Wideband CELP Speech Coding

by

Karim Abboud

B. Eng.

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

Department of Electrical Engineering
McGill University
Montreal, Canada
November, 1992

Acknowledgements

I would like to thank my supervisor, Dr. Peter Kabal, for his help and guidance throughout my years as an undergraduate and graduate student at McGill. The financial support provided by my supervisor and by the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (Fonds FCAR) were very much appreciated. All the research was conducted at l' Institut National de Recherche Scientifique (INRS) - Télécommunications and the facilities provided were excellent and came as a great help to my work.

I would like to give special thanks to my parents and brother for their love and understanding. I am also very grateful to Nabih Maroun and for his consistent support and companionship.

Abstract

The purpose of this thesis is to study the coding of wideband speech and to improve on previous Code-Excited Linear Prediction (CELP) coders in terms of speech quality and bit rate. To accomplish this task, improved coding techniques are introduced and the operating bit rate is reduced while maintaining and even enhancing the speech quality.

The first approach considers the quantization of Linear Predictive Coding (LPC) parameters and uses a three way split vector quantization. Both scalar and vector quantization are initially studied; results show that, with adequate codebook training, the second method generates better results while using a fewer number of bits. Nevertheless, the use of vector quantizers remain highly complex in terms of memory and number of computations. A new quantization scheme, split vector quantization (split VQ), is investigated to overcome this complexity problem. Using a new weighted distance measure as a selection criterion for split VQ, the average spectral distortion is significantly reduced to match the results obtained with scalar quantizers.

The second approach introduces a new pitch predictor with an increased temporal resolution for periodicity. This new technique has the advantage of maintaining the same quality obtained with conventional multiple coefficient predictors at a reduced bit rate. Furthermore, the conventional CELP noise weighting filter is modified to allow more freedom and better accuracy in the modeling of both tilt and formant structures. Throughout this process, different noise weighting schemes are evaluated and the results show that the new filter greatly contributes in solving the problem of high frequency distortion.

The final wideband CELP coder is operational at 11.7 kbits/s and generates a high perceptual quality of the reconstructed speech using the fractional pitch predictor and the new perceptual noise weighting filter.

Sommaire

L'objectif de ce mémoire est d'étudier le codage de la parole en bande élargie, ainsi que d'améliorer les résultats obtenus par les codeurs précédents de type CELP. Cette recherche porte essentiellement sur la réduction du débit binaire tout en preservant un niveau de qualité relativement supérieur de la parole. Plusieurs techniques de codage ont donc été étudiées et développées pour atteindre ce but.

La première approche examine la quantification des coefficients de prédiction linéaire (CPL) et utilise une quantification de vecteur éclatés combiné à une mesure de distance pondérée. Cette quantification vectorielle consiste à diviser un vecteur de paramètres représentant les coefficients CPL en trois sous-vecteurs et à quantifier ces différents sous-vecteurs. La supériorité de cette methode est ensuite evaluée à l'aide d'une comparaison avec la méthode de quantification scalaire. Les résultats concluent qu'avec un taux de transmission inférieurs, la quantification vectorielle éclatée achève une distortion spectrale similaire à celle d'une quantification scalaire.

La deuxième approche améliore la performance du filtre prédicteur de la fréquence fondamentale avec l'utilisation d'une périodicité de haute résolution. L'étude poursuivie dans ce domaine analyse les filtres prédicteurs d'un ou de plusieurs coefficients de périodicité et les compare à ce nouveau filtre. Cette technique sera directement responsable d'un rehaussement perceptuel accentué de la qualité de la parole. Néanmoins la qualité perceptuelle obtenue n'est pas encore optimisée, et ce n'est qu'après l'introduction d'un filtre de pondération perceptuelle que les problèmes du codage des signaux en large bande sont résolus, notamment les difficultés dues à la largeur de leur bande spectrale dynamique.

Le codeur CELP final est opérationel à 11.7 kbits/s et produit une qualité de parole supérieure à l'aide du filtre prédicteur de haute résolution et du filtre de pondération perceptuelle.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Digital speech coding has become an essential part of many speech processing applications because it maintains efficient and secure transmission of data. The speech signal is coded into a bit stream, transmitted over a channel, and finally converted back to a signal that is the closest to the input signal. Economical digital signal representation and minimal quality loss are the two factors that will distinguish good from bad coders.

For the past few years, we have witnessed continuous breakthroughs in the development of speech coding techniques, but most of the research accomplished was related to narrowband speech signals where the bandwidth is limited to 200–3400 Hz. Consequently, the speech coding community is placing greater emphasis on the need for high quality speech and therefore a larger speech bandwidth especially for applications such as teleconferencing, videophones, digital cordless telephony, digital mobile radio, high quality voice-mail services and wideband telephone intended for the ISDN (Integrated Service Digital Network) network.

With a bandwidth of 50–7000 Hz corresponding to wideband speech, the bandwidth limitation at 3.4 kHz is eliminated and a substantial increase in perceived

quality is observed. The added low frequencies increase the voice naturalness while the added high frequencies make the speech sound sharper and more intelligible especially in fricative sounds. Obviously, a larger number of bits is required to code the additional information which leads us to the trade-off between preserving acceptable speech quality of the reconstructed signal and maintaining a relatively low operating bit rate.

Speech quality can be grouped into four different classes: (1) *commentary* or *broadcast* quality that corresponds to wideband speech with no perceptible noise; (2) *toll* quality that refers to narrowband speech that can be heard over the telephone network; (3) *communication* quality describes speech that is highly intelligible but more distorted when compared to toll quality speech; and finally (4) *synthetic* quality that remains intelligible but loses naturalness. These deficiencies can be measured using either subjective or objective measures. A commonly applied subjective indicator is the Mean Opinion Score (MOS) where a scale of 1 to 5 is used to refer to the level of speech quality, while objective indicators include the Signal-to-Noise Ratio (SNR). Segmental SNR (segSNR), Spectral Distortion Measures and Perceptual Noise Measures.

Two different classes of coders exist: *waveform coders*, *source coders*. Waveform coders encode the speech directly and reconstruct it as accurately as possible at the receiver on a sample-by-sample basis. Source coders, on the other hand, model the speech production mechanism and identify the key elements of the speech. Waveform coders remain the best choice to encode speech while preserving naturalness and maintaining a low level of distortion. Specifically, the *Code-Excited Linear Prediction* (CELP) scheme is now the most commonly used analysis-by-synthesis scheme.

This algorithm which falls under the Linear Predictive Coding (LPC) category was first introduced in 1984 by Atal and Shroeder [27], and has proved to be one of the most efficient coding schemes. Today, it provides excellent narrowband speech production combined with a relatively low bit rate, high quality reconstructed nar-

rowband speech is now available at 8 kbits/s [3] and at 4.8 kbits/s [9], but the quality of the synthetic speech degrades rapidly below 5 kbits/s. New methods using *prototype waveforms* [1] are being investigated to overcome this degradation and produce high quality speech at 4 kbits/s.

1.2 Wideband CELP Speech Coding

In wideband speech, an efficient 64 kbits/s algorithm has been already developed primarily for ISDN teleconferencing and loudspeaker telephony, this algorithm is known as the G.722 CCITT (Consultative Committee for Telephone and Telegraph) standard and is based on split-band coding using *adaptive differential pulse code modulation* or ADPCM in each subband as shown in Fig. 1.1. In this scheme, unequal bit allocation is used to provide more control over subband coding of speech: six bits per sample are allocated to the lower subband while two are allocated to the higher subband where frequencies are less perceptible. Nevertheless, the bit rate obtained by the G.722 coding algorithm remains relatively high making some of other techniques more efficient.

In recent years, different wideband coding schemes were introduced as alternatives for the G.722 standard with bit rates almost similar to those found in high quality narrowband telephony systems: a 32 kbits/s low-delay CELP was introduced by Ordentlich and Shoham [20] as well as another 32 kbits/s wideband coder designed by Quackenbush [23] and intented for the ISDN network. Several coders were also implemented at 16 kbits/s, notably the split-band CELP structure introduced by Roy and Kabal [26], the algebraic CELP scheme proposed by the Communication Research Center at the University of Sherbrooke [14] and the multipulse coding (MPLPC) method studied by Montagna, Jacovo, Perosino and Sereno [19]. Fig. 1.2 shows the aim of most of the research being accomplished in this field, the figure shows the expected speech quality to be attained over the G.722 algorithm.

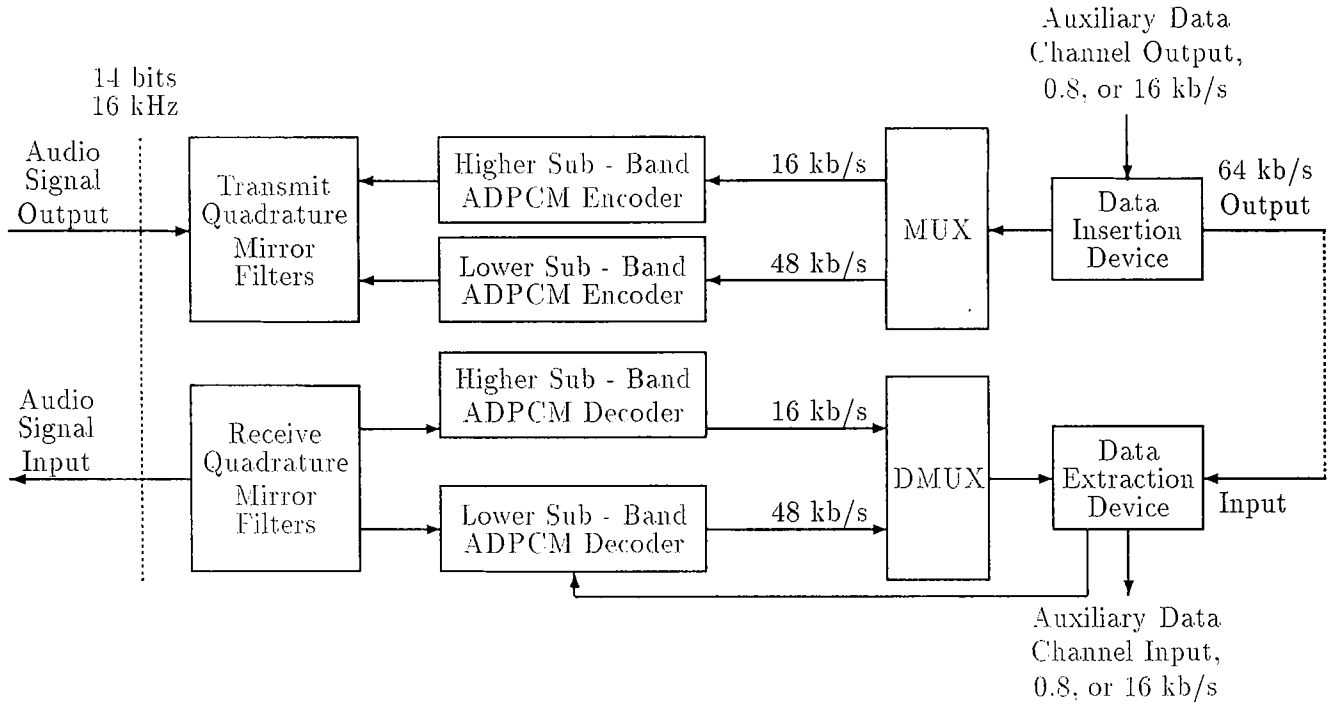

Fig. 1.1: Two-band subband coder for 64-kb/s coding of 7-kHz audio.

Three approaches were analyzed for the implementation of the wideband CELP coder: the *full-band* scheme, the *split-band* scheme and a hybrid scheme taking advantage of both structures. The full-band approach analyzes and codes the speech signal with all its frequency contents and other parameters considered together, while the split-band approach divides the speech signal into a low (0 to 4 kHz) and a high (0 to 8 kHz) band signal and each band is dealt with separately. Since most of the perceptual importance (approximately 80%) of the speech lies in the lower band (0.2 to 3.2 kHz) compared to the higher band, unequal bit allocation allows better and more flexible control over the coding resolution given to the low and high frequency components of the speech.

The principal goal of this thesis was to improve on previous work done in the field of wideband CELP and more specifically the work of Roy [25] and mainly to lower the bit rate to approximately 12 kbits/s while keeping a speech quality comparable to the 16 kbits/s coders designed by Roy and Kabal [26]. Different coding schemes

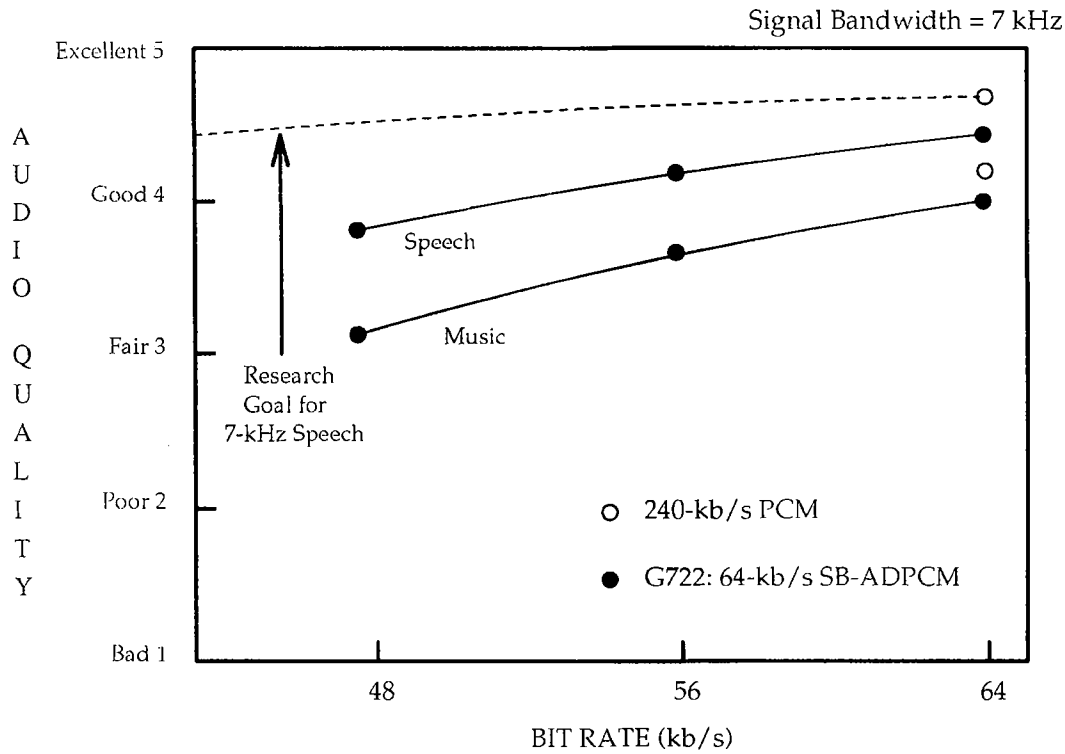

Fig. 1.2: 7-kHz digital audio quality as a function of the G.722 algorithm bit rate.

were taken into consideration to accomplish this task:

- **Vector quantization of Line Spectral Frequencies (LSF):**

  In wideband spectral envelope coding, the LSF's are divided into either two or three subgroups and each subgroup is quantized separately with the lower subgroup receiving the highest number of quantization bits. In this method, unequal bit allocation is accomplished according to the frequency band position.

- **Pitch prediction with high temporal resolution:**

  A first-order pitch filter with fractional delays is incorporated the CELP coder structure. Three different pitch resolutions are used to implement the pitch filter. This approach solves the problem of the destruction of harmonic structures in the high frequency regions usually obtained with integer pitch periods.

- **Perceptual noise weighting:**

  While coding of the low-frequency region seems to be easy, coding of the high frequency components remains a very difficult task. This asymmetry creates audible high frequency distortion. A perceptual noise weighting filter is used to allow better control over noise weighting in both the lower and higher frequency regions.

- **Comparative study of full-band and split-band coders:**

  This study consists of a performance evaluation of both a split-band and a full-band CELP. Perceptual and ordinary objective indicators are used to determine the overall speech quality.

1.3 Organization of the Thesis

This thesis is divided into seven chapters. The second chapter deals with background material and parameter definitions are given on wideband CELP coders. The third chapter is a study of wideband envelope coding with special emphasis put on vector quantization (VQ) of LSF's, other techniques are also discussed and compared to the VQ techniques. The fourth chapter introduces the notion of fractional pitch predictors for wideband speech, and the importance of higher pitch resolution at low bit rates is investigated. The fifth chapter studies the concept of perceptual noise weighting and the improvements that can be obtained on modeling the formant structure and the spectral tilt concurrently; this chapter also expands on certain techniques used in the algebraic CELP and the possible use of these techniques to improve the performance of the wideband CELP. The sixth chapter is a synthesis of the performance of all previous techniques combined in one coder. A comparison is also established between the split- and full-band approaches. Finally, the last chapter concludes with a summary of the results and new recommendations for future research.

Chapter 2

Wideband CELP Speech Coding

This chapter is divided into three sections. The first section covers background material including linear predictive coding, formant filtering and pitch filtering. The second section introduces the concept of Code Excited Linear Predictive Coding, while the third section describes the full-band implementation of our wideband CELP coder.

2.1 Linear Predictive Coding

Linear predictive coding (LPC) is a very popular speech modeling technique and is used in many speech coders including the generalized ADPCM. Its success is due to accurate representation of the speech spectral magnitude and to its low level of complexity. It essentially uses linear combinations of past speech values to predict future values. this operation is performed with the help of a prediction filter $F(z)$ and a quantizer Q. Only the difference between the predicted value and the original input value is transmitted as shown in Fig. 2.1. The power of the resulting signal $\tilde{s}(n)$ can then be compared to the power of the input signal $s(n)$ to determine the quality of the coder.

Therefore, the LPC scheme tries to extract the best set of parameters that would describe the speech. In turn, this translates into more efficient transmission

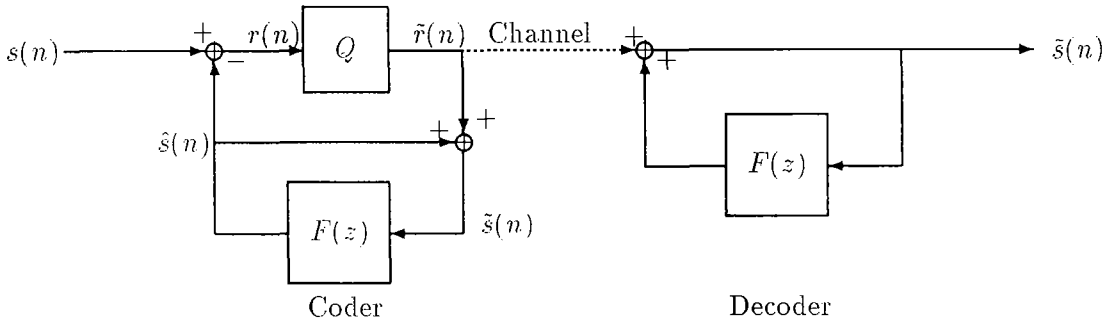

Fig. 2.1: Prediction coder block diagram.

systems where the model parameters, rather than the signal itself, are coded and sent. In human speech characterized by the vocal tract shape and vocal cord vibrations, we can use two forms of linear prediction filters to extract both the shape and the vibration parameters from the speech:

- *short term* or *formant* filtering
- *long term* or *pitch* filtering

The two schemes are used to construct different coders including RELP (Residual Excited Linear Prediction) coders and CELP coders. In the RELP configuration, introduced by Un and Magill [31], a prediction filter is used to extract the formant information from the speech, while in CELP coding both formant and pitch prediction filtering are used.

2.1.1 Formant filtering

The human vocal tract can be modeled as an acoustic tube with resonances known as formants. By changing the shape of the vocal tract, we alter the frequency response and therefore the formant frequencies. In formant prediction, the formant structure of an input frame of speech samples is determined. The operations are carried out by

a linear prediction filter $F(z)$ where

$$\begin{aligned} F(z) &= a_1 z^{-1} + ... + a_{N_p} z^{-N_p} \\ &= \sum_{k=1}^{N_p} a_k z^{-k} \end{aligned} \tag{2.1}$$

The LPC coefficients a_k are determined with the *inverse formant filter* or *error formant prediction filter* $A(z)$ defined as

$$A(z) = 1 - F(z) = 1 - \sum_{k=1}^{N_p} a_k z^{-k} \tag{2.2}$$

During the LPC analysis operation, an input speech waveform $s(n)$ is passed through the filter $A(z)$ to generate an error signal $d(n)$ where

$$d(n) = s(n) - \sum_{k=1}^{N_p} a_k s(n-k) \tag{2.3}$$

The LPC coefficients are then determined by minimizing, in the mean square sense (MS), the error signal $d(n)$ also known as the *formant residual*. this leads to the following for an N samples frame of speech:

$$\epsilon(a_1, ..., a_k) = \sum_{n=0}^{N-1} [d(n)]^2 = \sum_{n=0}^{N-1} [s(n) - \sum_{k=1}^{N_p} a_k s(n-k)]^2 \tag{2.4}$$

The optimal solution will strip the input speech signal from most of the short term redundancies. and will be determined by setting the gradient of the error ϵ to zero and solving a set of k equations:

$$\nabla \epsilon(a_1, ..., a_k) = 0 \tag{2.5}$$

The resulting equation is

$$\sum_{n=0}^{N-1} s(n-i)s(n) = \sum_{k=1}^{N_p} a_k \sum_{n=0}^{N-1} s(n-i)s(n-k), \qquad i = 1, 2, ..., N_p. \tag{2.6}$$

Only one approach is considered in this research to determine the solutions of Eq. (2.5). This approach known as the *autocorrelation* method generates a set of stable

LPC coefficients where we have the autocorrelation function $R(i)$ of the signal $s(n)$ defined as

$$R(i) = \sum_{n=i}^{N-1} x(n)x(n-i), \qquad i = 1, 2, ..., N_p. \tag{2.7}$$

Consequently, Eq. (2.6) becomes

$$\sum_{k=1}^{N_p} a_k R(i-k) = R(i), \qquad i = 1, 2, ..., N_p. \tag{2.8}$$

With the use of the autocorrelation $R(i)$ and its properties mainly the fact that $R(i)$ is an even function where $R(i) = R(-i)$, the optimal solution of Eq. (2.5) or the minimum residual energy becomes

$$E_{min} = R(0) - \sum_{k=1}^{N_p} a_k R(k) \tag{2.9}$$

where $R(0)$ is equal to the energy of the signal $s(n)$.

In formant synthesis, a spectral shaping filter $H(z)$ is used with an input that has a flat spectral envelope and a uniform amplitude distribution. The choice of a flat spectrum for the input is important because it confines all relevant spectral details to the filter $H(z)$.

In order to simplify the computations in obtaining $H(z)$, we consider an input speech that is stationary during a window or frame of N samples (typically a frame of 20 ms or 320 samples is used). The formant synthesis filter can now be modeled with constant coefficients updated with each new frame of data. In general prediction, $H(z)$ is assumed to have p poles and q zeros known as an *autoregressive moving average* (ARMA) model, therefore generating a reconstructed speech by a linear combination of p previous output samples and $q+1$ previous input samples. However, in most applications of speech LPC analysis, an all-pole model (also know as an *autoregressive* or AR model) is used, this substantially reduces the amount of computations required to derive the LPC parameters. This simplification can be a drawback since the actual speech spectrum has zeros from the vocal tract response and the glottal

source. Nevertheless, human ear sensitivity is high at spectral peaks (poles) and low at spectral valleys (zeros) making the all-pole model an appropriate choice.

Fig. 2.2 shows the effects of filtering a Gaussian waveform with a flat spectrum through a formant filter, the generated output signal has now multiple formants.

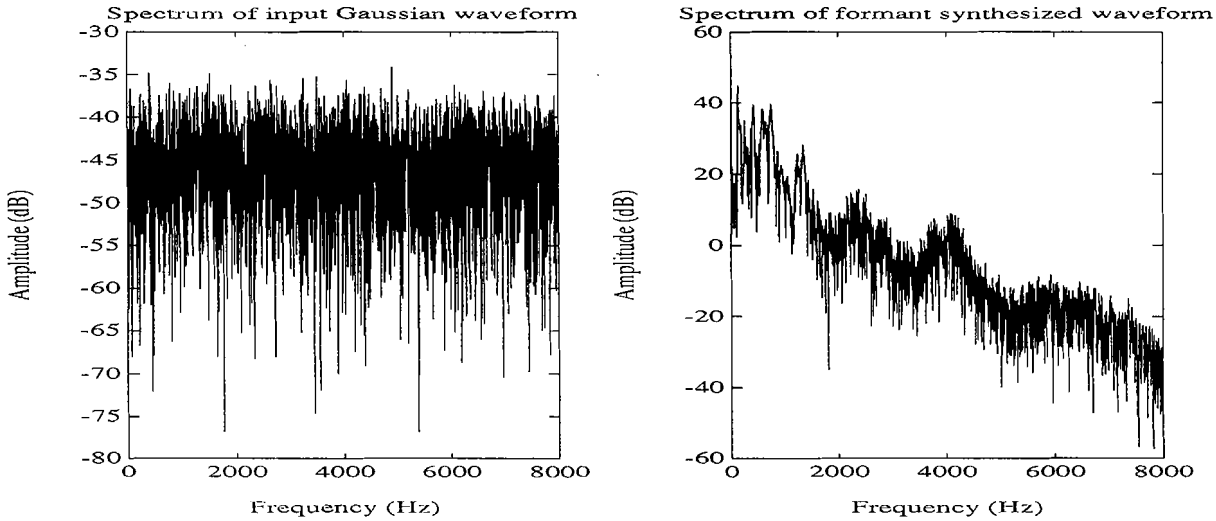

Fig. 2.2: Effect of formant filtering on a Gaussian waveform.

The selection of the order p of $H(z)$ is a tradeoff between accuracy and complexity. In general, the number of poles directly corresponds to the number of formants (2 poles per formant), and 2–4 poles are used in addition to approximate possible zeros in the spectrum and general spectral shaping.

The general expression for the formant synthesis filter $H(z)$ is

$$H(z) = \frac{1}{1 - F(z)} = \frac{1}{1 - \sum_{k=1}^{N_p} a_k z^{-k}} \tag{2.10}$$

where the a_k are the LPC coefficients and N_p is the number of poles (usually N_p ranges between 16 and 32 for wideband speech).

Given the fact that most low bit rate linear prediction systems transmit the prediction coefficients as side information, efficient quantization of these coefficients is crucial to the quality of the speech coder. These coefficients are actually not well

suited for transmission because a bit error in any one can cause the synthesis filter to become unstable [11]. To overcome this difficulty, multiple transformation methods have been introduced such as reflection coefficients, log-area ratios, autocorrelation coefficients of the input samples, direct form predictor coefficients and line spectral frequencies (LSF). Recently, the usage of LSF's as an efficient transformation method has become very popular and they are investigated in Chapter 3.

Once the transformation of the coefficients is performed, quantization can take place. Again, the choice is between two different quantization techniques. The first, known as *scalar quantization*, quantizes each LPC transformed coefficient individually, while the second, known as *vector quantization*, quantizes all the LPC transformed parameters as a group. Comparison between these two methods is also discussed in Chapter 3.

2.1.2 Pitch filtering

In pitch prediction filtering, the pitch period of the glottal excitation is estimated. During unvoiced speech segments, no clear pitch period can be detected; therefore, the pitch prediction filter has to be disabled; while in voiced speech segments, the pitch filter is enabled to generate the optimal pitch period. The expression of the pitch prediction filter $P(z)$ is

$$P(z) = \sum_{i=-L}^{L} \beta_i z^{-M+i} \tag{2.11}$$

where β_i are the pitch coefficients, M is the *pitch lag* or *tap delay* (it is usually of the order of 40 to 320 samples for a 16 kHz sampled signal) and finally L is related to the number of pitch coefficients (typical values are 0 or 1 for one or three tap pitch predictors respectively). Single-tap filters are still very common, but three-tap filters provide better performance at the expense of an increased bit rate. Fractional pitch delays are a good option to solve the dilemma of bit rate and speech quality, a full discussion is provided in Chapter 4.

The previous described LPC analysis removed most of the near sample redundancies from the speech signal. Far sample redundancies are dealt with during the pitch prediction operation. The formant residual signal $d(n)$ obtained during the LPC analysis is passed through the *error pitch prediction filter* or *pitch inverse filter* $B(z)$:

$$B(z) = 1 - P(z) = 1 - \sum_{i=-L}^{L} \beta_i z^{-M+i} \tag{2.12}$$

The resulting error signal $r(n)$ also known as the pitch residual is defined as

$$r(n) = d(n) - \sum_{i=-L}^{L} \beta_i d(n - M + i) \tag{2.13}$$

To compute the pitch filter parameters β_i and M, we have to minimize the error for an N samples frame of speech:

$$\epsilon(\beta_{-L}, ..., \beta_{-L}, M) = \sum_{n=0}^{N-1} [d(n) - \sum_{i=-L}^{L} \beta_i d(n - M + i)]^2 \tag{2.14}$$

The error is first minimized over the pitch coefficients β_i, the resulting optimum for the β_i's is in terms of the tap delay M where

$$\frac{\delta \epsilon}{\delta \beta_{-L}} = 0 \Rightarrow \beta_{-L} = f(M) \tag{2.15}$$

$$\vdots$$

$$\frac{\delta \epsilon}{\delta \beta_{+L}} = 0 \Rightarrow \beta_{+L} = f(M) \tag{2.16}$$

These solutions are then substituted back into Eq. (2.14) and the error is minimized over all the range of the tap delay M.

Once all the pitch parameters are determined, they can be used to construct the pitch synthesis filter $G(z)$ defined as

$$G(z) = \frac{1}{1 - P(z)} = \frac{1}{1 - \sum_{i=-L}^{i=L} \beta_i z^{-M+i}} \tag{2.17}$$

The filter is then used in speech reconstruction to add far sample redundancies to Gaussian waveforms as shown in Fig. 2.3. This filter models the periodic vibrations of the vocal cords.

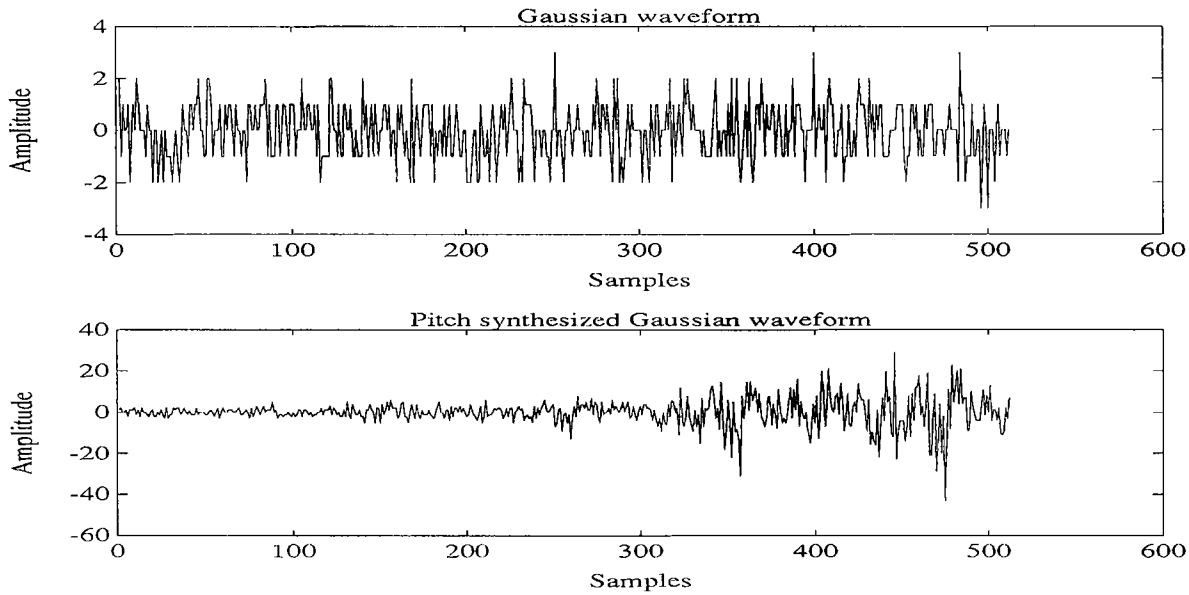

Fig. 2.3: Effect of pitch filtering on a Gaussian waveform.

2.2 CELP coder

In CELP coding, the reconstruction of the input speech signal involves the use of a pitch synthesis filter, a formant filter and a residual codebook. An excitation waveform $\hat{r}_i(n)$ is first selected from the codebook as shown in Fig. 2.4 and then goes through a cascade of the two filters to give an initial reconstructed speech signal, the operation is repeated until the best match to the original signal is determined. This operation falls under the *analysis-by-synthesis* category of linear predictive systems and is therefore divided into two stages: the analysis and synthesis stages.

During the analysis stage, the input speech is divided into equal length blocks of samples or frames (e.g. 20ms). The input speech frame $s(n)$ is then passed through the inverse formant filter $A(z)$ and the LPC coefficients a_k are determined by performing a standard autocorrelation LPC analysis on the input speech. Then, the pitch parameters β_i and M are derived with the inverse pitch filter $B(z)$. Formant and pitch parameters are used to construct both formant and pitch synthesis filters in the next stage.

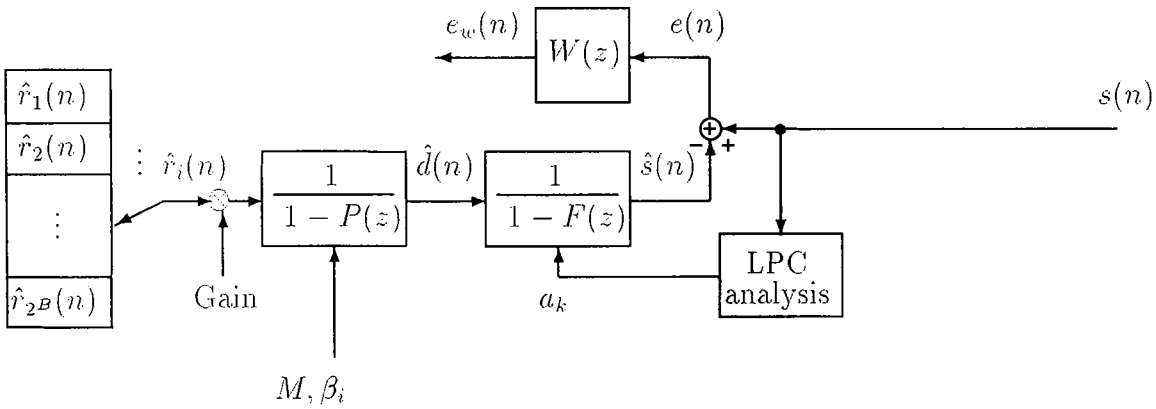

Fig. 2.4: Basic CELP coder.

During the synthesis stage, the reconstructed speech frame is generated using pre-determined synthesis parameters on speech subframes (i.e. excitation waveforms, gain values, lag values, pitch and LPC coefficients). The codebook is populated with normalized Gaussian sequences. A Gaussian waveform is first selected from the codebook and is scaled by a gain factor G. Periodic components are then added during voiced speech to the excitation waveform after its passage through the pitch synthesis filter $G(z)$. Next, the formant resonances are added to the resulting signal $\hat{d}(n)$ after the formant synthesis filter $H(z)$ to obtain the initial synthesized frame of speech. This speech is then subtracted from the original speech and the result is weighted with $W(z)$ so as to cover the coding noise by the formant regions. The noise weighting filter $W(z)$ is defined as

$$W(z) = \frac{H(\gamma z)}{H(z)} = \frac{1 - F(z)}{1 - F(\gamma z)} \tag{2.18}$$

where γ, the bandwidth expansion factor, has a value between zero and unity (usually 0.75) and controls the level of coding noise in the formant regions. As shown in Eq. (2.18), the weighting filter is directly related to the LPC coefficients and should be updated at every frame. The perceptual improvements obtained with the spectral noise weighting filter are significant where the noise is now covered by the formant peaks.

Finally, this weighted difference $\epsilon_w(n)$ is minimized in a mean-square (MS) sense

for each gain G, pitch lag M, pitch coefficients β and excitation waveform $\hat{r}_i(n)$. The index of the waveform yielding the lowest error energy is sent to the decoder along with the other synthesis parameters.

A sub-optimal procedure is used for the codebook search where only two parameters are considered: index and gain. The optimal scheme would be to perform a joint optimization of both pitch and codebook parameters, but decoupling these parameters eliminates the computational burden induced by nesting exhaustive lag and codebook searches and makes the sub-optimal approach the best candidate for efficient optimization [18].

2.3 Full-band coder system configuration

This section gives a detailed description of the actual CELP coder that has been used during the course of this work. Both pitch and codebook search optimization techniques are derived and discussed. The algorithm described below was the starting point of the research into wideband CELP coding.

The following system configuration was used to accomplish all different simulations throughout the course of this research. The full-band CELP coder configuration shown is similar to the one in Fig. 2.4 except that the weighting filter $W(z)$ is moved ahead of the summation. In the upper branch of the block diagram, the input speech signal is now filtered by $W(z)$, while in the lower branch, the formant synthesis filter $H(z)$ is combined with $W(z)$ to form the bandwidth expanded version $H_\gamma(z)$ of $H(z)$.

The impulse response of $H_\gamma(z)$ is identical to the impulse response of the filter $1/a(\gamma z)$ and is given by

$$h_\gamma(n) \;=\; \gamma^n h(n), \qquad n = 0, 1, 2, ... \tag{2.19}$$

where $h(n)$ is the impulse response of the synthesis filter $1/a(z)$. For $\gamma = 1$, $h_\gamma(n)$ is identical to $h(n)$. For values of γ less than 1, the impulse response is exponentially weighted and decays rapidly. The desired transfer function can be obtained by

inserting a multiplier with a multiplication factor γ before each delay element.

This new system configuration, shown in Fig. 2.5, simplifies some of the computations involved in generating a high quality reconstructed speech. The new optimization procedure now involves the unweighted mean square error between the weighted input and the synthesized/weighted signal.

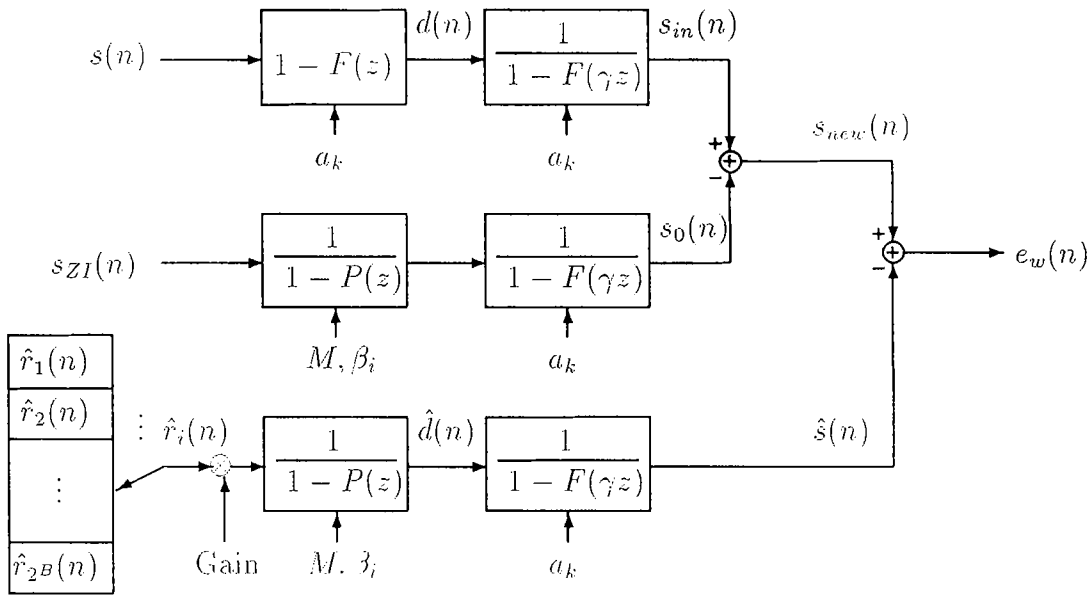

Fig. 2.5: CELP system configuration for the full-band coder.

An additional section has been added to the analysis stage of the basic CELP coder. This section subtracts out the effect of the past excitation from the filtered reference waveform. An input $s_{ZI}(n)$ with all its samples set to zero is passed through the pitch filter and the bandwidth expanded formant filter, both with zero memories, to generate the zero input response $s_0(n)$. This response is then subtracted from the filtered reference waveform $s_{in}(n)$:

$$s_{new}(n) = s_{in}(n) - s_0(n) \tag{2.20}$$

where $s_{new}(n)$ is the new signal used as a reference for the synthesis stage. All the

other filter memories are updated for every new frame of speech.

The gain G, codebook index i and pitch parameters β_i and M are updated for every subframe while the LPC coefficients are updated for every frame.

2.3.1 Pitch search

This subsection studies the pitch optimization procedure related to the CELP coder. This procedure that is based on a closed loop analysis searches over all lags and all pitch coefficient values to find the combination resulting in the best synthesized output. As described earlier, the pitch search involved the minimization of the energy ϵ_p defined as:

$$\epsilon_p = \sum_{n=0}^{N-1} (s_{new}(n) - \hat{s}_0(n))^2 \tag{2.21}$$

where N is the length of the subframe, $s_{new}(n)$ is the new reference signal as shown in Eq. (2.20), and $\hat{s}_0(n)$ is the zero input response ($G = 0$) of the cascade of the pitch and bandwidth expanded filters.

$$\hat{s}_0(n) = \sum_{k=0}^{\infty} d(k) h_\gamma(n-k) \tag{2.22}$$

In this description of the pitch search, a one-tap pitch filter is used. Two different cases corresponding to the size of the pitch lag are considered in solving for the pitch coefficient over all the lag range:

- **Short lags:**

  This is the case where the tap delay M is shorter than the subframe of length N,

  $$\hat{d}(n) = \begin{cases} \beta \hat{d}(n-M) & 0 \le n < M \\ \beta^2 \hat{d}(n-2M) & M \le n < N \end{cases} \tag{2.23}$$

  The new expression for the energy ϵ_p is

  $$\epsilon_p = \sum_{n=0}^{N-1} [s_{new}(n)]^2 - 2\beta \sum_{n=0}^{N-1} s_{new}(n) \dot{d}_{[0,M)}(n, M) + \beta^2 \sum_{n=0}^{N-1} [\dot{d}_{[0,M)}(n, M)]^2$$

$$-2\beta^2 \sum_{n=M}^{N-1} s_{new}(n)\dot{d}_{[M,N)}(n,2M) + 2\beta^3 \sum_{n=M}^{N-1} \dot{d}_{[0,M)}(n,M)\dot{d}_{[M,N)}(n,2M)$$
$$+\beta^4 \sum_{n=M}^{N-1} [\dot{d}_{[M,N)}(n,2M)] \quad (2.24)$$

where

$$\dot{d}_{[M,N)}(n,2M) = \sum_{k=M}^{N-1} \hat{d}(k-2M)h_\gamma(n-k) = \frac{1}{\beta^2}\sum_{k=M}^{N-1} \hat{d}(k)h_\gamma(n-k)$$
$$\dot{d}_{[0,M)}(n,M) = \sum_{k=0}^{M-1} \hat{d}(k-M)h_\gamma(n-k) = \frac{1}{\beta}\sum_{k=0}^{M-1} \hat{d}(k)h_\gamma(n-k) \quad (2.25)$$

- **Long lags:**

  In this case, the tap delay M is longer that the subframe of length N,

$$\hat{d}(n) = \beta\hat{d}(n-M) \quad 0 \le n < M \quad (2.26)$$

  The expression for the energy ϵ_p becomes:

$$\epsilon_p = \sum_{n=0}^{N-1} [s_{new}(n)]^2 - 2\beta \sum_{n=0}^{N-1} s_{new}(n)\dot{d}_{[0,N)}(n,M) + \beta^2 \sum_{n=0}^{N-1} [\dot{d}_{[0,N)}(n,M)]^2 \quad (2.27)$$

  where

$$\dot{d}_{[0,N)}(n,M) = \sum_{k=0}^{N-1} \hat{d}(k-M)h_\gamma(n-k) = \frac{1}{\beta}\sum_{k=0}^{N-1} \hat{d}(k)h_\gamma(n-k) \quad (2.28)$$

The technique used in minimizing the error will depend on whether the pitch coefficient is quantized or not. When the quantization is used, each quantized value of β is substituted into Eq. 2.24 or 2.27 accordingly, and the minimum error is determined for all the lag range. When the quantization is turned off, the derivative $\frac{\delta\epsilon_p}{\delta\beta}$ is set to zero and the equation is solved.

2.3.2 Codebook search

In this subsection, the codebook search procedure used for the CELP's configuration is introduced. This procedure is also based on a closed loop analysis where both the index i of the optimal codeword and the corresponding gain G are determined.

By analyzing the synthesis stage of Fig. 2.5, we have:

$$\hat{d}(n) = G\hat{r}_i(n) + \beta\hat{d}(n-M) \tag{2.29}$$

and the output of the bandwidth expanded formant filter is:

$$\hat{s}(n) = \sum_{k=0}^{\infty}[\beta\hat{d}(k-M) + G\hat{r}_i(k)]h_\gamma(n-k) \tag{2.30}$$

also,

$$\hat{s}_0(n) = \sum_{k=-\infty}^{-1}[\beta\hat{d}(k-M) + G\hat{r}_i(k)]h_\gamma(n-k) \tag{2.31}$$

where $\hat{s}_0(n)$ represents the effect of past codewords.

The new weighted reference signal is:

$$s_{new}(n) = [s(n) - s_0(n)]_W \tag{2.32}$$

where $s(n)$ is the original input speech signal and W indicates that the signal has been passed through the noise weighting filter $W(z)$.

The overall energy of the codebook search procedure is:

$$\epsilon_c = \sum_{n=0}^{N-1}[s_{new}(n) - \hat{s}(n)]^2 \tag{2.33}$$

and by setting the derivative $\frac{\delta\epsilon_c}{\delta G}$ to zero. we have:

$$G = \frac{\sum_{n=0}^{N-1} s_{new}(n)\tilde{r}^i_{[0,N)}(n)}{[\sum_{n=0}^{N-1}\tilde{r}^i_{[0,N)}(n)]^2} \tag{2.34}$$

where $\tilde{r}^i_{[0,N)}(n)$ is the synthetic speech generated by passing the excitation waveform through the bandwidth expanded synthesis filter as shown in Eq. (2.35).

$$\tilde{r}^i_{[0,N)}(n) = \sum_{n=0}^{N-1}\hat{r}_i(k)h_\gamma(n-k) \tag{2.35}$$

Once the gain G is determined, the value is substituted back into Eq. (2.33), and the optimal excitation waveform is determined by minimizing the energy ϵ_c over all the codewords $r_i(n)$.

This concludes a detailed description of the CELP's system configuration. All simulations performed throughout this research used this configuration. Subsequent chapters and sections will refer to this configuration with specific changes to the pitch filter $P(z)$, the weighting filter $W(z)$ and LPC parameter coding.

Chapter 3

Wideband LPC Parameter Coding

3.1 Introduction

In this chapter, we are concerned with determining efficient transformation and quantization methods of the formant filter parameters a_k's. These parameters are not well suited for transmission because an error in any one coefficient can cause the filter to become unstable and their wide dynamic range makes an efficient quantization practically impossible.

In retrospect, one of the most significant factors in contributing to a successful implementation of the current wideband LPC was the choice to transmit reflection coefficient rather than prediction coefficients. The weakness of these new coefficients was that changes in one coefficient caused speech spectral changes in the entire passband. To overcome this weakness, frequency domain parameters were introduced and more specifically line spectral frequencies (LSFs) [7].

The use of LSFs is highly recommended because they allow filter coefficient quantization in accordance with properties of auditory perception, their quantization is more efficient due to a band-limited dynamic range (50–7000 Hz for wideband speech). Another advantage of using LSFs is that an error in one LSF only affects the synthesized spectrum near that frequency. Another LSF feature that is useful is

the one-to-one correspondence between LSF and LPC parameters. this feature will make any ordered set of LSFs correspond to a stable synthesis filter.

Extensive research has been accomplished in LSF coding, but most of these studies dealt with narrowband applications where only 8 to 10 poles were transformed and coded. Paliwal and Atal [22] were able to implement a 10 poles system with a rate of 1200 bits/sec (20 ms frame) while maintaining a high standard of speech quality. In this research, higher order systems are required (16 to 20 poles), consequently increasing the bit rate as high as 2500 bits/sec to code 16 LSFs.

In this chapter, we first define the LSFs and describe some of their properties, we then study the effects of both scalar and vector quantization of the LSF on the performance of the coder; finally, we look at the advantages of using LSF interpolation to improve speech quality.

3.2 LSF representation and properties

The inverse formant N_p-th order filter $A_{N_p}(z)$ that models the shape of the vocal tract transforms speech samples into prediction residual samples where

$$A_{N_p}(z) = 1 - \sum_{k=1}^{N_p} a_k z^{-k} \tag{3.1}$$

The transfer function of the LPC analysis inverse filter can also be expressed in lattice form, thereby corresponding to an acoustical tube model of the vocal tract. The recursive relationship of $A_{n+1}(z)$ in terms of $A_n(z)$ $(n = 1, \ldots, N_p)$ is established as

$$A_{n+1}(z) = A_n(z) - k_{n+1} z^{-(n+1)} A_n(z^{-1}) \tag{3.2}$$

where k_{n+1} is the $(n+1)$-th reflection coefficient for the $(n+1)$-th tube.

The lattice form of the filter is decomposed into a sum of an even $Q_{n+1}(z)$ and an odd $P_{n+1}(z)$ function and we have

$$A_n(z) = \frac{1}{2}[P_{n+1}(z) + Q_{n+1}(z)] \tag{3.3}$$

where $P_{n+1}(z)$ represent the complete opening of the glottis with k_{n+1} set to +1 and $Q_{n+1}(z)$ represent the complete closure of the glottis with k_{n+1} set to -1

$$\begin{aligned} P_{n+1}(z) &= A_n(z) - z^{-(n+1)}A_n(z^{-1}) \\ Q_{n+1}(z) &= A_n(z) + z^{-(n+1)}A_n(z^{-1}) \end{aligned} \tag{3.4}$$

The LSFs are determined by solving the above two polynomials. According to Soong and Juang [28], all zeros of $P_{n+1}(z)$ and $Q_{n+1}(z)$ lie on the unit circle, roots of $P_{n+1}(z)$ and $Q_{n+1}(z)$ alternate between the two polynomials as the angle ω increases:

$$0 = \omega_0 < \omega_1 < ... < \omega_{N_p} < \omega_{N_p+1} = \pi \tag{3.5}$$

The LSFs correspond to these angular positions. Both ω_0 and ω_{N_p+1} induced by the $(N_p + 1)$st stage are implicit LSFs and consequently are not transmitted.

The first important characteristic of the LSFs is that peaks in the spectral envelope (formant frequencies) are identified by the closeness of neighbouring LSFs as shown in Fig. 3.1 where the dotted lines represent LSFs for a 16-th order formant filter.

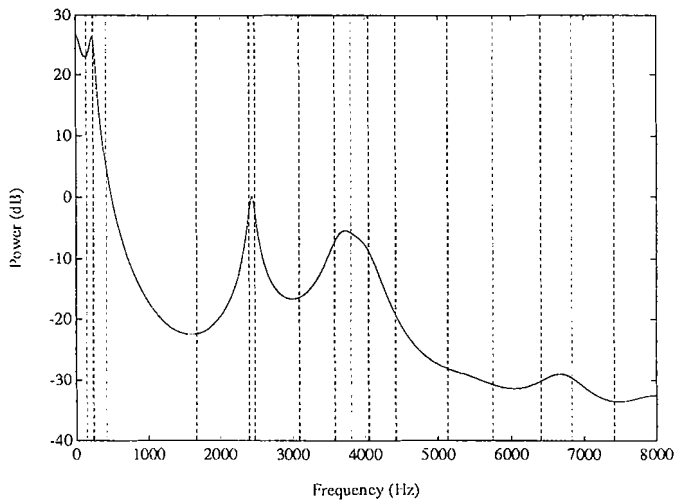

Fig. 3.1: LPC power spectrum and associated LSFs.

Also, the spectral sensitivities of LSFs are localized, a change in one LSF results in local change of the LPC power spectrum in the neighbourhood of this LSF. Fig.

3.2 shows the effects of modifying the 12th LSF on a 16-th order formant filter from a value of 5127 Hz to 5085 Hz, the changes in the spectrum only appear in the neighbourhood of 5127 Hz, the dotted curve represent the modified LPC spectrum.

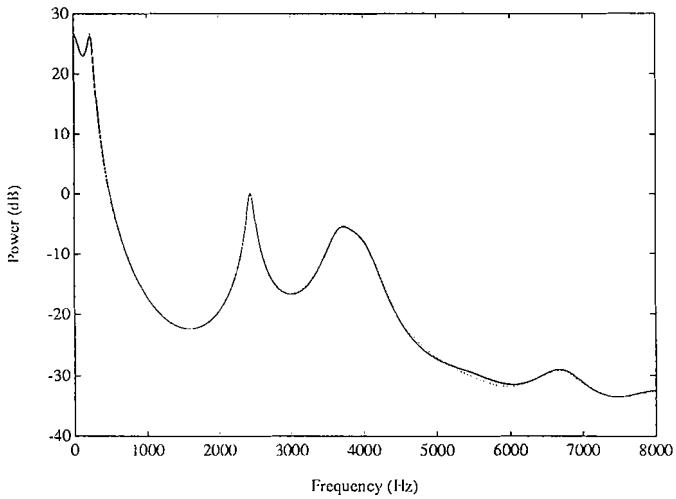

Fig. 3.2: Spectral sensitivities of LSFs.

With these properties, an efficient numerical computation of the LSFs from the two polynomials can be established.

- Kang and Fransen [10] first introduced an approach where the roots on the unit circle are determined by an all-pass ratio filter $R(z)$

$$R(z) = \frac{z^{-(N_p+1)}A(z^{-1})}{A(z)} \tag{3.6}$$

  The phase function of the filter is then determined, and the LSFs coincide with the phase response that takes on a value that is a multiple of π.

- Soong and Juang [28] developed a method where the implicit roots ω_0 and ω_{N_p+1} were removed and the new polynomials $P'(z)$ and $Q'(z)$ evaluated. A discrete cosine transformation to the coefficients of $P'(z)$ and $Q'(z)$ is then applied. The roots are finally found by searching for the changing sign of the polynomials along the $\omega = [0, \pi]$ range.

- Kabal and Ramachadran [24], similar to the previous method, used the Chebyshev transformation $x = \cos\omega$ to map the upper semicircle in the z-plane to the $[-1, +1]$ range. The polynomials $P'(\omega)$ and $Q'(\omega)$ are expanded into $P'(x)$ and $Q'(x)$ using the Chebyshev polynomials. The roots are determined again by searching for sign changes in the interval $[-1, +1]$ and the LSFs are computed by performing an inverse transformation $\omega = \arccos x$ on these roots.

During this research, the last method was selected to convert the predictor coefficients into normalized line spectral frequencies.

3.3 LSF quantization

In the past, two basic approaches for the quantization of LPC coefficients were used. The first, *scalar quantization*, quantized each LPC coefficient individually, while the second, *vector quantization*, quantized all the LPC coefficient as a group. The first suffered from a high number of bits required for quantization while the second faced the misfortune of being highly complex in terms of the amount of training data needed, the memory and the number of computations. In this section, both scalar quantization and a modified vector quantization are investigated and their respective performances are compared.

In both methods, the quantizers were designed using a set of 4800 non-silent frames from the wideband speech database described in Appendix A. The design was performed for both a 16-th and 20-th order LPC filter. The update rate in the following simulations uses a frame of 320 samples and a subframe of 40 samples (320:40 mode). All the other quantizers (gain and pitch) were turned off, and the codebook search was accomplished with 1024 Gaussian waveforms.

3.3.1 Scalar quantization

The use of LSFs as efficient representation of LPC coefficients was essentially chosen for its narrower dynamic range. Nevertheless, this new dynamic range is still difficult to quantize. Instead of performing a direct quantization of LSFs, Soong and Juang [29] introduced a differential coding scheme where the spectral distance $d(\omega_i, \omega_{i+1})$ between neighbouring LSFs is encoded. The main advantage of this method is that it preserved the ascending order of the LSFs.

This method is better known as differential non-uniform quantization. An M-level quantizer is used to quantize the first LSF ω_1 into $\hat{\omega}_1$, and this quantizer is designed by minimizing the average square error distortion D:

$$D = \int_{\omega_{min}}^{\omega_{max}} (\omega - q(\omega))^2 p(\omega) d\omega \tag{3.7}$$

where $q(\omega)$ and $p(\omega)$ are respectively the quantizer and the density functions of ω.

Subsequent LSFs are then determined with ω_1 and $d(\omega_i, \omega_{i+1})$. For instance, to regenerate ω_k as $\hat{\omega}_k$ at the receiver, we have:

$$\hat{\omega}_k = \hat{\omega}_1 + \sum_{i=1}^{k-1} d(\omega_i, \omega_{i+1}) \tag{3.8}$$

Two performance measures are used to determine the efficiency of scalar quantization. The first one, *Segmental Signal to Noise Ratio*, measures the SNR in dB of the reproduced speech with respect to the original; while the second one, *average spectral distortion*, measures the distortion level in dB^2 between coded and original spectral envelopes:

$$SD = \frac{1}{N_{fr}} \sum_{n=1}^{N_{fr}} [\frac{1}{\pi} \int_0^{\pi} (\log[E_n(\omega)] - \log[E_n(\hat{\omega})])^2] \tag{3.9}$$

where $E_n(\omega)$ and $E_n(\hat{\omega})$ are respectively the unquantized and quantized LPC spectra for the n-th frame, and N_{fr} is the total number of frames.

Table 3.1 shows the performance of scalar quantizers for the two LPC orders with different quantization levels and no quantization (noq). The performance measures

were carried out on 48 speech files (24 male and 24 female) described in Appendix A with an update mode of (320:40).

Order	Bits	Ave. SD	Outliers(in %)		SegSNR
		(dB)	2–4 dB	>4 dB	(dB)
16-th	noq	–	–	–	14.47
	30	2.0896	22.30	10.11	14.09
	50	0.8578	1.57	0.01	14.31
20-th	noq	–	–	–	13.41
	46	1.7079	17.41	5.82	13.20
	63	0.7290	0.97	0.02	13.27

Table 3.1: Spectral distortion and SegSNR measures for scalar quantization.

Note that the performance in terms of SegSNR of these quantizers deteriorate as the number of poles increases. this is direct consequence of a lower number of bits being allocated per pole. The SegSNR figures show also that with a higher number of poles (20) with an increase in the number of bits will still generate a lower quality speech when compared to a lower number of poles (16). In general, quantization effects become negligible for spectral distortion measures that fall under 1 dB^2 making the 50 bits/frame 16-th order quantizer the appropriate candidate for scalar quantization. Fig. 3.3 and 3.4 show examples of LPC power spectral envelopes (quantized and unquantized) for this configuration.

3.3.2 Vector quantization

In vector quantization, three parameters control the quality and performance of the coder:

- Size of the codebook
- Method used to generate the codebook

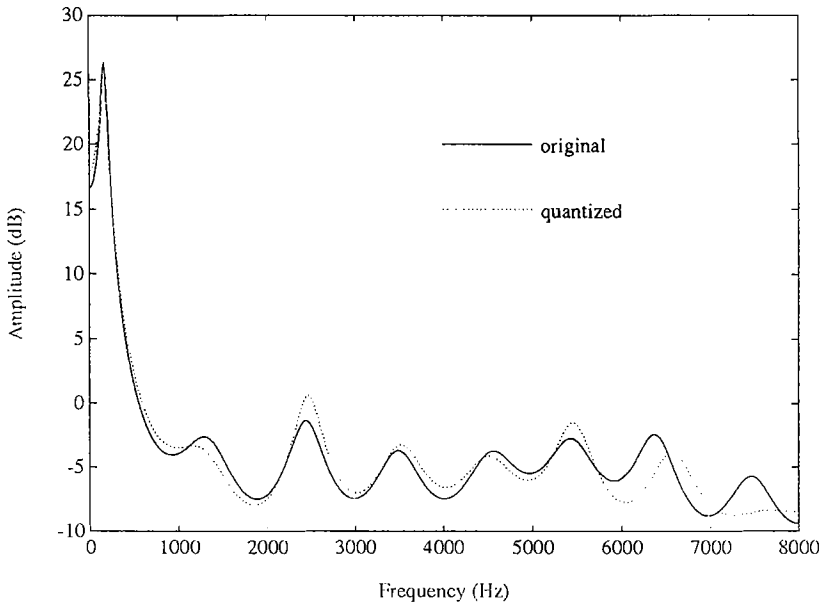

Fig. 3.3: Male LPC power spectral envelopes for SQ.

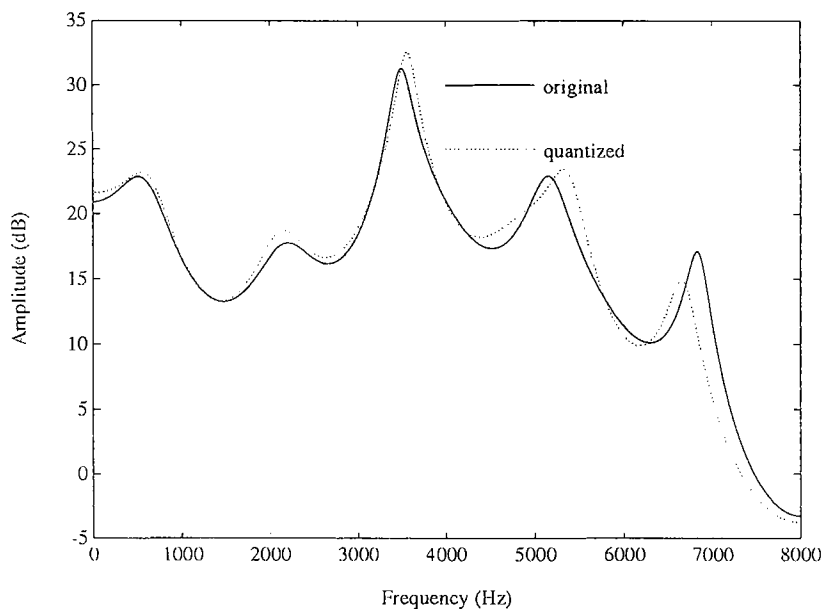

Fig. 3.4: Female LPC power spectral envelopes for SQ.

• Distance measure used to select the optimal vector

In this section, a three way split vector quantization on wideband LSF parameters is introduced. The reference LPCs are first transformed into LSFs and then divided into three subgroups. A training data of LSF vectors is used to construct different codebook sets with varying levels of complexity (e.g. 30–33 bits used). This operation is performed with the use of the Linde Buzo Gray (LBG) algorithm [15]. This algorithm designs vector quantizers in the following manner:

1. Data files are generated containing the LSFs from 4800 frames of speech.
2. Weights are assigned to LSFs: low band LSFs get high weights while high band LSFs get low weights.
3. Centroid of the LSF data with weighting is determined.
4. Centroid is split into two centroids.
5. LSF data is clustered to the closest centroid using the difference measure $\frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^2$ with weighting.
6. New centroid of clustered data is determined.
7. Distortion of centroids is measured, if low continue and if not repeat step 3.
8. Repeat step 2 until the required codebook size is reached (1 to 14 bits).

During the training of these codebooks, coefficients of LSF vectors can loose their order and result in an unstable LPC filter. After several splitting, the LSF centroid can cause the LSF vector to loose its well-orderness. To correct this instability of the codebooks, ill-conditioned vectors should be either removed or corrected from the codebook.

Once the codebook design process is over and the optimal codebooks generated, the selection process can take place where codebook LSF vectors are compared to a

reference LSF vector by minimizing a distortion measure. The optimum codebook LSF vector is found by minimizing this distance measure.

Weighted LSF distance measures

In this section, two new weighted Euclidean LSF distance measure are introduced. For a given reference LSF vector $\vec{v}_{ref}$, these two measures determine the best matching spectral envelope $\vec{v}_{cod}$ from a vector quantization codebook. We have,

$$d(\vec{v}_{ref}, \vec{v}_{cod}) = \sum_{k=1}^{N_{lsf}} w_k (f_k - \hat{f}_k)^2 \tag{3.10}$$

where f_k and $\hat{f}_k$ are the k-th LSFs in the reference and codebook vector, respectively, while w_k is the k-th LSF weighting factor that considers both the frequency sensitivity and distance between LSFs for the first measure $d_1(\vec{v}_{ref}, \vec{v}_{cod})$, or the frequency sensitivity and the position of the LSF for the second measure $d_2(\vec{v}_{ref}, \vec{v}_{cod})$:

$$w_k = w_k^{(i)} w_k^{(ii)} \tag{3.11}$$

The first weighting factor $w_k^{(i)}$ models the hearing sensitivity to frequency differences curve as shown in Fig. 3.5. This curve shows our hearing sensitivity to frequency difference as function of frequency. Specific weights are assigned to the LSFs according to their position in the frequency spectrum.

The second weighting factor depends on the distance measure used. For the first measure, $w_k^{(ii)}$ refers to the distance between LSFs. The closer they are together, the more likely they are to fall near a formant.

$$w_k^{(ii)} = 0.05 + [1 - \frac{d_k}{d_{max}}]^2 \tag{3.12}$$

where d_k is the distance between LSF f_{ii} and its closest neighbour f_{ii-1} or f_{ii+1}, and d_{max} is the maximum distance between the LSFs.

For the second measure, the weighting factor $w_k^{(ii)}$ refers to the position of the LSF in the LPC spectrum. Higher weighting is assigned to LSFs in the formant

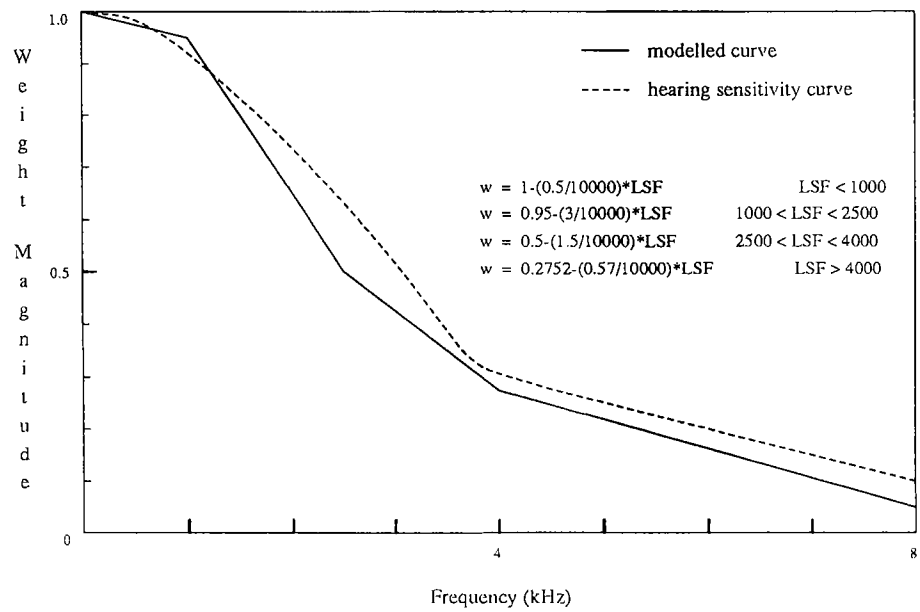

Fig. 3.5: Human and modeled hearing sensitivity to discriminating frequency differences.

regions than those outside these regions. More importance is also given to LSFs corresponding to high amplitude formants than to the ones corresponding to lower amplitude formants. The weighing factor is defined as

$$w_k^{(ii)} = [P(f_{ii})]^r \tag{3.13}$$

where $P(f)$ is the LPC power spectrum and r a constant that control the weighting assigned to different LSFs. It is set to 0.15 which is a satisfactory value for this study.

These two measures were thoroughly investigated, and the results in terms of SegSNR and SD were very similar. the second method outperforming the first only by 0.002 dB^2 in SD and 0.01 dB in SegSNR. The only real difference between the two distance measures was in the level of complexity. In the second method, where an LPC spectrum was generated for every analysis frame, a large amount of computations was needed to carry out a 512 point FFT (Fast Fourier Transform) on the LPC signal. Consequently, the first distance measure was selected for use in the LSF vector quantization scheme because of its low level of complexity and good performance.

Split vector quantization

The split vector quantization scheme was first introduced by Paliwal and Atal [22], but its application was limited to narrowband speech. In this subsection, we describe the application of this algorithm to wideband speech. Initially, the research was conducted on splitting the LSF vector into two subvectors, but it turned out that the bit assignment, required to yield an acceptable level of distortion was still too high. A decision was then made to split the LSF vector into three subvectors with varying configurations:

- For the 16-th order, three configurations were investigated:
  - The first 4, the middle 4 and the last 8 LSFs (4-4-8)
  - The first 8, the middle 4 and the last 4 LSFs (8-4-4)
  - The first 4, the middle 6 and the last 6 LSFs (4-6-6)

- For the 20-th order, two configurations were investigated:
  - The first 5, the middle 5 and the last 10 LSFs (5-5-10)
  - The first 10, the middle 5 and the last 5 LSFs (10-5-5)

Different size codebooks (2 to 14 bits) were generated for all these configuration and simulations to determine the best configurations were carried out. Two search techniques in conjunction with the distance measure selected were studied. The first search techniques conducts an independent search for every LSF subgroup. While, the second technique performs a nested search where priority is given to the first LSF subgroup where most of the perceptual information is stored; the optimal first vector is combined with the second LSF codebook to generate the second LSF vector; finally, the optimal first and second vectors are combined with the third LSF codebook to obtain the overall LSF vector.

Fig. 3.6 shows the block diagrams of these two techniques. The nested search technique was very effective for first subgroups containing the highest number of LSFs

(8-4-4 mode and 6-4-4 for the 16-th order and 10-5-5 mode for the 20-th order). The overall performance of the nested search technique over the independent technique was a gain of 0.09 dB in SegSNR and 0.02 dB^2 in SD.

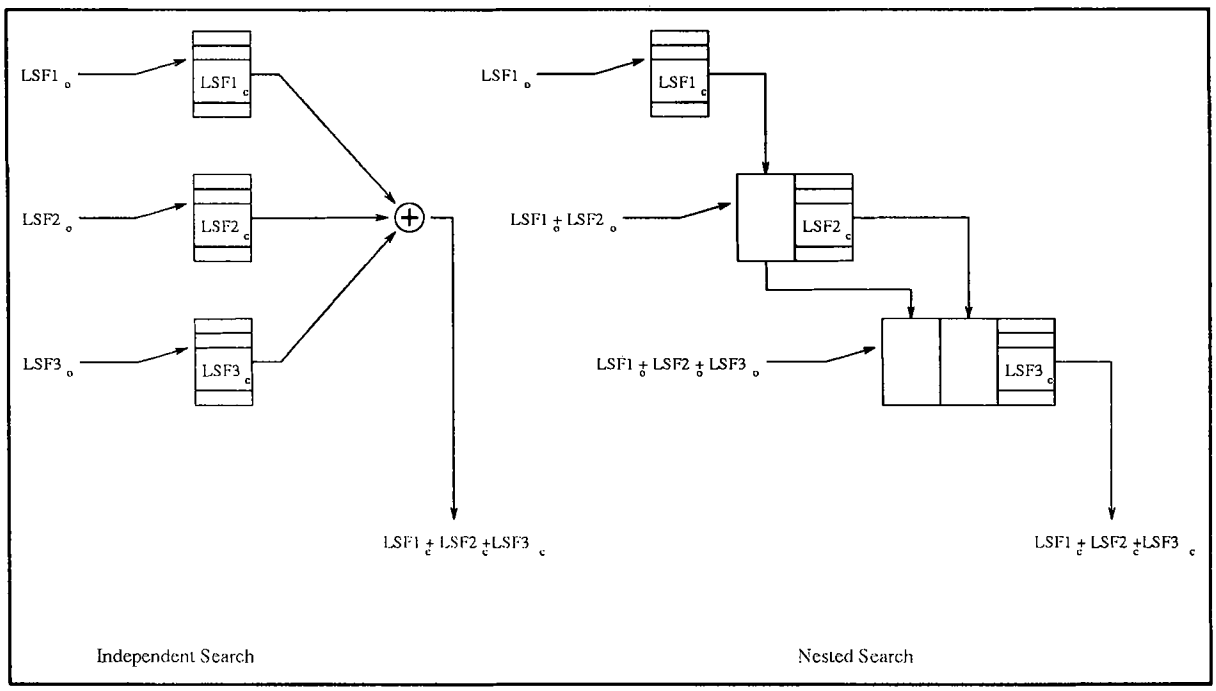

Fig. 3.6: LSF codebook search techniques.

Table 3.2 shows the overall performances in terms of spectral distortion. For every configuration, two bit assignments were used, one at 30 bits/frame and the other at 33 bits/frame. The 20-th order configuration were ruled out because of their high level of spectral distortion. From Table 3.2, the best candidate for split vector quantization is the 16-th order third entry (4-4-8 mode for 30 and 33 bits/frame). Figures 3.7 and 3.8 show examples of LPC power spectral envelopes (quantized and unquantized) for the 30 bits/frame configuration.

Finally, Table 3.3 shows the SegSNR and SD performance for the 4-4-8 split vector quantization scheme using the nested search method and the second distance measure for both a 30 bits/frame and 33 bits/frame configuration. The simulations were performed on the same speech files used to evaluate the scalar quantization

Order	Splits (number of LSFs and bits used)						SD
	Part 1	Bits	Part 2	Bits	Part 3	Bits	(dB.)
16-th	8	13	4	11	4	9	0.921
	8	13	4	9	4	8	0.996
	4	11	4	11	8	11	0.840
	4	10	4	10	8	10	0.934
	4	13	6	12	6	8	0.870
	4	13	6	10	6	7	0.960
20-th	10	13	5	11	5	9	1.052
	10	13	5	9	5	8	1.055
	5	11	5	11	10	11	0.933
	5	10	5	10	10	10	0.994

Table 3.2: Spectral distortion (SD) measures.

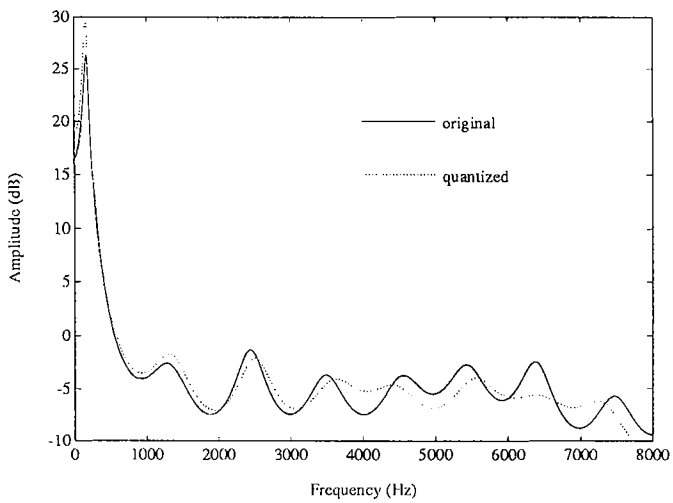

Fig. 3.7: Male LPC power spectral envelopes for VQ.

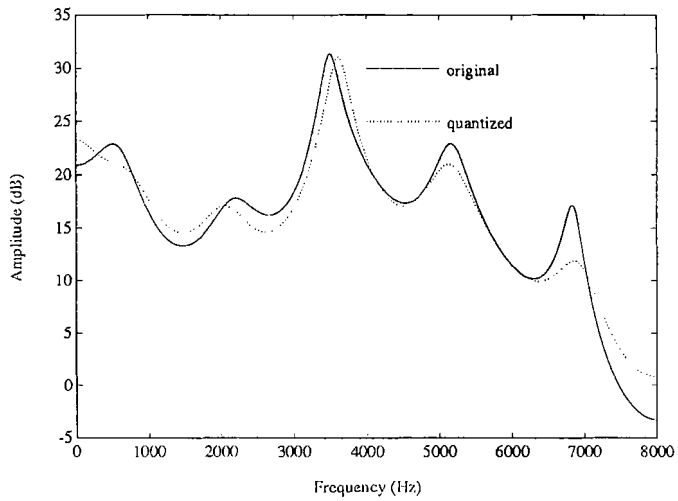

Fig. 3.8: Female LPC power spectral envelopes for VQ.

approach with an update mode of (320:40).

Order	Bits	Ave. SD	Outliers(in %)		SegSNR
		(dB)	2–4 dB	>4 dB	(dB)
16-th	noq	–	–	–	14.47
	30	0.9342	1.45	0.04	14.28
	33	0.8403	1.36	0.02	14.32

Table 3.3: Spectral distortion and SegSNR measures for vector quantization.

By comparing Table 3.3 and 3.1, we can conclude that split vector quantization gave a comparable performance to scalar quantization but with fewer bits (20 bits/frame less), therefore reducing the bit rate while keeping a high level of speech quality.

3.4 LSF cross-overs

The stability of the LPC analysis filter of order N_p is only maintained when the LSFs are in ascending order and do not cross-over. The condition is that

$$\omega_{i+1} > \omega_i \quad i = 1, ..., N_p - 1 \tag{3.14}$$

Cross-overs are more frequent in scalar quantization that in vector quantization. In scalar quantization, every LSF is quantized independently, and when the bit assignment per LSF is insufficient, cross-overs become a serious concern. In vector quantization, this problem is less acute because of the fact that LSFs are quantized in groups therefore maintaining ordered sequences, the only cross-overs that might arise in this case are between LSF subgroups and more specifically between the endpoint LSFs of these subgroups.

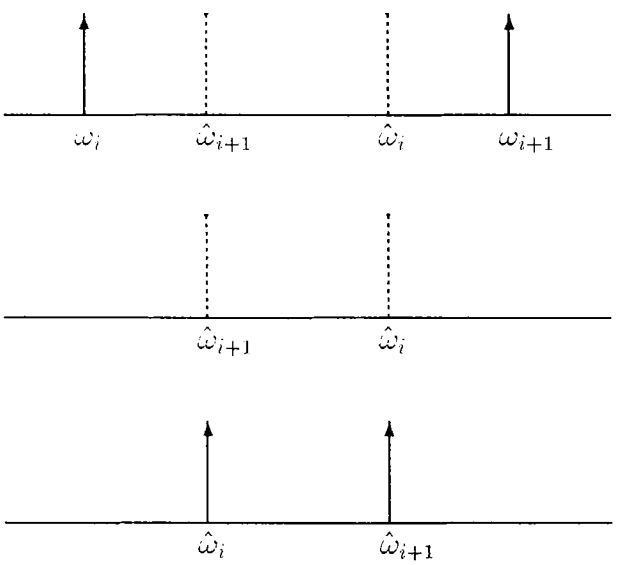

Fig. 3.9: LSF cross-over correcting scheme.

A simple approach is used to correct LSF cross-overs where the positions of the concerned LSFs are inverted so as to stabilize the LPC filters as illustrated in Fig.

3.9. The dotted lines represent the best quantized LSF candidates for the two original ω_i and ω_{i+1}. their positions are then switched to preserve the ascending order of the LSFs.

3.5 LSF interpolation

A further improvement that can be added to wideband spectral envelope coding is the use of interpolation of LSF parameters. By studying successive LSF frames, a strong correlation between neighbouring frames can be established and motivate the use of an interpolation scheme to increase the update rate of LSFs. Initially, LSFs were updated every frame, but with interpolation they are now updated every subframe, therefore improving the quality of the reproduced speech.

In this interpolation scheme, each LSF subframe of a given frame is assigned specific weights with a combination of the previous and present LSF frames where

$$\mathrm{LSF}_{sub}(i) \;=\; w_{sub1}\mathrm{LSF}_{fr}(i-1) + w_{sub2}\mathrm{LSF}_{fr}(i) \tag{3.15}$$

The weights used for three different update modes (frame:subframe) are shown in Table 3.4. These weights were determined experimentally by trial and error techniques. For example, if LSF 1 in the 320:40 mode is 150 Hz and the previous LSF 1 was 185 Hz, the value used for subframe 2 of the present frame is

$$\mathrm{LSF2}_{sub} \;=\; 0.80 \times 185 + 0.20 \times 150 \;=\; 178 \text{ Hz}$$

The main advantage of using LSF interpolation is a noted increase in SegSNR figures by approximately 0.1–0.4 dB with no additional bit requirements as shown in Table 3.5 with all quantizers turned off including LSF quantization.

In this chapter, we studied both scalar and split vector quantization, and we showed that with the use of a new perceptually weighted Euclidean distance measure and a nested search technique the bit rate was reduced by 20 bits/frame when compared to scalar quantization. Using this new split VQ techniques, we were able to achieve

Mode	Subframe	Previous LSF	Present LSF
160:40	1	0.875	0.125
	2	0.625	0.375
	3	0.375	0.625
	4	0.125	0.875
250:50	1	0.80	0.20
	2	0.70	0.30
	3	0.50	0.50
	4	0.30	0.70
	5	0.20	0.80
320:40	1	0.85	0.15
	2	0.80	0.20
	3	0.70	0.30
	4	0.55	0.45
	5	0.45	0.55
	6	0.30	0.70
	7	0.20	0.80
	8	0.15	0.85

Table 3.4: LSF weighted averaging figures for three modes.

Mode	Interpolation	SegSNR (dB)
160:40	off	15.06
	on	15.25
250:50	off	13.40
	on	13.46
320:40	off	14.47
	on	14.86

Table 3.5: SegSNR figures for LSF interpolation.

an almost transparent quantization of LPC information (i.e. with less than 1 dB^2 average spectral distortion, less that 2% outliers in the range 2-4 dB, and almost no outliers having spectral distortion > 4 dB).

Chapter 4

Improved Pitch Filtering

4.1 Introduction

The addition of a pitch filtering stage to the CELP coder constitutes a major part of its success especially at lower bit rates. At high bit rates, a substantial number of bits is assigned to the excitation signal to enable the coder to reconstruct the harmonic structure that the long term predictor fails to model. However, at lower bit rates, the excitation signal does not have so much variability and the speech quality is much more dependent on the performance of the pitch predictor.

The long term predictor excitation signal and the pitch filtering parameters are determined for every subframe. The computational load to generate these parameters is considerably reduced when the subframe length is set to be always smaller than the minimum pitch delay. In this case, an adaptive codebook is used to offset the degrading effects of this constraint, the codebook contains past excitation signals starting from a minimum lag value of 40 samples back in time (2.5 ms) to a maximum lag value of 320 samples back in time (20 ms).

According to Moncet and Kabal [18], the pitch delay tends to vary smoothly in voiced segments, and only occasionally departs from its smooth trajectory. However, in unvoiced segments, the pitch delay tends to jump around. Reducing the resolu-

tion of the pitch delay results in problems locking onto the correct pitch during the transition from silence to voiced speech. Consequently, a good pitch delay resolution should be maintained at all times during the analysis and synthesis stages of the CELP coder.

The outline of the present chapter is as follows. First, we provide a description of a basic one-tap pitch predictor. Then, we discuss the use of multi-tap pitch predictors and their performances. Finally, we study the impact of single-tap pitch predictors with fractional delays.

4.2 Basic one-tap pitch filter

The system configuration used to perform the simulations for a one-tap pitch predictor is already described in Section 2.3. In the following, we consider the generation and quantization of both the pitch coefficient and pitch lag.

Only a single pitch coefficient is used here. Fig. 4.1 shows the histogram of pitch coefficient values where a total of 38400 subframes (3.125 ms each) including both male an female speech utterances were used. As shown in Fig. 4.1, the pitch coefficient values detected tend to be mainly positive, the negative ones are less important to encode because they usually occur in speech regions with low energy and therefore contribute much less to the overall speech quality.

We also computed the prediction gain for every pitch coefficient to analyze the relative importance of different pitch values as shown in Fig. 4.2. Again, the pitch filtering gain tends to be higher in the positive pitch coefficient region and lower in the negative region.

The pitch coefficients were quantized with 5 bits and the resulting optimal quantizer values are shown in Table 4.1. Note that the number of positive pitch coefficient quantizer output levels is higher that the number of negative ones. The adaptive pitch lag codebook as described earlier contains 256 values (coded with 8 bits). Fig.

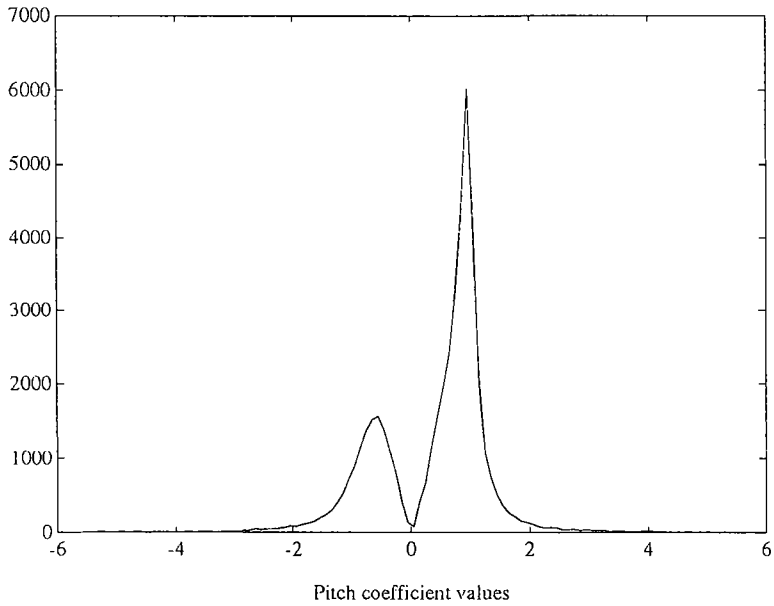

Fig. 4.1: Histogram of a single pitch coefficient.

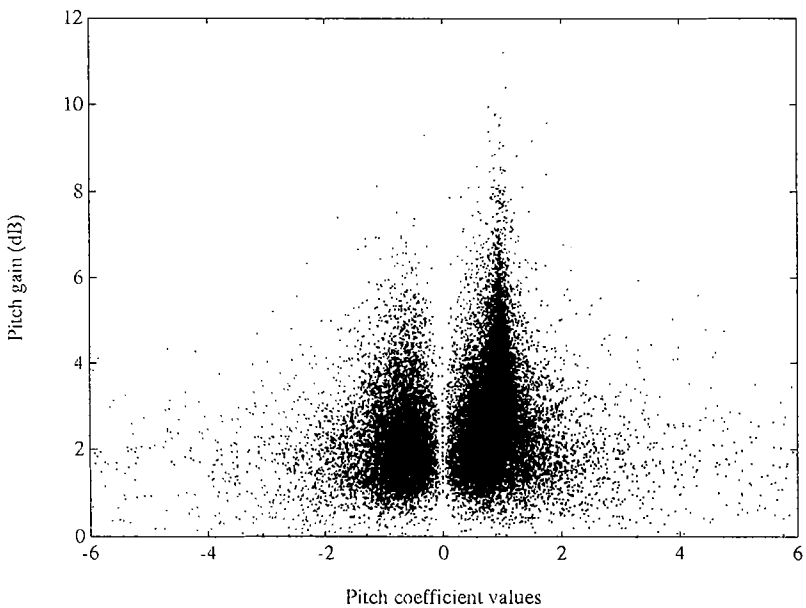

Fig. 4.2: Prediction gain vs pitch coefficient value.

Output	Decision	Output	Decision	Output	Decision
-2.638	-2.453	-0.297	-0.231	1.180	1.252
-2.269	-2.091	-0.164	-0.082	1.324	1.402
-1.914	-1.769	0.000	0.084	1.480	1.562
-1.623	-1.497	0.168	0.234	1.644	1.731
-1.371	-1.260	0.300	0.363	1.818	1.916
-1.148	-1.058	0.425	0.489	2.014	2.111
-0.968	-0.894	0.552	0.613	2.207	2.300
-0.819	-0.751	0.674	0.736	2.393	2.474
-0.682	-0.617	0.799	0.858	2.555	2.644
-0.552	-0.490	0.918	0.979	2.733	
-0.428	-0.363	1.041	1.110		

Table 4.1: Optimal quantizer for pitch predictor coefficient.

4.3 shows the parameter tracks for pitch filter coefficient values, pitch lag values and pitch filtering gain values.

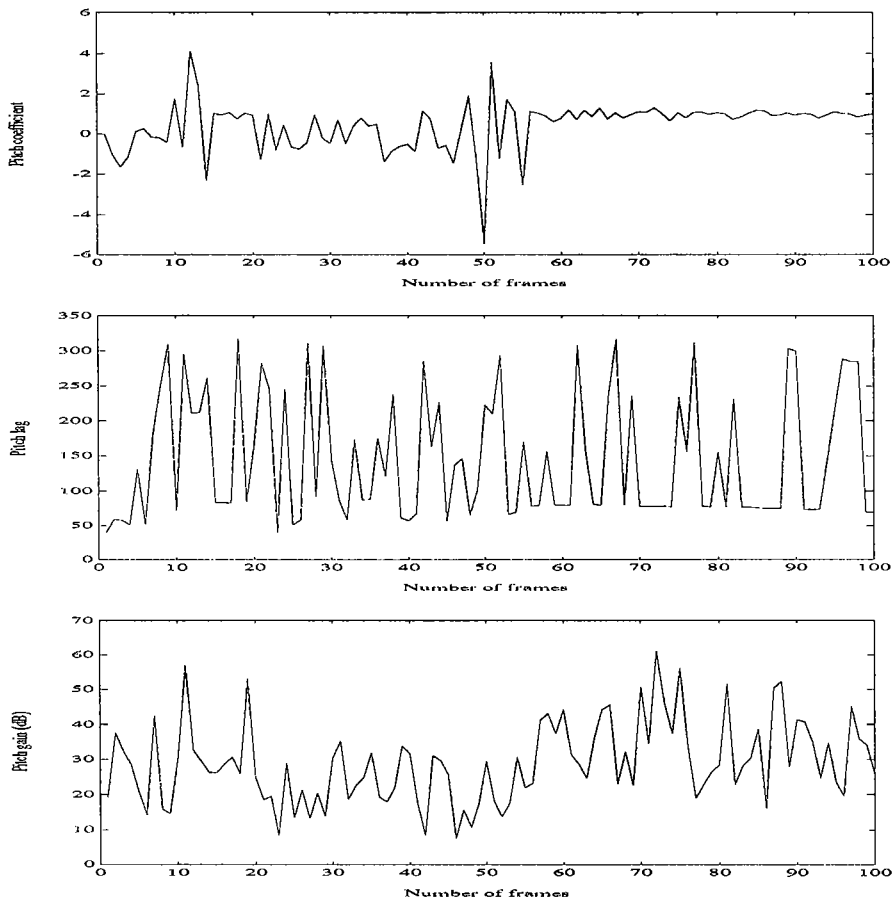

Fig. 4.3: Parameter tracks.

The sub-optimal procedure where the codebook parameters (gain and index) are decoupled from pitch parameters (coefficient and lag) is used here as described in Section 2.2. The pitch coefficient is quantized first and quantization errors are compensated by an adequate selection of the gain factor.

The performance figures of the one-tap prediction filter are shown in Table 4.2. All other quantizers are turned off and simulations are carried out on the 48 speech files of Appendix A.

Mode	Quantization	SegSNR (dB)	
		female	male
250:50	off	13.88	12.92
	on	13.71	12.91
320:40	off	14.72	14.21
	on	14.66	14.12

Table 4.2: SegSNR figures using a one-tap pitch predictor.

4.3 Multi-tap pitch filtering

So far, we have only discussed the behaviour of single tap pitch predictor, but multi-tap pitch predictors, especially three-tap pitch predictors, are now frequently used in CELP coders because of the improved speech quality they produce when compared to single tap pitch predictors. Nevertheless, the improvement will come at the cost of an increased bit rate needed to encode the additional pitch parameters.

The energy of the weighted error signal using a multi-tap pitch filter with N_p pitch coefficients is now defined as

$$\epsilon_{pit} = \sum_{n=0}^{N-1} e_w^2(n) = \sum_{n=0}^{N-1} (s_{new}(n) - \hat{s}_0(n))^2 \tag{4.1}$$

where

$$\hat{s}_0(n) = \sum_{i=1}^{N_{pit}} \beta_i y_i(n) = \sum_{k=0}^{N-1} [\sum_{i=1}^{N_{pit}} \beta_i \hat{d}(k-M-i)] h_\gamma(n-k) \tag{4.2}$$

and

$$y_i(n) = \sum_{k=0}^{N-1} \hat{d}(k-M-i) h_\gamma(n-k) \tag{4.3}$$

Differentiating the expression of ϵ_{pit} with respect to the pitch coefficient where β_i $i = 1, ..., N_{pit}$, we have:

$$\frac{\delta \epsilon_{pit}}{\delta \beta_j} = -2 \sum_{n=0}^{N-1} s_{new}(n) y_j(n) + 2 \sum_{n=0}^{N-1} \sum_{i=1}^{N_{pit}} \beta_i y_i(n) y_j(n) \tag{4.4}$$

By setting the derivatives to zero, we can solve the system of N_{pit} equations and determine all the pitch coefficients β_i.

This optimization procedure is applied to a three-tap pitch filter. the simulations are carried out with a total of 11 bits for pitch coefficient quantization (5 for β_1, 3 for β_2 and 3 for β_3). Again all other quantizations are turned off and the performance figures are shown in Table 4.3.

Mode	Quantization	SegSNR (dB)	
		female	male
250:50	off	14.33	14.02
	on	14.25	13.97
320:40	off	15.22	15.11
	on	15.13	15.00

Table 4.3: SegSNR figures using a three-tap pitch predictor.

4.4 Fractional pitch filtering

In most coding applications, the pitch period is restricted to integer multiples of the sampling interval as described in the last two sections. This restriction has more pronounced effects on high pitch sounds, resulting in the partial destruction of the harmonic structure, especially in the high frequency regions [17]. The use of fractional pitch delays has proved to be a very efficient method to overcome this problem in CELP coders [12], but so far studies were only made on narrowband speech. Considering the fact that the periodicity of a wideband signal is almost nonexistent in the 4–8 kHz band and that doubling the sampling frequency to 16 kHz meant improved resolution, the need for fractional delays in wideband speech is reduced.

In this section, we investigate the actual impact of fractional pitch delays on

wideband speech. In fact, the use of non-integer delays could be more beneficial in terms of lower bit rates (10 bits/subframe) when compared to a multiple tap integer delay predictor (11 bits/subframe for 3 pitch taps). High temporal resolution for pitch delays can be achieved by specifying the delay as an integer number of samples plus a fraction of a sample $\frac{l}{D}$ where $l = 0, 1, ..., D-1$, and l and D are integers.

The pitch delay in wideband speech ranges from $M = 40$ to $M = 320$ samples with some delays occurring more often than others, therefore it would be beneficial to assign finer resolution to these delays while leaving the others at a lower resolution level. With the use of interpolation and polyphase filters [4, Section 6.3], fractional delays can be efficiently implemented for a first order pitch predictor.

In fractional pitch filtering, the elements of the adaptive codebook have to be shifted by the desired fractional sample. Fig. 4.4 shows the different steps involved in performing a fixed delay of l/D.

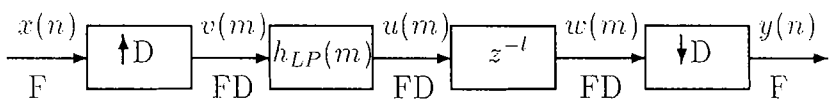

Fig. 4.4: Multirate structure for a delay of l/D samples.

Note that a delay of l/D samples at a rate F is equivalent to a delay of l samples (i.e. an integer delay) at a rate FD. Therefore, the following steps for the realization of a delay of l/D are:

1. increase the sampling rate F of the input signal by D.

2. filter the signal with a low-pass filter $h_{LP}(m)$ to prevent aliasing.

3. delay the signal by l integer samples.

4. decimate the sampling rate FD down to F.

By examining the block diagrams in Fig. 4.4, we have:

$$V(e^{j\omega}) = X(e^{j\omega}) \tag{4.5}$$

$$U(e^{j\omega}) = H_{LP}(e^{j\omega})V(e^{j\omega}) = H_{LP}(e^{j\omega})X(e^{j\omega}) \tag{4.6}$$

$$W(e^{j\omega}) = U(e^{j\omega})e^{-j\omega l} = H_{LP}(e^{j\omega})X(e^{j\omega})e^{-j\omega l} \tag{4.7}$$

and the output is

$$\begin{aligned} Y(e^{j\omega}) &= \frac{1}{D}\sum_{r=0}^{D-1} W(e^{-j2\pi r/D}e^{j\omega/D}) \\ &= e^{-j\omega I}e^{-j\omega I/D}X(e^{j\omega}) \end{aligned} \tag{4.8}$$

by taking into account three assumptions:

- $H_{LP}(e^{j\omega})$ sufficiently attenuates the images of $X(e^{j\omega})$, therefore only the $r = 0$ is considered.

- $H_{LP}(e^{j\omega})$is an FIR filter with exactly linear phase whose delay at high rate FD is $(N-1)/2$ samples and this value is chosen to be an integer multiple of D

$$\frac{N-1}{2} = ID \tag{4.9}$$

- $H_{LP}(e^{j\omega})$ has a magnitude response approximately equal to D in the passband.

Therefore, the overall structure will result in a fixed integer delay of I samples and a variable non integer delay of l/D samples. The previous structure can be realized with a network of polyphase filters as illustrated in Fig. 4.5.

The polyphase filter is defined as:

$$p_\rho(n) = h_{LP}(nD + \rho - 1) \quad 1 \leq \rho \leq D \tag{4.10}$$

The above expression already takes into consideration the sampling rate increase by D, the low-pass filtering and the delay of l samples by the selection of the corresponding polyphase filter $p_\rho(n)$. Finally, the downsampling by D is accomplished

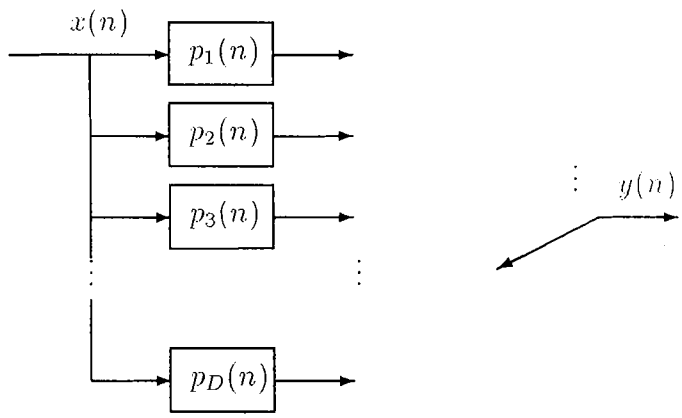

Fig. 4.5: Polyphase network implementation of a fractional sample delay network.

by moving the arm of the commutator (it is back to its original position for every M samples).

The polyphase filters $p_\rho(n)$ can directly implement the operations of sampling rate increase and low-pass filtering. For each value of the delay l/D, a corresponding ρ-th polyphase filter branch is used. With a delay I for the low-pass filter, the expression for the new pitch predictor with a fractional delay of $M + l/D$ is:

$$P(z) = 1 - \beta \sum_{n=0}^{b-1} p_\rho(n) z^{-(M-I+n)} \tag{4.11}$$

where b is the number of coefficients of the polyphase filter and β is the pitch predictor coefficient.

The polyphase filters used in the simulations are a $\sin(x)/x$ function weighted with a Hamming window. For each value D, the length of the filter was chosen such that the delay I at the lower sampling frequency is equal to 16 (refer to Eq. (4.9)).

With the use of 38400 pitch subframes of 3.125 ms each, a pitch delay distribution is generated shown in Fig. 4.6 as well as a pitch filter gain versus pitch lags shown in Fig. 4.7. From these two figures, we conclude that the pitch lags occur more oftenly and with higher pitch gain in the 71–100 range than any other range.

A nonuniform distribution of non-integer delays can then be set up to construct

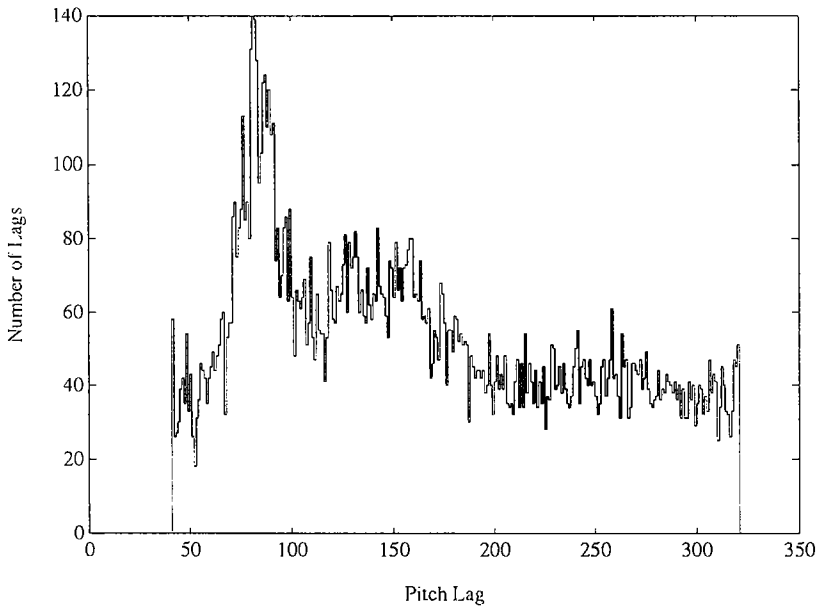

Fig. 4.6: Distribution of pitch delays.

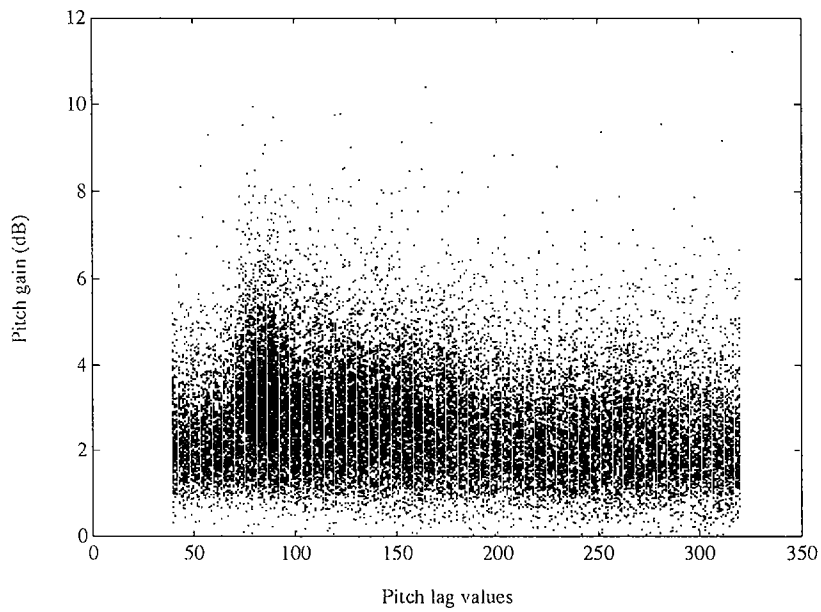

Fig. 4.7: Pitch filter gain vs pitch delays.

the pitch delay codebook. Two configurations are set up accordingly with two levels of complexity giving higher resolution to frequently occurring pitch delays as shown in Table 4.4. The highest resolution is given to pitch lags in the range of 71–100 while

Coder	Pitch Range	Resolution
I	40–70	1/3
	71–100	1/4
	101–140	1/3
	141–320	1
II	40–70	1/4
	71–100	1/6
	101–197	1/4
	198–302	1/3
	302–320	1

Table 4.4: Configuration of the pitch delay codebook.

the lowest resolution is given to the end of the lag range. In order to test these two configurations, eight speech files (4 female and 4 male speakers), a formant frame of 15.62 ms, a pitch subframe of 3.125 ms, a Gaussian codebook with 1024 codewords were used. The quantization of both the gain and LPC parameters was turned off.

As shown in Table 4.5, non-integer delays improved the quality of the reconstructed speech by 0.2–1.1 dB in terms of segmental SNR when compared to the performance of the one-tap pitch predictor and a substantial increase in perceived quality was observed. Note also that the SegSNR figures of the 10 bits fractional pitch predictor and the three-tap integer pitch predictor were very similar, yet the first used 10 bits codebook while the second used an 11 bits codebook. Therefore, the use of fractional pitch reduces the bit rate while maintaining a comparable level of quality to a three tap pitch predictor.

Mode	Pitch Predictor		SegSNR (dB)	
	Order	Delays(bits)	female	male
250:50	1	non-integer (9)	13.90	13.42
	1	non-integer (10)	14.22	14.17
	1	integer (8)	13.71	12.91
	3	integer (11)	14.25	13.97
320:40	1	non-integer (9)	14.94	14.36
	1	non-integer (10)	15.02	14.50
	1	integer (8)	14.66	14.13
	3	integer (11)	15.13	15.00

Table 4.5: Effect of high resolution pitch filtering.

Chapter 5

Improved Noise Weighting

5.1 Introduction

A commonly used error criterion in speech coding is the mean-squared error, which provides satisfactory performance with an appealing simplicity. However, at lower bit rates it becomes cumbersome to match closely the original speech waveform, and the mean-squared error between the original and the reconstructed speech looses significance as illustrated in Fig. 5.1 where the noise level is almost flat.

A model of auditory perception must be incorporated with the speech coder's error criterion to better control the noise bursts. By doing so, the synthetic speech can follow the natural speech in those aspects that are perceptually important.

In discussing auditory perception, we are concerned with what sounds are perceptible and how different components of those sounds affect and interfere with one another. Hearing perceptibility depends on its intensity and spectrum; the ear is capable of hearing sounds over a wide dynamic range (from about 16 Hz to 18 kHz); therefore, depending on the sound location in the speech spectrum it will require either more or less energy to be heard as shown in Fig. 5.2 [21]. Again, sounds in the higher spectrum (5 kHz and up) are less perceptible, this justifies the use of a reduced number of bits to code higher band speech as demonstrated previously with

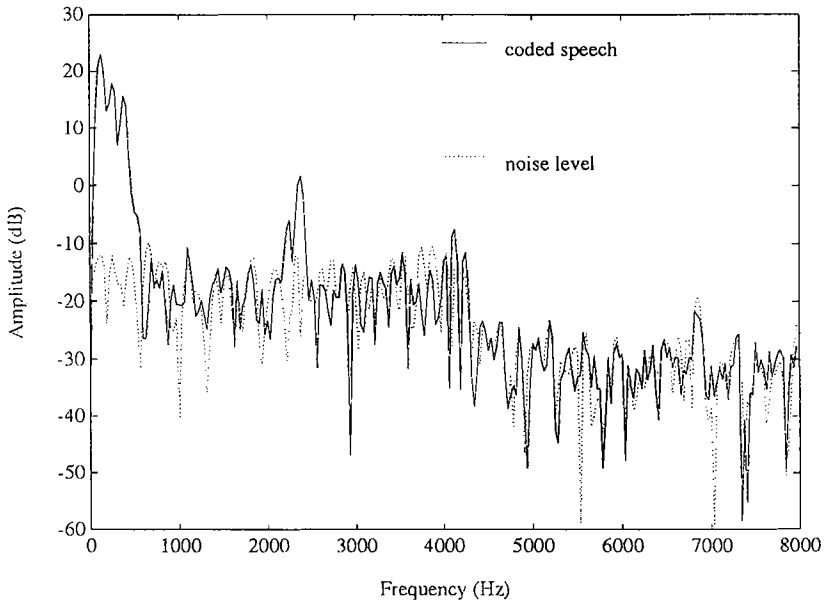

Fig. 5.1: Comparison of noise level with respect to coded speech.

the use of line spectral frequencies.

The physiological behavior of the ear in response to simple tones is relatively straightforward, but most sounds are time varying and give many spectral components. The hearing system has only a limited capability to detect small errors in the frequency bands where the speech signal has high energy (as in the case of formant regions). Consequently, the perception of one sound could be obscured by the presence of another: this phenomenon, better known as *masking*, takes place when one sound raises the hearing threshold of another. Different techniques have been developed to take advantage of the masking theory in speech coding. Quantization noise that arises in speech coders can, therefore, be covered by high speech energy in formant regions.

In this chapter, we discuss the advantages of this phenomenon in speech coding. Section 5.2 reviews the use of a simple noise weighting filter $W(z)$. Section 5.3 studies the effects of using a shaping filter in combination with the excitation codebook. Section 5.4 studies the notion of perceptual noise weighting with a modified filter

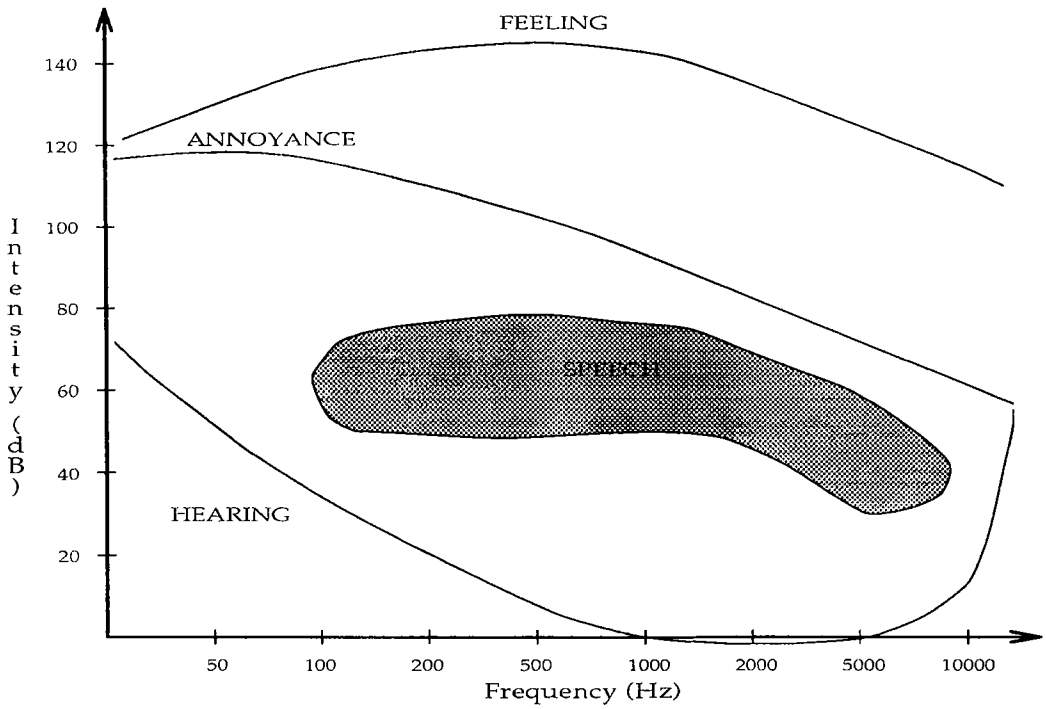

Fig. 5.2: Areas of speech perception inside the limits of overall perception.

$W'(z)$. The last section evaluates the performances of all three approaches.

5.2 Simple noise weighting

To make use of the masking effect, the quantization noise has to be distributed in relation of the speech power over different frequency bands. This task is accomplished by minimizing a weighted error with the noise shaping filter $W(z)$ as described in Chapter 2 (Section 2.2) is defined as

$$\begin{aligned} W(z) &= \frac{A(z)}{A(z/\gamma)} \\ &= \frac{1-\sum_{i=1}^{N_p} a_i z^{-i}}{1-\sum_{i=1}^{N_p} a_i \gamma^i z^{-i}} \end{aligned} \tag{5.1}$$

where $A(z)$ is the short-time predictor as defined in Eq. (2.2) . The value of γ is determined by the degree desired to de-emphasize the formant regions in the error spectrum. Decreasing the value of γ moves the poles of the filter $\frac{1}{A(z/\gamma)}$ inward and therefore increases the bandwidth of the poles of $W(z)$. The increase in bandwidth $\Delta\omega$ is given by the relation [13]

$$\Delta\omega \;=\; -\frac{f_s}{\pi}\ln\gamma \;\; \text{Hz} \tag{5.2}$$

where f_s is the sampling frequency in hertz. The optimum value of γ, determined by listening tests, is set to 0.75. This corresponds to an increase in bandwidth of about 1465 Hz.

Referring back to Section 2.2.3, the final unweighted error signal $E(z)$ is expressed in terms of the weighted error signal $E_w(z)$ where

$$E(z) \;=\; S(z) - \hat{S}(z) \;=\; \frac{E_w(z)}{W(z)} \tag{5.3}$$

Consequently, the resulting noise level has the spectral shape of $W^{-1}(z)$ and will therefore be concentrated in the formant peaks and attenuated in the formant valleys.

Note the weighting filter $W(z)$ is responsible for changes in SegSNR performance measures. In certain frequency regions, the SegSNR will improve but this will come at the expense of a significantly reduced SegSNR in other regions. The overall SegSNR will therefore drop with the use of the weighting filter but the perceptual quality of the speech will improve.

Figures 5.3 and 5.4 show the effect of using noise weighting filter in the reconstruction process of the input speech.

As shown in the two figures, the quantization noise level curve is no longer flat and is better distributed over the frequency spectrum. When $\gamma = 0.51$, noise becomes more or less audible in the upper frequency band as shown in Fig. 5.4; while for the optimal γ, the noise level is well covered in all the formant regions.

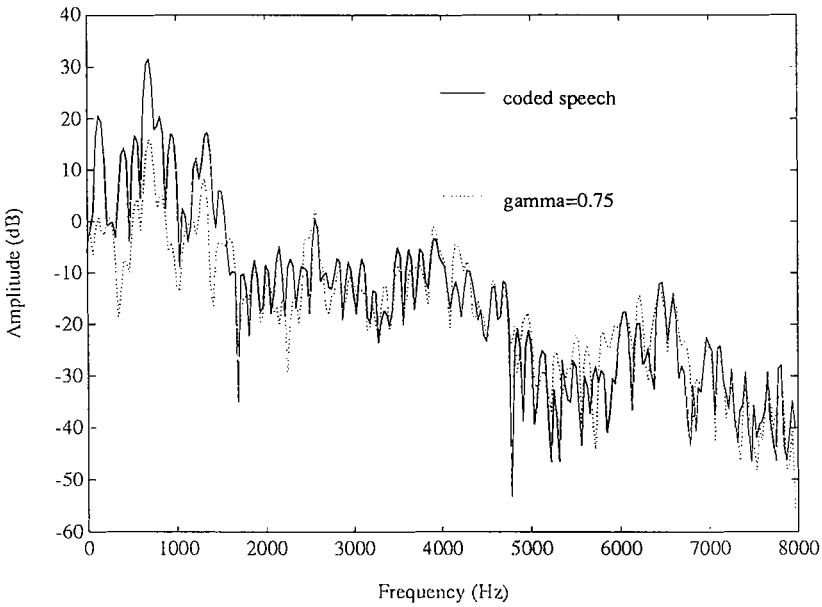

Fig. 5.3: Noise weighting with $\gamma = 0.75$.

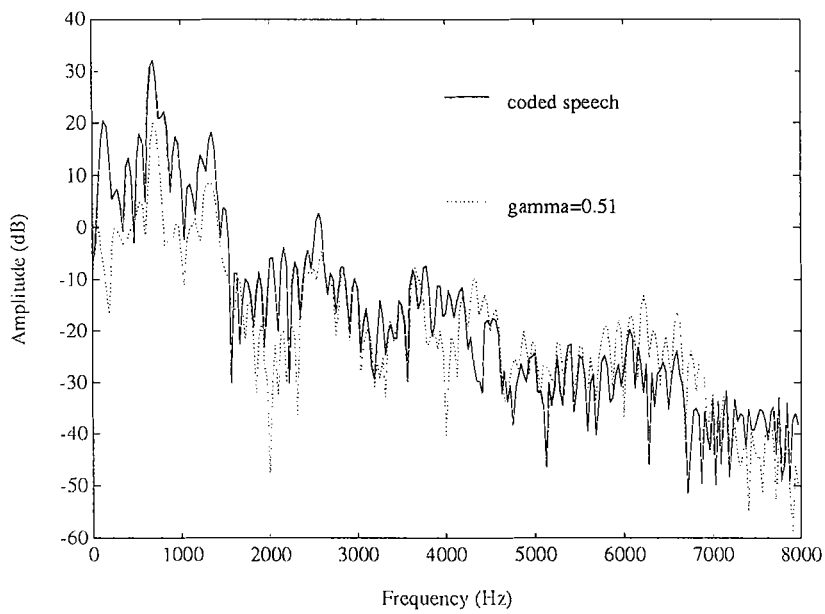

Fig. 5.4: Noise weighting with $\gamma = 0.51$.

5.3 Codebook shaping filter

A new approach for codebook design is studied in this section where an excitation codebook is combined with a shaping matrix $\mathbf{F_c}$ to form an enhanced codebook structure. This method was investigated by the Communication Research Center Group of the University of Sherbrooke [14] and was used in combination with a 20 bits codebook and an improved codebook search technique to yield a high quality coder at low bit rates. In this research, we only consider the codebook shaping part where an excitation vector is given by

$$\hat{\mathbf{r}}'_{\mathbf{k}} = \mathbf{F_c}\hat{\mathbf{r}}_{\mathbf{k}} \tag{5.4}$$

The shaping matrix $\mathbf{F_c}$ is dynamically changed to control the statistical properties of the codebook in time and in frequency. The corresponding shaping filter $F_c(z)$ used is actually a function of the LPC model $A(z)$. Its main role is to shape the excitation codewords in the frequency domain so that their energies are concentrated in the important frequency bands.

Fig. 5.5 shows the overall structure of the frequency shaped excitation codebook.

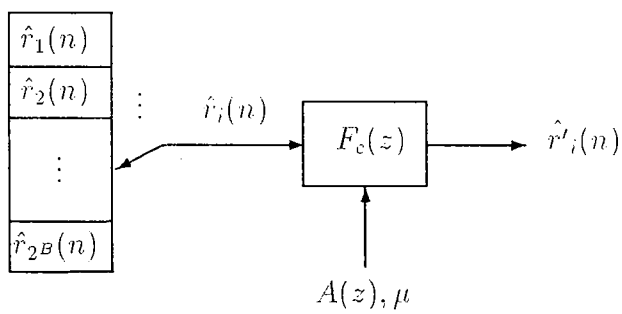

Fig. 5.5: Frequency shaped excitation codebook.

The filter $F_c(z)$ is defined in the following

$$F_c(z) = (1 - \mu z^{-1})\frac{A(z/\gamma_1)}{A(z/\gamma_2)} \tag{5.5}$$

where $A(z)$ is the LPC inverse filter, γ_1 and γ_2 are constants, and μ is a factor that controls the spectral tilt and varies in every excitation frame.

This filter is actually a combination of a first-order preemphasis filter of the form $1 - \mu z^{-1}$ and a weighting filter $\frac{A(z/\gamma_1)}{A(z/\gamma_2)}$ in cascade together. The value of μ has to be optimized; a differencer would use $\mu = 1$, but the optimum preemphasis filter which maximizes the output spectral flatness measure will have $\mu = \frac{r_{\hat{r}}(1)}{r_{\hat{r}}(0)}$, where $r_{\hat{r}}(n)$ represents the autocorrelation sequence for the input speech data sequence $\hat{r}(n)$. To show that, we consider $\hat{r}'(n)$ as the time sequence of the preemphasis filter's output, then we have

$$r_{\hat{r}'}(0) = (1 + \mu^2) r_{\hat{r}}(0) - \mu r_{\hat{r}}(1) \tag{5.6}$$

The two autocorellations $r_{\hat{r}'}(i)$ and $r_{\hat{r}}(i)$ can be determined with the following:

$$\begin{aligned} r_{\hat{r}'}(i) &= \sum_{j=i}^{N} \hat{r}'(j)\hat{r}'(j-i) \\ r_{\hat{r}}(i) &= \sum_{j=i}^{N} \hat{r}(j)\hat{r}(j-i) \end{aligned} \tag{5.7}$$

where N is the number of samples per speech frame.

The optimal spectral flatness will occur at the minimum value of $r_{\hat{r}'}(0)$ and this value according to Eq. (5.6) is $\mu = \frac{r_{\hat{r}}(1)}{r_{\hat{r}}(0)}$. For unvoiced sounds, this fraction is relatively small and the effect of the preemphasis filter becomes negligible; while for voiced sounds where $r_{\hat{r}}(1)$ is very close to $r_{\hat{r}}(0)$, the preemphasis greatly affects the spectral flatness and the preemphasis filter behaves almost as a differencer.

We experimented with different values of γ_1 and γ_2 and we found that $\gamma_1 = 0.80$ and $\gamma_2 = 0.95$ gave the best results. Fig. 5.6 shows the spectrum of the reconstructed frame of speech compared with its noise level spectrum using the above codebook shaping technique and the simple noise weighting scheme.

The preemphasis filter coefficient can be severely quantized, since any value of μ between zero and twice the value $\frac{r_{\hat{r}}(1)}{r_{\hat{r}}(0)}$ will enhance the spectral flatness [6]. The weighting filter, varied from frame to frame will directly be responsible for dampening

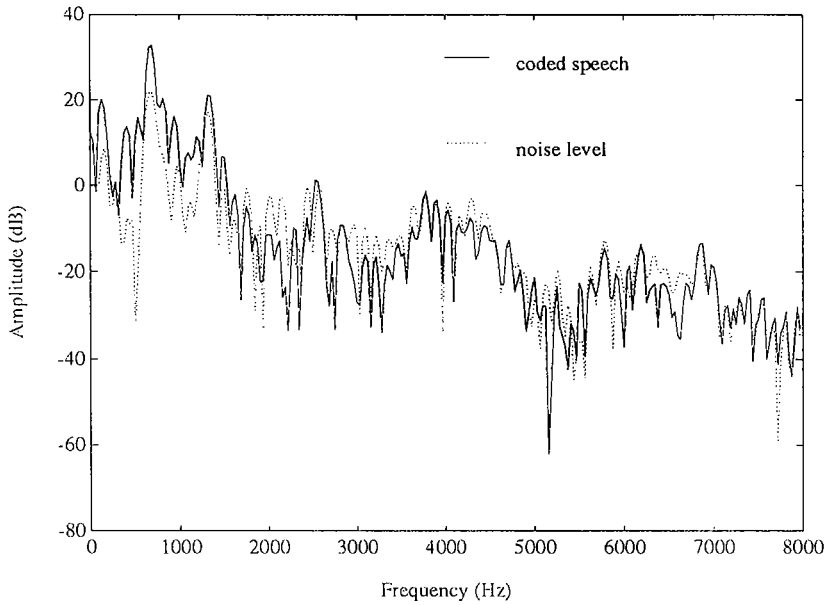

Fig. 5.6: Noise level using frequency codebook shaping.

frequencies most annoying to the ear, the resulting performance figures are shown in Table 5.1 for the 320:40 update mode.

Weighting	Shaping	SegSNR (dB)	
		female	male
off	off	15.33	14.84
off	on	13.72	13.27
on	on	9.54	8.82

Table 5.1: SegSNR figures using a shaping filter and simple noise weighting.

The performance figures show a drop in SegSNR when the weighting filter is used even though the perceptual quality of the reconstructed speech is significantly improved, this is due to the fact that the performance measure used here, the SegSNR, is an objective measure and it does not take into consideration the perceptual aspect

of the reconstructed speech. this is furthermore discussed in Section 4.4.

5.4 Perceptual noise weighting

A further improvement that can be added to the CELP coder is the use of an enhanced noised weighting scheme introduced by Shoham and Ordentlich [20]. This perceptual noise weighting technique solves the problem of high frequency distortion. This method is now extensively used because of the increased perceptual quality it adds to the reconstructed speech.

The major disadvantage of a normal noise weighting filter $W(z)$ is inadequate balancing of low and high frequency coding. This asymmetry is mainly due to the interdependency of both tilt and formant parameters. Modeling one accurately requires sacrifices in modeling the other. This difficulty is more acute in wideband speech since there is no appreciable spectral tilt, and this problem becomes more significant at lower bit rates where the noise shaping technique must be maximized to overcome the additional quantization noise.

The tilt is controlled by the difference $1-\gamma$, and we are faced with two difficulties while trying to control this parameter:

- The tilt is global over all the speech spectrum and it is impossible to emphasize it separately for high frequencies.
- The tilt affects the shape of the formants of $W(z)$, for instance a pronounced tilt results in higher and wider formants which entails an increased level of noise at low frequencies and in between formants.

This enhanced noise weighting approach introduces a decoupling factor that results in an independent control of the tilt with respect to the formants. An additional filter $P(z)$ is used and is responsible for the tilt only.

We first started by using a adaptive three pole filter $P_3(z)$ with the weighting filter defined as:

$$W_3'(z) = W(z)P_3(z) = \frac{H(\gamma z)}{H(z)} \frac{1}{1 + \sum_{k=1}^{3} p_k \delta^k z^{-k}} \tag{5.8}$$

where the coefficients p_k are determined by an LPC analysis on the first four correlation coefficients of the inverse filter $A(z)$ and δ is a spectral tilt controlling parameter and is set to 0.5.

Fig. 5.7 shows the effect of using the additional three pole filter. The solid curve represents a spectrum of the conventional inverse filter $W^{-1}(z)$ while the dashed curve displays the spectrum of the enhanced weighting filter $W^{-1}(z)P_3^{-1}(z)$. The

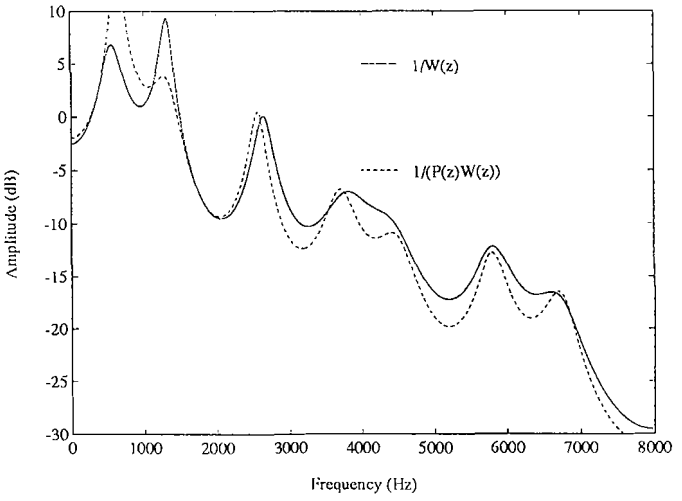

Fig. 5.7: Performance of the three pole weighting filter.

adaptive three pole filter comes in very handy in boosting high frequencies near the half sampling rate due to the presence of a real pole but this is obtained at the expense of a broad band increase in the level of distortion at lower frequencies as shown in Fig. 5.7.

Because of this limitation, we then switched to an adaptive two pole filter $P_2(z)$

with:

$$W_2'(z) = W(z)P_2(z) = \frac{H(\gamma z)}{H(z)} \frac{1}{1 + \sum_{k=1}^{2} p_k \delta^k z^{-k}} \tag{5.9}$$

where again the coefficients p_k are determined by an LPC analysis on the first three correlation coefficients of the inverse filter $A(z)$ and δ is set to 0.7.

By getting rid of the real pole, we were able to obtain lower level of distortion at lower frequencies while maintaining an acceptable level of high frequency noise in the upper band as shown in Fig. 5.8.

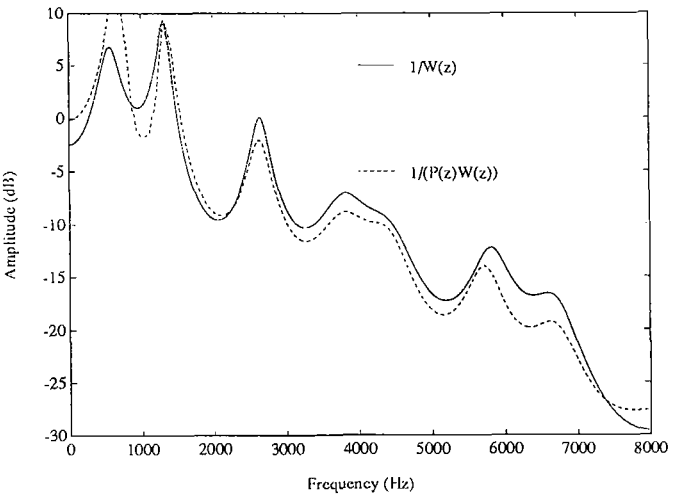

Fig. 5.8: Performance of the two pole weighting filter.

The addition of perceptual noise weighting to the CELP coder did not improve the segmental SNR figures but the perceptual quality of the coded speech was enhanced with no additional bit requirements.

5.5 Performance measures

Throughout the previous chapters, we have used the SegSNR measure to evaluate the degree of distortion in our speech coder. This measure was very helpful in determining the quality of the reconstructed speech. Nevertheless, it failed to give us a measure

of the perceptual quality of this same speech, and in this chapter using the SegSNR measure becomes meaningless in evaluating the performance of the different weighting filters used.

In this section, we use a new distortion measure that was introduced by De and Kabal [5] where both the original and coded speech are transformed from the time domain to a perceptual domain using a cochlear model. With this cochlear model, the basic features of the hearing process were simulated in order to give a good perceptual evaluation of the coded speech when compared to its original version.

Three basic features are studied and simulated: the outer ear, the middle ear and the inner ear (cochlea). In the outer ear, the eardrum first senses speech pressures variations, these variations are then transformed into mechanical vibrations by the middle ear. Finally, the cochlea turns these mechanical vibrations into electrical excitations. The last feature remains the most difficult to simulate. In De's work, the cochlea role is thoroughly investigated.

In this process, the electrical activity generated in the cochlea is due to the presence of nerve cells, these cells *fire* in response to the mechanical vibrations of the middle ear. These neurons activity patterns, that contain information about the pitch and formants, are presented in the perceptual domain where firing probabilities values can be obtained. Finally, the probabilities generated for both the original and coded speech can be compared, in an information-theoretic sense, to obtain the desired distortion measure.

Let $p_{1|k}$ and $p_{2|k} = 1 - p_{1|k}$ be the firing and non-firing probabilities of the original speech at a certain time t in the k-th neural channel. Similarly, $q_{1|k}$ and $q_{2|k} = 1 - q_{1|k}$ are the probabilities used for the coded speech. The distortion measure used to discriminate between the original and synthetic speech is defined in the following:

$$D_\alpha(P;Q|k) = \sum_{j=1}^{2} p_{j|k} \log(\frac{p_{j|k}}{q_{j|k}}) \qquad \text{for } \alpha = 1 \tag{5.10}$$

$$= \frac{1}{1-\alpha} \log(\sum_{j=1}^{2} \frac{p_{j|k}^{\alpha}}{q_{j|k}^{\alpha-1}}) \qquad \text{for } \alpha \neq 1 \tag{5.11}$$

This measure is known as *cochlear directed divergence measure* .

The measurement, used with $\alpha = 1$, was applied on a total of eight coded speech files each using a different noise weighting scheme. All quantizers were turned on and an update mode of 320:40 was used. The following eight configuration were used:

1. No weighting used

2. Conventional noise weighting used with

$$W(z) = \frac{A(z)}{A(\gamma z)} \tag{5.12}$$

3. Conventional noise weighting used and codebook shaping used with a shaping filter

$$F(z) = (1 - \mu z^{-1}) \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \tag{5.13}$$

4. Perceptual noise weighting used with three poles where

$$W_3'(z) = W(z) P_3(z) \tag{5.14}$$

5. Perceptual noise weighting used with two poles where

$$W_2'(z) = W(z) P_2(z) \tag{5.15}$$

6. Perceptual noise weighting used with three poles where and codebook shaping used with a shaping filter

$$F_3(z) = (1 - \mu z^{-1}) \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \tag{5.16}$$

7. Perceptual noise weighting used with two poles and codebook shaping.

The results are shown in Table 5.2 where both the perceptual distortion measure and SegSNR figures are indicated.

Note that the higher is the value of the cochlear directed divergence measure (more added new information) the worst is the coded speech quality when compared to the original one. The results shown in Table 5.2 indicate that:

Configuration	File	SegSNR (dB)	Dir. Div.
1	nowem2-01.aud	16.06	2.73
2	wem2-01.aud	12.89	2.68
3	wealm2-01.aud	11.00	2.65
4	we3m2-01.aud	11.90	2.54
5	we2m2-01.aud	11.59	2.52
6	we3alm2-01.aud	9.69	2.61
7	we2alm2-01.aud	9.95	2.62

Table 5.2: Distortion measures for different noise weighting schemes.

- The use of a simple noise weighting filter $W(z)$ improves the perceptual quality of the reconstructed speech while causing a deterioration in the SegSNR measure.

- The use of a codebook shaping filter $F_c(z)$ in combination with $W(z)$ will yield even better results in the perceptual domain.

- The two pole adaptive weighting filter $W_2'(z)$ out-performs the three pole adaptive weighting filter $W_3'(z)$ as expected.

- The best candidate for an improved noise weighting scheme is the adaptive two pole weighting filter $W_2'(z)$.

Chapter 6

Enhanced Wideband CELP

6.1 Introduction

The basic structures for the full-band coder were first introduced back in Chapter 2 and the initial performance analysis was very promising. However, the bits assigned for the transmission of the speech coding parameters remained relatively high. Modifications were needed to lower the bit rate while keeping a high standard of quality in the reconstructed speech. These modifications that were studied in previous chapters took place in the main building blocks of the CELP coder, and they included:

- The application of split vector quantization on the LPC parameters instead of scalar quantization.
- Interpolation of LPC parameters to enable updating for every subframe.
- The use of a fractional pitch predictor for the long term analysis stage instead of multi-tap pitch predictor.
- The selection of an adaptive two pole weighting filter as the best candidate for efficient noise weighting.

The main goal of this research was to build a high quality low bit rate wideband CELP coder that could outperform most of its predecessors and more specifically the split-band CELP coder of Roy and Kabal [26]. In this chapter, the first section describes the final configuration of the full-band coder; a detailed description of parameter selection and quantization is given and experiments were carried out to determine the overall performance of the coder. The second section studies the split-band CELP structure introduced by Roy; and finally the last section compares the performances of the split- and full-band coder.

6.2 Final configuration of the full-band CELP

The modified version of the CELP coder is shown in Fig. 6.1 where the additional improvements are depicted.

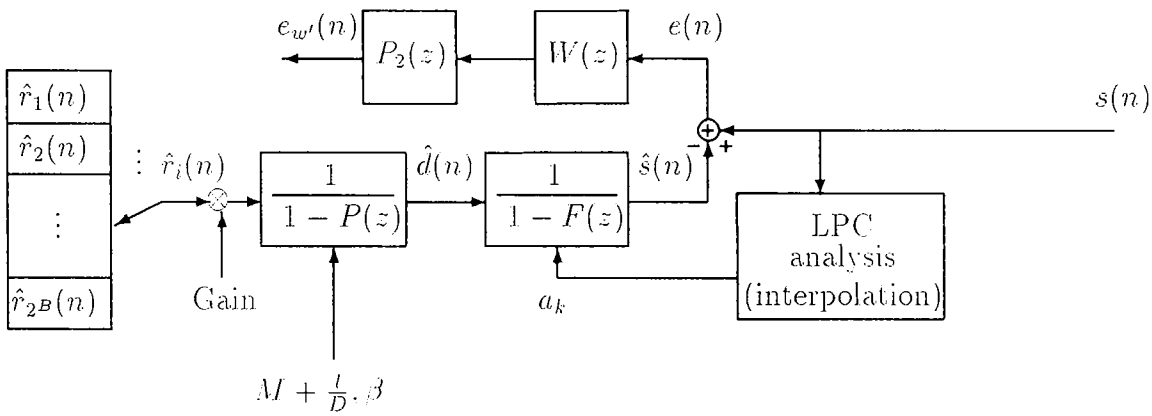

Fig. 6.1: Enhanced CELP coder.

6.2.1 Parameter selection and quantization

This section describes the design and coding of each parameter used in our CELP coder. As described earlier, the simulations were carried out with speech signals sampled at 16 kHz and bandlimited to 7500 Hz.

- **Frame and subframe sizes**

  The frame and subframe sizes control the update rate of the coder's parameters. Both pitch (fractional lag and coefficient) and codebook (gain and index) parameters are updated for every subframe, while only the LPC parameters a_k are updated for every frame. Three update modes were investigated. With a sampling frequency of 16 kHz, the first (320:40) has a formant frame of 320 samples (50 Hz) and a pitch subframe of 40 samples (400 Hz). The second (250:50) uses a frame of 250 samples (64 Hz) and a subframe of 50 samples (320 Hz). Finally, the third mode (160:40) has a frame of 160 samples and a subframe of 40 samples. this mode is used to test transparent speech quality but at higher bit rates ($\sim$ 24 kbits/s).

- **LPC coefficient coding**

  A 16-th order formant filter is used, although studies were made on a 20-th order filter but quantization proved to be too costly. The LPCs are first transformed into LSFs, split into three subgroups with the 4-4-8 mode. Vector quantization is then applied to each subgroup with the weighted Euclidean distance measure as the selection criterion. The number of quantization bits is set to 33 bits/frame. Interpolation of the LSFs is used to provide interpolated LPC coefficients in every subframe.

- **Pitch coefficient and lag coding**

  One pitch tap is used and coded with a 5 bit non-uniform scalar quantizer. The lag value ranges from 2.5 ms to 20 ms or from 40 to 320 samples. The lag selection process uses fractional pitch prediction where both an integer lag value and a fraction are determined. The configuration uses a total of 10 bits to store these values into a codebook.

- **Gain coding and codebook design**

  A 4 bit differential quantizer with a leaky predictor is used to code the differences in successive subframes magnitudes. An extra bit codes the sign. The codebook consists of normalized *iid* (independent identically distributed) Gaussian sequences, the size of the codebook is varied between 128 and 1024 codewords.

6.2.2 Performance

With all the previous parameter settings described, two operating rate were established for the full-band CELP coder as listed in Table 6.1 and Table 6.2.

Parameter	Bits	Update rate(Hz)	Bits/sec
LPC coefficients	33	50	1650
β	5	400	2000
gain G	5	400	2000
lag M	10	400	4000
codebook	10	400	4000
		Total	13650

Table 6.1: Full-band CELP coder operating rate for 320:40 mode.

The resulting segmental SNR figures for four speech files (2 males and 2 females) are shown in Table 6.3. All the parameters are quantized accordingly. The target operating rate for this research was set to 12 kbits/sec, that makes our second configuration appropriate.

Note that the use of smaller pitch frames induces faster subframe rates, this consequently results in better speech quality as observed between the two modes (320:40) and (250:50). Nevertheless, the performance of the second configuration (250:50) mode tends to improve much faster than the first when larger codebooks are

Parameter	Bits	Update rate (Hz)	Bits/sec
LPC coefficients	33	64	2112
β	5	320	1600
gain G	5	320	1600
lag M	10	320	3200
codebook	10	320	3200
		Total	11712

Table 6.2: Full-band CELP coder operating rate for 250:50 mode.

Codebook size	320:40 mode	250:50 mode
128	13.43	11.64
256	13.78	12.05
512	14.02	12.79
1024	14.37	13.09

Table 6.3: Full-band SegSNR performance.

used as shown in Table 6.3. The gap of 1.8 dB present at a codebook size of 128 is now reduced to a 1.3 dB difference between the two operating rate.

6.3 Split-band CELP

The split-band configuration shown in Fig. 6.2, introduced by Roy and Kabal [26], is very similar to the full-band case, the difference is that the codebook is now split into a lower and upper band Gaussian codebooks with two distinct gains (G_L and G_H) and two pitch synthesis filters ($G_L(z)$ and $G_H(z)$). The resulting pitch excitation signal $\hat{d}(n)$ becomes the addition of the lower band part $\hat{d}_L(n)$ and the higher band part $\hat{d}_H(n)$, and we have:

$$\begin{aligned} \hat{d}(n) &= \hat{d}(n)_L + \hat{d}_H(n) \\ &= G_L r_{Li}(n) + \beta_L \hat{d}_L(n - M_L) + G_H r_{Hi}(n) + \beta_H \hat{d}_H(n - M_H) \end{aligned} \quad (6.1)$$

An additional step is used also where the pitch parameters (gain, lag and coefficients) are re-optimized. The optimization procedures used to minimize the energy ϵ are similar to the ones described in Eq. (2.27) and (2.33) except that a joint optimization of both pitch and codebook parameters is now performed with an increased number of parameters (pitch coefficient β, pitch lag M and codebook gain G) for the additional band.

- The new codebook weighted error $e_w(n)$ for every codeword with index i is:

$$e_w(n) = s_{new}(n) - G_L \tilde{r}^i_{L[0,N)}(n) - \beta_L \tilde{d}^i_{L[0,N)}(n) - G_H \tilde{r}^i_{H[0,N)}(n) - \beta_H \tilde{d}^i_{H[0,N)}(n) \quad (6.2)$$

where s_{new} is the new reference signal for the synthesis stage and

$$\begin{aligned} \tilde{r}^i_{L[0,N)}(n) &= \sum_{n=0}^{N-1} \hat{r}_{Li}(k) h_\gamma(n - k) \\ \tilde{r}^i_{H[0,N)}(n) &= \sum_{n=0}^{N-1} \hat{r}_{Hi}(k) h_\gamma(n - k) \\ \tilde{d}^i_{L[0,N)}(n) &= \sum_{n=0}^{N-1} \hat{d}_{Li}(k - M_L) h_\gamma(n - k) \end{aligned} \quad (6.3)$$

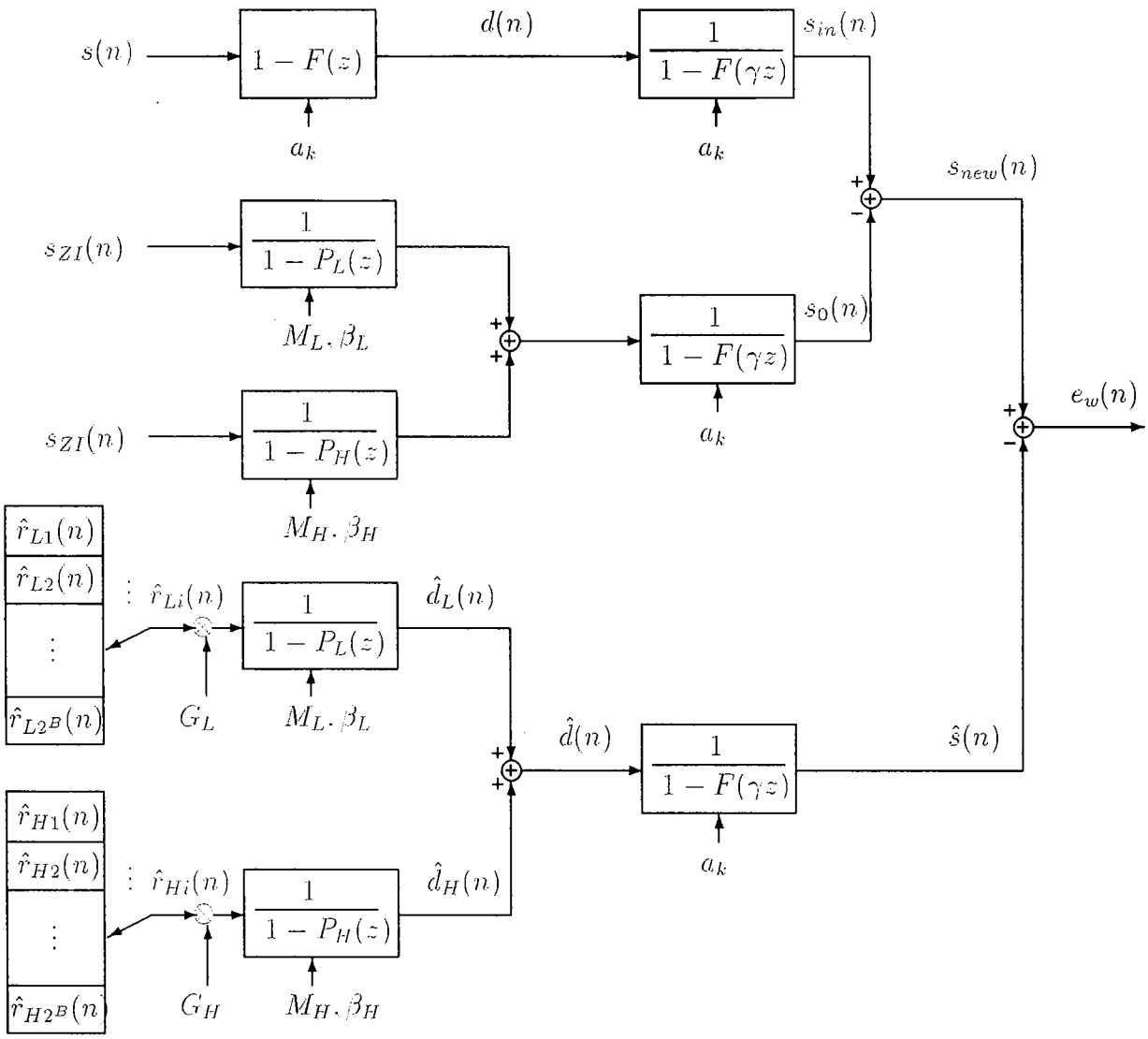

Fig. 6.2: CELP system configuration for the split-band coder.

$$\hat{d}^{i}_{H[0,N)}(n) = \sum_{n=0}^{N-1} \dot{d}_{Hi}(k - M_H) h_\gamma(n-k)$$

The energy of the weighted error signal in the pitch subframe is

$$\epsilon_c = \sum_{n=0}^{N-1} e_w^2(n) \tag{6.4}$$

The weighted error is minimized in a mean square sense by differentiating Eq. (6.4) with respect to the gains and pitch coefficients of both the low and high band signals. The resulting expressions are then set to zero. This yields a set of linear equations that are solved using the Cholesky algorithm.

In the pitch synthesis filters, different pitch coefficients β_L and β_H are used while the pitch lags M_L and M_H are equal. According to Roy [25], simulations with separate lag values showed that most of the time, both lag values were either close to each other, or close to a common multiple. To reduce the large amount of computations involved in searching for the optimal codewords in both bands, the search was limited to the lower band and the optimal index found is shared by the two equal size codebooks. This sub-optimal method significantly reduces the bit rate with little effect on the reconstructed speech [25].

6.3.1 Parameter selection and quantization

This section deals with parameter selection and quantization for the split-band coder structure. Again, the speech signal used are the ones described in Appendix A. The parameter update rates and gain estimate used in this part are exactly the same as one studied in the full-band configuration.

- **LPC coefficients coding**

  The LPC coefficients a_k are first coded using Line Spectral Frequencies. Again, 16 coefficients are used to model the spectral envelope. Nevertheless, the difference here is that instead of using split vector quantization, we go back to

scalar quantization. The quantization scheme, better known the non-uniform differential scalar quantization [30], uses a total of 48 bits to code the 16 LSF coefficients. These 48 bits are unequally allocated across the frequency band, more bits are assigned to lower LSFs to emphasize the perceptual importance of lower frequencies.

- **Pitch coefficient coding**

  In this split-band configuration, two pitch taps have to be used for the higher and lower bands. The coding is performed with non-uniform scalar quantizers. The lower pitch tap β_L is coded with 5 bits while the higher one β_H uses 3 bits considering the fact that pitch information is crucial at lower frequencies. A single lag value is used for both bands and is coded with 7 bits.

- **Codebook design**

  Separate excitations are needed in each band. The two codebooks are designed by bandlimiting normalized *iid* Gaussian sequences. Experimental results showed that the best configuration consists of a full-pass codebook for the low-band and high-pass codebook for the high-band. This scheme prevents the high-band excitation from affecting the low-band perceptually important regenerated speech. Codebook sizes are set to 1024 codeword, but only one index is used for both codebooks as described earlier. Gain coding is accomplished with 6 bits assigned to G_L and 4 bits assigned to G_H.

6.3.2 Performance

Two operating rate were established using both the (250:50) and (320:40) update modes. As shown in Table 6.4 and 6.5. the resulting operating rates are 16 kbits/sec and 14 kbits/sec for the (320:40) mode and (250:50) mode respectively.

The resulting performance figures are shown in Table 6.6. The simulations were performed on the same four speech files used in the full-band configuration and with

Parameter	Bits	Update rate(Hz)	Bits/sec
LPC coefficients	48	50	2400
β_L	5	400	2000
β_H	3	400	1200
gain G_L	6	400	2400
gain G_H	4	400	1600
lag M	7	400	2800
codebook	9	400	3600
		Total	16000

Table 6.4: Split-band CELP coder operating rate for 320:40 mode.

Parameter	Bits	Update rate(Hz)	Bits/sec
LPC coefficients	48	64	3072
β_L	5	320	1600
β_H	3	320	960
gain G_L	6	320	1920
gain G_H	4	320	1280
lag M	7	320	2240
codebook	9	320	2880
		Total	13952

Table 6.5: Split-band CELP coder operating rate for 250:50 mode.

all parameters quantized.

Codebook size	320:40 mode	250:50 mode
128	12.89	11.17
256	13.12	11.85
512	13.58	12.24
1024	13.96	12.69

Table 6.6: Split-band SegSNR performance.

6.4 Comparison of both the split- and full-band CELP

The main goal of this research was to improve on the wideband CELP coder build by Roy [25]. In the last section, this split-band coder was investigated to establish the differences that were brought with our enhanced full-band CELP coder.

The idea of a split-band structure first came with the introduction of the G.722 64 kbits/s 7 kHz audio codec. Roy's coder made use of this concept, but instead of using an ADPCM structure, the work was accomplished with a CELP coder. Some of the advantages offered by this coder were:

- Higher flexibility in constraining the quantization noise in the band where it is produced.
- Better representation of individual subbands in terms of weighting the ones that are the most significant from a perceptual point of view.

Nevertheless, some disadvantages were also present, mainly:

- A higher system complexity brought by the increase in the number of bands. This complexity would translate in a heavier computational load where parameters for both the higher and lower band are to be calculated. This would make a real-time implementation of this coder difficult.

- A higher operating rate caused by a higher number of bits required to code the additional pitch and gain parameters.

Throughout this research, we tried to preserve some of the advantages of the split-band coder while eliminating the disadvantages. With the introduction of the two pole perceptual noise weighting filter, the flexibility over the control of quantization noise was maintained. Also, the use of the multi-stage split VQ technique for LSFs reinforced the control of bit allocation for different frequency bands. This method also helped in reducing the overall operating bit rate. In terms of overall performance, the results obtained with our coder were significantly better than those obtained with Roy's coder:

- The SegSNR figures improved by 0.4 dB over the split-band coder.

- The operating rate was reduced from 16 kbits/s to 11.7 kbits/s.

- The perceptual quality of the speech improved by 0.08 on the directed divergence measure described in Section 5.5.

Chapter 7

Conclusion

The purpose of this thesis was to examine and improve the coding of wideband speech by using analysis-by-synthesis coders. To accomplish this task, different existing wideband coders were investigated as well as specific narrowband coding schemes that could be adapted to a wideband environment.

In particular, known spectral coding techniques using vector quantization were extended to deal with a higher number of LPC coefficients, other methods dealing with higher resolution in pitch prediction and better perceptual noise weighting were also evaluated and adapted to a larger frequency band. All these different schemes were assembled into a new wideband CELP coder that could encode the different parameters more efficiently and produce a reconstructed speech with high quality.

Adequate short time spectral envelope coding was crucial in the synthesis stage of the CELP coder. An all-pole filter was implemented to model the behaviour of formants in human speech. Parameters of this all-pole filter were obtained using a 16-th order LPC analysis. The resulting LPC coefficients were not well suited for transmission because a bit error in any one coefficient could destabilize the all-pole filter. To overcome this weakness, we transformed these coefficients into line spectral frequencies (LSFs) that are good representatives of the formant frequencies.

The LSFs still needed to be quantized prior to their transmission. An adequate

quantization was then developed in accordance with the objectives set at the beginning of the research. The aim was to establish transparent quantization of the LSFs by maintaining a relatively low operating bit rate. Two quantization techniques were investigated: scalar and vector quantization. The first suffered from a high number of bits ($\sim$ 50 bits/frame) required to achieve transparent quality, while the second presented a highly complex structure in terms of memory requirements, excessive computational load and long training process.

Because of these problems, a sub-optimal VQ had to be used. A different quantization scheme, known as split vector quantization, was then investigated and turned out to be a good candidate for LSF quantization. This method substantially reduced the complexity usually present with vector quantizers by splitting the LSFs into three subgroups where each subgroup was quantized as a separate entity. The best matching spectral envelope was then selected by minimizing an Euclidean weighted distortion measure. This measure was dependent on two weigting factors. The first one approximated the human hearing sensitivity curve and the second measured the proximity of LSF coefficients to establish the presence of a formant frequency. Both scalar and split vector LSF quantization results were compared and the new scheme showed a bit saving of 20 bits/frame while keeping the same level of reconstructed speech quality.

The addition of a pitch filter to the CELP coder's synthesis stage contributed to a major part of its success especially at low bit rates. Single-tap pitch predictors were initially studied, followed by three-tap pitch predictors that produced better perceptual speech quality but suffered from a larger number of bits assigned to the quantization of the pitch coefficients. An alternative scheme using fractional delay prediction was then investigated. The studies made in the fractional delays field were limited to narrowband speech, so the described method had to be extended to wideband speech (e.g. wider pitch lag range). Different experimental results showed that the performance of the coder improved when compared to the three-tap pitch

predictor and the bit allocation was reduced from 11 bits/subframe (three-tap) to 10 bits/subframe (fractional delays).

A further improvement that was added to the CELP coder structure was the use of the perceptual noise weighting filter. This new noise weighting technique solved the problem of high frequency distortion by, in effect, establishing a better control over both tilt and formant parameters. Specifically, the existing $W(z)$ spectral noise weighting filter was modified by cascading it with a lower order adaptive three pole filter. This enabled the decoupling of formant weighting from spectral tilt weighting. The new filter gave a wider range of achievable noise shapes and thereby allowed the coder to better exploit the masking properties of wideband speech.

Finally, the subjective performance of our wideband CELP coder was assessed by comparing it to the 16 kbits/s coder implemented by Roy [25]. Different simulations were conducted and the results showed that our coder operating at a lower bit rate 11.7 kbits/s generated a better reconstructed speech quality than the split-band algorithm. Nevertheless, the proposed scheme can still be further improved by studying the effects of training the excitation codebook on the quality of the reconstructed speech.

Appendix A

All the speech files used throughout this research were taken from the wideband audio database (Tables A.1 and A.2) containing 48 sentences taken, these files are generated with phonetically balanced sentences. The sampling frequency is set to 16 kHz, and the spectrum information is preserved up to the Nyquist rate (8 kHz). There are four different speakers (2 males, 2 females), and each file identifies the speaker (e.g. M2-07 : Sentence #7 in list of Male #2).

File	Sentence
F1-01	The dark pot hung in the front closet.
F1-02	Carry the pail to the wall and spill it there.
F1-03	The train brought our hero to the big town.
F1-04	Tin cans are absent from store shelves.
F1-05	Slide the box into that empty space.
F1-06	The rude laugh filled the empty room.
F1-07	The plant grew large and green in the window.
F1-08	Tea served from the brown jug is tasty.
F1-09	A dash of pepper spoils beef stew.
F1-10	A zestful food is the hot-cross bun.
F1-11	The cold drizzle will halt the bond drive.
F1-12	The mute muffled the high tones of the horn.
F2-01	He wrote down a long list of items.
F2-02	A siege will crack a strong defense.
F2-03	Grape juice and water mix well.
F2-04	There is a lag between thought and act.
F2-05	Seed is needed to plant the spring corn.
F2-06	The drip of the rain makes a pleasant sound.
F2-07	Draw the chart with heavy black lines.
F2-08	Serve the hot rum to the tired heroes.
F2-09	Much of the story makes good sense.
F2-10	The sun came up to light the eastern sky.
F2-11	The desk was firm on the shaky floor.
F2-12	Nudge gently, but wake her now.

Table A.1: Female speech files.

File	Sentence
M1-01	The small pup gnawed a hole in the sock.
M1-02	The fish twisted and turned on the bent hook.
M1-03	Press the pants and sew a button on the vest.
M1-04	The swan dive was far short of perfect.
M1-05	The beauty of the view stunned the young boy.
M1-06	Two blue fish swam in the tank.
M1-07	Both lost their lives in the raging storm.
M1-08	The colt reared and threw the tall rider.
M1-09	It snowed, rained and hailed the same morning.
M1-10	Use a pencil to write the first draft.
M1-11	The wrist was badly sprained and hung limp.
M1-12	The frosty air passed through the coat.
M2-01	The young kid jumped the rusty gate.
M2-02	Guess the results from the first scores.
M2-03	A salt pickle tastes fine with ham.
M2-04	The just claim got the right verdict.
M2-05	These thistles bend in a high wind.
M2-06	Pure bred poodles have curls.
M2-07	Add the store's account to the last cent.
M2-08	The spot on the blotter was made by green ink.
M2-09	Mud was spattered on the front of his white shirt.
M2-10	Fairy tales should be fun to write.
M2-11	The pencils have all been used.
M2-12	Steam hissed from the broken valve.

Table A.2: Male speech files.

References

[1] B. Caspers and B. Atal, "Beyond multipulse and CELP: towards high quality speech at 4 Kb/s," *Advances in Speech Coding*, Kluwer Academic Publishers, 1990.

[2] J. R. Crosmer and T. P. Barnwell III, "A low bit rate segment vocoder based on line spectrum pairs," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 7.2.1–7.2.4, Tampa. March 1985.

[3] M. Copperi and D. Sereno. "CELP coding for high-quality speech at 8 kbits/s," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 1685–1688, Tokyo, May 1986.

[4] R. E. Crochiere and L. R. Rabiner, "Multirate Digital Signal Processing," Prentice Hall. Englewood Cliffs, NJ, 1983.

[5] A. De and P. Kabal, "Cochlear discrimination: an auditory information-theoretic distortion measure for speech coders," *Bicennial Symposium on Communications*, pp. 1–4. Kingston, Ont., 1992.

[6] A. H. Gray and J. D. Markel, "A spectral flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. on Acoust. Speech and Sign. Process.*, ASSP-22, pp. 207–217, 1974.

[7] F. Ikatura. "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.* 57, Supplement No.1, S35, 1975.

[8] N. S. Jayant, "High-quality coding of telephone speech and wideband audio," *IEEE Commun. Mag.*, pp. 10–20, Jan. 1990.

[9] P. Kabal, "Code excited linear prediction coding of speech at 4.8 kbits/s," *Rapport technique de l'INRS-Télécommunications*, No. 87-36, July 1987.

[10] G. S. Kang and L. J. Fransen, "Application of line spectrum pairs to low bit rate speech encoders," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 7.3.1–7.3.4, Tampa, March 1985.

[11] G. S. Kang and L. J. Fransen, "Low-bit rate speech encoders based on line spectrum frequencies (LSFs)," *Naval Research Laboratory Report 8857*, Nov. 1984.

[12] P. Kroon and B. S. Atal, "Pitch predictors with high temporal resolution," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 661–664. Albuquerque, 1990.

[13] P. Kroon and E.F. Deprette, "A class of Analysis-by-Synthesis Predictive Coders for High Quality Speech Coding at Rates Between 4.8 and 16 kb/s," *IEEE Journal on Selected Areas in Comm.*, vol. 6 (2), pp.353–362, 1988.

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

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

[16] J. S. Marques, I. M. Trancoso, J. M. Tribolet and L. B. Almeida "Improved pitch prediction with fractional delays," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 665–668, Albuquerque, 1990.

[17] I. M. Trancoso, J. M. Tribolet and L. B. Almeida "A study on the relationships between stochastic and harmonic coding." *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 1709–1712, Tokyo, May 1986.

[18] J. L. Moncet and P. Kabal "Codeword selection for CELP coders," *Rapport technique de l'INRS-Telécommunications*, No. 87-35, July 1987.

[19] R. D. de Jacovo, R. Montagna, F. Perosino and D. Sereno "Some experiments of 7 kHz audio coding at 16 kbit/s," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 192–195, Glasgow, 1989.

[20] Y. Ordentlich and Y. Shoham, "Low-delay code-excited linear-predictive coding of wideband speech at 32 kbps," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 9–12, Toronto, May 1991.

[21] D. O'Shaughnessy, "Speech Communication, Human and Machine," *Addison-Wesley*, pp. 142, 1987.

[22] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 661–664, Toronto, May 1991.

[23] S. Quackenbush, "A 7 kHz bandwidth, 32 kbps speech coder for ISDN," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 1–4, Toronto, 1991.

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

[25] G. Roy "Low-rate analysis-by-synthesis wideband speech coding," *Rapport technique de l'INRS-Télécommunications*, No. 90-26, August 1990.

[26] G. Roy and P. Kabal, "Wideband CELP speech coding at 16 kbits/sec," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 17–20, Toronto, May 1991.

[27] B. Atal and M. R. Schroeder, "Code-excited linear prediction (CELP): high quality speech at very low bit rates," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 937–940, San Diego, March 1984.

[28] F. K. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech data compression," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 1.10.1–1.10.4, San Diego, March 1984.

[29] F. K. Soong and B. H. Juang, "Optimal quantization of LSP parameters," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 394–397, New York, April 1988.

[30] N. Sugamura and N. Favardin, "Quantizer design in LSP speech analysis-synthesis," *IEEE Journal on Selected Areas in Communication*, Vol. 6, No. 2, pp. 432–440, February 1988.

[31] C. K. Un and D. T. Magill, "The residual excited linear predictive vocoder with transmission rate below 9.6 kbits/s," *IEEE Trans. Commun.*, Vol. COM-23, pp. 1466–1474, December 1975.