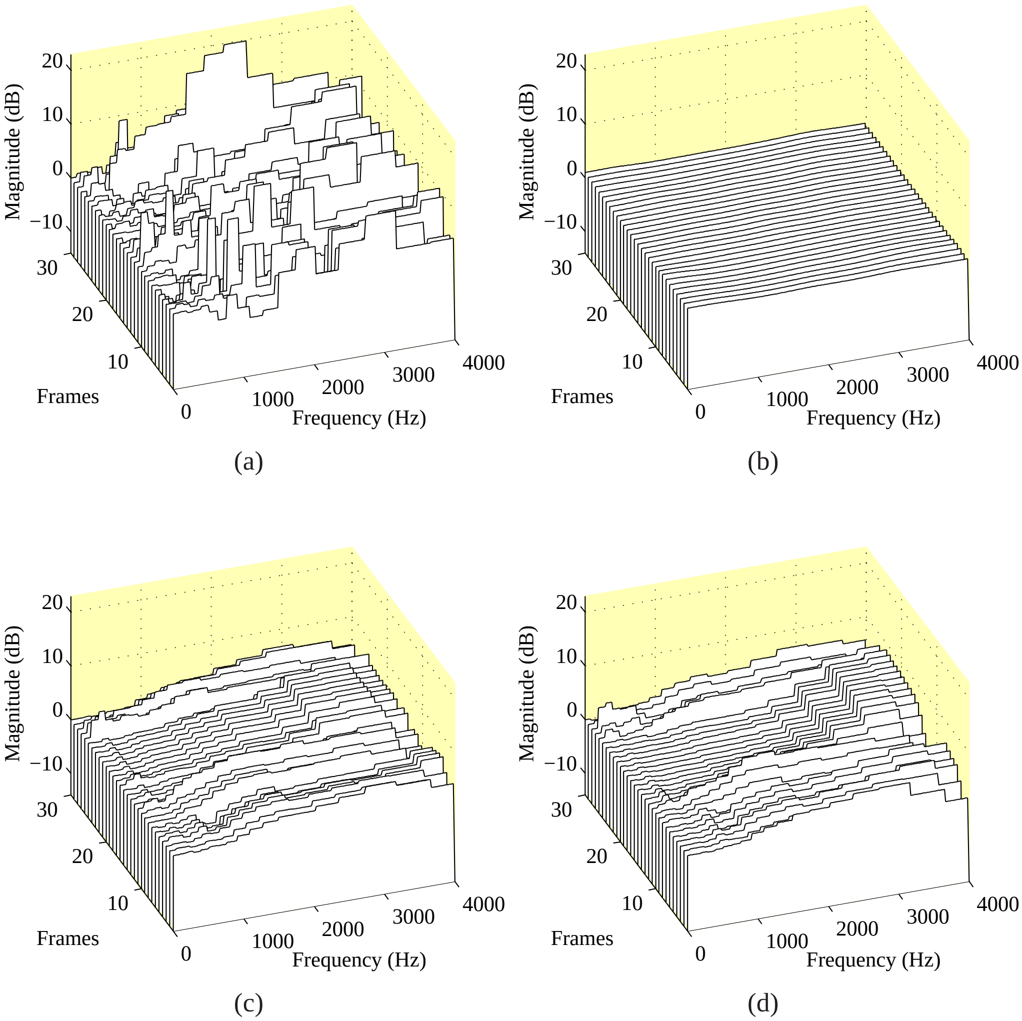
Fig. 5.4 shows the spectrograms of clean speech (*First*), G.723.1 coded speech with standard postfiltering (*Second*), G.723.1 coded speech with the new perceptual postfilter using static features only (*Third*), and G.723.1 coded speech with the new perceptual postfilter using both static and dynamic features (*Fourth*), respectively.

Low bit rate LPAS coding emphasizes the high energy parts (generally formants at low frequencies) and loses some naturalness at high frequencies. From Fig. 5.4, it can be seen that the perceptual postfilters recover some of the high frequency loss from encoding, while the difference between postfilter estimation with only static features and that with both static and dynamic features is not obvious.

Subjective Evaluation

Although there are some objective quality measures (See Section 2.6.2) to evaluate the performance of our algorithm, we found that our proposed perceptual postfilter has lower scores than the conventional adaptive postfilter with those measures. The best measure of perceptual quality of speech is the *mean opinion score* (MOS), which is obtained from a formal subjective listening test. Since it is difficult to gauge the effectiveness of postfiltering quantitatively by objective measures [15] and formal MOS tests are not available in our research environment, we evaluated the perceptual-quality-improving capability of the perceptual postfilter by informal listening tests.

In order to measure the subjective performance of the perceptual and the conventional postfilters, informal tests were used with 6 untrained listeners. 8 sentences pairs for 4 speakers (2 male and 2 female speakers in the evaluation speech set) were processed by the G.723.1 codec at rate of 5.3 kbps. The decoded speech signals were modified by the proposed perceptual postfilter and



(a)

(b)

(c)

(d)

Fig. 5.3 Postfilter Gains: (a) Ideal Perceptual Postfilter Gains; (b) ITU-T G.723.1 Rate 5.3 kbps Formant Postfilter Gains; (c) Estimated Perceptual Postfilter Gains by 80 GMM with Static Features Only; (d) Estimated Perceptual Postfilter Gains by 80 GMM with Static and Dynamic.

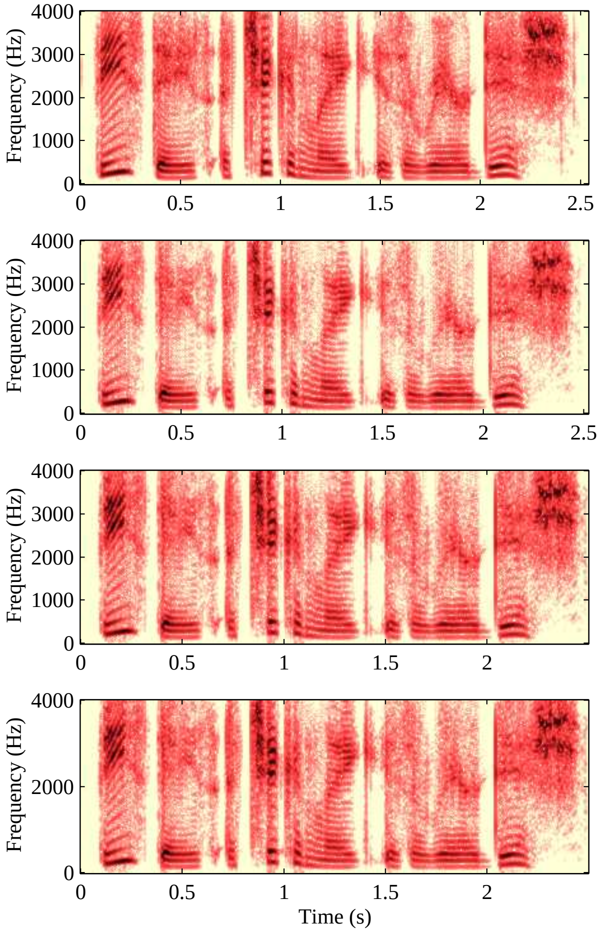


Fig. 5.4 Spectrograms. First: Original speech; Second: ITU-T G.723.1 coded speech with the standard postfiltering; Third: ITU-T G.723.1 coded speech with the perceptual postfiltering estimated with static features only; Fourth: ITU-T G.723.1 coded speech with the perceptual postfiltering estimated with both static and dynamic features.

the standard conventional postfilter, respectively. For each sentence, the two postfiltered versions were compared according to the original clean speech. The listeners were asked to pick one of the two postfiltered speech which they preferred and give the reason.

The test speech signals were presented over both headphones and loudspeakers to the listeners. Informal listening tests showed that the proposed perceptual postfilter gives much more natural sound than the conventional postfilter for most of the tested speech. This is because the high frequency distortion from coding is lessened by the proposed perceptual postfilter. Postfiltered speech with filter estimation based on both static and dynamic features gave a bit smoother sound than that based only on static features. With static and frame-differential features, the perceptual postfilter captures some dynamic information of speech.

With loudspeakers, the listeners all preferred the perceptual postfiltered speech signals to the standard postfiltered ones. The improved naturalness with the proposed postfilter delivered significantly better quality, while the standard postfilter still sounded thin and a bit muffled (which is common in low bit rate speech coding). However, a little degradation was audible in some perceptual postfiltered speech when listened with headphones. The standard postfiltered one sounded smoother. This may be caused by the fact that the G.723.1 speech coder is a LPAS coder and not primarily perceptual-based, while the ideal perceptual postfilter is totally perceptual motivated and could change abruptly from frame to frame. The conventional postfilter is based on pitch and LSFs which change slowly. The ideal conventional postfilter itself has a smooth contour and small dynamic range, and the conventional postfilter is a closer resemblance to it. However, the nature of the ideal conventional postfilter made it hard to improve the other aspects of the speech quality (for instance, naturalness and intelligibility) other than some coding noise reduction. Our proposed perceptual postfilter improves the decoded speech quality by recovering some information in the original speech, but introduces some unforeseeable distortion at the same time as well. In some part of consecutive processing blocks, the estimates of the perceptual postfilter cannot catch the fast change of the ideal perceptual postfilter. This may be the reason why the objective measure scores of our perceptual postfiltered speech were worse than those with the conventional postfilter.

Chapter 6

Conclusion

Psychoacoustic principles have been widely used in low bit rate speech and audio signal processing. Bit rate reduction can be achieved without coding the perceptually irrelevant information. Active research has been increasingly concentrated on exploiting human auditory properties in speech and audio coding in the past decade. The masking phenomena for noise reduction are the key theory for its practical applications. Quantization noise control, speech enhancement and objective quality measurements are the major applications. The examples are noise suppressor (speech enhancement) in Enhanced Variable Rate Codec (EVRC) [22], coding noise shaping with auditory models (quantization noise control) in Moving Pictures Experts Group (MPEG) standards—MPEG-1 [66], MPEG-2 [67] and MPEG-4 [68], and PEAQ [13] for audio perceptual quality measure.

The main goal of this thesis has been to improve perceptual quality of low bit rate coded speech. This work has focused on design and implementation of a perceptual postfiltering technique based on perceptual models. A novel perceptual postfilter for low bit rate LPAS speech coders has been introduced in this thesis. The proposed postfilter is perceptually based and is an add-on part at the receiver just as a conventional adaptive postfilter. It has shown that, with the proposed postfilter, speech quality is improved with a more natural sound than the conventional postfiltered speech.

6.1 Summary of Our Work

After a brief introduction about speech coding methods, especially utilization of the masking concept, Chapter 1 outlined motivation and objective of the work in this thesis. Chapter 2 started off by introducing the bases of the modern LPAS speech coders. In order to reduce perceptual distortion from LPAS coders, noise shaping and adaptive postfiltering (both based on masking properties) are exploited in the encoder and the decoder, respectively. Emphasis was placed on methods of adaptive postfiltering. Due to the theory behind LPAS coding and masking properties, conventional adaptive postfiltering has two parts: long-term postfiltering and short-term postfiltering. Various realizations of adaptive postfiltering were included. Speech quality measurements were also described at the end of Chapter 2.

Chapter 3 concentrated on presenting three psychoacoustic models. The description of the models begined with Johnston's masking model in Section 3.2. This model is used to control the coding noise in a perceptual transform coder. In Section 3.3, two models from PAQM and PEAQ were presented. These models are parts of the original objective quality evaluation procedures. The calculations of these psychoacoustic models are similar, and the difference is in the detailed operations. Also, the intermediate models, such as the masking model and the excitation model, are very useful in speech and audio signal processing.

In Chapter 4, a novel postfiltering method combining perceptual properties and statistical estimation together was proposed. Specific perceptual properties was applied to the postfiltering other than the masking threshold concept used in the conventional postfilter. First, the proposed postprocessing structure was given in Section 4.1. A perceptual postprocessor model is easily applied to current low bit rate narrowband LPAS coders. The proposed system diagram was given in Section 4.1.1. Section 4.2 developed a perceptual postfilter scheme. The idea was motivated by a generalized perceptual filter by Lam and Stewart in Section 4.2.1: perceived coding noise is suppressed by setting internal representation of the modified coded signal to that of the original signal. Under the assumption that the psychoacoustic model is an invertible auditory model, Section 4.2.2 derived a new perceptual postfilter which is based on equalizing the critical band intensities between the original and the coded signals.

At the decoder, the original signal is unavailable. Section 4.3 builded a MMSE estimator of the perceptual postfilter given information at the decoder with a GMM-modelled pdf. The GMM was trained at the encoder where the perceptual filter gains are easy to get with the availability of both the original and the coded signals. The EM algorithm for GMM training was described in Section 4.3.1. Derivation of the MMSE estimators with given features were presented in Section 4.3.3. Both static and dynamic features were taken into consideration.

Chapter 5 described how the algorithm was utilized in a real speech coder and the resulting performance. ITU-T G.723.1 speech codec at rate of 5.3 kbps was examined. The LSFs, pitch and LTP gains from the decoder were used to estimate the perceptual postfilter gains with a MMSE estimator using a GMM. Low bit rate coding emphasizes the high energy parts (generally formants at low frequencies) and loses some naturalness at high frequencies. The perceptual postfilter recovered some of the high frequency loss. Informal listening tests have shown an improved speech quality with a more natural sound.

6.2 Future Research Directions

This section provides possible future research in perceptual postfiltering. The design of the perceptual postfiltering scheme mainly depends on the speech enhancement of speech corrupted with speech-correlated noise. Furthermore, perceptual postfiltering can be applied in other aspects in speech coding.

- Better perceptual postfiltering methods

  Our perceptual postfilter is derived from the internal representations of a basilar membrane model. At the same time, the linear error spectra analysis would yield some additional information about the distortions [25]. Kleijn [34, 69] studied some methods of improving the speech periodicity to get better perception. While using a FFT-based perceptual model, it is possible to incorporate the perception of fundamental frequency in postfiltering. Also, our perceptual postfilter is based on a level-independent spreading function. Further research can consider level-dependent spreading functions which is closer to how the human ear works.

- Enhancement for noisy speech

  A speech coder is designed to work with clean speech. Accurate estimation of the coder parameters is impossible with ambient noise. The performance of a speech coder can be very bad under noisy environments. A noise suppressor is usually applied before encoding in a practical speech coder to reduce additive noise, for instance in [22]. This part is independent of coding. A noise suppressor could be built with the proposed perceptual postfilter to enhance noisy speech. The GMM parameters will be adjusted according to SNR estimation and the perceptual filter gains will be estimated with the noisy input and the GMM parameters.

- Preprocessing for Speech Coders

LPAS codecs update the pitch information on a block-by-block basis. This pitch distortion makes the coded speech noisy, especially obvious at low bit rates. Generalized Analysis-by-Synthesis Coding has been studied [69] and implemented in [7]. It preprocesses speech before encoding to improve the pitch prediction. [70] removed the perceptually irrelevant simultaneously masked frequency components of a speech signal by a masking model to get a more efficient coding than the original signal without significant degradation of the speech quality. While the perspective of perceptual information implementation is different from the masking model, a method similar to perceptual postfiltering can be applied to improve the quality of a speech coder by preprocessing speech with a perceptual model.

- Embedding perceptual postfiltering in the speech encoder

  Operating at the decoder end, postfiltering is not considered at the encoder. Its well-known shortcomings are both speech distortion and noise enhancement. If we can incorporate perceptual filtering in encoding, the problem of speech distortion will be lessened. Most low bit rate speech coders work in the time domain. During encoding, the excitation signal, which gives the least weighted MSE of the speech signal, is chosen by passing each candidate excitation through the LP synthesis filter. Adding a frequency domain technique within the time domain analysis-by-synthesis loop is not easy. A possible solution is to get a time-domain filter from its counterpart in the frequency domain. For example, a time-domain all-pole filter can be derived from the magnitude spectrum of a perceptual postfilter. Its power spectral density is approximated with the periodogram which is directly calculated from the magnitude spectrum of the postfilter [71]. Therefore, the all-pole filter coefficients is easily obtained with inverse Fourier transformed power spectral density by the Levinson-Durbin algorithm [71].

Appendix A

Derivation of the General Perceptual Filter from PAQM

A general perceptual filter is given in [52] based on Eq. (4.1), Eq. (4.4) and Section 3.3.1. The perceptual filter is applied to enhance the quality of the coded audio signal.

An assumption[1] is used in the calculation of the excitation intensity in Eq. (3.15)

$$1 + 0.2dz(i - v) \approx 1. \tag{A.1}$$

This approximation ignores the level dependency of the upper slope of the frequency spreading function. Therefore, for the original signal frame $s(m, n)$, its excitation level function satisfies

$$\begin{aligned}
S_e^{\alpha/2}(i,n) &= \sum_{v=i}^{B-1} \left[10^{-S_l(v-i)dz/10} S_t(v,n)\right]^{\alpha/2} + \sum_{v=0}^{i-1} \left[10^{S_0(v)(i-v)dz/10} S_t(v,n)\right]^{\alpha/2} \\
&= \sum_{v=i}^{B-1} S_t^{\alpha/2}(v,n)\left[10^{-S_l(v-i)dz/10}\right]^{\alpha/2} \\
&\quad + \sum_{v=0}^{i-1} S_t^{\alpha/2}(v,n)\left[10^{S_0(v)(i-v)dz/10}\right]^{\alpha/2} \\
&= \sum_{v=0}^{B-1} S_t^{\alpha/2}(v,n) C_{i,v}, \quad 0 \le i \le B-1,
\end{aligned} \tag{A.2}$$

[1] The assumption Lam and Stewart [52] used is $1 + 0.02dz(i - v) \approx 1$.

where $S_t(v,n)$ is the time-domain smeared pitch representation of $s(m,n)$. The definition and derivation of $S_t(v,n)$ is given in Section 3.3.1. It is clear that $C_{i,v}$ is only frequency dependent. The excitation level values of the original signal are calculated by Eq. (A.2). Similarly, the excitation function of the coded signal $\hat{s}(m,n)$ is

$$\hat{S}_e^{\alpha/2}(i,n) = \sum_{v=0}^{B-1} \hat{S}_t^{\alpha/2}(v,n) C_{i,v}, \quad 0 \le i \le B-1. \tag{A.3}$$

Given Eq. (4.1), the modified signal has an excitation function as

$$\begin{aligned}
\tilde{S}_e^{\alpha/2}(i,n) &= \sum_{v=0}^{B-1} \tilde{S}_t^{\alpha/2}(v,n) C_{i,v} \\
&= \sum_{v=0}^{B-1} C_{i,v} \left\{ \sum_{j=n-1}^{n} T_f(v,j) \tilde{S}_a(v,j) \right\}^{\alpha/2} \\
&= \sum_{v=0}^{B-1} C_{i,v} \left\{ \sum_{j=n-1}^{n} T_f(v,j) a_0(v) \sum_{k \in b_v} \tilde{S}_p(k,j) \right\}^{\alpha/2} \\
&= \sum_{v=0}^{B-1} C_{i,v} \left\{ \sum_{j=n-1}^{n} T_f(v,j) a_0(v) \sum_{k \in b_v} H(v,j) \hat{S}_p(k,j) \right\}^{\alpha/2}.
\end{aligned} \tag{A.4}$$

The time varying filter is assumed to vary very slowly $H(v,j) = H(v)$. Combining Eq. (4.4) and Eq. (A.4), we set the excitation value of the modified signal to that of the original signal

$$\begin{aligned}
S_e^{\alpha/2}(i,n) &= \sum_{v=0}^{B-1} H^{\alpha/2}(v) C_{i,v} \left\{ \sum_{j=n-1}^{n} T_f(v,j) a_0(v) \sum_{k \in b_v} \hat{S}_p(k,j) \right\}^{\alpha/2} \\
&= \sum_{v=0}^{B-1} H^{\alpha/2}(v) \hat{S}_t^{\alpha/2}(v,n) C_{i,v}, \quad 0 \le i \le B-1.
\end{aligned} \tag{A.5}$$

The perceptual filter coefficients $H(v)$ are obtained by solving these equations in Eq. (A.5).

References

[1] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. John Wiley & Sons, 2003.

[2] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*. Elsevier, 1995.

[3] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, “MELP: The new federal standard at 2400 bps,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (Munich, Germany), pp. 1591–1594, Apr. 1997.

[4] US Federal Standards, *Specifications for the Analog to Digital Conversion of Voice by 2,400 Bit/Second Mixed Excitation Linear Prediction*. Draft, May 1998.

[5] ITU-T, *Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s*. ITU-T Recommendation G.723.1, International Telecommunication Union, Mar. 1996.

[6] ITU-T, *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Predicition (CS-ACELP)*. ITU-T Recommendation G.729, International Telecommunication Union, Mar. 1996.

[7] 3GPP2, *Selectable Mode Vocoder Service Option for Wideband Spread Spectrum Communication Systems*. 3GPP2 - Speech Service, Jun. 2001.

[8] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, Inc., 2nd ed., 2001.

[9] L. Robles and M. A. Rugerra, “Mechanics of the mammalian cochlea,” *Physiological Reviews*, vol. 81, pp. 1305–1352, Jul. 2001.

[10] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 5th ed., 2003.

[11] J. D. Gordy and R. A. Goubran, “On the perceptual performance limitations of echo cancellers in wideband telephony,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 33–42, Jan. 2006.

[12] B. C. J. Moore, *Frequency Selectivity in Hearing*. Academic Press, 1986.

[13] ITU-R, *Method for Objective Measurements of Perceived Audio Quality*. ITU-R Recommendation BS. 1387, International Telecommunication Union, Dec. 1998.

[14] R. V. Cox and P. Kroon, “Low bit-rate speech coders for multimedia communication,” *IEEE Commun. Mag.*, pp. 34–41, Dec. 1996.

[15] J.-H. Chen and A. Gersho, “Adpative postfiltering for quality enhancement of coded speech,” *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 59–71, Jan. 1995.

[16] A. Mustapha and S. Yeldener, “An adaptive post-filtering technique based on a least squares approach,” in *Proc. IEEE Speech Coding Workshop*, (Porvoo, Finland), pp. 156–158, 1999.

[17] J. Makhoul, “Linear prediciton: A tutorial review,” *Proc. IEEE*, vol. 63, pp. 124–142, Apr. 1975.

[18] F. Itakura, “Line spectrum representation of linear predictive coefficients of speech signals,” *JASA*, vol. 57, p. S35, Apr. 1975.

[19] J. S. Marques, I. M. Trancoso, J. M. Tribolet, and L. B. Almeida, “Improved pitch prediction with fractional delays in celp coding,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (Hongkong, China), pp. 665 – 668, 1990.

[20] P. Kroon, E. Deprettere, and R. Sluyter, “Regular-pulse excitation–a novel approach to effective and efficient multipulse coding of speech,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 1054–1063, Oct. 1986.

[21] ITU-T, *Coding of Speech at 16 kbit/s Using Low-delay Code Excited Linear Prediction*. ITU-T Recommendation G.728, International Telecommunication Union, Sept. 1992.

[22] TIA/EIA/IS-127, *Enhanced Variable-Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*. Draft, Feb. 1996.

[23] J. Skoglund and W. B. Kleijn, “On time-frequency masking in voiced speech,” *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 361–369, Jul. 2000.

[24] P. Kroon and B. S. Atal, “Strategies for improving the performance of CELP coders at low bit rates,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (New York, USA), pp. 151–154, 1988.

[25] T. Thiede, *Perceptual Audio Quality Assessment Using a Non-Linear Filter Bank*. PhD thesis, Technical University of Berlin, 1999.

[26] B. S. Atal and M. R. Schroeder, “Predictive coding of speech signals and subjective error criteria,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 247–254, Jun. 1979.

[27] V. Ramamoorthy and N. S. Jayant, “Enhancement of ADPCM speech by adaptive postfiltering,” *Bell Syst. Tech. J.*, vol. 63, pp. 1465–1475, Oct. 1984.

[28] N. S. Jayant and V. Ramamoorthy, “Adaptive postfiltering of 16 kb/s-ADPCM speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 829–832, Apr. 1986.

[29] V. Ramamoorthy, N. S. Jayant, R. V. Cox, and M. M. Sondhi, “Enhancement of ADPCM speech coding with backward-adaptive algorithms for postfiltering and noise feedback,” *IEEE J. Sel. Areas Commun.*, vol. 6, pp. 364–382, Feb. 1988.

[30] Y. Yatsuzuka, S. Iizuka, and T. Yamazaki, “A variable rate coding by APC with maximum likelihood quantization from 4.8 kbit/s to 16 kbit/s,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 3071–3074, Apr. 1986.

[31] J.-H. Chen, *Low-Bit-Rate Predictive Coding of Speech Waveforms Based on Vector Quantization*. PhD thesis, University of California, Santa Barbara, Mar. 1987.

[32] O. Ghitza and J. L. Goldstein, “Scalar LPC quantization based on formant JNDs,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 697–708, Aug. 1986.

[33] H. W. Lee, J. J. Kim, K. A. Jang, and M. J. Bae, “The speech enhacement of the G.723.1 vocoder using multi-order formant postfilter,” in *TENCON 99. Proc. IEEE Region 10 Conf.*, pp. 1710–713, 1999.

[34] W. B. Kleijn, “Enhancement of coded speech by constrained optimization,” in *Proc. IEEE Speech Coding Workshop*, (Tsukuba City, Ibaraki, Japan), pp. 163–165, Oct. 2002.

[35] P. Kabal, “ITU-T G.723.1 speech coder: A Matlab implementation,” Tech. Rep., McGill University, Aug. 2004. Available: `http://www-mmsp.ece.mcgill.ca/Documents/Reports/2004/KabalR2004a.pdf`.

[36] V. Grancharov, *Human Perception in Speech Processing*. PhD thesis, Royal Institute of Technology, Jun. 2006.

[37] A. Mustapha and S. Yeldener, “An adaptive post-filtering technique based on the modified Yule-Walker filter,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, pp. 197–200, Mar. 1999.

[38] B. Friedlander and B. Porat, “The modified Yule-Walker method of ARMA spectral estimation,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 20, pp. 158–173, Mar. 1984.

[39] Global IP Sound, *Internet Low Bit Rate Codec at 13.3 kbit/s and 15.2 kbit/s*. IETF RFC3951, Dec. 2004.

[40] M. M. A. Khan, “Coding of excitation signals in a waveform interpolation speech coder,” Master’s thesis, McGill University, Jul. 2001.

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

[42] ITU-T, *Objective quality measurement of telephoneband (300-3400 Hz) speech codecs*. ITU-T Recommendation P.861, International Telecommunication Union, Feb. 1998.

[43] ITU-T, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Vodecs*. ITU-T Recommendation P.862, International Telecommunication Union, Feb. 2001.

[44] J. D. Johnston, “Transform coding of audio signals using perceptual noise criteria,” *IEEE J. Sel. Areas Commun.*, vol. 6, pp. 314–323, Feb 1988.

[45] J. G. Beerends and J. A. Stemerdink, “A perceptual audio quality measure based on a psychoacoustic sound representaion,” *J. Audio Eng. Soc.*, vol. 40, pp. 963–978, Dec 1992.

[46] 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.

[47] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer-Verlag, 2nd ed., 1999.

[48] P. Kabal, “An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality,” Tech. Rep., McGill University, May 2002. Available: `http://www-mmsp.ece.mcgill.ca/Documents/Reports/2002/KabalR2002v2.pdf`.

[49] R. Der, P. Kabal, and W.-Y. Chan, “Towards a new perceptual coding paradigm for audio signals,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (Hongkong, China), pp. 457–460, 2003.

[50] D. Sen, *Perceptual Enhancement of Low Rate Speech Coders*. PhD thesis, University of New South Wales, 1994.

[51] L. Lin, W. H. Holmes, and E. Ambikairajah, “Speech denoising using perceptual modification of Wiener filtering,” in *Electron. Lett.*, pp. 1486–1487, 2002.

[52] Y. H. Lam and R. W. Stewart, “Perceptual suppression of quantization noise in low bitrate audio coding,” in *Conf. Rec. 31st Asil. Conf. Signals, Syst., Comput.*, (Pacific Grove, California, U.S.A.), pp. 49–53, 1997.

[53] D. E. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, “Perceptual filters for audio signal enhancement,” *J. Audio Eng. Soc.*, vol. 45, pp. 22–35, Jan./Feb. 1997.

[54] W. Chen, P. Kabal, and T. Z. Shabestary, “Perceptual postfilter estimation for low bit rate speech coders using Gaussian mixture models,” in *InterSpeech 2005*, (Lisbon, Portugal), pp. 3161–3164, Sept. 2005.

[55] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 72–83, Jan. 1995.

[56] I. T. Nabney, *NETLAB: Algorithms for Pattern Recognition*. Springer, 2002.

[57] P. Hedelin and J. Skoglund, “Vector quantization based on Gaussian mixture models,” *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 385–401, Jul. 2000.

[58] A. D. Subramfaniam and B. D. Rao, “Pdf optimized parametric vector quantization of speech line spectral frequencies,” *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 130–142, Mar. 2003.

[59] Y. Qian and P. Kabal, “Dual-mode wideband speech recovery from narrowband speech,” in *EUROSPEECH*, (Geneva, Switzerland), pp. 1433–1436, Sept. 2003.

[60] Y. Qian and P. Kabal, “Combining equalization and estimation for bandwidth extension of narrowband speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (Montreal, Canada), pp. 713–716, May 2004.

[61] L. Deng, J. Droppo, and A. Acero, “Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features,” *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 218–233, May 2004.

[62] S. Kotz, N. Balakrishnan, and N. Johnson, *Continuous Multivariate Distributions*, vol. 1. John Wiley & Sons, 2000.

[63] X. D. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Prentice Hall, 2001.

[64] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*. Prentice-Hall, 3rd ed., 1996.

[65] M. Nilsson, H. Gustafsson, S. V. Andersen, and W. B. Kleijn, “Gaussian mixture model based mutual information estimation between frequency bands in speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 525–528, 2002.

[66] ISO/IEC 11172-3, *Coding of Moving Pictures and Associated Audio for Digital Storage Media at Up to About 1.5 Mbit/s—Part 3: Audio*. Draft, Sept. 1991.

[67] ISO/IEC 13818-7, *Generic Coding of Moving Picture and Associated Audio Information—Part 7: Advanced Audio Coding (ACC)*. Draft, Apr. 1997.

[68] ISO/IEC 14496-3, *Coding of Audiovisual Objects—Part 3: Audio*. Draft, May 1998.

[69] W. B. Kleijn, "Improved pitch prediction," in *IEEE Workshop Speech Coding Telecom.*, pp. 19–20, 1993.

[70] M. Lähdekorpi, J. Nurminen, A. Heikkinen, and J. Saarinen, "Perceptual irrelevancy removal in narrowband speech coding," in *EUROSPEECH*, (Geneva, Switzerland), pp. 1081–1084, 2003.

[71] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 3rd ed., 1996.

[72] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time Processing of Speech Signals*. IEEE Press, 2000.

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

[74] D. O'Shaughnessy, *Speech Communications: Human and Machine*. Wiley-IEEE Press, 2nd ed., 1999.

[75] H. Najafzadeh-Azghandi, *Perceptual Coding of Narrowband Audio Signals*. PhD thesis, McGill University, Apr. 2000.

[76] M. Tammi, *Techniques for Low Bit Rate Speech Coding – Speech Modeling in WI and RCELP Coders*. PhD thesis, Tampere University of Technology, 2002.

[77] K. K. Paliwal and S. So, "Multiple frame block quantisation of line spectral frequencies using Gaussian Mixture Models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (Montreal, Canada), May 2004.

[78] T. H. Falk, W.-Y. Chan, and P. Kabal, "Speech quality estimation using Gaussian mixture models," in *InterSpeech 2004*, (Jeju Island, Korea), Oct. 2004.

[79] Y. Qian and P. Kabal, "Wideband speech recovery from narrowband speech using classified codebook mapping," in *Proc. 9th Austr. Int. Conf. on Speech Sci. Technol.*, (Melbourne, Australia), pp. 106–110, Dec. 2002.

[80] P. Kabal, F.-M. Wang, D. O'Shaughnessy, and R. P. Ramchandran, "Adaptive postfiltering for enhancement of noisy speech in the frequency domain," in *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 312–315, 1991.

[81] J. I. Lee and C. K. Un, "Improving speech quality of CELP coder," *Electron. Lett.*, vol. 25, pp. 1275–1277, Sept. 1989.

[82] R. A. McDonald and P. M. Schultheiss, “Information rates of Gaussian signals under criteria constraining the error spectrum,” *Proc. IEEE*, vol. 55, pp. 415–416, 1964.

[83] A. A. Azirani, R. L. B. Jeannes, and G. Faucon, “Optimizing speech enhancement by exploiting masking properties of the human ear,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 800–803, 1995.

[84] Q. Dai, Y. Chen, and Z. Bian, “Optimizing speech enhancement based on noise masked probability,” in *Proc. Int. Conf. Signal Process.*, pp. 448–451, 2002.

[85] Y. Hu, M. Bhatnagar, and P. C. Loizou, “A cross-correlation technique for enhancing speech corrupted with correlated noise,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (Salt Lake City, Utah, U.S.A.), pp. 673–676, 2001.

[86] M. Klein and P. Kabal, “Signal subspace speech enhancement with perceptual post-filtering,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 537–540, 2002.

[87] T.-W. Lee and K. Yao, “Speech enhancement by perceptual filter with sequential noise parameter estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (Montreal, Canada), pp. 693–696, 2004.

[88] W. B. Kleijn, R. P. Ramachandran, and P. Kroon, “Generalized Analysis-by-Synthesis coding and its application to pitch prediction,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (San Francisco, California, U.S.A.), pp. 337–340, 1992.

[89] W. B. Kleijn, “Signal processing representations of speech,” *IEICE Trans. Inf. Syst.*, vol. E86-D, pp. 359–376, Mar. 2003.