Toll-Quality Speech Coding at 8 kb/s

by

Nabih Maroun

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
Montréal, Canada
February, 1993

Abstract

There has been an ongoing effort to achieve very high quality speech coding at medium transmission bit rates. Consequently, the TIA has chosen the Vector Sum Linear Predictive (VSELP) implementation of an 8 kb/s coder to be the standard for North-American cellular digital telephony. However, it was only recently that, in view of the increased research focus on developing toll-quality speech coding at such bit rates, the CCITT has imposed a set of specifications for standardizing low-delay coders operating at 8 kb/s. The Low-Delay Code Excited Linear Predictive (LD-CELP) suggested by Chen is presently the only potential candidate for CCITT standardization, achieving a one-way coding delay of 10 ms. However, just like the VSELP coding algorithm, the 8 kb/s LD-CELP version does not quite yield toll-quality reconstructed speech. The purpose of the work in this thesis is to establish the minimum requirements for a coding structure capable of generating toll-quality coded speech at 8 kb/s. The purpose of this thesis is to show that, by slightly relaxing the coding delay constraint, perceptual enhancement techniques yield toll-quality coding after redesigning and fine-tuning the optimization and quantization procedures of a CELP coder.

Issues in forward adaptive linear prediction analysis such as windowing and prediction order are studied. Once a suitable analysis method is chosen, the attention is directed toward the quantization of the LPC parameters. With transparent quantization of those parameters being a must for toll-quality coding, an LSF split vector quantization scheme endowed with an improved perceptual distortion measure overcomes the challenge. Joint optimization of the CELP synthesis parameters is then shown to yield improved results when compared to the usual sequential approach. Due to the limited bit resources for quantizing the synthesis parameters, a performant gains (pitch and codebook) vector quantizer is developed. Nevertheless, perceptual enhancement techniques of the coded speech quality remain the major contributors to toll-quality coding. The speech periodicity is improved by both increasing the resolution the the long term predictor delays and by combining the spectral noise weighting with an adaptive harmonic weighting scheme. Coded speech quality comparable to that of a 7-bit log PCM is however only attained with the introduction of a delayed-decision coding technique, extending the CELP parameter selection process beyond the subframe boundary with no extra cost in coding delay.

Sommaire

Beaucoup d'efforts ont été dernièrement accumulés afin d'obtenir un système de codage de la parole de très haute qualité opérant à taux moyens de transmission. La TIA a ainsi selectionné la version VSELP d'un système de codage d'un taux de 8 kb/s comme standard pour la téléphonie cellulaire digitale en Amérique du Nord. Mais ce n'est que récemment, à cause de l'augmentation de l'intérêt que porte la recherche à l'accomplissement d'une qualité téléphonique de parole codée à de tels taux, que le CCITT a introduit un ensemble de spécifications pour la standardisation de systèmes de codage de petits retards à des taux de 8 kb/s. Présentement, le seul candidat potentiel qui se conforme aux recommendations du CCITT est le LD-CELP suggéré par Chen, avec un retard de codage unidirectionnel de 10 ms. Néanmoins, la version du LD-CELP opérant à 8 kb/s, tout comme le système VSELP, n'atteint pas encore la qualité téléphonique. L'objectif de ce mémoire est de montrer qu'en assouplissant les contraintes imposées sur la durée du délai de codage, des techniques de rehaussement perceptuel de la qualité peuvent engendrer un codage de qualité téléphonique au terme d'une nouvelle conception et d'une fine mise au point des procédures d'optimisation et de quantification d'un système de codage CELP.

Plusieurs sujets concernant l'analyse prédictive linéaire adaptée de manière directe, tel le choix de fenêtres et d'ordre de prédiction, sont soulevés. A l'issue d'un choix judicieux de la méthode d'analyse, l'attention est redirigée vers la quantification des paramètres LPC. Une quantification transparente de ces paramètres étant de rigueur pour obtenir une qualité téléphonique, un quantificateur vectoriel partagé des paramètres LSF se révèle être à la hauteur du défi. Une optimisation conjointe des paramètres de synthèse du système CELP est ensuite présentée, exhibant une meilleure performance que l'approche séquentielle habituelle. Une structure de quantification vectorielle des gains (du fondamental et du dictionnaire) est construite, afin de détourner les limites imposées par l'insuffisance du nombre de bits disponible. Les techniques de rehaussement perceptuel de la qualité restent néanmoins les raisons majeures de l'obtention de parole codée de qualité téléphonique. La périodicité de la parole est accentuée par l'augmentation de la résolution du délai du filtre de prédiction à long terme, et par l'adjonction d'une procédure de pondération harmonique adaptative de l'erreur de quantification. Une qualité de parole comparable à celle d'un système 7-bit log PCM n'est finalement obtenue qu'avec l'introduction d'une technique de codage à décision retardée au delà des limites d'une sous-fenêtre de parole.

Acknowledgements

I would like to thank my supervisor Dr. Peter Kabal for his guidance throughout my graduate studies. The feedback and motivation I received from Huan-Yu Su of Bell Northern Research proved also to be very helpful. The major part of the research was conducted at Institut National de la Recherche Scientifique (INRS)-Télécommunications laboratories. All the facilities provided by INRS contributed greatly to the accomplishments of this work. The financial support provided by my supervisor and from the National Science and Engineering Research Council (NSERC) was infinitely appreciated.

This thesis could not have been completed without the constant support of my parents, my brothers and my sister. Special thanks go as well to Karen for her affection and her understanding. I am finally very grateful for the companionship provided by Karim Abboud and my many friends at INRS, McGill and back home in Lebanon.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Digital coding of speech signals is being increasingly used for transmission of speech over long distances. The most straightforward way of carrying out such a coding system is to sample the speech signal at a fixed rate and assign to each sample value a binary number. From the well-known sampling theorem [1], one can recover the analog signal exactly from the above created digital signal if the original analog speech signal is bandlimited to at most half the sampling frequency [2]. The advantages of such a digital representation of speech signals is the ease of its manipulation, of its regenerative amplification and the lack of significant degradation during transmission. Some undesired distortion, due to the transmission channel, can however affect the perceived quality of the speech signal at the receiver end. Reducing the distortion requires often increasing the bit rate which results in higher transmission bandwidth. The choice of bit rate does not only depend on bandwidth constraints, but also on transmission cost. Cheap copper wires and optic fibers allowing larger bandwidth in terrestrial communication networks (such as the telephone network) have handled well rudimentary amplitude compression techniques. However, with the introduction of mobile telephony and satellite communications, bandwidth restrictions have acquired a greater importance. This, of course, has led to the development of more sophisticated techniques for bit rate reduction. The tradeoff between bit rate and coded speech quality is still the main issue in the speech coding research area, while other problems such as computational complexity and real-time implementation are next in line.

Speech is commonly sampled at either 8 kHz or 16 kHz. Prior to obtaining the first sampled version, the original speech waveform is lowpass filtered to guarantee a bandwidth of 0–3400 Hz. The 8 kHz sampled speech is then known as *narrowband* speech. For the second sampled version, the speech working bandwidth before sampling is limited to 0–7000 Hz, in which case the speech is known as *wideband* speech. Narrowband speech preserves the structure of the first three, possibly four formants (resonances), and thus the essential characteristics of the speech signal. Wideband speech can accomodate up to seven formants which guarantees a clearer audible speech quality.

Measuring the speech quality has always been a difficult problem. While some rely on objective measures such as the Signal-to-Noise Ratio (SNR) and the segmental SNR (segSNR), other definitely prefer subjective measures of which the most common is the Mean Opinion Score (MOS). The MOS quality measure is a subjective rating between 1 and 5, from unacceptable to excellent going up the scale. *High-quality* and *near-transparent* attributes are given to speech scoring above 4.0. *Network quality* often replaces the term near-transparent quality in Low-Delay coding applications. Other terms are often used to qualify speech. *Toll-quality* or *telephone quality* for example denote narrowband speech with no perceptible noise, similar to what is heard over telephone networks. *Communications quality* is an attribute to speech with perceivable distortion but highly intelligible, scoring around 3.5 on the MOS scale. *Synthetic quality* is used for unnatural sounding speech but still highly intelligible.

Research in speech coding focuses on either minimizing the perceived distortion of the reconstructed speech signal (at the decoder end) at a given bit rate, or to minimize the bit rate at a given distortion. Two classes of coding systems can usually be discernable: *waveform coders* and *source coders*. Waveform coding is a sample-by-sample based procedure, where the coded signal tries to match the incoming signal as accurately as possible. Source coders exploit the human speech production mechanism and the human auditory system. Such coders derive a speech model characterized by key parameters which are transmitted to the receiver so that the speech can be reconstructed using the same model. Evaluation of the coded speech quality is more perceptually justified, as sample-by-sample reconstruction is virtually impossible in this class of coders. However, large bit rate reductions are possible in source coding

while maintaining a given perceptual quality. In effect, if the speech production model is considered, some parameters that characterize the model might have a limited dynamic range or vary slowly with time. Fewer quantization levels, less frequent updates and interpolation between successive time intervals, all allow bit rate savings if those parameters are sent to the receiver instead of the quantized speech signal itself. On the other hand, given a certain bit rate, the perceptual quality of the coded speech can be dramatically improved if the properties of the human auditory system are exploited. The most common perceptual improvement technique is the inclusion of *spectral masking* in the distortion criterion [3]. The masking phenomenon is a well-known property of the auditory system. Since the ear is less sensitive to distortions located in the high energy regions of the speech spectrum, most of the quantization noise can be moved to less critical regions of the coded speech spectrum if the distortion criterion is appropriately modified.

A coding system relying on a production model based on the physiology of the human speech organs has been proven to offer high quality reconstructed speech with substantial bit rate economies. It is useful however, before setting the scope and requirements of the desired optimal coding system that makes the object of this thesis, to give a general overview on the performance of the existing coders and the scheme the standardization process is following. Rough estimates of bounds bit rates can attain in speech coding can be derived. Defining the upper bound for the bit rate required in speech transmission is equivalent to determining the maximum rate at which information can be transmitted in a signal having the same bandwidth as that of toll-quality speech at low noise levels [4]. The validity of such an equivalence stems from the fact that the transmission of speech implies the transmission of information, with the added assumption that each symbol is independent of the other symbols being transmitted (the structure of the signal is ignored). The bandwidth of speech over the telephone network, denoted by W, is of 4 kHz, and the slight distortion is assumed to be due to white additive Gaussian noise. With an SNR of 30 dB corresponding to subjectively excellent speech quality, the ratio of the average power of the speech signal, P, to the power of the additive noise, G, is P/G=1000. The classic paper on Information Theory by Shannon [5] provides the mean to compute the maximum information rate C which can be decoded from the signal containing

the additive noise:

$$C = W log_2(1 + \frac{P}{G}). \tag{1.1}$$

Hence a coding system capable of yielding reconstructed speech at such high quality is likely to operate at bit rates around this informal bound of about 40 kb/s of information rate. Remembering that the structure of the speech signal was not taken into consideration, it is in fact possible to do significantly better by exploiting the correlation among adjacent samples in the sampled speech signal.

On the other hand, in deriving the informal lower bound on bit rates, the signal structure must be overestimated. However, since the information rate computation becomes much more complicated with the received symbols being interdependent, an alternative estimation method is proposed in Kleijn's work [4]. If English speech is considered, the speech signal can be described in terms of a sequence of high level linguistic symbols, known as phonemes, independent from each other. Spoken at a rate of about 10 phonemes a second, a set of 42 phonemes constitute the entire language. The information content per phoneme is approximately of 5 bits, in reference to a table of the relative probability of the occurence of the phonemes, which yields an information rate of 50 b/s. However, in such a lower bound estimation, only the phonemic information is considered, which results in a loss of the speaker identity (intonation, rate of speaking, etc...).

As depicted in Fig. 1.1, the current performance of speech coders is given as the operating bit rate versus the subjective MOS scale. The conventional *Pulse Code Modulation* (PCM) with μ-law or A-law companding schemes are currently commonplace in the telephone network, operating at a bit rate of 64 kb/s, approaching indeed the estimated upper limit of 40 kb/s. Few assumptions on the speech signal structure are made in these non-uniform quantization schemes, known as log-PCM since the quantizer levels are logarithmically distributed. Exploiting the redundancies in the speech signal, waveform coders allow significant bit savings at the cost of an introduced coding delay, while preserving very high speech quality. *Differential Pulse Code Modulation* (DPCM) and *Adaptive DPCM* (ADPCM) schemes belong to the set of differential coders, a subclass of waveform coding. In these schemes, a predictor filter estimates the upcoming speech sample to be reconstructed. The actual difference between the original speech sample and the estimated speech sample is

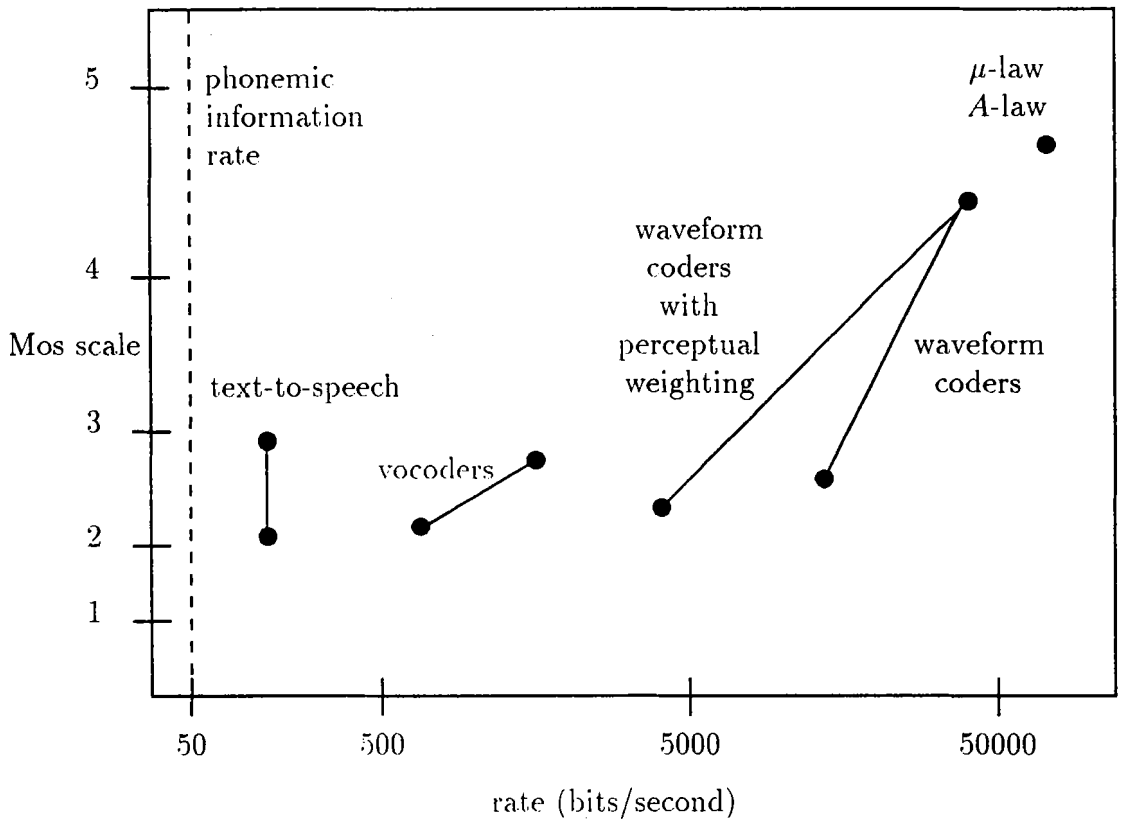


Figure 1.1: Current performance of speech coders

quantized, and the coding scheme might incorporate quantizer level and gain adaptation techniques. As a result, coding rates down to 32 kb/s are capable of yielding the equivalent toll-quality of 64 kb/s log-PCM coders. As the bit rate in Fig. 1.1 is decreased, the efficiency of perceptually-weighted waveform coders becomes more apparent when compared to the simple class of waveform coders. By appropriately modifying the error criterion, the distortion is displaced to high energy content regions in the frequency spectrum and thus rendered less audible. Exploiting this property of the human auditory system improves the subjective quality of waveform coders in all bandwidth constrained applications. The other class of coders formed by vocoders allows substantial bit rate savings by dispensing the speech residual waveform (the speech signal that is left over after all redundancies removal) from transmission, but pay the price in quality: the naturalness of toll-quality in vocoders has not been

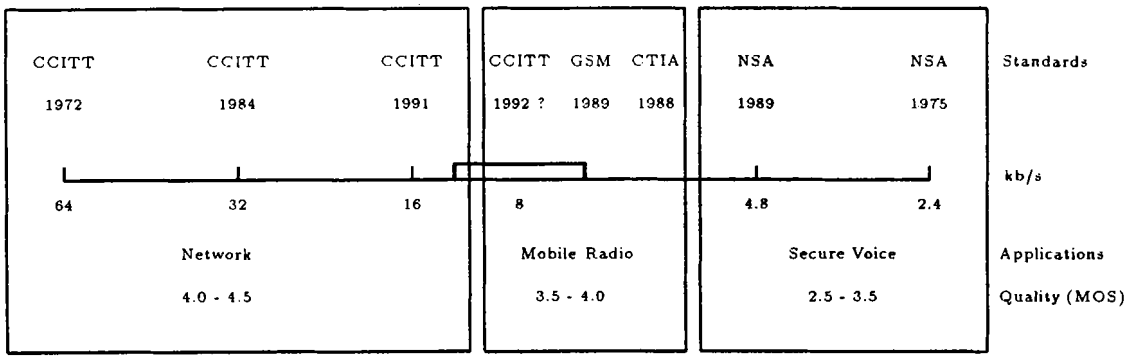


Figure 1.2: Digital telephony standards [6] (CCITT:Consultative Committee for Telephone and Telegraph, GSM: Group Special Mobile, CTIA: Cellular Technology Industry Association, NSA: National Security Agency)

reached yet, and only hybrid models combining waveform coding and vocoding have attained communication-quality. Vocoders in effect rely on speech-specific models, exploiting the usual redundancies and transmit almost all the side information used by waveform coders (pitch, voicing, formants, etc...) but lack the essence of speech contained in the residual. Early vocoders were based on transformations between time and frequency domains like *Adaptive Transform Coding* (ATC) and harmonic coding. The recent years, however, have witnessed an increased dedication toward Linear Predictive Coding vocoders, which will be the object of the next chapter.

The goal of achieving toll-quality coding schemes has been attained so far by coders operating at bit rates starting from 16 kb/s and up. As can be seen from the summary of the current state of digital telephony standards in Fig. 1.2, the CCITT has standardized the 64 kb/s log-PCM and the 32 kb/s ADPCM (G.721) coders encountered previously. Recently, the Low-Delay 16 kb/s high-quality coder based on linear prediction techniques has also been standardized.

As expected, the next CCITT aim is the achievement of near-transparent quality coding at 8 kb/s. Table 1.1 summarizes the CCITT specifications for the 8 kb/s coder standardization. The low-delay requirement is somewhat more loose than the 2 ms objective of the 16 kb/s CCITT standard, but is still very demanding when compared to existing high-quality 8 kb/s coders, recording coding delays between 16

ms and 20 ms. Some of the requirements in Table 1.1 can vary depending on the coding application. Channel error rates can be much more severe in Mobile Radio or indoor wireless applications. The only present candidate for the 8 kb/s CCITT standardization is a Low-Delay Code-Excited Linear Predictive (LD-CELP) coder [7].

The work in this thesis is dedicated toward the achievement of toll-quality speech coding at 8 kb/s. The near-toll quality barrier has already been crossed by two versions of linear prediction based analysis-by-synthesis coders: the Vector-Sum Excited Linear Predictive (VSELP) coder [29] and the Low-Delay Code Excited Linear Predictive (LD-CELP) coder. Both coders registered scores around 3.95 on the MOS scale [7,29]. Toll quality was previously defined to describe reconstructed speech scoring above 4.0 in mean opinion, perceptually comparable to 7-bit log PCM coding quality. Before fullfilling all the requirements of Table 1.1, it seems fundamental to reach toll-quality at an operating rate of 8 kb/s with no restrictions on the coding delay, the computational complexity or any other issue relevant to real-time hardware implementation. However, some of the above mentioned issues will be discussed.

The next chapter strongly arguments the fact that the CELP coding algorithm is the most qualified candidate for undertaking such a challenge. Starting from the foundations set by a conventional CELP coder, all of the components will then be redesigned and optimized either individually or jointly depending on their subjective and objective performances, before being integrated in the coder. Quantization procedures for the prediction parameters, the excitations, the gains and the pitch lags will all be addressed. Perceptual weighting techniques and subtleties enabling the coder to bridge the gap between communication-quality and toll-quality speech will also be described. It will be clear that finer quantization of the coder parameters is not sufficient to obtain the results sought after. Techniques enhancing the perceptual quality of the reconstructed speech by either masking or removing the objectionable distortions seem to be the path to follow.

PARAMETERS	REQUIREMENTS	OBJECTIVES
Speech quality in error free condition	Not worse than that of G.721	
Speech performance with bit errors: BER $< 10^{-3}$ random errors	Not worse than that of G.721 under similar conditions	Equivalent to 16 kb/s CCITT coder under evaluations
One way coder/decoder delay in ms, frame sizes total CODEC delay	$\leq$ 16 ms $\leq$ 32 ms	$\leq$ 5 ms $\leq$ 10 ms
Capability to transmit voice-band data	Not needed	
Quality dependency on speakers	Not worse than that of G.721	
Capability to transmit music		No annoying effects have to be generated
Gross bit rate, kb/s	8	
Tandeming capability for the speech Tandeming with other CCITT coding standards	2 asynchronous with a total distortion $\leq$ 4 G.721 $\leq$ 4 G.721	3 asynchronous $\leq$ 4 asynchronous G.721 Synchronous tandeming property
Capability to operate at different bit rates (9.6 kb/s to 6.4 kb/s)	Needed	Graceful degradation at 6.4 kb/s and improved performance at 9.6 kb/s
Complexity	To be defined	As low as possible

Table 1.1: CCITT standardization requirements for 8 kb/s high-quality coders.

1.1 Organization of the Thesis

With the ultimate aim of implementing a toll-quality speech coder operating at 8 kb/s, the present thesis is structured as follows. The various components that constitute a CELP coder are separately considered and either redesigned or fine-tuned, then they are assembled in such a way to operate efficiently in the whole coder environment. Chapter 2 reviews the theoretical background of linear prediction and introduces the basic concepts of analysis-by-synthesis based linear predictive coders, the general class to which the CELP coding algorithm belongs. The quantization of the LPC parameters is addressed in Chapter 3, where an efficient LPC parameters vector quantizer operating in the Line Spectral Frequencies domain is detailed and evaluated. Pitch prediction techniques are investigated in Chapter 4. Extensive comparisons between various pitch and codebook parameters optimization schemes are performed. The chapter also includes discussions on increased resolution pitch predictors, either by increasing the number of predictor taps or by allowing subsample resolution of the predictor delay. The investigations lead to the elaboration of an efficient fractional pitch prediction scheme that will consistently improve the periodicity and thus the quality of the reconstructed speech.

Realizing that even an optimized performance of a basic 8 kb/s CELP coding scheme is not sufficient to bridge the gap between communication-quality and toll-quality coding, Chapter 5 presents all the methods employed to perceptually enhance the quality of coded speech. With the final judge being the ear, many of the human auditory system properties will be exploited to elaborate improvement techniques such as adaptive postfiltering and harmonic noise weighting. In addition, after justifying the suboptimality of the CELP parameters transmitted on a subframe basis, an approach that delays transmission of those parameters until they have been optimized over several subframes is introduced. This delayed-decision coding technique will have a tremendous impact on the perceptual coding quality as well as on the objective quality measurement criteria. Finally, the performance of the implemented coding scheme is evaluated in Chapter 6, followed by the presentation of future toll-quality speech coding trends and concluding remarks.

Chapter 2

Linear Prediction based Analysis-by-Synthesis Coding

2.1 Introduction

The end purpose of this Chapter is a formal introduction of the Code-Excited Linear Prediction (CELP) coding algorithm. Retracing the historical evolution of speech coding, methods that exploit the characteristics of the human speech production system are presented and evaluated. Modeling the vocal tract in a more cursory way, linear prediction based coding schemes have gained popularity over the more accurate physiological models for their well-established parameter computation procedures, their bit-rate reduction capabilities while maintaining at the same time high coded speech quality. Linear prediction techniques are discussed in detail with some experimental and theoretical results presented. The general class of linear prediction based analysis-by-synthesis coders to which the CELP algorithm belongs is then outlined, going from the basic analysis-by-synthesis structure to finally develop an efficient CELP coding scheme. Very good speech quality results from the original CELP algorithm when operating at intermediate rates, but the extremely high computational complexity is a major drawback. Moreover, accounting for some of the properties of the human auditory system, modifications of the CELP algorithm contribute to increasing the perceptual speech quality in a coding environment at the expense of an even heavier computational load. To make the implementation of the

modified CELP algorithm possible in practical coding applications, fast procedures adapted to the dynamic character of the algorithm were developed. Some of these methods are briefly mentioned at the end of the Chapter.

2.2 Physiology of Speech Production

Exploiting the natural redundancies that exist in speech signals is of prime concern when bit rate reduction in coding is sought. It is hence very instructive to briefly investigate the nature of those redundancies before attempting any kind of speech modeling. Speech redundancies are a direct consequence of the human vocal tract structure and the auditory perception properties. The most common way of characterizing speech production is by a mechanism consisting of three separable entities: an excitation signal generator, an acoustic tube of non-uniform cross sections and radiation walls. Excitation of the acoustic tube results into space radiation and sound waves creation. Thus, the speech signal can be represented in the z-domain by $S(z)$, a product of the excitation signal $X(z)$, the transfer function of the acoustic tube $H(z)$, and the radiation transfer function $P(z)$:

$$S(z) = X(z)H(z)P(z). \tag{2.1}$$

In the human speech production apparatus, the excitation signal is generated by forcing air from the lungs through the vocal cords into the vocal tract. If *voiced* speech is intended (such as /a/, /i/, /o/), the vocal cords will vibrate rapidly inward and outward, shutting and opening sequentially the passage of air between the trachea and the vocal tract. The change of vocal cords vibration rate (fundamental frequency F0, or pitch) is relatively slow: few tens of milliseconds of a vowel incorporate 5 or 6 pitch periods. Furthermore, this generated excitation signal of strongly periodic nature has smooth glottal waveform (pitch cycle) transitions most of the time. The vocal tract, located between the lips and the nostrils on one end, and the vocal cords on the other, acts as the acoustic tube. Different shapings of the vocal tract (with the aid of the tongue, the lips, the jaw and the velum) result in different sounds. The shaping of the vocal tract characterizes the speech spectrum which varies relatively in time when compared to the vibration rate of the vocal cords. Moreover, most of the speech energy is located at low frequency (below 1 kHz with a falloff of -6

dB/octave in frequency for vowels). Some sounds are the result of noisy excitation signals, in which case the vocal cords do not vibrate, but airflow is rushed instead through vocal tract constrictions (lips, teeth etc...). The produced sounds are then classified as *unvoiced* speech.

The intent of the brief description of the human speech production apparatus [8] is a motivation for predictive coding. By appropriately modeling the glottal excitation and the vocal tract system function $H(z)$ with few parameters to be transmitted, substantial bit savings can be achieved. Furthermore, by exploiting the limitations and properties of the human auditory system such as masking phenomena, increasing sensitivity to lower frequency and insignificance of spectral zeros, the perceived speech quality can dramatically be improved. The object of the following subsections is to introduce efficient models of the vocal tract transfer function and the glottal excitation to be used in the scope of linear prediction, the key element of analysis-by-synthesis coding techniques. It is worth mentioning, however, that physiologically-based models for the excitation generation and vocal tract shape suffer from limitations in speech coding applications. The main reasons for the lack of effectiveness of such models is the difficulty of extracting the model parameters from the speech signal and the poor exploitation of the auditory perception properties [4].

2.3 The Purpose of Prediction in Speech Coding

Achieving toll-quality speech coding at rates from 32 kb/s downward was only possible with the introduction of linear prediction. Prior to this, only log-PCM coding techniques reached such quality, with coding rates attaining 64 kb/s. By incorporating a linear predictor in the coding scheme, speech signal redundancies could be exploited, at the expense of an introduced coding delay. Differential Pulse Code Modulation (DPCM) methods have managed to bring down toll-quality speech coding at rates below 32 kb/s by generating a predicted speech sample value from prior speech samples for each speech sample to be quantized. The difference between the original sample and the predicted sample is then quantized. If the prediction filter parameters are only considered to be stationary for small speech segments and are adapted for successive segments, coding rates can be further reduced. In addition, by adapting

the quantizer levels to the prediction error signal (the difference between the original and the predicted speech samples) dynamic range, Adaptive Differential Pulse Code Modulation schemes result, upon which the 32 kb/s CCITT standard is based.

Before discussing the implementation of the predictor which in fact roughly models the vocal tract, a preliminary on closed-loop and open-loop prediction techniques and their role in predictive coding is necessary. Let $P(.)$ be a predictor of order N that attempts to predict an original speech sample $s(n)$ from N past samples. In the case of open-loop prediction, the N past samples are original speech samples, and the *open-loop residual* $x(n)$ is defined as the difference between the original sample $s(n)$ and the predicted sample $\hat{s}(n)$:

$$x(n) = s(n) - \hat{s}(n) \tag{2.2}$$

with

$$\hat{s}(n) = P(s(n-1), s(n-2), ..., s(n-N)). \tag{2.3}$$

If in a coding system prediction is based on past reconstructed speech samples $\bar{s}(n-1), ..., \bar{s}(n-N)$, the *closed-loop residual* $\tilde{x}(n)$ is obtained by:

$$\tilde{x}(n) = s(n) - P(\bar{s}(n-1), \bar{s}(n-2), ..., \bar{s}(n-N)). \tag{2.4}$$

By optimizing the closed-loop predictor, the energy of the closed-loop residual $\tilde{x}(n)$ is minimized allowing smaller quantization bin width, which in turn minimizes the quantization errors. Fig. 2.1 depicts the closed-loop configuration of predictive coding.

The speech signal can be reconstructed from the transmitted quantized version of the residual, $\bar{x}(n)$, provided that the receiver employs the same predictor found in the encoder. For this purpose, either the predictor is kept with fixed parameters, those parameters can be transmitted as side information with the quantized residual, or they can be computed from past reconstructed speech. The current reconstructed speech sample $\bar{s}(n)$ is obtained, as seen in the receiver of Figure 2.1, by:

$$\bar{s}(n) = \bar{x}(n) + P(\bar{s}(n-1), \bar{s}(n-2), ..., \bar{s}(n-N)). \tag{2.5}$$

Two points are of interest in the case of closed-loop prediction. By comparing Eqs. (2.4) and (2.5), the quantization error $s(n) - \bar{s}(n)$ is found to be identical to $\tilde{x}(n) -$

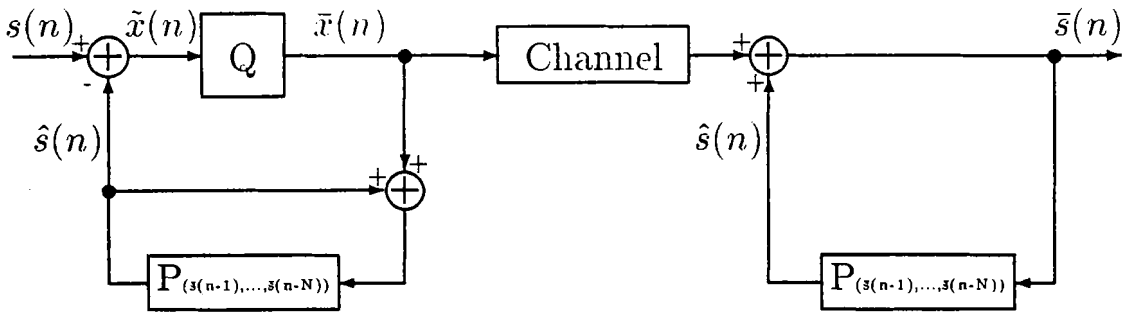


Figure 2.1: ADPCM coder with error free transmission

$\tilde{x}(n)$. Knowing that the quantization error is directly proportional to the signal energy, it is then much more advantageous to quantize $\tilde{x}(n)$ since its energy is less than that of $s(n)$. Defining the *Prediction Gain* as the ratio of the energy of the speech signal to that of the residual signal with both energies averaged over a defined segment,

$$PG = \frac{\sum_{n=1}^{N} s(n)^2}{\sum_{n=1}^{N} \tilde{x}(n)^2}, \tag{2.6}$$

the filtering operation that yields the reconstructed speech according to (.) scales the residual energy by a factor approximately equal to PG [4]. If open-loop prediction is used in the encoder as shown in Fig. 2.2, the quantization error on the open-loop residual will be magnified by this factor. Hence, the larger the prediction gain is, the more justified the quantization of the closed-loop residual is, rather than quantizing either the original speech or the open-loop residual.

Finally, it is important to note that the predictor parameters have to be optimized for open-loop prediction since the prediction gain decreases with decreasing quantization accuracy. The closed-loop prediction structure is only used in the encoder for the quantization of the residual.

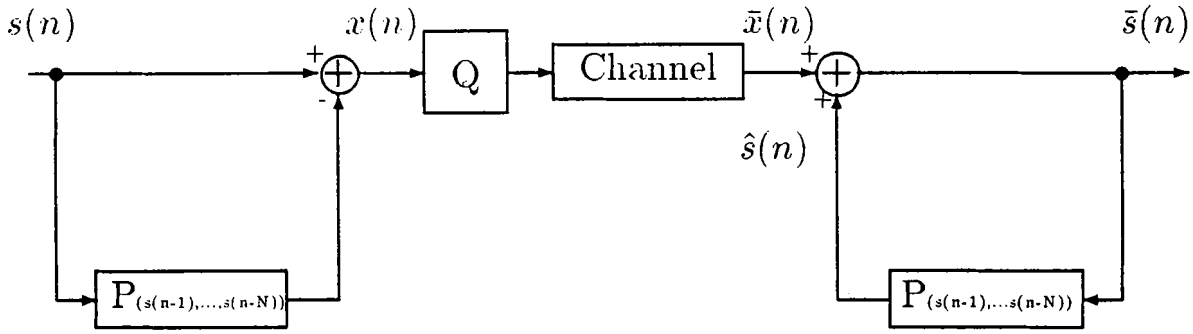


Figure 2.2: Open-loop prediction in predictive coding

2.4 Linear Prediction

The purpose of prediction in speech coding was defined in the previous section to exploit the redundancies that exist in speech signals. Seen from a different viewpoint, a predictor can be considered as a generic model for the vocal tract. *Linear Predictive Coding* has become over the past decade the most popular coding scheme at medium and low bit rates, and has been used in almost exclusively all predictive coders. The next subsection will briefly justify the validity of a linear model for the predictor, and the remaining parts will introduce formal means to estimate the parameters of prediction filters and discuss relevant issues in linear prediction.

2.4.1 Validity of Linear Prediction

Linear prediction of a speech sample from previous samples is optimal in the least-squares sense if the samples of the speech signal are assumed to be random variables with Gaussian distribution [9]. Experiments have shown that, taken over short time segments, speech signal samples can be assumed to have a Gaussian distribution [10].

On a physiological basis on the other hand, a lossless vocal tract can be described by an all-pole filter (any lossless tube model is equivalent to an all-pole filter). Lattice filters are also used to model the vocal tract because of the similarity of both structures, the efficient recursive procedures that exist for parameters computation, the simple stability properties of the filters and the smoothness of the filter characteristics change as a function of the coefficients. However, limitations of the lattice filter

model have been addressed by [4]; the confusion stems from the fact that the lattice filter configuration of all-pole filters corresponds to the transfer function of airflow through a concatenation of tubes of various cross-sectional areas, which is not always an adequate model of the vocal tract. Nevertheless, significant prediction gains were reached with linear prediction that assumes an all-pole model for the speech signal.

A brief word on nonlinear prediction is worth mentioning. Serious research in this field has only started recently [11] and higher prediction gains than those of linear prediction were recorded. The optimality of linear speech prediction can therefore be questionned, but the practical significance of the new results has not been formalized yet.

2.4.2 Linear Prediction and Speech Spectra

The most general predictor form in linear prediction is the Auto-Regressive Moving Average (ARMA) model where a speech sample $\hat{s}(n)$ is predicted from N past predicted speech samples $\hat{s}(n-1), \ldots, \hat{s}(n-N)$ with the addition of an excitation signal $u(n)$ according to:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \hat{s}(n-k) \; + \; G \sum_{l=0}^{q} b_l u(n-l) \tag{2.7}$$

with G being a gain factor and $\{a_k\}$ and $\{b_k\}$ being sets of filter coefficients. Very often, the Auto-Regressive (AR) model corresponding to an all-pole predictor is preferred to the pole/zero ARMA model in which case the prediction operation is written as:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \hat{s}(n-k). \tag{2.8}$$

The major drawback in this model is the absence of representation of the spectral zeros due to the glottal source and the vocal tract response in the nasal portion. In addition, unvoiced sounds are poorly predicted. One common remedy is the addition of 2 or 3 extra poles that can approximate the zeros contribution closely in the predictor frequency response.

Considering the AR model of the predictor, the open-loop residual can be written as:

$$e(n) = s(n) \; - \; \sum_{k=1}^{p} a_k s(n-k). \tag{2.9}$$

Seen in a reverse manner, a speech production model can be elaborated, where an excitation signal $X(z)$ (the z-transform of the sequence $x(n)$) is passed through a shaping filter $H(z)$ to produce reconstructed speech $\hat{S}(z)$. By letting $F(z)$ be the system response of the linear prediction process, the shaping filter $H(z)$, also known as the *synthesis filter* is expressed as:

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

The residual $X(z)$ is obtained by passing a speec signal $S(z)$ through the *inverse filter* $A(z)$.

In reference to Eq. (2.9), the energy of the residual when the speech signal is considered to be deterministic can be expressed according to Parseval's theorem as:

$$\sum_{k=n}^{n+L-1} x(k)^2 = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \frac{|S(e^{jw})|^2}{\frac{1}{|1-F(e^{jw})|^2}} dw = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \frac{|S(e^{jw})|^2}{|H(e^{jw})|^2} dw. \tag{2.11}$$

The objective of linear prediction is well-known to be the minimization of the residual energy. As can be seen from the above equation, this amounts to minimizing the integral of the ratio of the speech signal power spectrum to that of the all-pole synthesis filter. In other words, the power spectrum of the synthesis filter should be an approximation to the power spectrum of the original signal. Thus, the methods that will be investigated next for the computation of the predictor coefficients $\{a_k\}$ can be viewed as methods for fitting the power spectrum of the associated all-pole synthesis filter to the power spectrum of the speech signal, with Eq. (2.11) being the distortion measure.

When the speech signal is assumed to be a stochastic process, the linear prediction procedure still provides an estimate of the spectral envelope with the Itakura-Saito measure being used now as a distortion criterion [12]:

$$d_{is} = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \left(\ln \left[\frac{|S(e^{jw})|^2}{|H(e^{jw})|^2} \right] + \frac{|S(e^{jw})|^2}{|H(e^{jw})|^2} - 1 \right) dw. \tag{2.12}$$

Both Eqs. (2.11) and (2.12) lead to predictor coefficients describing the spectral envelope of the speech signal. The errors in the spectral estimate, as can be seen from those equations, are weighted most heavily in frequency regions where the speech power spectrum $|S(e^{jw})|^2$ is large.

2.4.3 Estimation of the Linear Prediction Coefficients

A speech signal is not stationary and its statistics are not explicitely known. The predictor must therefore be adapted to the changing signal characteristics in LPC coding applications. It is of common practice to consider the speech signal as stationary over short time intervals (of about 20 ms). The predictor coefficients can thus be estimated from a sequence of speech samples obtained from an interval over which the signal is considered to be stationary. Windowing the sampled signal is therefore the first step in linear prediction parameters estimation. Choosing the appropriate window is a whole issue in itself that will be brought up in a subsequent section. Now depending on the linear predictor form to be employed, the parameters to be estimated differ. If a transversal structure is selected (direct-form digital filter), the least-squares method is used to estimate the prediction coefficients $\{a_k\}$: the *Autocorrelation* procedure is employed if windowing is performed on the speech signal whereas the *Covariance* method results when windowing is applied on the residual (error) signal. Open-loop prediction is normally considered in the optimization procedure of the predictor coefficients. Recent work, however, have shown that estimation of the prediction parameters based on closed-loop predictors can lead to significant improvement of the predictor performance at the expense of increased computational complexity [13]. On the other hand, if the linear prediction filter is implemented in a lattice form, both open-loop residual and closed-loop residual energies have to be minimized in order to estimate, in this case, a set of reflection coefficients $\{k_i\}$. All three computation procedures are detailed next.

The Autocorrelation Method

A speech signal is sampled over a time segment where it is considered to be a stationary random signal with for the time being known statistics. Fig. 2.3 describes how to obtain the open-loop residual from the windowed speech samples ($w_s(n)$ is the data window).

The open-loop prediction residual is:

$$\epsilon(n) = s_w(n) - \sum_{k=1}^{p} s_w(n-k). \tag{2.13}$$

Minimizing the energy of the residual amounts to minimizing the expectation value

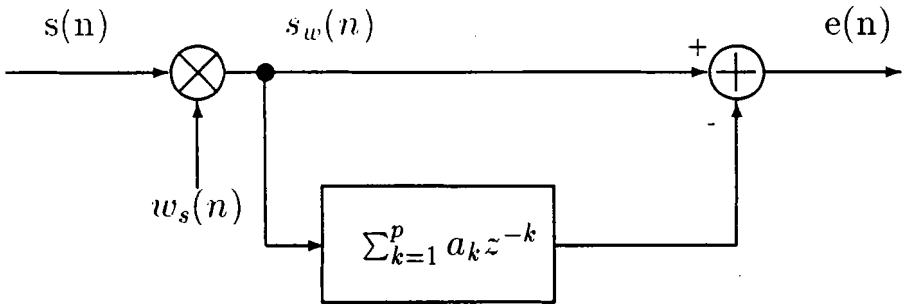


Figure 2.3: Residual $e(n)$ for the Autocorrelation method.

of the square of the prediction residual, $E\left[\epsilon(n)^2\right]$, which can be written as:

$$\begin{aligned} E\left[\epsilon(n)^2\right] = \; & E\left[s_w(n)^2\right] - 2\sum_{k=1}^{p} a_k E\left[s_w(n)s_w(n-k)\right] \\ & + \sum_{k=1}^{p}\sum_{l=1}^{p} a_k a_l E\left[s_w(n-k)s_w(n-l)\right] \end{aligned} \tag{2.14}$$

By taking the partial derivatives of Eq. (2.14) with respect to every a_k and setting the result equal to zero, the Autocorrelation linear system of p equations, $\mathbf{Ra} = \mathbf{r}$, is obtained. The expanded form of the system with autocorrelation matrix $\mathbf{R}$ is:

$$\begin{bmatrix} R(0) & R(1) & \dots & R(p-1) \\ R(1) & R(0) & \dots & R(p-2) \\ \vdots & \vdots & & \vdots \\ R(p-1) & R(p-2) & & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix}, \tag{2.15}$$

where each entry R_{ij} in the autocorrelation matrix is given by $R_{ij} = R(|i-j|)$ and the autocorrelations are defined as:

$$R(i-j) = E\left[s_w(i)s_w(j)\right]. \tag{2.16}$$

The system of Eqs. (2.15) is in fact the Yule-Walker equations with the autocorrelation matrix $\mathbf{R}$ being symmetric and Toeplitz. A fast method for solving the Yule-Walker equations is the Levinson-Durbin recursion [9,12].

Even if a speech signal is considered to be stationary over a short time interval such as $s_w(n)$, its statistics are not explicitely known. In using the Autocorrelation

method to compute the set of linear prediction coefficients $\{a_k\}$, one must estimate the autocorrelations $R(i-j)$ from the windowed speech sample sequence, and then insert those estimates into the Yule-Walker Eqs. (2.15). Thus, if the used window $w_s(n)$ is of length L, the estimated autocorrelations from the sample sequence are usually chosen to be:

$$\hat{R}(k) = \sum_{n=0}^{L-1-k} s_w(n)s_w(n+k) \tag{2.17}$$

and the resulting linear system is $\hat{\mathbf{R}}\mathbf{a} = \hat{\mathbf{r}}$ where the estimated autocorrelation matrix $\hat{\mathbf{R}}$ preserves its symmetric and Toeplitz properties.

The Covariance Method

Minimization in the Covariance method is performed on the windowed error as shown in Fig. 2.4. The window $w_r(n)$ has L non-zero samples. Applying the least-squares method, the mean energy of the error,

$$E = \sum_{n=-\infty}^{+\infty} e_w(n)^2 \tag{2.18}$$

is minimized by taking the derivative of Eq. (2.18) with respect to all the a_k's, and setting the results equal to zero. Once again, a linear system of equations $\mathbf{\Phi a} = \mathbf{\Psi}$ results. The expanded covariance system has the form:

$$\begin{bmatrix} \Phi(1,1) & \Phi(1,2) & \dots & \Phi(L,p) \\ \Phi(2,1) & \Phi(2,2) & \dots & \Phi(2,p) \\ \vdots & \vdots & & \vdots \\ \Phi(p,1) & \Phi(p,2) & & \Phi(p,p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \Phi(0,1) \\ \Phi(0,2) \\ \vdots \\ \Phi(0,p) \end{bmatrix}, \tag{2.19}$$

with the covariance given by:

$$\Phi(i,j) = \sum_{n=0}^{L-1} x(n-i)x(n-j)w_r(n)^2 \quad i = 0,\dots,p \quad j = 1,\dots,p. \tag{2.20}$$

The Covariance matrix preserves its symmetric property but is not necessarily Toeplitz, which makes the Covariance method computationally less efficient. Cholesky decomposition is usually used to solve for $\mathbf{a}$ in the linear system of Eqs. (2.19). Note that in this method, windowing was applied to the error signal which in fact imposes on the speech segment to be of length $L+p$, running from $x(-p)$ to $x(L-1)$. The choice of the error window $w_r(n)$ will also be discussed in a subsequent section.

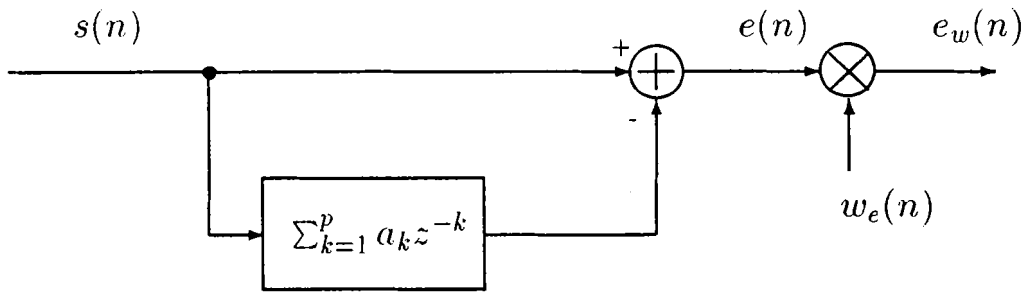


Figure 2.4: Windowed residual for the Covariance method.

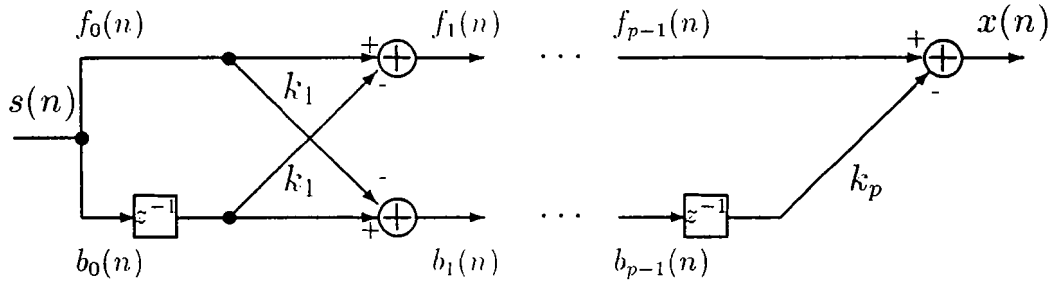


Figure 2.5: Lattice configuration of the inverse filter.

The Lattice Method

The inverse filter $A(z)$ of order p is represented in lattice form in Fig. 2.5. The set of parameters to be estimated in this method are the reflection coefficients $\{k_i\}$. The estimation procedure takes advantage of both the *forward* residual (resulting error after predicting the present sample from the delayed one) $f_i(n)$, and the *backward* residual (resulting error after predicting the delayed sample from the present one) $b_i(n)$. The output of the filter is the usual open-loop residual $x(n)$ corresponding to the forward prediction error $f_p(n)$ at the last stage of the lattice structure. The forward and backward residual samples are recursively obtained by:

$$\begin{aligned} f_{i+1}(n) &= f_i(n) - k_{i+1} b_i(n-1) \\ b_{i+1}(n) &= b_i(n-1) - k_{i+1} f_i(n) \end{aligned} \tag{2.21}$$

with the initial and final conditions being:

$$\begin{aligned} f_0(n) &= b_0(n) = s(n) \\ f_p(n) &= f_{p-1}(n) - k_p b_{p-1}(n-1) = x(n). \end{aligned} \tag{2.22}$$

From the inherent recursive structure in Fig. 2.5, it is obvious that recursive techniques will be used to compute the reflection coefficients. The two most popular techniques are the Itakura and the Burg methods [14,15]. The first method follows directly from the Levinson-Durbin recursion when windowing is applied to the speech signal. It exploits in the computation of the reflection coefficients the partial correlation between the forward and the backward error signals normalized by their energies. The Burg technique is based upon minimizing the weighted sum of the forward and backward residuals within an analysis window $w_e(n)$.

If once again, the speech signal is assumed to be stationary with known statistics, the correlation and energies can be written as:

$$\begin{aligned} F_i(n) &= E\left[f_i(n)^2\right] \\ B_i(n) &= E\left[b_i(n)^2\right] \\ C_i(n,k) &= E\left[f_i(n) b_i(k)\right]. \end{aligned} \tag{2.23}$$

The Itakura method defines the reflection coefficients as:

$$k_i = \frac{C_{i-1}(n, n-1)}{\left[F_{i-1}(n) B_{i-1}(n-1)\right]^{1/2}}. \tag{2.24}$$

The Burg technique minimizes the following recursive windowed error energy:

$$E_i(n) = \sum_{k=-\infty}^{n} w_e(n-k) e_i(k)^2 \tag{2.25}$$

where $e_i(k)^2$ is the barycentre of the forward and backward residual sample energies:

$$e_i(k)^2 = (1-\gamma) f_i(k)^2 + \gamma b_i(k)^2 \qquad 0 \le \gamma \le 1\ . \tag{2.26}$$

By minimizing $E_i(n)$ with respect to the k_i's, the resulting update expressions are:

$$\begin{aligned} k_{i+1} &= C_i(n) \ / \ D_i(n) \\ C_{i+1}(n) &= \sum_{k=-\infty}^{n} w_e(n-k) f_i(k) b_i(k-1) \\ D_{i+1}(n) &= \sum_{k=-\infty}^{n} w_e(n-k) \left[\gamma f_{i-1}(k)^2 + (1-\gamma) b_{i-1}(k-1)^2\right]. \end{aligned} \tag{2.27}$$

The choice of γ and the error window $w_e(n)$ have repercussions on the all-pole synthesis filter corresponding to the set of reflection coefficients $\{k_i\}$, which will be justified shortly. More computationally efficient procedures, such as the Covariance-Lattice method [15] as well as techniques to guarantee better numerical stability introduced by Cumani [16], have also been developed but will not be detailed in this thesis since complexity of the intended coding scheme is not the major target.

2.4.4 Synthesis Filter Stability

In a coding scheme, the predictor coefficients are used in both all-zero filtering operations (inverse filtering) to obtain residual signals and in all-pole filtering operations (synthesis filter) to reconstruct speech signals. Stability of the synthesis filter is of premium importance if performance degradation of the coder is to be avoided in noisy channel conditions. Indeed, any channel error can result in diverging outputs at the receiver if the all-pole filter is unstable. Stability of the synthesis filter is guaranteed by having all the zeros of the inverse filter $A(z)$ reside inside the unit circle in the z-domain. The usual method for stability checking is to convert the direct-form filter prediction coefficients $\{a_k\}$ to the reflection coefficients $\{k_i\}$ of the equivalent lattice-form filter. Stability is ensured if all the reflection coefficients are less than unity in magnitude. The Burg solving technique in the lattice-form filter will yield a stable synthesis filter provided that the lattice stability constant γ is chosen to be 0.5, and the error window $w_e(n)$ is causal [17]. A magnitude larger than one for the k_i's is in fact a physically impossible situation as those k_i's represent the reflection coefficients for fluid flow at the junction of two tube sections when the vocal tract is modeled by concatenating sections of acoustic tubes of different areas.

On the other hand, the Autocorrelation method will always result in a stable synthesis filter associated with the predictor coefficients $\{a_k\}$. This property is motivated by the bias introduced in the autocorrelations of Eq. (2.17), as the decreasing window size with increasing lag guarantees a positive-definite autocorrelation matrix $\mathbf{R}$ [18]. The Covariance method, unfortunately, does not guarantee stability despite the fact that in many cases it results in higher prediction gains.

2.4.5 Backward and Forward Adaptation of LPC Coefficients

In a coding structure, the LPC coefficients should be made available to the decoder every time they are determined for a given segment of speech, in order to enable the reconstruction of one speech sample or a group of samples. Those parameters are usually transmitted as side information to the receiver, along with the quantized residual. Adaptation of the predictor coefficients is then said to take place in a *forward* manner. As mentioned previously, estimation of the coefficients is performed on a frame-by-frame basis in order to comply to stationarity assumptions and to facilitate transmission. The inherent advantage is that the parameters are optimized for the frame in which reconstruction will take place, but a delay can in some applications result in audible echoes during transmission. *Backward* adaptation uses a block of past reconstructed speech samples up to the present one in order to estimate the prediction parameters. The coding delay is therefore suppressed since no buffering of future samples is needed. The other advantage is that the predictor coefficients do not have to be transmitted to the receiver, since the latter has the past reconstructed speech samples available, from which those LPC coefficients can be computed. It seems at first glance that backward adaptation permits substantial bit rate reductions as no bits have to be allocated to quantize the parameters to be transmitted. One must not, however, overlook the fact that the parameters that are being optimized for a block of reconstructed speech samples will only be used for reconstructing speech in the following block (present and future samples). Due to the non-stationary characteristics of speech signals, the *analysis frame*, i.e the frame over which the LPC parameters are estimated, has to be made short looking back at past samples, and thus the update rate of the predictor must be greater than the one adopted in forward adaptation. This will evidently impose constraints on bit rate reduction. In addition, the analysis frame should be highly overlapped for a better tracking of spectral changes in the speech signal. The other major drawback in backward adaptation is the fact that the LPC parameter estimation is based on the reconstructed past speech samples which incorporate quantization errors. Prediction gain values are slightly lower than those obtained in forward adaptation. However, with a frequent update of the predictor and a good quantization scheme of the residual, backward adaptation exhibits excel-

lent performance in applications where low coding delay is a necessity. The zero-delay ADPCM 32 kb/s CCITT standard and the low-delay 16 kb/s standard are both based on backward adaptation. Chen [7] has demonstrated that a high-quality low-delay 8 kb/s coder cannot rely solely on backward adaptation since the backward predictor order and the update rate imposed by the bit rate do not permit a full exploitation of the long term redundancies (periodicity). In the work carried out in this thesis to achieve toll-quality coding at rates around 8 kb/s, the delay constraints will be overlooked and forward adaptation schemes will be adopted, thus guaranteeing the highest open-loop prediction gains possible.

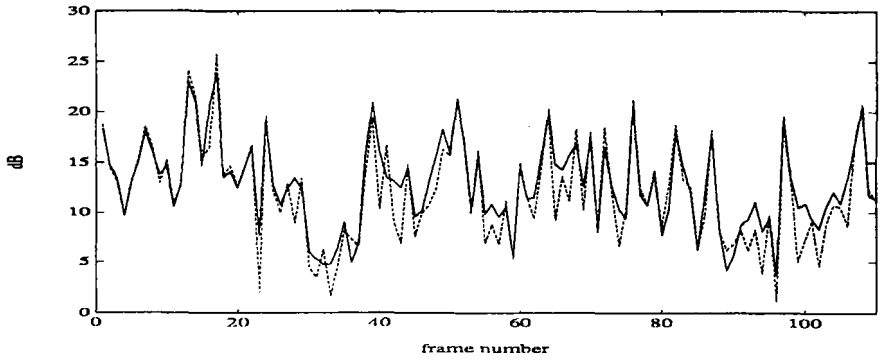
2.4.6 Windowing and Predictor Order Considerations

The LPC parameter optimization and predictor adaptation, alltogether known as *LPC analysis*, relied on windowing of the speech signal in order to preserve quasi-stationarity, but also on methods to solve a linear system of order p, which is actually the order of the linear predictor (*cf* Section 2.4.3). The choice of a suitable window and prediction order is in fact very crucial in the analysis stage of a coding process, affecting many issues such as redundancy removal, numerical stability, minimum-phase property of the inverse filter, computational complexity and real-time implementation. Some of these issues will now be considered.

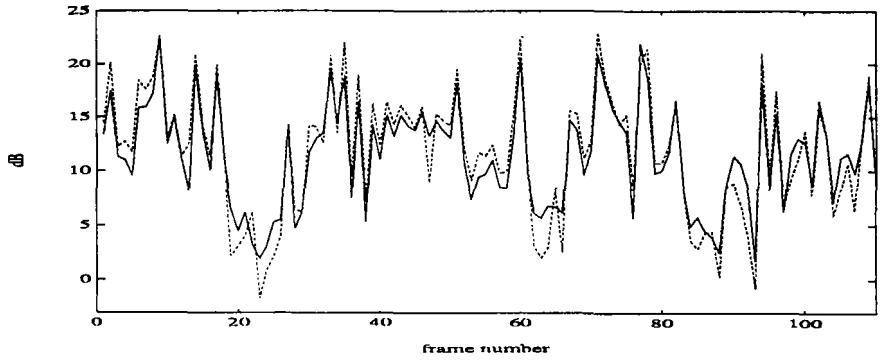
Two types of redundancies are usually treated in speech signals. *Near-sample* redundancies are due to the formant structure of the speech, allowing the prediction af a sample from its immediate predecessors. *Far-sample* redundancies are accounted for the pitch structure that manifests itself mostly in voiced segments of the speech signal. The past samples that are located around one or two pitch lags (periods) from the present sample contribute to the prediction of this latter. The range of pitch and formant redundancies actually overlap, especially in the case of female speech. In natural speech, the pitch lies between 61 Hz and 400 Hz, with a fundamental frequency ($F0$) range of 80 Hz to 160 Hz (a period of 100 samples to 50 samples in 8 kHz sampled speech) for male speakers and an average pitch range of 132 Hz to 223 Hz (60 to 35 samples period) for female speakers. The chosen window in the analysis stage must, *a priori*, be of large size in order to take the long-term redundancies into account. With a maximum and average pitch lags (distance between pitch peaks) of 120 and

60 samples, the window time duration must not be less than **16 ms** (128 samples times the sampling period of 0.125 ms) and the order p of the predictor should be high enough to include the past samples around the lowest pitch lag. Such selections however incur drawbacks if one aimed at capturing within the window the total pitch range or at increasing the predictor order until the male speech pitch lags are within reach. Referring back to equation (2.17), it is clear that the LPC analysis stage relies on an accurate estimate of the autocorrelations (or the covariances). The selected data window must therefore include enough samples to yield a valid estimation of the long-term correlations, and thus be of a length corresponding to two or three times the maximum pitch lag. Nevertheless, such a window would attain 400 samples in length, violating the formant structure stationarity assumption, valid for speech segments of around 100 samples. In fact, the non-stationarity of near-sample redundancies is more harmful to the prediction gain than the accurate tracking of long-term redundancies. On the other hand, the predictor order selection is restricted by the computational expenses of the LPC analysis. Orders up to 50 have been well handled in coding schemes [19]. Predictors of that order exploit quite well the female speaker pitch range, but the male speaker pitch range is only partly captured. It would be then logical to increase the prediction order up to 60 or 70 for a better coverage of the male speech pitch lags. Experiments in [20] have however concluded that only very slight increases result in the prediction gain (lower than 0.5 dB) when the predictor order is varied between 20 an 70. Numerical problems (ill-conditioning of the autocorrelation or the covariance matrix) that arise from high prediction orders are, along with the drawbacks of large window sizes, a major cause of this behaviour. Fig. 2.6 gives an idea on the prediction gain improvement for male and female speech when the predictor order is considered to be 10 and 50 respectively. As can be seen from this figure, the long-term redundancies are better exploited in female speech rather than in male speech with an order 50 predictor, as the average pitch lag of 30 samples for female speakers falls well within the coverage range of the predictor.

Many attemps of better pitch tracking configurations were studied in the past years. One such configuration is the use of a direct-form transversal prediction filter with arbitrary spacing of the taps. 20 taps could be for example allocated for formant tracking (near-sample redundancies) and 30 taps for pitch tracking. The first 20 taps



(a) Male



(b) Female

Figure 2.6: Analysis stage of male and female speech using LPC of order 10 (solid line) and of order 50 (dashed line). The Autocorrelation method with a 20 ms Hamming window is used. SegSNR values for frames of 160 samples are shown.

would assume a fixed position while the remaining 30 are repositionned in such a way to cover the whole lag range (20 to 150 samples). They could be either equally spaced, located around average pitch lags or even around current pitch lags in which case long-term redundancies tracking becomes adaptive. The major disadvantage in such prediction schemes is that the autocorrelation matrix looses its Toeplitz property which renders the LPC parameter computation less efficient. Also, the numerical problems encontered in these methods were worse than those of the regularly spaced taps predictor configuration [20].

The most popular way of alleviating the windowing problems while still conserving prediction gains comparable to those of high-order predictors is to use two separate predictors in a sequential configuration. Two analysis windows of different lengths and different adaptation methods can be adopted in this case. One predictor would be employed for pitch tracking while the other for modeling the formant structure. Joint optimization of the pitch and formant predictor parameters [21] will usually yield a higher prediction gain at the cost of higher computational complexity. A formant predictor of order 10 following a pitch predictor of an order up to 3 configuration results in a performance almost equivalent to that of a single high-order (order 50) predictor [20]. Increases in prediction gain (up to 2 dB increase) due to the pitch predictor stage manifest themselves in voiced regions of the speech signal. Unvoiced segments of speech do not contain any periodic structure and therefore the pitch predictor contribution is useless. As a last remark for this paragraph, the pitch structure in speech signals varies much more rapidly than the formant structure, the update rate for the pitch predictor is hence at least three to four times greater than that of the formant predictor. Extensive description of pitch prediction techniques including pitch predictor optimization will be presented in Chapter 4.

The effect of the selected window size in LPC analysis was introduced earlier. Along with the size, the shape of the data or error windows plays an important role in the predictor optimization. Rectangular and Hamming windows are commonly used in forward adaptation, whereas exponential windows seem to be more efficient in backward adaptation. While the definition of the length L (in samples) is clear for finite windows (Hamming, Raised Cosine, ...), a common way to define the effective

length L_e for semi-infinite causal windows $w(n)$ (exponential windows) is:

$$L_e = \frac{\sum_{n=0}^{\infty} w(n)}{\sum_{n=0}^{\infty} w(n)^2}. \tag{2.28}$$

As it was mentioned previously, the length (or effective length) of the window are chosen so that enough samples are gathered to make the correlation estimates valid, without violating the stationarity assumptions of the speech signal. The best compromise between accuracy of long-term correlation estimations and short-term correlations minimum smoothing results from the selection of windows of length around **20 – 22 ms** (160 - 180 samples). Such lengths also satisfy the condition for an accurate correlation estimate, stipulating that the window length should be much larger than the predictor order. Increasing the window length will degrade the predictor performance rather than yielding more accurate correlation estimates, as the speech stationarity assumption breaks down for segments longer than 22 ms.

Barnwell [22] has extensively used 1, 2 and 3-pole exponential windows in his derivation of recursive windowing methods for generating autocorrelation lags. Barnwell autocorrelation methods proved to be very useful thereafter in real-time implementations which took advantage of the recursive feature. Fig. 2.7 displays the time series of rectangular, Hamming and Barnwell autocorrelation windows. Better prediction gains and subjective ratings are obtained when Barnwell autocorrelation windows are used instead of Hamming windows in backward adaptive LPC analysis [20]. The main reason for the better performance of exponential windows is the heavier emphasis applied to immediate past samples, in opposition to a broader range of "sample capture" for the Hamming window. Up to 1 dB prediction gain improvements can be reached with exponential windowing in backward prediction, especially when the prediction order is relatively high. In fact, the "long tail" of the exponential window adds accuracy to the correlation estimates, more precisely in the large lags correlations where the finite size Hamming window has limitations. Nevertheless, the use of Hamming error windows in the Covariance method for forward prediction is more efficient than exponential windowing [23].

Many of the predictor order and windowing issues have been left out in the previous discussion. For example, the number of bits available for quantizing the LPC

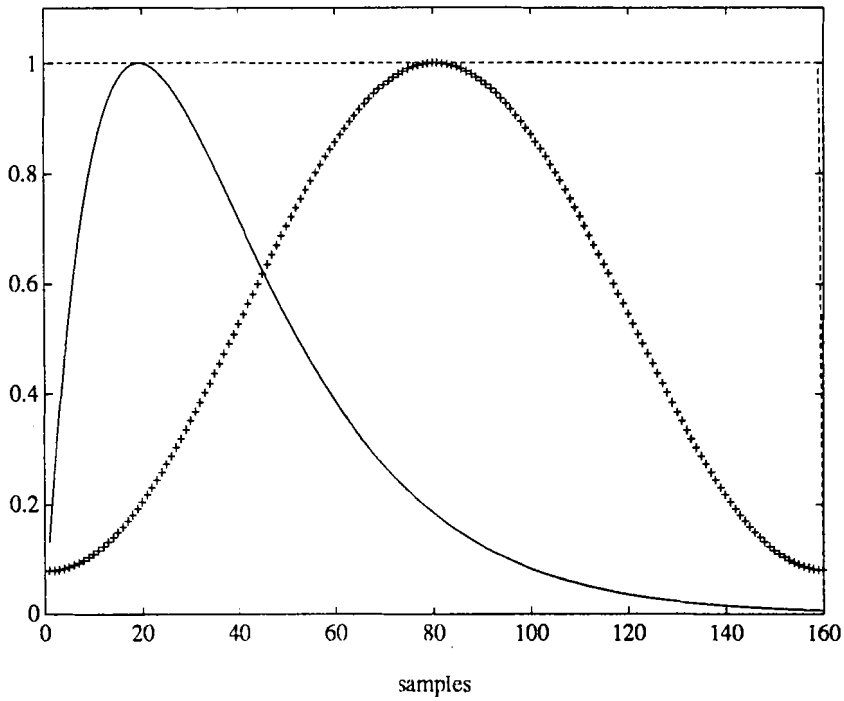


Figure 2.7: 2-pole exponential window (effective length of 158 samples)(solid line), rectangular window (dashed line) and Hamming window (crossed line).

parameters imposes constraints on the predictor order. The causality of the selected windows affects the stability of the all-pole synthesis filter. Numerical problems and computational complexity issues have been left out, but many other will be addressed in their appropriate chapters. With LPC analysis grounds being formally defined, it is now possible to introduce a class of coders that has gained the leading edge in speech coding research for its outstanding bit rate reduction capabilities and high reconstructed speech quality: *linear prediction based analysis-by synthesis* coding.

2.5 Analysis-by-Synthesis Coding based on Linear Prediction

As it was emphasized in Section 2.3, the greatest advantage of linear predictive coders is the quantization of the speech residual rather than the signal itself, allowing a finer

quantization due to the lower energy content of the residual. Now supposing that in the hope of reducing the bit rate, one decides to apply the principles of linear predictive coding to encode a speech signal on a *frame-by-frame* basis, naturally at the expense of a certain coding delay. The closed-loop residual $\tilde{x}(n)$ would then have to be quantized on a blockwise basis. Recalling that the reconstructed speech is obtained by all-pole filtering the quantized residual,

$$\bar{s}(n) = \bar{x}(n) + \sum_{k=1}^{p} a_k \bar{s}(n-k), \tag{2.29}$$

the above operation can be recursively used to obtain a trial block of reconstructed speech samples. More clearly, instead of directly quantizing the closed-loop residual, trial excitation vectors (blocks) are successively selected from a book of all possible excitation vectors and passed through the synthesis filter to yield a trial reconstructed speech frame. The selection of the best matching reconstructed speech vector to the original speech vector should rely on minimizing a certain error criterion. The quantization error in the residual is not a powerful distortion measure when a coding scheme is operating on a blockwise basis, since $\tilde{x}(n)$ depends on the previous reconstructed speech samples $\bar{s}(n)$. Instead, a selection criterion based on the quantization error in the speech signal $s(n) - \bar{s}(n)$, taken on a frame-by-frame basis, seems to be more appropriate. Once the excitation vector that yields the reconstructed speech vector matching the original signal best is determined, its codebook index is transmitted to the receiver. The decoder contains the exact replica of the excitation codebook and thus speech can be reconstructed upon receiving the indices. This qualitatively describes the basic principles of analysis-by-synthesis coding based on linear prediction.

The all-pole filter of Eq. (2.29) with the corresponding z-transform $\frac{1}{A(z)}$ is very often approximated by an all-zero filter of finite impulse response $h_0, h_1, \ldots, h_L$ to simplify the computations in the trial of all excitation sequences. A vector notation can be adopted, as indicated in Fig. 2.8, to describe the operation of the analysis-by-synthesis coder on a blockwise basis. Let $\mathbf{H}$ be the matrix corresponding to the FIR

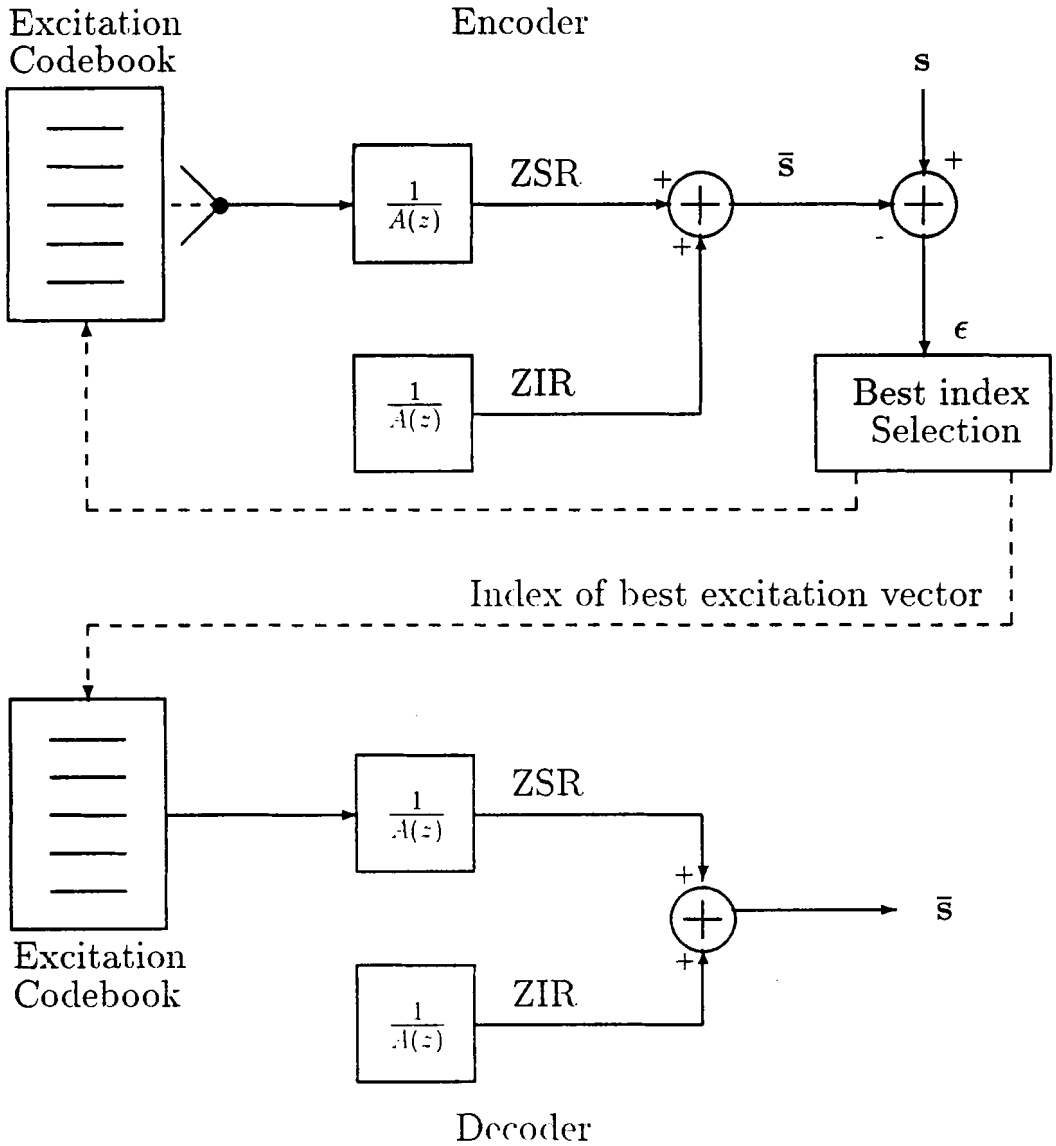


Figure 2.8: Analysis-by-synthesis coding scheme. For every input speech frame **s**, all the excitation entries in the codebook are synthesized by $\frac{1}{A(z)}$. The zero-input response (ZIR) resulting from the previous frame is added to each zero-state response (ZSR) obtained to yield a trial reconstructed speech vector $\bar{s}$. Based on the minimization of an error criterion, the best index is selected and transmitted to the receiver to enable speech reconstruction.

filter $\{h_0, \ldots, h_L\}$ given by:

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & \ldots & & 0 \\ h_1 & h_0 & \ldots & & \\ \vdots & \vdots & & & \vdots \\ & & & h_0 & 0 \\ h_{L-1} & h_{L-2} & & h_1 & h_0 \end{bmatrix} \tag{2.30}$$

With $\bar{\mathbf{s}}$ being a vector representing a frame of L reconstructed speech samples and $\bar{\mathbf{x}}$ a vector denoting a frame of L quantized residual samples,

$$\begin{aligned} \bar{\mathbf{s}} &= [\bar{s}(n), \bar{s}(n+1), \ldots, \bar{s}(n+L-1)]^T \\ \bar{\mathbf{x}} &= [\bar{x}(n), \bar{x}(n+1), \ldots, \bar{x}(n+L-1)]^T \end{aligned} \tag{2.31}$$

the filtering operation of (2.29) is approximated by

$$\bar{\mathbf{s}} = \mathbf{H}\bar{\mathbf{x}} + \bar{\mathbf{z}} \tag{2.32}$$

where $\bar{\mathbf{z}}$ is the zero-input response of the all-pole synthesis filter used to reconstruct the speech at the current frame.

The error criterion ϵ is conventionally chosen to be the least-squares criterion expressed as:

$$\epsilon = (\mathbf{s} - \bar{\mathbf{s}})^T (\mathbf{s} - \bar{\mathbf{s}}). \tag{2.33}$$

It can be rewritten with the help of Eq. (2.32) as:

$$\epsilon = (\mathbf{x} + \mathbf{q} - \bar{\mathbf{x}} - \bar{\mathbf{q}})^T \mathbf{H}^T \mathbf{H} (\mathbf{x} + \mathbf{q} - \bar{\mathbf{x}} - \bar{\mathbf{q}}) \tag{2.34}$$

where $\mathbf{x}$ denotes a frame of unquantized open-loop residual samples, $\bar{\mathbf{q}} = \mathbf{H}^{-1}\bar{\mathbf{z}}$, and $\mathbf{q}$ its unquantized counterpart. The codebook excitation entry $\bar{\mathbf{x}}$ that yields the minimum error ϵ is selected for synthesis. The above error criterion is the basis of the class of analysis-by-synthesis coders based on linear prediction. Some attempts of directly quantizing the residual vector instead of the codebook selection procedure have been made, but they proved not to be as efficient as analysis-by-synthesis techniques.

The design of the excitation codebook is closely related to the characteristics of the speech residual. Excitation sequences in the early coders were generated stochastically, assuming a Gaussian distribution (white noise) for the residual samples [10].

Since then, much more structured sequences have been created, stemming from the increased knowledge about speech signals. The most advanced coders today exploit the pitch structure remaining in the residual signal, and some make use of codebooks trained on a large speech database.

A sample of the quantized residual, $\bar{x}(n)$, can be viewed as consisting of two contributions: a periodic component at a lag d, $\bar{x}(n-d)$ scaled by a gain β, and a non-periodic component $\bar{e}(n)$:

$$\bar{x}(n) = \beta\bar{x}(n-d) \; + \; \bar{e}(n). \tag{2.35}$$

Each contribution has been redesigned and optimized over the years, every time improving the perceptual quality of the encoded speech. The periodic and non-periodic components are now introduced separately.

2.5.1 Pitch Contribution to the Excitation

The periodic contribution to the excitation signal is the outcome of a pitch prediction (long-term prediction) filter. The simplest model for the predictor is a single tap transversal filter, of tap delay d and a filter gain β. The tap delay d and the coefficient β are usually adapted on a blockwise basis for the optimal predictor for a frame of original speech samples. The tap delay range corresponds more or less to the pitch lag range in natural speech (20 to 140 samples for 8 kHz sampled speech). Optimization of the pitch prediction filter parameters will be extensively discussed in Chapter 4, but it is informal to mention for the time being that pitch prediction is efficient the most when it is performed in a closed-loop fashion. The periodic contribution to the excitation signal is thus a scaled version of a frame of past reconstructed residual samples. Closed-loop prediction is in many cases modeled as an *adaptive codebook* containing overlapping excitation sequences [24]. The best adaptive codebook vector is the one closest in the least squares sense (or any other variation to this criterion) to a target residual. Significant bit savings can be achieved if one has an approximate estimation of the target location in the multi-dimensional space the codebook defines (i.e. an estimation of the coming lag value).

Multiple tap pitch prediction filters provide higher prediction gains and better overall subjective quality than one-tap filters in full coders. Three-tap pitch prediction

filters are of common usage in coders that can afford to allocate extra bits for the filter coefficients. The harmonic structure and the spectral envelope of the reconstructed speech signal are better controlled with multiple tap pitch predictors. Fractional delay pitch predictors record almost the same performance as three-tap pitch prediction filters [25]. The tap delay in such filters is allowed to assume non-integer sample values of speech. Very efficient interpolation procedures for non-integer sample resolution are described in Chapter 4, along with stability issues and predictor optimization methods.

In the speech reconstruction stage either in the encoder or the decoder, the excitation vector is passed through a pitch prediction synthesis filter (all-pole) to add a periodic structure to the signal, then the outcome is fed to the linear prediction synthesis filter that takes care of the formant structure. Reversing the filtering order has also been tried, but due to the discontinuous changes of the tap delays of the pitch prediction filter, discontinuous waveforms resulted. Minor clicks were heard in the reconstructed speech and lower overall prediction gains were obtained [21]. It is therefore more effective to place the short-term prediction synthesis filter after the long-term synthesis filter when reconstructing the speech signal, with the order being naturally reversed in the analysis stage.

2.5.2 Non-periodic Excitation Contribution Generation

Once the pitch structure of the speech frame to be reconstructed has been determined, the periodicity is removed from the open-loop residual with the help of a long-term prediction error filter. The remaining target signal has characteristics very close to white Gaussian noise, and can be used for the determination of the non-periodic excitation component $\bar{e}(n)$. Some of the methods to be described have originally been implemented with no prior pitch prediction, but all current coders include long-term prediction techniques.

Multipulse Linear Prediction coding (MPLP) was the pionneer in the class of analysis-by-synthesis coders. This technique searches in each target frame for the best location and amplitude of a pulse in a single pulse excitation vector, subtracts this vector from the target frame to form a new target vector, then recursively repeats all the previous steps until the number of allowed pulses in the excitation vector $\bar{e}(n)$ is

reached. Fast algorithms to determine the multipulse excitation vector and to jointly reoptimize the amplitude of all pulses determined so far in the iterative procedure [26] have made this technique attractive in many practical applications.

A derivative of the MPLP technique is the *Regular-Pulse excited Linear Prediction* (RPLP) method. The excitation vector in this case consists of a train of regularly spaced pulses. The offset of the pulse train is determined first by matching as closely as possible the target vector, then the individual amplitudes of the pulses are optimized and encoded [27].

Nevertheless, the codebook lookup procedure remains the most widely used technique to encode the non-periodic excitation component in analysis-by-synthesis coders. All new speech coding standards (16 kb/s and below) make use of the algorithm developed by Atal [28], known as *Code Excited Linear Prediction* (CELP) coding. Described in simple words, this method searches in a fixed codebook of excitation sequences for the best vector that minimizes a least squares based error criterion between original and reconstructed speech frames. With the elaboration of fast computation methods for the CELP algorithm, this latter very quickly became the most efficient and economical coding technique, yielding good quality speech at around 4.8 kb/s and near-toll quality speech at 8 kb/s, upon which secure and mobile communications systems rely [29,30]. A good bet for achieving toll quality at 8 kb/s is to minimize all the objectionable perceptual distortions incurred by the CELP algorithm, starting with the application of the masking properties of the human auditory system. In view of the critical importance that the coding scheme developed in this thesis places on the CELP coding technique, the basic algorithm will be detailed in the last section of this chapter.

2.6 Auditory Perception in Coding

The ultimate judge of the coding quality is after all the human ear. An increased knowledge of the speech signal processing that takes place in the auditory system will certainly help devising techniques to reduce noticeable distortions in reconstructed speech. The trend in high quality coders has been to move away from objective distortion criteria such as the least squares or mean squared error to adaptive criteria

putting more emphasis on the human auditory perception characteristics.

The first level of speech signal processing by the ear is done at the basilar membrane level. The processing is equivalent to passing the signal through a bank of filters of increasing bandwidth with frequency. Each bandpass filter selects a portion of the signal spectrum and the strengths of the signal are translated into firing patterns. The firing rates of the auditory nerve are highly non-linear and vary for different frequency bands [8]. Due to the overlapping bandpass filters preprocessing in the auditory periphery, the *masking* phenomenon occurs frequently. Two types of masking are encountered; spectral masking is said to happen when a louder signal renders another signal close to it in frequency inaudible. Also, in some frequency bands, the sensitivity of the ear to the signal strength decreases with increasing signal energy [8]. On the other hand, a signal can be masked in the time-domain if it immediately follows the end of a louder signal.

The first conclusion that can be made from the spectral masking phenomenon is that the human auditory system has access to only a part of the information contained in the speech signal. This has been thoroughly exploited in speech coding. Subband coders [8] for instance exploited the reduced resolution of the ear in certain frequency bands by allocating different bit rates to a set of linear prediction based coders spread along a set of distinct frequency bands on the speech spectrum range. The highest bit rates were assigned to the lower frequency bands where the ear is most sensitive. Other methods taking advantage of spectral masking will be introduced in what follows. Time-domain masking, on the other hand, was never exploited in coding techniques.

2.6.1 Spectral Perceptual Weighting

The CELP coding algorithm operates on the full signal energy band. Rather than splitting the signal spectrum into distinct energy bands, a form of spectral weighting can be incorporated in the error criterion derived in Eq. (2.34) emphasizing thus certain frequency regions more than others. The perceptual distorion due to quantization errors is less perceivable in high energy regions of the speech spectrum. Thus, larger quantization errors can be allowed to occur in formant regions of the spectrum. When *perceptually weighted* versions of the original and reconstructed speech frames are com-

pared instead of a direct error evaluation, a great deal of the noisy disturbances and reverberations in the reconstructed signal are suppressed. The noise-weighting filter is commonly a pole/zero filter based on the parameters (LPC coefficients) computed in the linear prediction analysis [3]. With $\frac{1}{1-F(z)}$ being the synthesis filter associated with the linear prediction filter $F(z)$, the adaptive noise-weighting filter $W(z)$ is given by:

$$W(z) = \frac{1 - F(z)}{1 - F(z/\gamma)} = \frac{A(z)}{A(z/\gamma)} \tag{2.36}$$

where γ (noise weighting or bandwidth expansion factor) assumes values between zero and unity. Changing the value of γ moves the poles of $W(z)$ radially in the z-domain (decreasing γ moves the poles inward). Perceptual noise weighting has proven to be so effective that it has been efficiently accomodated with the CELP algorithm, as will be shown in the next section.

2.6.2 Postfiltering

The perceptual noise present in the reconstructed speech signal can usually be attenuated or removed by postfiltering. All-pole and pole/zero postfilters have been used to enhance the formant structure of the transmitted speech. Adaptive postfilters [29,31] based on the LPC parameters have proven to be very effective in enhancing the perceptual quality of the coder although they resulted in lower objective measure values. Detailed description and performance of adaptive postfilters will be reported in Chapter 5. One must be careful, however, in tandeming situations where severe distortions might occur with postfiltering because of the modifications brought to the formant structure. Optimization techniques for the postfilter in multiple encodings schemes are detailed in [31].

2.6.3 Harmonic Noise Weighting

Spectral noise weighting methods introduced in Section 2.6.1 exploit the noise masking capacity of the speech signal due to the formant structure. This helps emphasizing some of the perceptually significant features of the signal. Enhancing the periodicity of the voiced regions in the reconstructed speech has also been the concern of many who looked into using a more perceptually accurate waveform matching criteria. This

is equivalent to accentuating the harmonic structure of the speech spectrum, thus removing the noise between harmonics. To this end, attempts of pitch postfiltering [32] and pitch prefiltering [29] were carried out on the reconstructed speech, after the selection of the optimal excitation vector. These techniques do not however take place in the analysis-by-synthesis iterations and do not contribute therefore to perceptually improving the matching criterion. Another approach known as the constrained excitation [33] treats the CELP excitation as a sum of an ideal excitation and an undesired noisy component. Improvements of the subjective quality resulted by lowering the scaling gain of the codebook excitation vector to a suboptimal value, achieving noise suppression. Such results clearly prove that even the incorporation of spectral weighting in the CELP error criterion is still insufficient. Pitch adaptive comb filtering of the excitation components [29] also helped remove the noise by attenuating the energy of the excitation spectrum between harmonics.

A very efficient way of attenuating the inter-harmonic noise was recently introduced by Gerson and Jasiuk [34]. On the same baseline of the spectral noise weighting methodology, they developed the *Harmonic Noise Weighting* (HNW) technique that exploits the noise masking potential of the harmonic structure of the speech signal. To fully take advantage of the noise masking phenomenon from both short-term and long-term correlations, a harmonic noise weighting filter $C(z)$ is cascaded to the spectral noise weighting filter $W(z)$. The HNW filter is an all-zero filter of the form:

$$C(z) = 1 \; - \; \varepsilon_p \sum_{k=-M}^{M} \beta_k z^{(-D+k)} \tag{2.37}$$

where D is the pitch period and β_k the pitch prediction filter coefficients, optimized in a closed-loop fashion for a frame of speech samples. ε_p is a parameter that specifies the amount of harmonic noise weighting to be applied. The error criterion reveals to be more perceptually accurate when spectrally and harmonically weighted versions of the reconstructed and the original speech are matched. An even better performance is achieved when subsample resolution is allowed in the HNW filter tap delays, especially when it is used in conjunction with a fractional delay pitch predictor. As expected, incorporating the harmonic noise weighting technique in the analysis-by-synthesis loop increases the CELP algorithm complexity, but suggestions to reduce this complexity, listed in [34], demonstrate that the implementation of the HNW technique can

combine affordability and efficiency of performance. A full description of the HNW design methodology and the corresponding perceptual coding improvements (despite lower objective measure scores) are postponed till Chapter 5.

2.7 The CELP Algorithm

The CELP algorithm was seen previously to belong to the same class of coders to which MPLP and RPLP coding schemes belong. These coders treat sampled speech on a frame-by-frame basis, transmitting to the decoder the index of the best codebook excitation signal suceptible of generating upon synthesis a reconstructed speech frame that matches best the original speech frame. The degree of matching is measured by a perceptually weighted error criterion and an analysis-by-synthesis iterative search determines the optimal index of the excitation that minimizes this error criterion. The speech synthesis is achieved by all-pole filtering the selected excitation vector, where the filter coefficients are determined in the LPC analysis stage (*cf* Section 2.4). In addition, all current coders based on the CELP algorithm with a coding delay exceeding 5 ms include long-term prediction filtering in their synthesis stage. Such filters can be viewed either as to be adding a scaled periodic structure to the selected codebook excitation, or as adaptive codebooks (for the 1-tap pitch prediction synthesis filter case) with a structure similar to that of the fixed excitation codebook. The filter coefficient is interpreted as a gain value that scales the adaptive codebook entries, which are in fact past "pitch" synthesized excitation vectors. The second representation of pitch synthesis will be adopted in the CELP configuration of this section. Codebook adaptation and transversal filter structure of the pitch synthesis operation will be discussed in Chapter 4. Fig. 2.9 shows a basic CELP coder with spectral and harmonic noise weighting of the original and reconstructed speech applied, as well as pitch prediction capabilities incorporated. Fig. 2.10 is a more efficient structure of the CELP coder with filtering reallocations and simplifications carried out.

As can be seen from Fig. 2.9 and Fig. 2.10, two codebook indices (i and the pitch predictor tap delay d) and two quantized gain values have to be transmitted along with the LPC coefficients in order to reconstruct the speech signal $\bar{\mathbf{s}}(\mathbf{n})$. Optimiza-

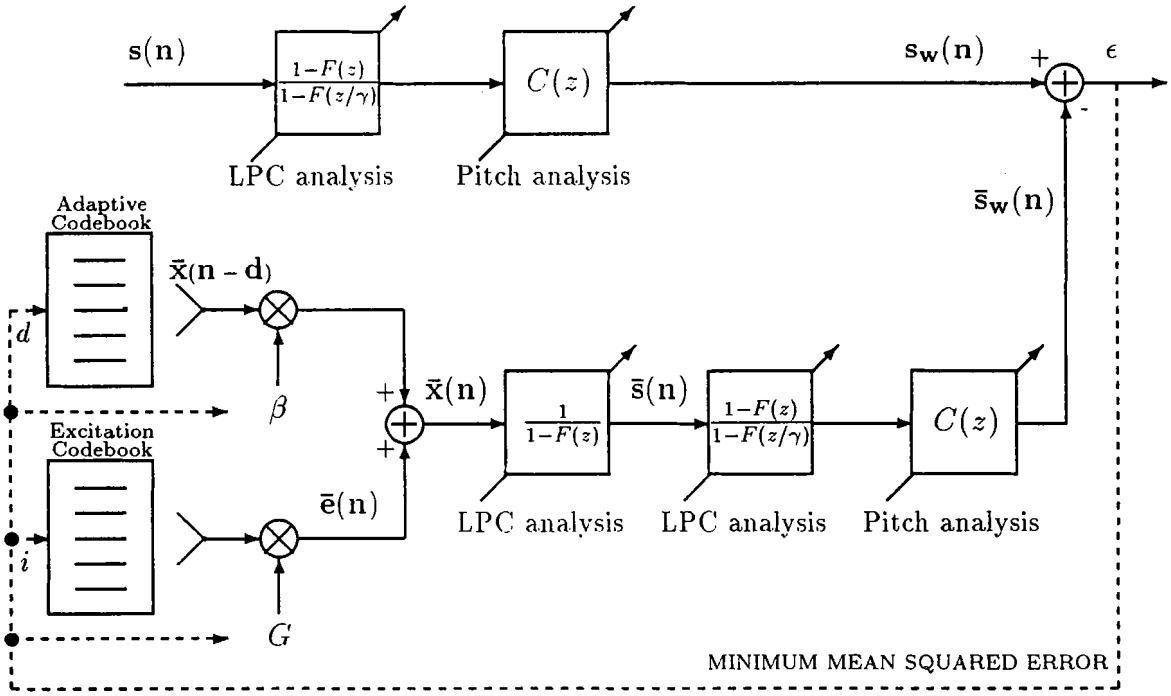


Figure 2.9: Basic CELP encoder including spectral and harmonic noise weighting. The pitch synthesis filter is modeled as an adaptive codebook with entries scaled by β.

tion of the codebook indices and gains can be performed jointly at the expense of a higher computational complexity [21]. However, in view of the resemblance between the adaptive and the excitation codebook structures, sequential optimization can be carried out, trading off optimality with complexity reduction, if the minor degradation that results in coding quality is acceptable. Supposing that the quantized excitation $\bar{\mathbf{x}}(\mathbf{n})$ consists only of a periodic component (excitation codebook entry set to zero), the optimal delay d and coefficient β can be selected in an analysis-by-synthesis procedure. A new target vector $\mathbf{x}(\mathbf{n}) - \beta_{\mathbf{opt}}\bar{\mathbf{x}}(\mathbf{n}-\mathbf{d})$ is computed and the excitation codebook elements (index i and gain G) can now be optimized using the same procedure for this new target. To keep the description of the CELP algorithm general, the excitation vector $\bar{\mathbf{x}}(\mathbf{n})$ will be considered to have a shape-gain structure [35], $\bar{\mathbf{x}}(\mathbf{n}) = \mu^{(i)}\mathbf{y}^{(\mathbf{i})}(\mathbf{n})$. The codebook entry $\mathbf{y}^{(\mathbf{i})}(\mathbf{n})$ can either be part of a stochastically generated set of vectors, a deterministic set of sequences or a trained set of

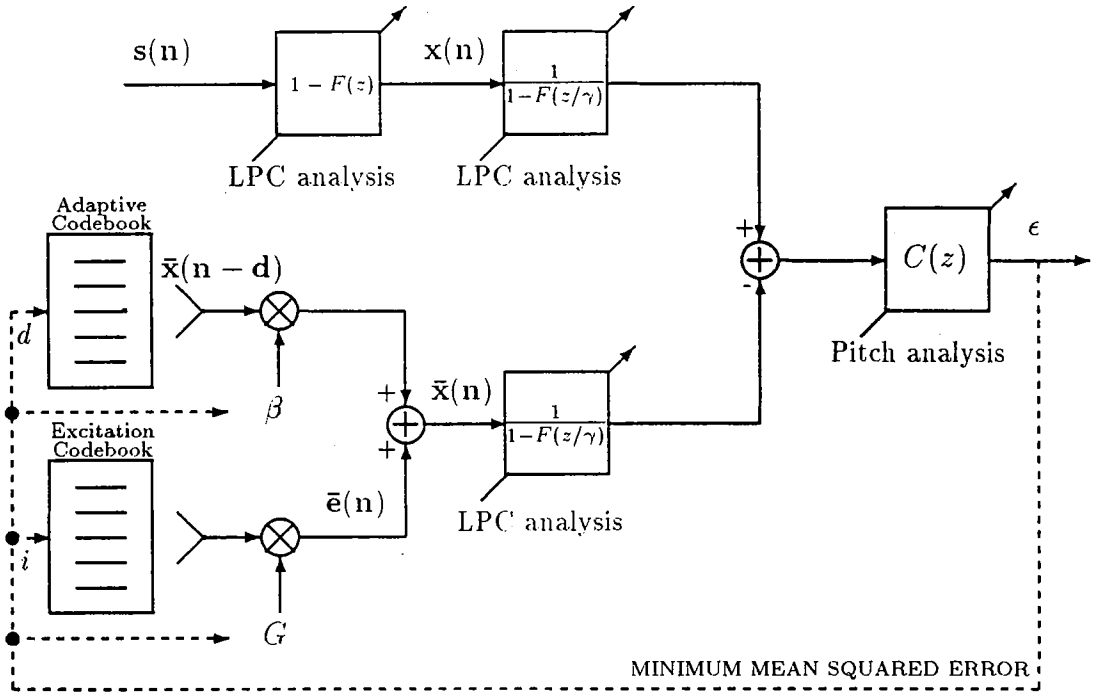


Figure 2.10: Improved CELP encoder. The spectral weighting is incorporated in the synthesis to form a weighted synthesis filter $\frac{1}{1-F(z/\gamma)}$, and the spectrally weighted quantization error is furthermore weighted by the HNW filter to yield the error to be minimized.

trial excitations. The gain $\mu^{(i)}$ belongs to the set of the gain quantization levels. It is very important to notice that the time index **n** used in the previous vector notations to indicate the beginning of a frame will be implicit in the coming derivations.

2.7.1 CELP Algorithm Description

Fig. 2.10 clearly indicates that both the residual vector **x** and the trial excitation $\mu^{(i)}\mathbf{y^{(i)}}$ are passed through the all-pole weighted synthesis filter. The coefficients of this filter (assuming a transversal structure) are the LPC coefficients computed in the analysis stage, $\{a_k\}$, multiplied by powers of the noise weighting factor γ: $\gamma a_1, \gamma^2 a_2, \ldots, \gamma^k a_k$. With the length of the current speech frame to be coded being N, the weighted synthesis filter $\frac{1}{A(z/\gamma)}$ can be approximated by an FIR filter of impulse

response $h_0, h_1, \ldots, h_{N-1}$. The all-pole filtering of a single residual frame $\mathbf{x}$ can be performed by a convolution of the impulse response $\{h_k\}$ of the approximation filter with the samples of $\mathbf{x}$. Written in a matrix form, the convolution becomes $\mathbf{Hx}$, with the matrix $\mathbf{H}$ being an N by N lower triangular with Toeplitz property:

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & \ldots & & 0 \\ h_1 & h_0 & \ldots & & \\ \vdots & h_1 & & & \vdots \\ & & & h_0 & 0 \\ h_{N-1} & h_{N-2} & & h_1 & h_0 \end{bmatrix} \tag{2.38}$$

It is critical to note that the filtering operation $\mathbf{Hx}$ yields the zero-state response (ZSR) of the weighted synthesis filter $\frac{1}{A(z/\gamma)}$. The weighted speech can actually be obtained by adding the zero-input response (ZIR) of the weighted synthesis filter (upper branch of Fig. 2.10). $\mathbf{z}$. to the outcome of the convolution:

$$\mathbf{s_w} = \mathbf{Hx} + \mathbf{z}. \tag{2.39}$$

The computational cost resulting from the addition of the ZIR of the weighted synthesis filter for each codebook excitation entry in the analysis-by-synthesis loop can be avoided by defining a new target vector to match, $\mathbf{t}$. It consists of the open-loop residual vector with the compensation for the quantization errors that occured in previous frames added:

$$\mathbf{t} = \mathbf{x} - \mathbf{H^{-1}z}. \tag{2.40}$$

The quantization of this new target vector follows the selection process of the shape-gain vector $\mu^{(i)}\mathbf{y^{(i)}}$ that minimizes the least squares dynamic error criterion:

$$\epsilon^{(\mathbf{i})} = (\mathbf{t} - \mu^{(i)}\mathbf{y}^{(i)})^T \mathbf{H^T H}(\mathbf{t} - \mu^{(i)}\mathbf{y}^{(i)}) \tag{2.41}$$

One can minimize $\epsilon^{(\mathbf{i})}$ with respect to the gain $\mu^{(i)}$ to obtain the following optimal scalar value,

$$\mu^{(i)} = \frac{\mathbf{t^T H^T H y^{(i)}}}{\mathbf{y^{(i)T} H^T H y^{(i)}}} \tag{2.42}$$

then use this value in the error criterion of Eq. (2.41). However it is more of a common practice to directly use the quantization level values for $\mu^{(i)}$ and select the one that yields the minimum error.

The introduction of perceptual weighting to the synthesis filter reduces the effective length of the finite impulse response approximation. The impulse response of the all-pole filter $\{h_i\}$ can thus be truncated after R samples, for a value of R less than N. The modified $\mathbf{H}$ matrix becomes:

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & \dots & & 0 \\ h_1 & h_0 & & & \vdots \\ \vdots & h_1 & & & 0 \\ & & & & h_0 \\ h_{R-1} & h_{R-2} & \dots & & h_1 \\ 0 & h_{R-1} & \dots & & \vdots \\ \vdots & & & h_{R-1} & h_{R-2} \\ 0 & \dots & & 0 & h_{R-1} \end{bmatrix} \tag{2.43}$$

The weighting matrix $\mathbf{H^T H}$ remains symmetric but becomes also a Toeplitz band matrix. The symmetry inferred to the error criterion is a major asset for the elaboration of fast algorithms in the scope of reducing computational complexity. If the $\mathbf{H}$ matrix of Eq. (2.38) is used in the error criterion, the CELP algorithm is said to be based on the covariance approach. On the other hand, the autocorrelation approach [26] results from using the modified error criterion with the band matrix $\mathbf{H}$ of Eq. (2.43); the symmetric matrix $\mathbf{H^T H}$ contains the autocorrelation of the truncated impulse response. The work in [4] shows that both approaches lead to sensibly the same subjective and objective performances. However, the additional Toeplitz property of the matrix $\mathbf{H^T H}$ that results from truncating the impulse response of the weighted synthesis filter after R samples leads to efficient computation techniques for the error criterion. Finally, the dynamic nature of the weighting matrix $\mathbf{H^T H}$ eliminates the possibility of using established fast search techniques from frame to frame, such as tree searches.

2.7.2 Computational Complexity

The assessment of the computational complexity for the CELP algorithm is obtained by counting the number of operations required to evaluate the error criterion of Eq. (2.41) for a speech frame of length N. Expanding the error criterion, a constant

term $\mathbf{t^T H^T H t}$ results along with a cross-correlation term $\mathbf{t^T H^T H y^{(i)}}$ and an energy term $\mathbf{y^{(i)T} H^T H y^{(i)}}$. The constant term does not interfere in the search for the best excitation vector, and thus does not need to be evaluated.

The constant vector $\mathbf{H^T H t}$ can be computed first in the evaluation of the cross-correlation term. Not including the overhead, N operations will be then required to compute the inner product $(\mathbf{H^T H t})^T y^{(i)}$. The constant vector is obtained by first computing the convolution $\mathbf{Ht}$ ($N(N+1)/2$ operations) then the "time-reversed" convolution $\mathbf{H^T(Ht)}$ ($N(N+1)/2$ operations) assuming that the lower triangular matrix $\mathbf{H}$ of the covariance approach is used. Similarly, $N(N+1)/2$ operations are required for the computation of the convolution $\mathbf{Hy^{(i)}}$ for each codebook vector, followed by N operations for the inner product to yield the energy term $\mathbf{y^{(i)} H^T H y^{(i)}}$. A total of $N(N+5)/2$ operations is therefore required for each iteration (codebook vector) in the analysis-by-synthesis loop, with the added overhead of the constant vector $\mathbf{H^T H t}$ computation. In a scenario where an adaptive codebook of 256 entries and a fixed codebook of 1024 entries are employed, a frame length of 40 samples yields about 230 million operations per second for a sampling rate of 8 kHz. The performance of today's general purpose digital signal processing devices reaches 50 million operations per second! The urge for computational expenses reduction is very serious in order to make real-time implementation of the CELP algorithm possible.

The design of fast techniques that reduce the computational effort of the CELP algorithm has been a major concern of researchers. Detailed description of these techniques will not be given out, but some of them will be briefly mentioned. The most common fast algorithms consist in redesigning the excitation codebook. Center clipping of the stochastic codebook reduces significantly the effort in computing the convolution $\mathbf{Hy^{(i)}}$; a zero sample in $\mathbf{y^{(i)}}$ allows the skipping of an entire column of $\mathbf{H}$. The 90% zero populated codebook with the remaining samples generated from independent identically distributed (iid) Gaussian processes yields the same speech quality obtained with a stochastic codebook [36]. Ternary codebooks, where all the non-zero samples were either set to 1 or -1 [30] provided improved speech quality when compared to iid Gaussian codebooks. Center clipping of the adaptive codebook resulted, however, in serious speech quality degradation and is therefore avoided.

Pre-selection techniques leading to multi-stage search procedures have also been

applied in order to reduce the set of candidate excitation vectors [36]. The non-weighted error criterion can for example be used to select a predetermined number of candidates from which minimization of the weighted error criterion will determine the optimal excitation vector. Pre-selection techniques are actually more effective on the adaptive codebook rather than on the fixed codebook, in view of the periodic structure of the speech frame and the codebook entries. Generalization of the two-stage vector quantization technique (adaptive and stochastic codebooks) leads to successive stochastic codebook quantization stages [37]. Multiple stage searches can noticeably reduce the complexity of the CELP algorithm, but less efficient encoding of the speech signal is suceptible.

A wide variety of other methods have been suggested since the introduction of the original CELP algorithm. Among those, transform methods such as Singular Value Decomposition (SVD) and Discrete Fourier Transform (DFT) techniques introduced in [38] are commonly employed. Coders that can afford a large amount of storage include lookup tables to store all the possible values of the weighting matrix $\mathbf{H^T H}$ and the vectors $\mathbf{H^T H y^{(i)}}$. Significant computational savings in the evaluation of the energy term $\mathbf{H^T H t}$ and the cross-correlation terms $\mathbf{t^T H^T H y^{(i)}}$ result. Algebraic codes [39] along with energy storage tables have also lead to fast algorithms for the computation of the cross-correlation terms. Lastly, recursive methods that rely on codebooks with overlapping entries have been investigated. The interest in such methods stems from the inherent structure of the adaptive codebook, where the shift between adjacent candidates is of one pitch cycle (for a pitch lag larger than the frame size). These procedures can be easily applied to the adaptive codebook entries and even extended to the case where the pitch lag is smaller than the frame length.

2.8 Conclusion

Although linear predictive schemes provide only a cursory model of the vocal tract, their performance in speech coding applications has been more consistent than physiologically more accurate models. Various techniques of estimating the predictor parameters were discussed extensively, leading to the conclusion that each method was appropriate for a certain coding environment. Low-delay coders would employ

for example backward prediction, in which case a single high-order (p=50) predictor could be used, and for which a sound procedure to optimize the LPC parameters would be the Barnwell autocorrelation method. If the delay requirement is more loose, a cascade of a 3-tap long term predictor followed by an order-10 short term predictor exhibits a satisfactory performance in both objective measurements (prediction gain) and subjective evaluation. The Covariance parameter estimation method yielded higher prediction gains than the Autocorrelation method, but turned out to be less numerically well-behaved and did not guarantee stable synthesis filters.

Linear prediction based analysis-by-synthesis coding exploits many of the advantages of linear predictive coding while allowing speech coders to operate on a frame-by-frame basis. At the expense of an increased coding delay, high-quality coded speech is maintained with further bit rate reductions. Among all coders belonging to this categoty, the CELP algorithm distinguishes itself for its conceptual simplicity , its high performance and its affordable implementation with the existing technology.

An increased knowledge of the human auditory perception contributes to enhancing the perceptual quality of CELP type speech coders, proving thus the suboptimality of the original least squares error criterion. Appropriate modification of the error criterion, namely by spectral and harmonic noise weighting of the original and reconstructed speech frames, suppresses much of the objectionable distortions that existed in earlier CELP versions. Postfiltering helps also enhancing the spectral structure of the reconstructed speech.

Finally, a good performance of the CELP algorithm inevitably requires open-loop or closed-loop pitch prediction. The latter form seems to be more efficient, considering that the closed-loop pitch predictor can be interpreted as an adaptive codebook of overlapping entries. Many of the fast algorithms can thus be applied to the adaptive codebook in order to regenerate the pitch structure in the reconstructed speech.

It is therefore only logical that, in view of its high speech quality and the existing fast computational algorithms, the CELP has become the most adopted technique for speech coding applications at rates ranging from 4 kb/s to 9.6 kb/s. Toll quality could very well be within reach in a CELP coding scheme operating at 8 kb/s.

Chapter 3

Quantization of LPC Parameters

3.1 Introduction

The LPC parameters computed at the analysis stage in a coding scheme represent the spectral envelope information for intervals where the speech signal is assumed to be stationary. These parameters are very often transmitted as side information along with the quantized residual. For medium and low bit rate coding applications, restrictions are imposed to the number of bits that can be allocated for LPC parameters quantization. Transparent quantization becomes then a harder task to achieve, even for moderate orders of linear prediction. Vector quantizers are known to be more efficient than scalar quantizers in view of their bit rate reduction capabilities. In addition the quantization distortion in vector quantization is smaller, as the existing correlation between the LPC parameters is exploited.

Using the CELP minimum energy criterion, the optimal set of quantized LPC coefficients can be obtained by searching exhaustively all the quantization levels. This procedure is however very expensive even if one considered scalar quantization of 8 or 16 levels per coefficient or a vector quantizer of 20 bits, mainly due to the synthesis filtering operation. Other distortion criteria for the quantization of the predictor coefficients can be derived, with most of them taking advantage of the human auditory perception properties. Such measures help decreasing substantially the computational complexity by bypassing the filtering operation while still yielding perceptually excellent reconstructed speech.

A mapping of predictor coefficients into another set of parameters to be quantized is very common in high-quality coding schemes. The intent of such transformations is to obtain a better-behaved set of parameters in the sense that the synthesis filter characteristics will vary smoothly as a function of those parameters. The set of prediction coefficients $\{a_k\}$ lack this behaviour, since a small change in a predictor coefficient (due to a channel error for example) can result in an unstable synthesis filter. The reflection coefficients $\{k_i\}$ are more often used as quantization basis, as they display a better behaviour. They are usually either quantized directly, their arcsine used, or transformed to log area ratios (LAR), $\log\left[\frac{1-k_i}{1+k_i}\right]$, to render quantization uniform.

Among all the existing LPC parameter representation domains, the line spectral frequencies (LSF's) are related to the speech spectrum characteristics in the most simple and straightforward way. They represent the phase angle of an ordered set of poles on the unit circle that describes the spectral shape of the inverse filter. With the benefit of many of their structural properties, especially their localized spectral sensitivity to quantization errors, many scalar quantization schemes and stability checking procedures for the LSF's have been developed. It was found, however, that simple Euclidean distances between unquantized and quantized LSF values is not a sufficient quantization distortion criterion. Sensitivity analysis of distortion measures yields an appropriate weighting of the LSF's in a modified error criterion.

Although vector quantization performs more efficiently than scalar quantization, computational complexity was initially a problem. A predictor coefficients vector quantizer requires at least 20 bits to exhibit acceptable distortion. The potential of vector quantization was later exploited, improving the performance of coding schemes at high distortion levels. Product codebooks is one way to overcome computational and storage inconveniences. This technique however is based on independent sets of parameters which are alltogether a one-to-one transform of predictor parameters. Splitting the spectrum into a high-frequency spectrum and a low-frequency spectrum by cascading two linear prediction filters is a direct approach to multi-codebook design [40]. Nevertheless, splitting the speech spectral information into a part related to the lower frequency regions and one corresponding to the higher frequency regions is simplest in LSF quantization since it only requires splitting the LSF's into

two groups with no need to evaluate any pole locations. Split vector quantization of LSF's has actually led to high-quality quantization of the LPC parameters at a rate of 24 bits/frame [41]. Further bit rate reduction is possible by increasing the number of splittings and exploiting intra-frame correlations, at the cost of a minor degradation in quality. At such rates, an efficient spectral distortion measure for vector quantization must be used. The one proposed is in direct relation with the LSF speech spectrum related properties. Moreover, it takes advantage of the human auditory system characteristics, which renders it more perceptually valid.

Large variations in filter coefficients from frame to frame can result in audible distortions. Thus, instead of updating and quantizing the LPC parameters on a frame-by-frame basis, the coefficients are interpolated before or after quantization for individual subframes of size varying between 2.5 ms and 7.5 ms. Interpolation of the predictor coefficients is generally avoided because of the unstable synthesis filters that might result. Transmitting interpolated values of the LAR, the arc-sine of the reflection coefficients or the LSF's then transforming them back to predictor coefficients allows on the other hand smoother variations of the synthesis filter characteristics (spectral shape and stability) and thus improved overall perceptual quality. The performance of the interpolation in the various transformation domains is essentially the same, with a preference going toward LSF interpolation for speech frames of 25 ms or longer [4].

3.2 Line Spectral Frequencies

The most popular set of transform parameters are the *Line Spectral Frequencies* (LSF) introduced by Itakura in 1975 [43]. The advantages of the LSF's will become very clear in view of their properties, providing easy stability checking procedures, spectral manipulations and convenient reconversion to predictor coefficients. Techniques for Line Spectral Frequencies computation are detailed first, then the LSF properties are illustrated.

3.2.1 LSF Computation Techniques

Conversion of the predictor coefficients $\{a_k\}$ to the LSF domain $\{l_i\}$ relies on the inverse prediction filter $A(z)$ of order p, defined here again for convenience:

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}. \tag{3.1}$$

The inverse filter is used to construct two polynomials $P(z)$ and $Q(z)$ in z:

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)}A(z^{-1}), \\ Q(z) &= A(z) - z^{-(p+1)}A(z^{-1}). \end{aligned} \tag{3.2}$$

If the synthesis filter is stable ($A(z)$ is minimum phase), all the roots of $P(z)$ and $Q(z)$ will lie on the unit circle, alternating between the two polynomials with increasing frequency. The LSF's correspond to the angular frequencies w_i of those poles, and can thus be converted to Hertz by a simple multiplication by $f_s/2\pi$, where f_s is the sampling frequency. As can be seen from the definition of $P(z)$ and $Q(z)$, two extraneous roots will lie on the unit circle at $w = 0$ ($z = 1$) and $w = \pi$ ($z = -1$). With the other roots occurring in complex conjugate pairs, p distinct LSF's can be therefore found between 0 and π.

The first approach for LSF computation is an iterative scheme developed by Kang and Fransen [44]. From the phase spectrum of the allpass filter $R(z)$ defined as:

$$R(z) = \frac{z^{-(p+1)}A(z^{-1})}{A(z)}, \tag{3.3}$$

the LSF's are found to be the frequencies where the phase response value is a multiple of π. The same authors proposed an alternate approach using the constructed polynomials $G(z)$ and $L(z)$:
for even values of p,

$$G(z) = \frac{P(z)}{1 + z^{-1}}, \quad L(z) = \frac{Q(z)}{1 - z^{-1}} \tag{3.4}$$

for odd values of p,

$$G(z) = P(z), \quad L(z) = \frac{Q(z)}{1 - z^{-2}}. \tag{3.5}$$

$G(z)$ and $L(z)$ can be rewritten in terms of their coefficients as polynomials of order

$2p$, yielding:

$$\begin{aligned} G(z) &= \sum_{i=0}^{2m} \left[g_i z^{-1} + g_{m-i} z^{-(m+i)} \right], \\ L(z) &= \sum_{i=0}^{2n} \left[f_i z^{-1} + f_{n-i} z^{-(n+i)} \right], \end{aligned} \tag{3.6}$$

with g_0 and f_0 being equal to unity, $m = n = p/2$ for p even and $m = (p+1)/2$, $n = (p-1)/2$ for p odd. Removing the linear phase of $G(z)$ and $L(z)$, the polynomials of Eq. (3.6) can be expressed as:

$$\begin{aligned} G(e^{jw}) &= e^{-jwm} G'(w), \\ L(e^{jw}) &= e^{-jwn} L'(w), \end{aligned} \tag{3.7}$$

where

$$\begin{aligned} G'(w) &= 2 \sum_{i=0}^{m} g_i \cos((m-i)w), \\ L'(w) &= 2 \sum_{i=0}^{n} f_i \cos((n-i)w). \end{aligned} \tag{3.8}$$

The local minima of the power spectra of the polynomials $G'(w)$ and $L'(w)$ correspond to the LSF's.

The other approach for finding the LSF's has been formulated by Soong and Juang [45]. It consists of transforming the coefficients of $G(z)$ and $L(z)$ by a Discrete Cosine Transform. The LSF's are then found by searching in the range $w = 0$ to $w = \pi$ for a sign change in the two polynomials.

The last method, upon which the LSF computation in this thesis is based, was proposed by Kabal and Ramachandran [46]. The polynomials $G'(w)$ and $L'(w)$ are expanded in terms of the Chebyshev polynomials $T_m(x)$. The Chebyshev polynomials are defined as:

$$T_m(x) = \cos(mw), \qquad x = \cos(w). \tag{3.9}$$

The Chebyshev expansion of $G'(w)$ and $L'(w)$ yields

$$\begin{aligned} G'(x) &= 2 \sum_{i=0}^{m} g_i T_{m-i}(x), \\ L'(x) &= 2 \sum_{i=0}^{n} f_i T_{n-i}(x). \end{aligned} \tag{3.10}$$

By tracking the sign changes of the above expansions along the interval $x = -1$ to $x = 1$, the roots are found iteratively. A simple inversion , $w = Arccos(x)$, of the roots results in the LSF set.

3.2.2 LSF Properties

Many of the LSF properties are directly exploited in the quantizer design procedure, stability checking routine and the spectral distortion measure. In addition, some LSF characteristics render them more robust to channel errors. All the properties are listed in this section, illustrated when possible, with the reference to their proofs added.

Starting with the polynomials $P(z)$ and $Q(z)$ given in Eq. (3.2), the following two properties are proved in [45]:

1. All zeros of $P(z)$ and $Q(z)$ lie on the unit circle.

2. The zeros of $P(z)$ and $Q(z)$ are interlaced.

The first property guarantees the uniqueness of the LSF's while the second ensures that the LSF's are in ascending order. It was seen that the efficient numerical computations of the LSF's briefly reviewed in the previous section make use of the above two properties. In addition, Soong and Juang [45] have shown that if the quantized and transmitted LSF's satisfy those properties, namely to be unique and in ascending order, then the inverse prediction filter $A(z)$ is guaranteed to have minimum phase (stable corresponding synthesis filter).

Fig. 3.1 displays the LPC spectrum of two 20 ms frames of speech with the corresponding LSF's, depicted here in Hertz. Two additional properties can be visualized in these two LPC spectra:

3. A cluster of two or three LSF's signals a formant frequency.

4. The bandwidth of a formant depends on the closeness of the corresponding LSF's.

It is well known that most of the speech energy is contained in the first three formants. Spectral distortion measures can make use of the fact that the set of LSF is ordered in frequency, along with these two properties, to assign perceptual weights to the LSF's. The lower LSF's will be naturally emphasized more than the higher order ones.

An additional important property of the LSF's is the localized spectral sensitivity. Small quantization errors due to a distorted channel can affect the quantized

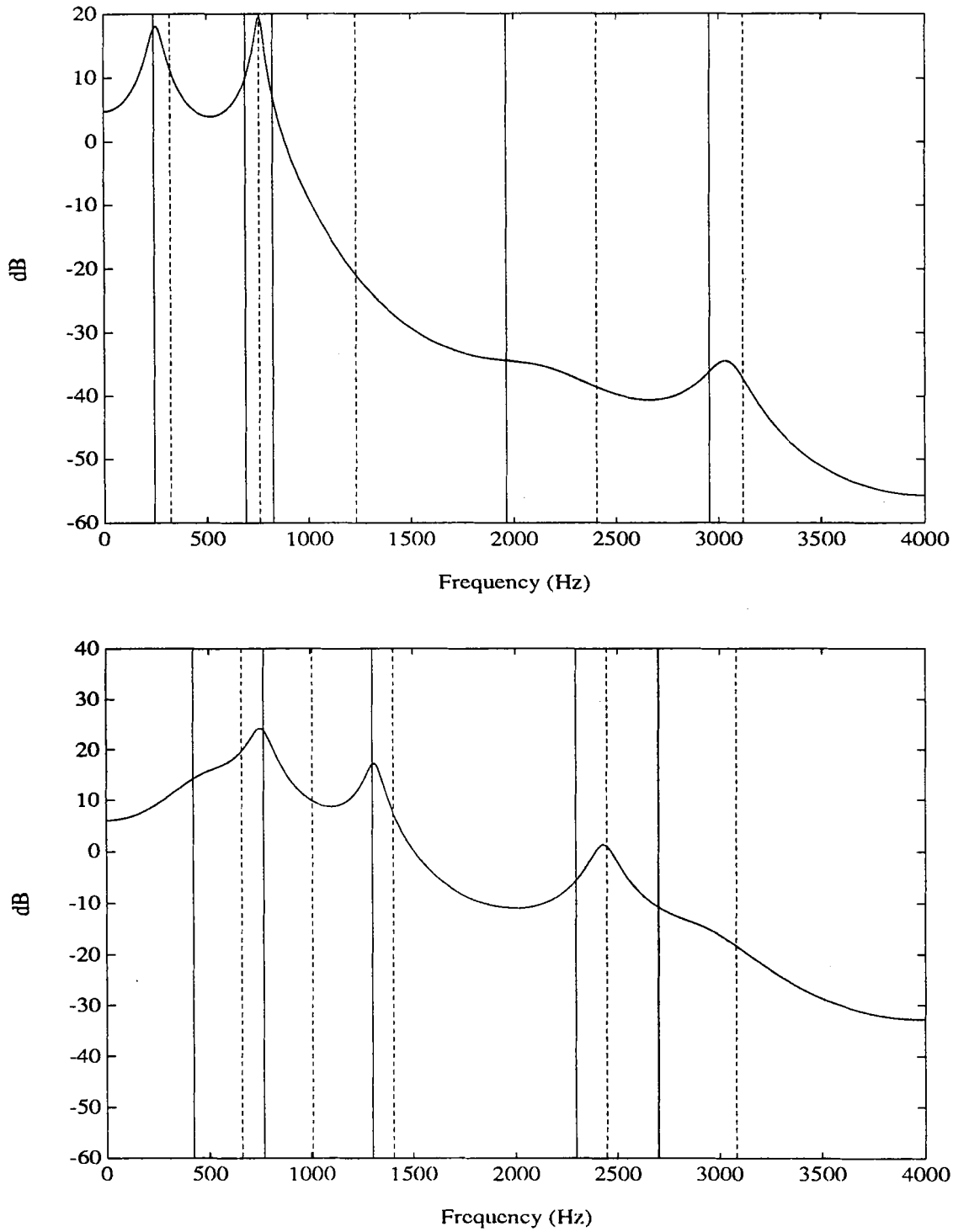


Figure 3.1: LPC spectra of two 20 ms speech frames with the corresponding LSF's displayed in Hertz (vertical lines).

LPC parameters. For predictor coefficients, a small variation in one coefficient can dramatically distort the spectral shape and even lead to unstable synthesis filters. Fig. 3.2 displays the LPC spectra of two 20 ms speech frames, where a distorted spectrum is overlayed on the original spectrum. In the first frame, the sixth LSF was slightly modified while in the second, the eighth LSF was increased by a small amount. It can be seen that the spectral distortion occurs only in the neighborhood of the modified LSF. The spectrum is modified around 1300 Hz in the first frame and around 2600 Hz in the second speech frame. Moreover, alteration of an LSF corresponding to a spectral valley results in less spectral distortion than a formant LSF. This localized spectral sensitivity of the LSF's is exploited in the design of product codebooks for vector quantization of the LSF's. Essentially, it allows one to split the LSF parameter set into subsets of independent parameters with almost no impact on the characteristics of the synthesis filter, and to assign different weights to each line spectral frequency according to its location. Later sections will demonstrate the utility of this last property.

3.3 Distortion Measures

3.3.1 Motivation

The techniques used to estimate the LPC parameters in the previous chapter were seen to be equivalent to attempts to fit the power spectrum of the associated synthesis filter to that of the speech signal (*cf* Section 2.4.2). In a similar manner, vector quantization of LPC parameters can be viewed as selecting from a quantization codebook the LPC vector that yields the best matching spectral envelope to the given spectrum of a short frame of speech. The matching criterion can be directly derived from the analysis-by-synthesis error criterion model, based in on minimizing the energy of the speech error incurred after quantizing the LPC parameters. However, even with moderate size codebooks or good scalar quantizers (rate around 30 bits/frame), the computational load is very large. Therefore quantitative distortion measures that directly attempt to match the trial LPC vectors to a set of original LPC parameters are needed. The Euclidean distance between original and trial LPC vectors have been widely used in early vector quantizers. The limitations of such a measure quickly

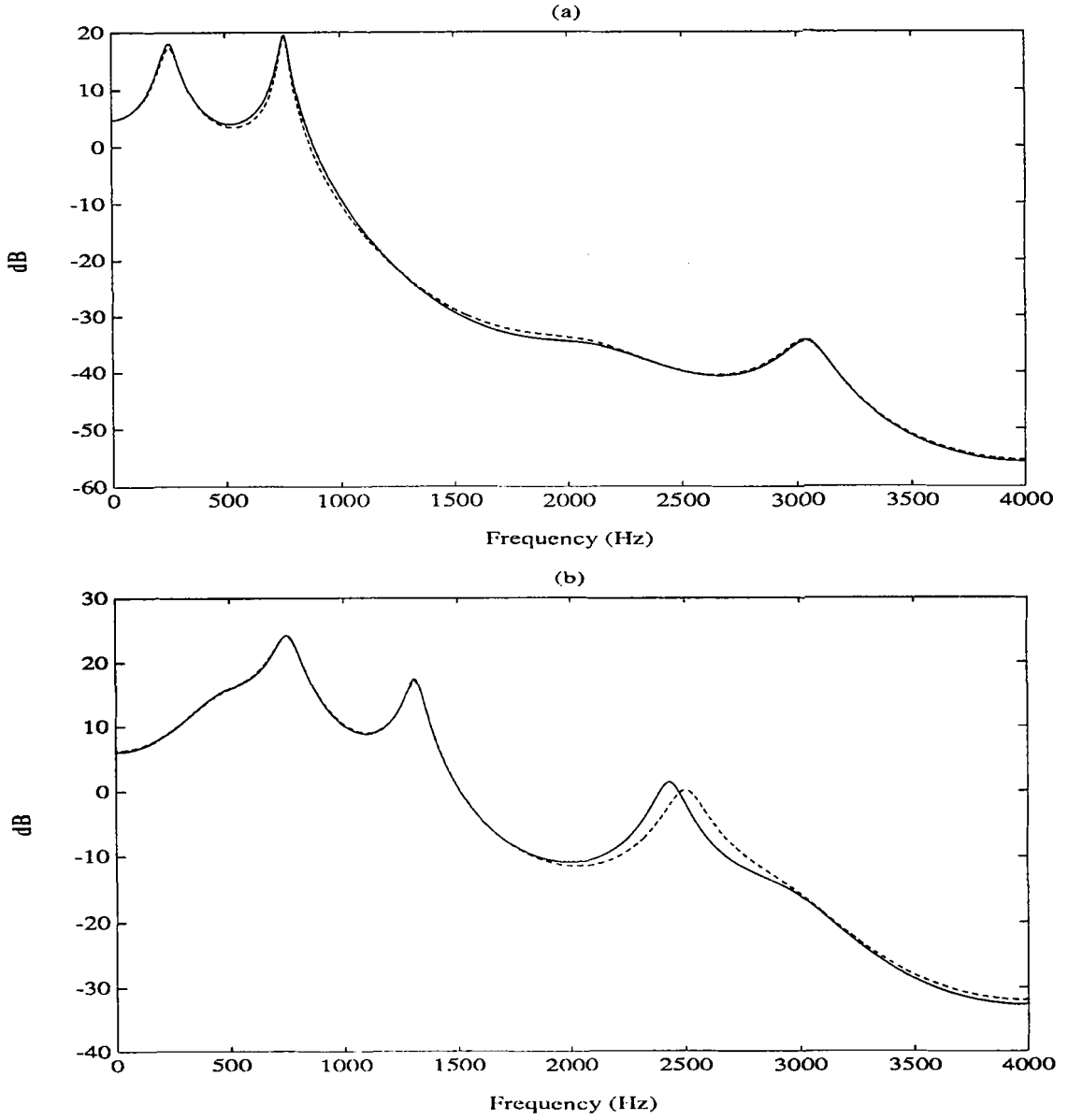


Figure 3.2: Impact of LSF variation on the LPC spectrum. The original spectra are displayed in solid lines and the distorted spectra in dashed lines. The sixth LSF was changed from 1230 Hz to 1300 Hz in (a) and the eighth LSF was changed from 2448 Hz to 2505 Hz in (b).

revealed themselves in unsatisfactory reconstructed speech quality. Taking perceptual considerations into account, the Euclidean measure can nevertheless be appropriately modified to achieve high-quality quantization, namely by appropriately weighting the individual components of the LPC parameter vector.

Since the sought distortion measures quantitatively compare the synthesis filter (LPC) spectrum and the speech frame energy spectrum, they are termed *spectral distortion measures*. Depending on the selected domain for quantization, an appropriate distortion measure is used as a selection criterion for the value or the vector index to be transmitted. The Itakura-Saito spectral measure, the log-area ratio measure and the Euclidean LSF distance will be briefly introduced. However, with the LSF's being the parameters that are quantized and transmitted in the coding scheme of this thesis, more effort is devoted to the design of perceptually weighted Euclidean LSF distances, exploiting the frequency discrimination characteristics of the human ear as well as the LSF properties.

There are other contexts in which those distortion measures can apply. The performance of speech coders can for example be evaluated when quick objective results are needed. Criteria based on the speech spectral envelope lead to a greater insight than the regular SNR objective criterion. The use of the detailed measures in such contexts is not attempted in this work, but results found in previous literature are reported.

3.3.2 Spectral Envelope Distortion Measures

The basis for defining and comparing the spectral envelope distortion measures is the comparison of the original speech LPC spectrum obtained from the synthesis filter $1/A(z)$ and the energy spectrum of the synthesis filter associated with the quantized LPC parameters, $1/A'(z)$. Both attempt to accurately model the energy spectrum of the speech signal taken on a frame-by-frame basis.

The **Itakura-Saito measure** IS is directly related to the logarithm of the original and quantized LPC spectra [42]. It is defined as:

$$IS = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[e^{V(w)} - V(w) - 1 \right] dw \tag{3.11}$$

where

$$V(w) = \ln\left[\frac{1}{|A(e^{jw})|^2}\right] - \ln\left[\frac{1}{|A'(e^{jw})|^2}\right]. \tag{3.12}$$

Denoting by α the energy of the residual signal obtained upon passing the speech signal through the inverse filter $A(z)$, and by α' the energy of the residual resulting from the inverse filter $A'(z)$, the integrals in Eq. (3.11) are evaluated to yield:

$$\begin{array}{ll} \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{V(w)}dw = & \frac{\alpha'}{\alpha} \\ \frac{1}{2\pi}\int_{-\pi}^{\pi} V(w)dw = & 0. \end{array} \tag{3.13}$$

The resulting Itakura-Saito measure in decibels is therefore:

$$IS_{dB} = 10\log\left[\frac{\alpha'}{\alpha} - 1\right]. \tag{3.14}$$

Weighting can be introduced to the Itakura-Saito measure to take advantage of the perceptual discrimination properties of the human ear. Weighting schemes $\omega(e^{jw})$are proposed in [47] and incorporated in the Itakura-Saito as follows:

$$IS_W = \frac{1}{2\pi}\int_{-\pi}^{\pi} \omega(e^{jw})\left[e^{V(w)} - V(w) - 1\right]dw. \tag{3.15}$$

The log-area ratio measure, naturally based on the set of reflection coefficients, is defined to be:

$$D_{LAR} = \sum_{n=1}^{p}\left[\log\left(\frac{1-k_i}{1+k_i}\right) - \log\left(\frac{1-k_i'}{1+k_i'}\right)\right]^2, \tag{3.16}$$

with $\{k_i\}$ being the set of p reflection coefficients and $\{k_i'\}$ their quantized counterpart.

The **Euclidean distance measure** can be employed in any quantization domain. It corresponds to minimizing the mean squared error between the LPC parameters and their quantized values. This simple distance measure, however, does not yield perceptually good LPC spectrum approximations. The complex weighting schemes introduced to LSF Euclidean distances contribute greatly to increasing the accuracy of the perceptual spectral envelope matching to original speech energy spectrum, as will be seen in the next paragraph.

To appropriately define a weighted Euclidean distance measure to be used in the vector quantization of the LSF's as a distortion criterion

$$DW_{LSF} = \sum_{i=1}^{p}[\omega_i(l_i - l_i')]^2, \tag{3.17}$$

where $\{l_i\}$ is the set of original LSF's and $\{l'_i\}$ their unquantized counterpart, the set of assigned weights $\{\omega_i\}$ should reflect the essential spectral properties of the LSF's. Looking back at Fig. 3.1, the LSF's are seen first to be spread out along the frequency range from 0 to 4 kHz. Moreover, each LSF value can vary in a limited frequency range. It is well-known on the other hand that the sensitivity of the human ear to speech sounds decreases with increasing frequency. High-frequency LSF's can therefore be given lower weights than those of the first ones. The other important observation is the clustering of the LSF's around the peaks in the spectral envelope. Those peaks characterize the speech formant frequencies which are much more perceptually significant than the spectral valleys. Relatively close LSF's can therefore be interpreted as modeling a formant frequency, and thus should be more emphasized than spaced LSF's which only control the spectral tilt.

The two previous observations are the basis of the weighting scheme upon which the LSF quantizer relies in this work. The weighting factor ω_i is separated into an *ear sensitivity* modeling part, ωes_i, and a spectral envelope *formant* characteristic part, ωf_i:

$$\omega_i = \omega es_i \, \omega f_i. \tag{3.18}$$

Curves that model weights for the human ear sensitivity to sound frequency have been studied in the past. Most of these studies were based on the Just Noticeable Differences (JND's) of a single tone. Defined formally, the JND is a subjective measure that detemines an acoustic distance threshold (loudness, frequency) above which two successive tones can be distinguished. These thresholds are based on a percentage of listeners distinction between successive tones as a function of acoustic parameters. The area of human perception for a sound is found to lie between 100 Hz to 8 kHz in frequency, for an intensity ranging between 30 dB and 80 dB [8]. While sensibly the same sound intensity is needed for speech to be heard when its spectral content varies between 200 Hz and 1 kHz. a sound at 4 khz needs almost 20 dB more intensity to be heard.

The detectability of a sound consisting of many spectral components is not a simple function of the detectability of its components. A weighting scheme guided by the ear frequency discrimination of tones can nevertheless provide a valid model. A piecewise linear model of the human hearing sensitivity to discriminating frequency

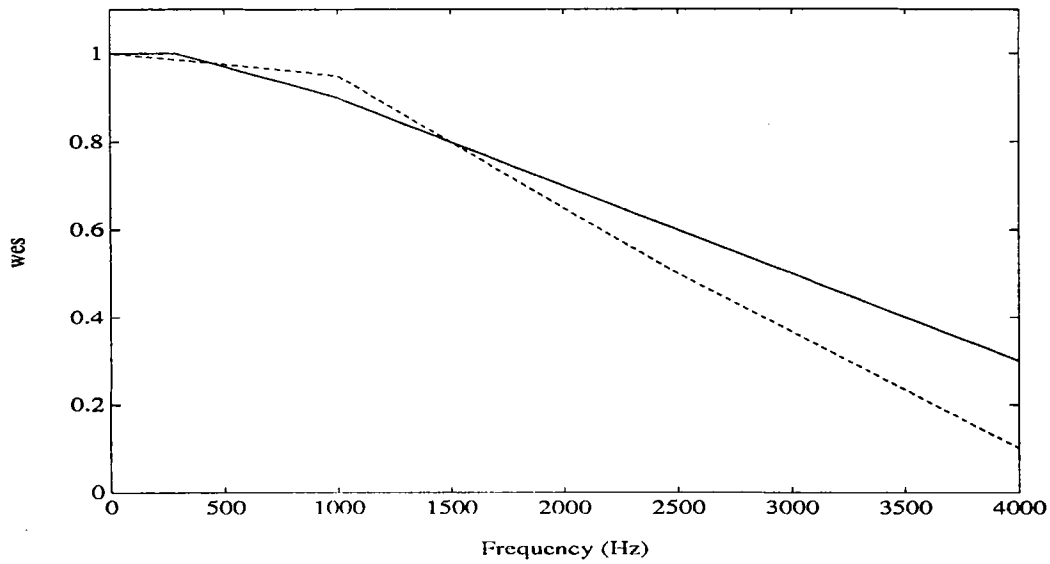


Figure 3.3: Ear sensitivity to discriminating JND based frequency differences. The solid line models the weighting scheme for the ear while the dashed line is the model piecewise approximation scheme.

differences based on the JND of a single tone was elaborated in [48]. Fig. 3.3 depicts this approximation with, superimposed, a three-component ear sensitivity weighting scheme applied to the intended LSF Euclidean distance measure. Fixed ear sensitivity weighting schemes have been proposed where a non-adaptive coefficient scales each squared difference for each LSF [41]. The linear approximation to the sensitivity curve is however a more theoretically accurate approach. The piecewise approximation model is in fact less accurate for the low frequencies and more exact for frequencies above 2 kHz. In this manner most of the emphasis is put on the first two formant frequencies. The form of the weighting scheme is the following:

$$\omega es_i = \begin{cases} 1 - 0.5/10000 l_i & \text{for } l_i < 1000 \\ 0.95 - 3/10000(l_i - 1000) & \text{for } 1000 \leq l_i < 2500 \\ 0.5 - 2.667/10000(l_i - 2500) & \text{for } l_i \geq 2500 \end{cases} \tag{3.19}$$

The relation between the formants and the LSF's is exploited in the second weighting component. In the distance measure, more weight should be given to the LSF's corresponding to higher amplitude formants than to those in non-formant regions.

Also, the LSF's corresponding to the spectral valleys are attributed the least weight. A direct formulation of this idea is to assign to the LSF's weights proportional to the value of the LPC power spectrum $|A(w)|^2$ at the LSF frequency [41]. The formant weight would therefore be computed as:

$$\omega f_i = \left[|A(2\pi l_i/f_s)|^2\right]^{\alpha}, \tag{3.20}$$

where l_i is the i^{th} original LSF, f_s is the speech sampling frequency, and α is an exponent usually chosen between 0 and 0.5 to control the relative weighting assigned to the LSF's. A simpler approach having sensibly the same impact on the weighting scheme is to use the closeness of neighbouring LSF's as a criterion for the characterization of the formant regions in the spectrum. Since the closer the LSF's are together the more likely they are to fall in a formant region, the distance between each LSF and its closest neighbour, d_i, can be found and normalized by the maximum distance found d_{max} to yield the quadratic weighting scheme proposed in [49]:

$$\omega f_i = 0.5 \; + \; 0.5\left[1 - \frac{d_i}{d_{max}}\right]^2. \tag{3.21}$$

3.3.3 Discussion

The performance of the different distortion measures for the different LPC parameter sets is evaluated in two contexts. The purpose of such measures is mainly an objective mean to assess the perceptual distortion in selecting a codebook entry of LPC parameters to match an original vector in vector quantization (VQ). The codebook consists in fact of a set of spectral envelopes from which one will perceptually match best the spectral envelope to be coded. The role of the distortion measures is thus to model the quantization error that would be perceived by the human auditory system. The other context of evaluation for the distortion measures is a complete coder environment. Studies in [49] have shown, however, that aside from the SNR and segSNR criteria the other measures perform poorly in discriminating coded speech sentences badness. One drawback of using such measures to evaluate speech coders is their averaging nature that disqualifies them from pinpointing isolated large errors in reconstructed speech that introduce considerable perceptual distortion.

The spectral envelope distortion measures are much more effective when employed for codebook vector selection. These measures are used in the domain in which they

are defined, although the LPC coefficients can be transformed from one representation to another. While keeping a given number of fixed spectral envelopes in a codebook, the different distortion measures were employed for LPC vector quantization in [49] for encoding a speech database. A difference of 2.5 dB in SNR was found between the best and the worst quantizing schemes. The most promising distortion measure was shown to be the weighted Euclidean LSF distance. On the basis of these conclusions, the next section will detail scalar LSF quantization schemes first, and then formally introduce the LSF vector quantizer implemented in this work.

3.4 Quantization of LPC Parameters

3.4.1 Transparent Quantization of Parameters

The aim of every LPC quantizer is to achieve *transparent* quantization of the parameters. By transparent quantization, it is meant that no additional audible distortion is added to the coded speech: the reconstructed speech using the unquantized LPC parameters and the version obtained by using the quantized parameters should be indistinguishible to the ear. Subjective evaluation, while being the most effective for evaluating the performance of quantizers, is not very convenient during the design stage. Objective criteria complying to the requirements of transparent quantization are needed. Spectral distortion measures have been conventionally used to evaluate the performance of quantization. It is defined as the root mean square difference between the original LPC log-power spectrum and the quantized LPC log-power spectrum. An average of 1 dB spectral distortion has been traditionally considered to be the threshold for transparent quantization. However, isolated outliers (speech frames recording spectral distortion greater than 1 dB) having large spectral distortion disrupted the perceptual quality of the coded speech, despite spectral distortion averages below 1 dB. An additional requirement along with the average spectral distortion will therefore be the minimization of the number of outliers. The formal characterization of transparent quantization, as suggested in [41], is to (a) guarantee an average spectral distortion of about 1 dB, (b) have absolutely no outlier frames having spectral distortion larger than 4 dB, (c) keep less than 2 % the number of frames having spectral distortion in the range 2-4 dB.

A 10-th order LPC analysis on 20 ms speech frames will be the common ground for all the quantizers that are evaluated in the remainder of this section. This yields a transmission rate of 50 frames/s of LPC parameters. A number of studies have used the LSF's as a representation for scalar quantization of LPC parameters [30,45]. It was found that 32 to 40 bits per frame were needed to achieve transparent quantization. This is prohibitively expensive at medium and low bit rates. The alternative for bit rate reduction is vector quantization (VQ) of the LPC parameters. A 10 bits/frame VQ scheme comparable in performance to a 24 bits/frame scalar quantizer was proposed in [50]. The average spectral distortion of this scheme was however of about 3.3 dB, clearly insufficient for high quality speech coding. Allocating more bits to vector quantization implies larger codebooks. A larger set of training data is also required which increases the complexity of the training process, not to mention the expensive storage and computational requirements in encoding the parameters. Transparent quantization has to be then reached using suboptimal VQ schemes.

Tree-search and product VQ are examples of suboptimal vector quantizers. Hybrid vector-scalar quantizers, with either cascaded or coupled vector and scalar quantizers, try to overcome the complexity of simple VQ schemes in [49] and [51]. At rates around 30–32 bits/frame, the average spectral distortion was reduced below 1 dB. Product-VQ, on the other hand, was efficiently exploited in [40] with the introduction of a cascaded VQ LPC quantizer. The LPC spectrum is decomposed in this scheme into a low-frequency and a high-frequency spectra. This is achieved by decomposing the LPC polynomial into one polynomial defined by the 6 lower frequency roots, and another one characterized by the remaining 4 higher frequency roots. The two resulting lower order LPC vectors were jointly quantized using a log likelihood ratio distance measure. A 26 bits/frame version of this cascaded VQ scheme yielded a 1.1 dB average spectral distortion.

The implemented scheme in this thesis is a variation of the Split VQ proposed in [41], also based on the product-VQ concept. The LPC parameters, in a suitable representation, are split into two or more lower order vectors and independently vector quantized. Splitting the LPC vector into 10 parts evidently results in scalar quantization of the parameters. For this scheme, a suitable parametric representation for the LPC coefficients has to be selected, as well as a proper distance measure. The

objective is the achievement of transparent quantization at a rate of 24 bits/frame.

3.4.2 Vector Quantization of LSF's

Rather than considering the parameters as separate quantities to be quantized (scalar), vector quantizers consider the entire set of LPC parameters for one frame as a single entity which enables a direct minimization of the spectral distortion. Smaller quantization distortions result in vector quantization when compared to scalar quantization at a given bit rate. Seen from another viewpoint, the existing correlation between the LPC parameters for one frame of speech is exploited in VQ, allowing thus bit rate reduction in LPC parameters quantization. Conceptually, vector quantization consists in finding from a codebook of pre-determined trial LPC coefficients vectors the vector that "matches" best the set of LPC coefficients computed for a frame of speech. Once this codevector is found, its index is transmitted to the decoder which contains the replica of the quantization codebook.

The composition and the size of the codebook are issues that largely affect the performance of VQ. The perceptual distortion measure used as a selection criterion can also greatly influence the accuracy of quantization. These vector quantizers design parameters are now briefly exposed before detailing the Split VQ quantization scheme and evaluating its performance.

Codebook Design

As more bits are allocated to the quantization of LPC parameters, larger codebooks can be designed, increasing thus the probability of finding better perceptual matches to a given original LPC vector. Large codebooks however entail, as mentioned previously, expensive computational and storage requirements in both their training and their use for encoding. A practical suboptimal VQ technique will be seen shortly.

A large database is usually required for the training of the codebook, at least several times larger than the intended codebook size. The codebook training in this work is based on the conventional Linde Buzo and Gray (LBG) algorithm [52]. Denoting the LPC vectors by $\mathbf{v}$ represented in a 10-dimensional space for our case, the flow of the algorithm is given below:

1. The centroid $\mathbf{c}$ of the training data is computed

2. The centroids (one centroid initially) are split in two by slightly perturbing their components

3. The training data is clustered around the closest new centroid, using the Euclidean distance measure $\frac{1}{2}(\mathbf{v}-\mathbf{c})^T(\mathbf{v}-\mathbf{c})$

4. The new centroid of the clustered data is determined

5. If the new centroids do not register a distortion below a given threshold, the data around the new centroids is re-clustered

6. Go back to step (2) until the desired codebook size is reached

Centroid splitting in the LBG algorithm can sometimes lead to LPC vectors that yield unstable synthesis filters. Examples of such unstable vectors can be reflection coefficients of magnitude greater than one, or not properly ordered LSF components. From a given centroid

$$\mathbf{c} = [c_1, c_2, \ldots, c_{10}], \tag{3.22}$$

the newly generated centroids are obtained by perturbing the components of $\mathbf{c}$ with the value ε, set around 0.005, according to:

$$\begin{aligned} \mathbf{c}' &= [(1+\varepsilon)c_1, (1-\varepsilon)c_2, \ldots, (1-\varepsilon)c_{10}], \\ \mathbf{c}'' &= [(1-\varepsilon)c_1, (1+\varepsilon)c_2, \ldots, (1+\varepsilon)c_{10}]. \end{aligned} \tag{3.23}$$

After several splittings the LSF centroid could loose the well-orderness principle if the i-th coefficient $c_i = l_i$ increases continually while $c_{i+1} = l_{i+1}$ keeps decreasing. Such unstable vectors should be removed from the codebook in order to guarantee stable reconstruction filters at the receiver. Some LPC parameter representations are not suited for vector quantization. The centroid of prediction coefficients can directly lead to unstable synthesis filters upon splitting. Furthermore, the Euclidean distance used for clustering the data around the centroids in the LBG algorithm does not display the same behaviour with the different parametric representations of the LPC parameters. Because of the limited frequency range of variance of each LSF and the direct relationship between the speech spectral energies and the LSF's spacing, such parameters lend themselves better to Euclidean distances than prediction or reflection coefficients. The choice of the LPC parametrization domain will affect the performance of the vector quantizer.

Selection Criterion

Evaluating the distortion between the energy spectral envelope of a speech frame and the trial codebook LPC spectral envelopes in the selection process is largely dependent on the domain of representation of the LPC parameters. While simple Euclidean distance measures exhibit a satisfactory performance with LSF's, they fail to emphasize the perceptual nature of the quantization error. The Itakura-Saito log-measure has a greater perceptual impact on the best LPC vector selection criterion with the inclusion of weights as functions of frequency. The inadequacy of the Euclidean distance as a distortion criterion in some LPC representations highlights the importance of using distortion measures relevant to the LPC parametrization domain.

Nevertheless, studies in [41] and [49] have selected the Line Spectral Frequencies to yield the best vector quantization performance under any distortion criterion. The reason for this distinction derives from the close relationship between the LPC spectral envelope and those parameters. The localized spectral sensitivity property displayed in Fig. 3.2 essentially allows one to weight the LSF's individually. A comparative study between a vector quantization scheme using the Euclidean LSF distance measure and one based on the weighted Euclidean LSF distance measure is completed in the next section. in the scope of the product-code vector quantizer implemented in this work.

Split Vector Quantization

Product-codebook techniques have contributed to the reduction of the computational complexity of vector quantizers. In such schemes, independent vector quantization of LPC sub-vectors using smaller size codebooks is carried out. Recent developements have followed this stream by splitting the LPC power spectrum into a lower frequency spectrum, more emphasized than a higher frequency spectrum [40]. Two lower order prediction coefficient vectors could then be vector quantized in any suitable parametric representation. Splitting the LPC spectrum is straightforward if the LPC parameters representation is based on the Line Spectral Frequencies. Each cluster of LSF's characterize a spectral frequency region (*cf* Section 3.2.2). With at most three LSF's corresponding to a spectral formant region, The first four LSF's could (the most important perceptually) constitute an LPC sub-vector of parameters modeling

the lower frequency part of the spectrum, including most of the time two formants. The remaining six LSF's would then be grouped together as another subvector taking care of the remaining LPC spectral characteristics. This product-codebook scheme, known as Split Vector Quantization (Split VQ), requires the design of two independent codebooks, in a four and six dimensional spaces successively.

Taking into consideration the fact that most of the speech energy is contained in the first spectral formants, the same number of bits will be allocated to the quantization of the first four components LSF vector and to that of the six components LSF vector. At a rate of 24 bits/frame, each codebook will have 4096 entries. The codebooks are designed through the use of the conventional LBG algorithm. The training database consists in about 10 minutes of english and french sentences spoken by different male and female speakers. The speech data is lowpass filtered at 3.4 kHz and sampled at 8 kHz. A 10-th order LPC analysis yields two sets of LSF sub-vectors, updated for every speech frame of 20 ms.

Instability of the all-pole speech reconstruction filters is avoided by ensuring that the first four LSF's in every entry of the first codebook are in increasing order, and similarly for the six remaining LSF's for the second codebook. However, the splitting procedure might lead to potential cross-overs of the fourth and the fifth LSF's belonging to the optimally selected codebook subvectors. Fig. 3.4 is a plot of the 4-th LSF values from the first codebook codesubvectors versus the 5-th LSF values from the second codebook codesubvectors. As can be seen most of the 5-th LSF values are above the $l_4 = l_5$ line. However, there are very few occasions where the first selected subvector has its fourth LSF component greater than the first component of the second selected subvector (5-th LSF). Many LSF cross-over correction methods have been proposed in previous works [48,49]. The simplest correction technique to avoid synthesis filter instability will be adopted in this implementation. It consists of swapping the values of the 4-th and the 5-th LSF's in order to reinstate the increasing orderness property of the LSF's. Subsequent checkings and swappings might be needed to ensure stability of the quantized LSF vector (3-rd LSF with the new 4-th LSF value for example).

The performance of the Split VQ scheme is evaluated from a set of english sentences spoken by male and female individuals not included in the training set, and

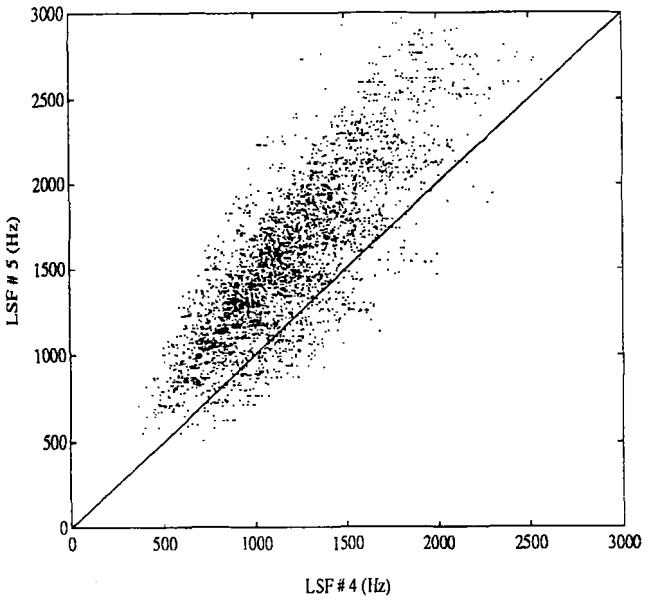
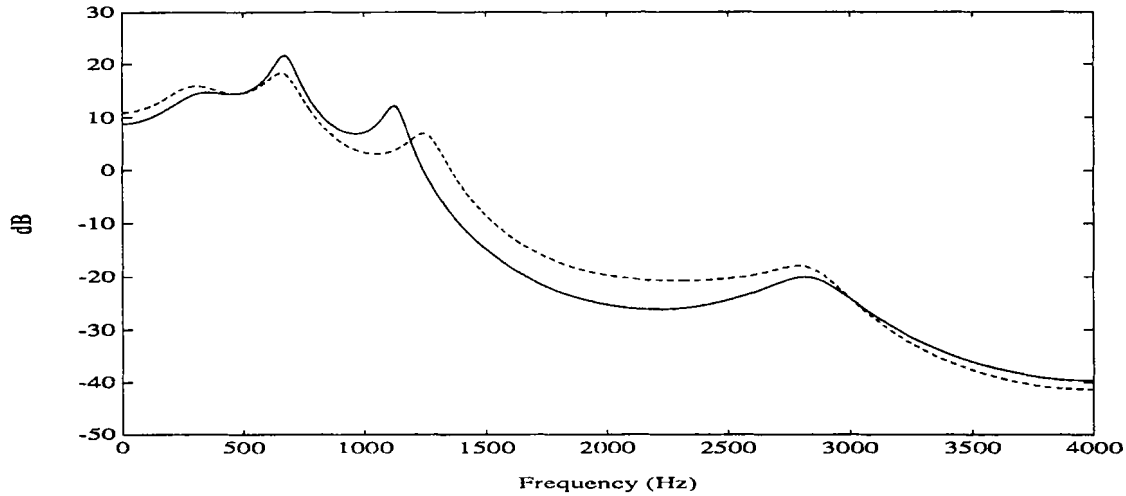


Figure 3.4: 4-th LSF values from the first codebook vs 5-th LSF values from the second codebook.

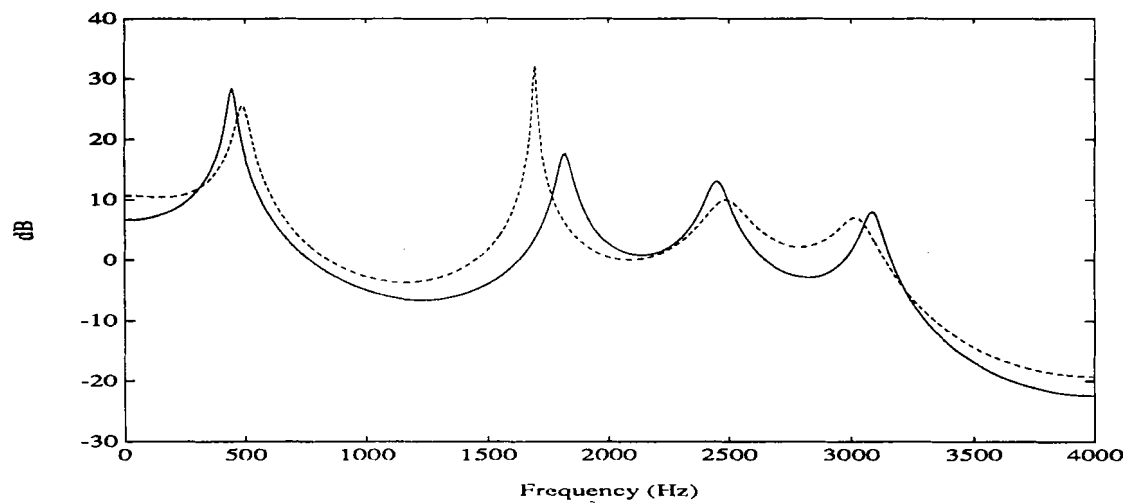
referred to as the test data. Both the Euclidean and weighted distance measures will be investigated.

Performance Evaluation

Objective and subjective evaluation of the performance of Split VQ is carried out in this section. Two versions of reconstructed speech will be compared, one based on the original LSF vector and the other on the selected entry from both codebooks, keeping this way similar the comparison conditions. Fig. 3.5 displays the original LPC power spectrum with, superimposed, the quantized LPC spectrum for two frames of male and female speech from the test data. The distortion criterion is selected here to be the Euclidean LSF distance measure, where all the LSF's are assigned equal weights. The LSF vectors for the same female and male speech frames are now quantized with the split VQ scheme using the weighted Euclidean LSF distance measure introduced in Section 3.3.2. Fig. 3.6 shows a reduced spectral distortion in the quantized LPC spectral envelopes. The emphasis that weighting puts on the LSF's corresponding to formant frequencies translates into a finer quantization around formants in both male

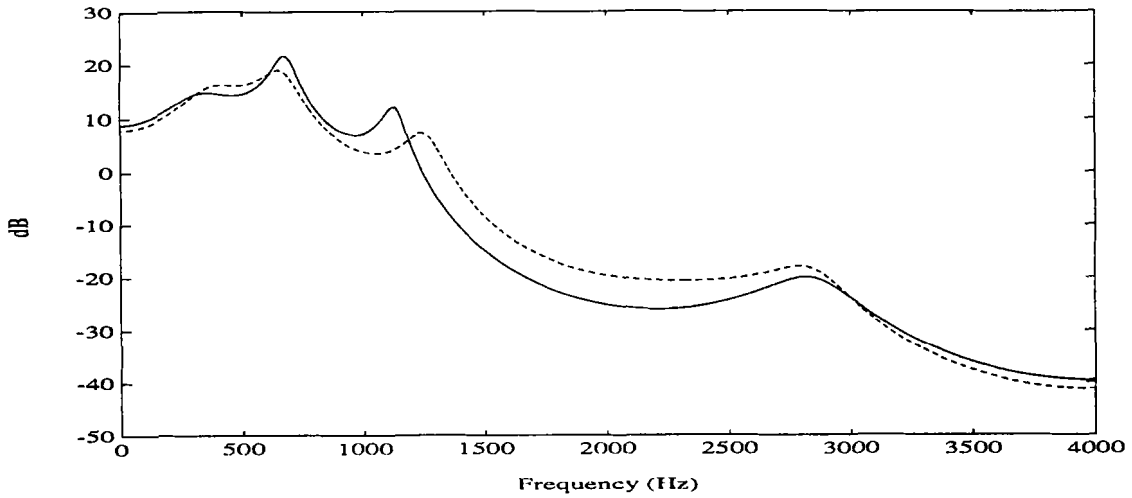


(a) Female

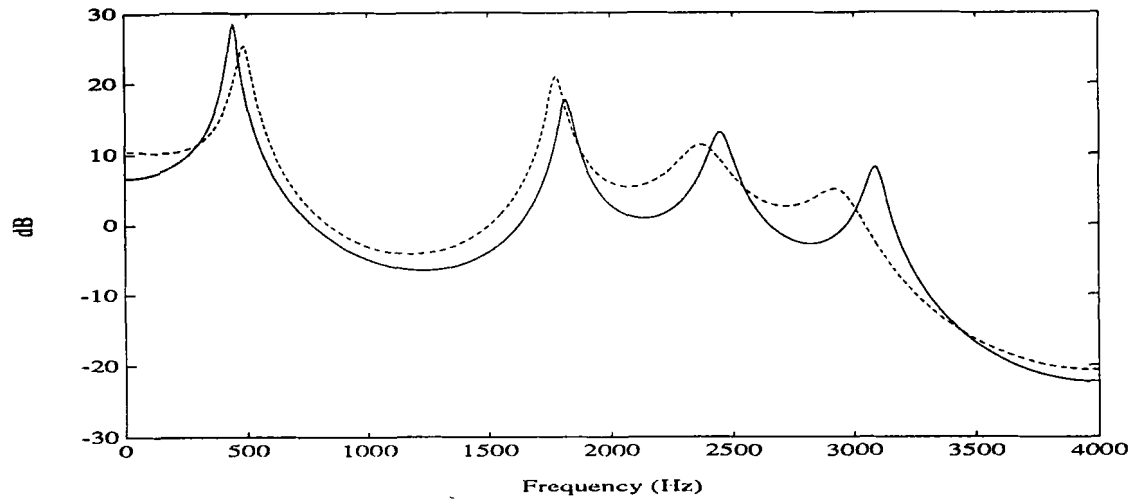


(b) Male

Figure 3.5: LPC power spectra for the set of unquantized LSF's (solid line) and quantized LSF's (dashed line). The illustrated split VQ spectral distortion is for two 20 ms female (a) and male (b) speech frames, with the Euclidean LSF distance used as a distortion measure.



(a) Female



(b) Male

Figure 3.6: LPC power spectra for the set of unquantized LSF's (solid line) and quantized LSF's (dashed line). The illustrated split VQ spectral distortion is for the 20 ms female (a) and male (b) speech frames of the previous figure, with the distortion measure being now the weighted LSF distance measure.

Distortion measure	AvSD (dB)	% of Outliers		AvmaxSD (dB)
		2-4 dB	> 4 dB	
Euclidean	1.26	3.3	0.13	3.61
Weighted Euclidean	1.09	2.6	0.07	3.42

Table 3.1: Performance of split VQ operating at 24 bits/frame. Spectral distortion values are given for the Euclidean and weighted Euclidean LSF distance measures used as distortion criteria.

and female speech spectra, at the expense of a coarser spectral matching in valleys. As can be seen, both Euclidean and weighted Euclidean measures record the same perfomance in spectral valleys, but the superiority of the weighted selection criterion reveals itself at the lower frequencies and around spectral peaks.

Table 3.1 reports the objective results necessary to characterize transparent LPC parameters quantization at an operating rate of 24 bits/frame. The average spectral distortion (avSD), the maximum average obtained spectral distortion (avmaxSD), the percentage of outlier frames recording spectral distortion between 2 and 4 dB and greater than 4 dB are all given for the split VQ scheme using successively the Euclidean and the weighted Euclidean LSF distance measures. With the help of the weighting in the LSF distance measure, transparent LPC parameter quantization is achieved, with an average spectral distortion around 1 dB. It is reported in [41] that the effect of weighting is to reduce the bit rate by 2 bits per frame, i.e a 26 bits/frame Euclidean distance measure based split VQ would yield the same performance of the split VQ used in this work. Also, from the performance of other LSF quantizers reported in literature, the recorded spectral distortion of 1.1 dB is attained by 32 bits/frame scalar quantizers, 30 bits/frame hybrid vector-scalar quantizers and 26 bits/frame cascaded VQ schemes.

The subjective performance of split VQ has also been tested. A set of four male and female sentences were coded using both the unquantized and the vector quantized LSF parameters. After listening to the coded pairs presented to the listener in random order, it was concluded that no difference could be distinguished. Transparent quantization quality is therefore achieved with a 24 bits/frame split LSF vector

quantizer.

3.5 Interpolation of LPC Parameters

In a speech coder, the LPC parameters are quantized and transmitted on a frame-by-frame basis, with an update rate usually around 20 ms. However, if those parameters are kept fixed for the frame duration, large changes in the filter coefficients at the frame boundaries can lead to audible distortions (clicks or pops). Unvoiced/voiced transition regions are examples of such discontinuities. In order to achieve a smoother transition of the filter characteristics at the frame boundaries, overlapped analysis frames and LPC parameters interpolation are widely employed.

A speech frame is usually divided into 4 or 5 subframes of duration ranging between 2.5 and 7.5 ms. Instead of performing the LPC analysis solely on the present frame, one or two subframes belonging to the past already encoded frame could be incorporated in the analysis data. In this manner the transition regions for the LPC parameters are rendered smoother. Look ahead techniques are also sometimes applied, taking into consideration a future subframe of data in the computation of the LPC parameters for the present frame [29]. Special care should however be taken in order not to induly increase the coding delay.

Interpolation of the LPC parameters provides a major contribution to the smoothness of the synthesis filter characteristics. Fixed interpolation schemes have been traditionally used where a weighted combination of the past frame LPC parameters and those of the present frame is individually assigned to every subframe in the present frame. A dynamic linear interpolation scheme has however recently been introduced in [53], where the LPC analysis is performed twice for one frame, and the extra LPC parameter set is used as middle values to determine the slope of the interpolation line. Extra bits must in this case be allocated for the transmission of the interpolation slope information to the decoder. along with the past and recent LPC parameter vectors. The interpolation is thus performed on a subframe basis, with new interpolated values generated every 2.5 to 7.5 ms.

Not many results have been published on the LPC parameter representation to be selected for satisfactory interpolation results. Experience has however concluded

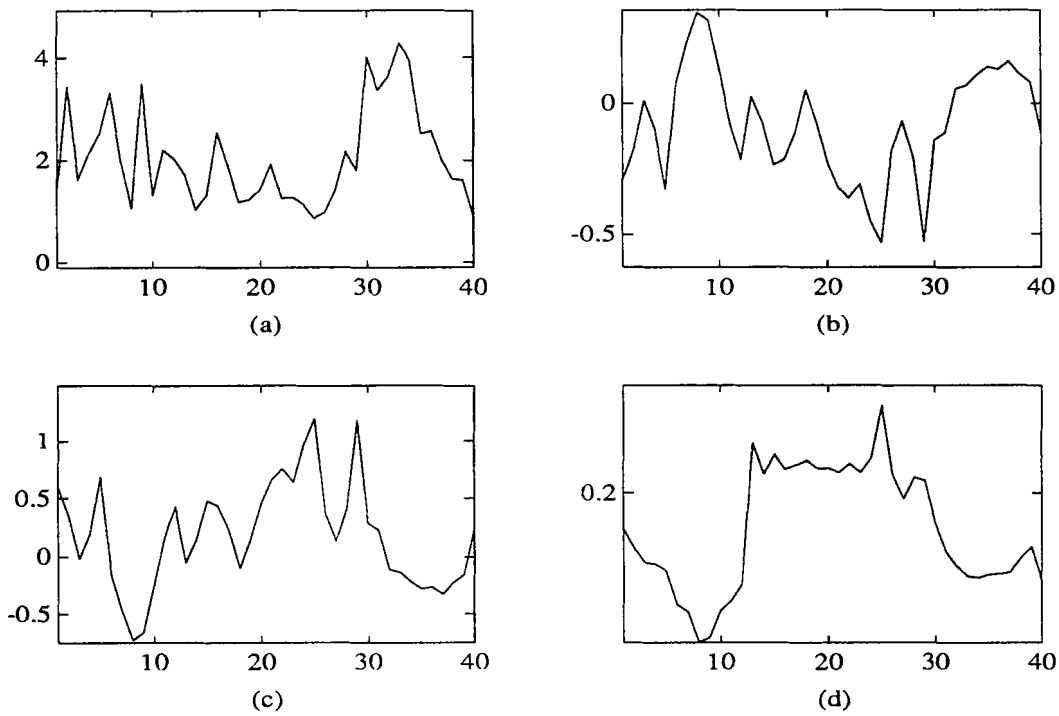


Figure 3.7: Frame-to-frame evolution of the fifth (a) predictor coefficient, (b) reflection coefficient, (c) log area ratio coefficient and (d) LSF for 40 analysis speech frames.

that direct interpolation on the predictor coefficients can lead to unstable filters and is usually avoided. The LPC parameters are generally transformed to other domains more suitable for fixed interpolation schemes. Fig. 3.7 displays the dynamic behaviour of four LPC parameter representations for 800 ms of speech. The subplots retrace sequentially the frame-to-frame variation of the fifth (a) predictor coefficient, (b) reflection coefficient, (c) Log area ratio coefficient and (d) LSF. It can be seen that the LSF representation yields the smoothest frame-to-frame variation and thus is very often the basis for LPC parameter interpolation. The log area ratios are also sometimes interpolated, as well as the autocorrelation coefficients. In fact experiments in [54] suggested that there was no significant perceptual quality difference in a CELP coder environment operating at 8 kb/s when interpolation was performed in the above various domains. The LPC parameter update interval of the coder was of 16 ms. For longer update intervals (25 ms and above), the LSF's tend to show a more smoother

evolution, precisely the reason for which they are employed for interpolation in the Federal standard (Fed-1016) 4.8 kb/s CELP coder [30] over 30 ms speech intervals. The implemented interpolation scheme is also based on the LSF representation of the LPC parameters.

The coding delay for the 8 kb/s CELP coder implemented in this thesis is 20 ms, corresponding to one speech analysis frame. However, in order to maintain the continuity of the LPC parameters, the analysis frame and the actual speech frame being coded do not match. Fig. 3.8 Shows how both the analysis frame and the speech frame being processed overlap. At a sampling rate of 8 kHz, 20 ms correspond to 160 samples. The frame to be coded is further divided into 5 subframes of 32 samples each (4 ms). The LPC parameters for the first subframe are entirely based on the analysis performed on the past frame, while the remaining four subframes carry interpolated LPC parameters. The analysis frame encloses the four last subframes of the frame being coded and one extra subframe from the next-in-line frame to be processed. The LSF's of the j^{th} encoded subframe are a weighted combination of the past analysis frame LSF's and the present analysis frame computed LSF's. With ω_j being the set of interpolation weights, the i^{th} LSF for the j^{th} subframe, $\tilde{l}_i^j$ is obtained by:

$$\tilde{l}_i^j = \omega_j l_i^{(k-1)} + (1-\omega_j) l_i^{(k)}, \tag{3.24}$$

where k represents the time index of the current analysis frame. Fig. 3.8 provides the subframe weights distribution. Since the decoder has available to it the transmitted LSF's for the past analysis frame, the first subframe can be reconstructed upon receiving the excitation codebooks index, even before the newly computed LPC parameters (present analysis frame) are sent. In this manner, the coding delay is not increased beyond 20 ms. For the remaining subframes, the present LSF's are required at the decoder end in order the compute the subframe interpolated LSF values. Both subjective and objective measures record an improved performance with the described interpolation scheme in a full coder environment.

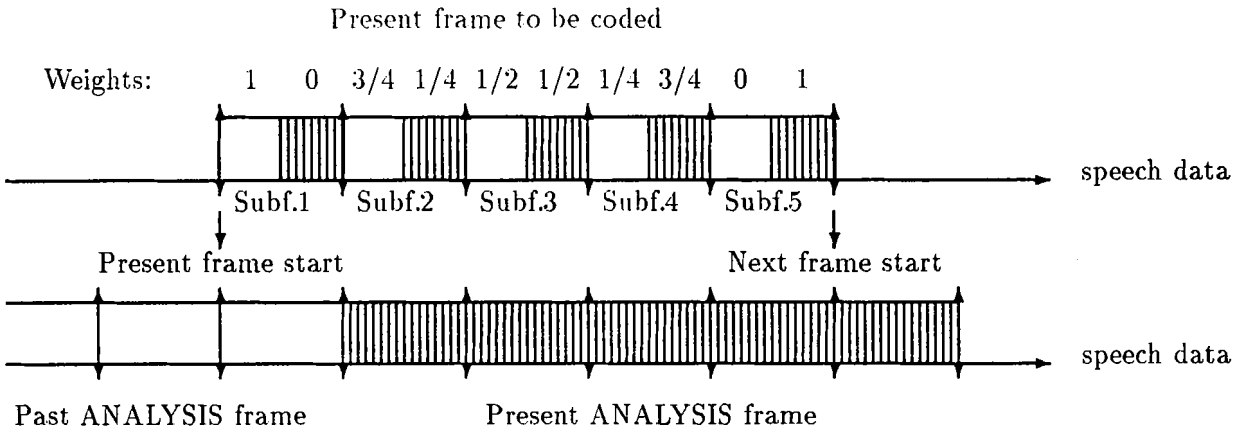


Figure 3.8: Analysis frame overlap and interpolation scheme applied to the subframe LSF's.

3.6 Conclusion

A suitable parametrization of the LPC coefficients was introduced in this chapter for both quantization and interpolation purposes. The line spectral frequencies (LSF's) exhibit many properties related to the LPC spectral envelope and to stability issues that make them attractive for quantization. Although other LPC parameter representations display a good behaviour in various quantization schemes, the LSF's are especially suited for vector quantization, in which the selection criterion amounts to minimizing a perceptually weighted LSF distance measure. This distortion measure takes advantage of both the frequency location of the LSF's determining the speech spectral peaks and valleys, and the human ear resolution along the frequency scale. Vector quantization of LPC parameters allows greater bit rates reduction than scalar quantization since it exploits the inter-correlation that exists among the LPC vector components, but the computational cost quickly grows with the codebook required size in order to achieve high-quality coding. A suboptimal vector quantization scheme yielding transparent coding of LPC parameters was presented and performance results were reported. Deriving from product-codebook vector quantization techniques, Split vector quantization decomposes the LPC energy spectrum into a lower frequency spectrum and a higher frequency spectrum. The original LSF vector to be quantized is split into a four-LSF subvector and a six-LSF subvector quantized independently

using the weighted LSF distance measure as a distortion criterion. Two codebooks are initially trained according to the conventional LBG scheme, each ending up with 4096 codevector entries. The operating rate of the split VQ scheme is therefore of 24 bits/frame. The average spectral distortion of the split VQ was around 1 dB, complying thus with the transparent quality coding requirements. Moreover, it turned out to be consistently better performing objectively and subjectively than a split vector quantizer using the simple Euclidean distance as a distortion criteron. At an update rate of 50 frames/second, the computational complexity of the implemented split VQ reaches 4 million operations per second. Further complexity reduction can be attained by splitting the LSF vector into more than two parts and reallocating the bits among the resulting codebooks. This however cannot be completed without any performance degradation.

Updating the LPC parameters on a frame basis can sometimes lead to discontinuities in the predictor coefficient values. To avoid such circumstances, the LPC parameters are interpolated on a subframe basis, i.e. for intervals shorter than the LPC update frame. Once again, a comparative study between different LPC parameter representations revealed that the LSF are very suitable for interpolation in view of their quasi-smooth frame-to-frame variation. The proposed interpolation scheme yields subframe LSF values obtained as a weighted combination of the past analysis frame and the present analysis frame LSF's. The weights are predetermined and kept fixed in the coding scheme. Moreover, to guarantee better speech spectral characteristics transitions, the analysis frame includes, in addition to a main portion of the present speech frame to be coded, a subframe of speech samples to be coded in the next frame. The fixed interpolation scheme combined with the look-ahead capabilities of the LPC analysis stage yields a better subjective and objective performance (higher SNR) of the overall coder.

Chapter 4

Pitch Prediction in CELP Coding

4.1 Introduction

Pitch prediction in linear predictive coding schemes is a powerful method to represent the periodicity in speech signals. Long term predictors are usually described by parameters representing the delay and by filter coefficients. In a CELP coding scheme, the pitch prediction parameters are more efficiently optimized in a closed-loop manner during the analysis-by-synthesis procedure. Along with the long term predictor parameters, the CELP codebook index and gain have to be also selected. Optimization schemes that jointly select the pitch predictor and the codebook parameters which minimize a weighted error criterion perform consistently better than a sequential optimization choosing first the prediction parameters and then the codebook index and gain. An overview of the synthesis parameters optimization schemes is presented in this chapter. From there, a combined optimization procedure joining both the reduced computational complexity of the sequential approach and the efficiency of the joint approach is proposed.

Multiple pitch predictor coefficients allow long term predictor delay interpolation in certain high energy regions of the speech spectrum for periodicities that are not an integer multiple of the sampling frequency. The coefficients of multiple-tap long term predictors are shown to be frequency dependent, emphasizing the lower frequency spectral regions and compensating for the prediction inaccuracies in the higher frequencies. They exhibit an improved performance over single-tap predictors, but their

transmission bit rate requirement is very expensive. Pitch prediction gain increases with increasing sampling rate. Fractional delay single-tap pitch predictors take advantage of this result to record performances higher than 3-tap pitch predictors. In such predictors, the delay is specified as an integer number of samples plus a fraction of a sample at the current sampling frequency. An efficient implementation technique for the interpolation between samples is presented here, followed by a fractional delay long term predictor design, with resolution up to 1/6 of a sample, operational in an 8 kb/s CELP coding scheme. Subjective and objective performance results are reported.

4.2 Pitch Prediction in CELP Coders

The incorporation of Long Term Predictors (LTP) in linear prediction based analysis-by-synthesis coders has greatly contributed to increasing the quality of the coded speech signal. The LTP, also known as the pitch predictor, was already introduced in Chapter 2 as a technique to generate periodicity in the reconstruction of voiced speech. A large part of the success of the CELP coding algorithm at rates between 4 kb/s and 10 kb/s can in fact be attributed to the linear pitch prediction capabilities included in the coder. Taking advantage of the analysis-by-synthesis configuration of the class of linear predictive coders considered, the parameters of the pitch predictor are usually updated in a closed-loop fashion. This closed-loop optimization procedure was originally introduced to enhance the performance of a multipulse linear predictive coding scheme [23]. The operation of a pitch predictor in a CELP coding environment follows the same model: a selected codebook excitation vector drives a *pitch reconstruction* filter (periodic structure is added) to yield a periodic excitation. This resulting excitation is fed in turn to the all-pole synthesis filter producing the reconstructed speech waveform by adding the formant structure of the speech frame being coded. A clear distinction must be made between the first excitation vector, termed LTP excitation, and the one used for synthesis, called LP (linear prediction) excitation.

The pitch reconstruction filter is commonly represented by an Auto-Regressive (AR) all-pole model with either one or multiple filter coefficients. The parameters of

a LTP are therefore the filter delay d, closely related to the pitch lag of the current speech frame, and its coefficients, $\{\beta_j\}$. If the all-pole pitch reconstruction filter is implemented as a transversal structure, one can then distinguish single-tap and multiple-tap LTPs. The latter form enhances the periodicity of the coded speech at the expense of a greater number of bits that have to be allocated for the quantization of the multiple coefficients. A very efficient variation of the single-tap pitch predictor is the *fractional delay* LTP, where the filter delay resolution is increased to less than a sample. The performance of such fractional delay filters is comparable to three-tap predictors for update intervals less than 10 ms, with the added advantage that no extra bits are needed to transmit more than one coefficient. The increased resolution of the delay must however be encoded and transmitted, requiring a smaller additional bit number.

In the framework of analysis-by-synthesis coders, such as CELP, the LTP excitation is determined on a blockwise basis. These blocks, called subframes, are usually much shorter than the LPC analysis update frame, since the pitch information varies more rapidly than the formant structure (*cf* Section 2.2). The analysis-by-synthesis loop in the CELP algorithm, detailed extensively in Section 2.7.1, proceeds in generating the LTP excitation by appropriately scaling an optimally selected signal vector from a codebook of fixed entries. This LTP excitation is then "pitch" synthesized to yield an LP excitation with periodic structure. The synthesis parameters that need to be optimized and transmitted are thus the fixed excitation codebook index i, the codebook scaling factor or gain G, the LTP delay d, and the LTP coefficients $\{\beta_j\}$. Joint optimization of all parameters gives the best coding performance, but the extensive fixed excitation codebook search while optimizing the LTP parameters is very expensive computationwise. An important complexity reduction is possible whenever the minimum LTP delay d is set to be greater than the subframe length. In this case, the LTP contribution to the current LP excitation depends only on the past LP excitation and therefore is independent of the current LTP excitation (scaled codebook selected entry). A sequential optimization procedure is then applied, where the periodic contribution to the LP excitation is determined first assuming a zero LTP excitation. Once the LTP optimal delay and coefficient values are obtained, the current LP excitation is further improved with the optimal LTP excitation selected

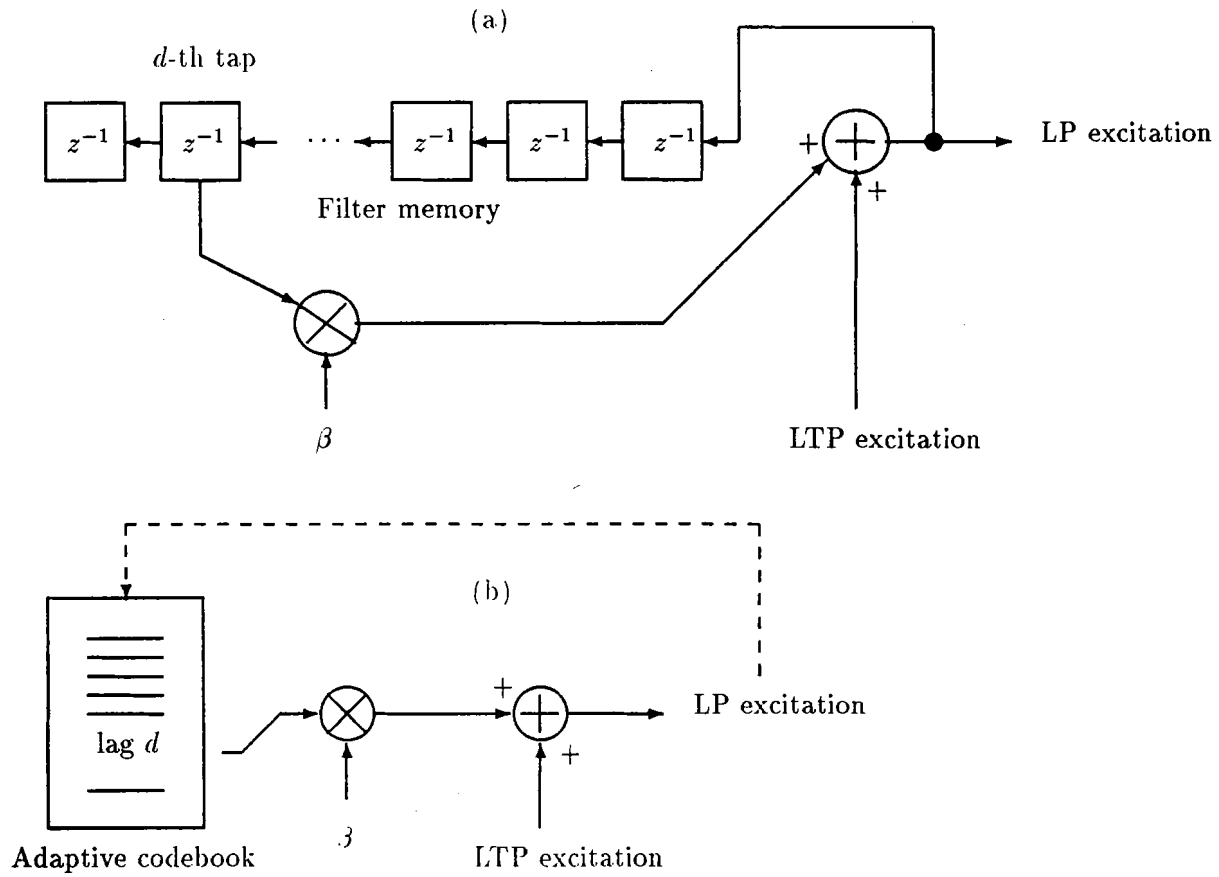


Figure 4.1: Transversal filter structure (a) and adaptive codebook (b) representations of the one-tap long term predictor.

from the codebook and scaled by G. In the case of a one-tap pitch predictor, the LTP contribution to the LP excitation can be viewed as a past delayed version of the LP excitation scaled by the filter coefficient β. The past LP excitations can be stored in an *adaptive* codebook where each entry differs by a shift of one sample. The long term predictor contribution is therefore obtained by selecting the optimal entry in the adaptive codebook and scaling that entry by the LTP coefficient. Fig. 4.1 displays both the transversal filter structure and the adaptive codebook representations of the one-tap pitch predictor. The optimization of the synthesis parameters becomes a two stage codebook entry selection, where the first codebook is adaptive and the second is fixed.

The LTP delay d gives in essence an estimate of the pitch lag of the recently coded speech subframe. Open-loop correlation techniques for pitch detection can be applied on the current subframe in order to obtain an estimate of the LTP delay. The analysis-by-synthesis procedure consisting of selecting out of all permissible delays the one that maximizes a certain periodicity measure is however much more efficient in a CELP coding scheme. Constraining the LTP minimum delay to be greater than the subframe length suffers from limitations especially in the case of female speech (average pitch period of 35 speech samples). The delay will in effect assume pitch doubled and tripled values on many occasions. Remedies to this problem consist in allowing the LTP delay to take values smaller than the subframe size and to recycle the current LP excitation through the pitch filter, or to include periodic extensions of a pitch cycle in the adaptive codebook. The adaptive codebook method, employed in [30], is however not equivalent to LTP filtering since the filter memory update is performed on recycled LTP excitations while the codebook update only uses concatenated LP excitation sequences. The degradations in perceptual quality are nevertheless minor and the latter technique will be employed.

At low coding rates, the LTP performance degrades as it becomes harder to recreate a smooth evolution of the pitch cycle waveform. The perceived periodicity in voiced segments of the reconstructed speech hence decreases. The simple AR model for the LTP might therefore be unable to reproduce with fidelity the pitch cycles of the original speech at bit rates dropping below 5 kb/s. Recent work has indeed addressed the limitations of the LTP by enhancing the periodicity of the coded speech either by increasing the correlation between adjacent pitch cycles [33] or by applying a harmonic noise weighting scheme to the CELP error criterion [34]. Some of these techniques are discussed in the next chapter.

The LTP parameters are transmitted at every subframe, requiring on average 12 bits per subframe, which corresponds to rates around 4 kb/s in the common CELP implementations. Bit savings can be obtained by encoding only the offset from the previous delay every other subframe [30] or by using differential encoding techniques [7]. A recent LTP delay interpolation technique described by piecewise linear delay contour trajectories, introduced by Kleijn [4], enabled the transmission of the LTP parameters once every few subframes and the interpolation parameters in between.

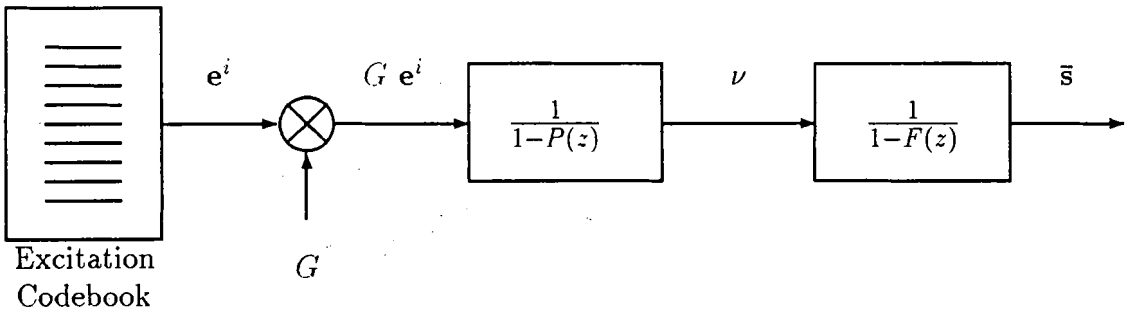


Figure 4.2: Basic CELP decoder.

An outline for the optimization of the synthesis parameters in a CELP analysis-by-synthesis loop is detailed in the next section, followed by implementational considerations for the case of one-tap LTP delay values smaller than the subframe length.

4.2.1 Analysis-by-Synthesis Model

An equivalent model to the two-codebook approach for the basic CELP decoder described in Chapter 2 is given in Fig. 4.2. The selected codebook excitation vector, $\mathbf{e}^i$, is scaled by an optimal gain G to form the LTP excitation vector. A periodic structure is added to this vector after passing it through the pitch all-pole synthesis filter:

$$\frac{1}{1-P(z)} = \frac{1}{1-\sum_{j=-M}^{M}\beta_j z^{-d-j}}, \tag{4.1}$$

where d and $\{\beta_j\}$ are respectively the optimal LTP delay and the set of $(2M+1)$ LTP coefficients. The outcome of the pitch synthesis filter, ν, forms the LP excitation vector, which is then passed through the all-pole synthesis filter $\frac{1}{1-F(z)}$ associated with the LPC linear predictor. Once the formant structure is added to the LP excitation, the reconstructed speech vector, $\bar{\mathbf{s}}$, is obtained.

In analysis-by-synthesis coders, a replica of the decoder is incorporated in the encoder in order to allow a direct comparison between the original and the reconstructed speech signals. In the CELP coding algorithm, it was however seen that instead of directly comparing the original and reconstructed speech signals, a per-

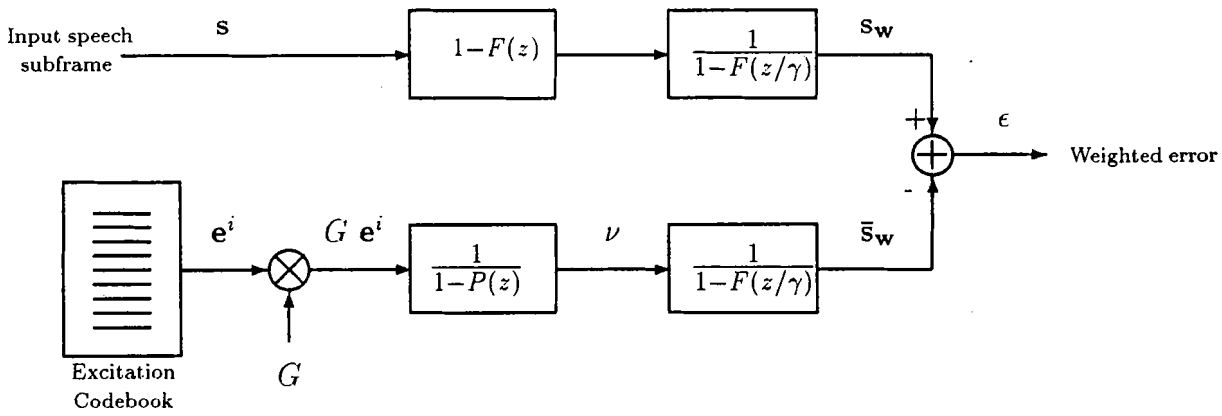


Figure 4.3: Analysis-by-synthesis loop in the CELP algorithm.

ceptually weighted version of these signals is used. The analysis-by-synthesis loop minimizes thus a spectrally weighted error criterion for optimizing the synthesis parameters. The spectral weighting filter $W(z)$, repeated here for convenience, is

$$W(z) = \frac{1 - F(z)}{1 - F(z/\gamma)}, \tag{4.2}$$

with γ being the bandwidth expansion factor (usually around 0.8), relocates the coding distortions around the formant regions where they are masked by the higher speech signal energy. Fig. 4.3 depicts the analysis-by-synthesis structure of the basic CELP coder after appropriately combining the spectral weighting filter with the formant reconstruction (synthesis) filter. The resulting spectrally weighted synthesis filter, $\frac{1}{1-F(z/\gamma)}$, is closely related to the speech reconstruction filter. It is thus of time-varying nature, with its coefficients $\{\gamma^k a_k\}$ being updated as discussed in Chapter 3. The other parameters to be optimized, namely the synthesis parameters (codebook index i and gain G, LTP delay d and coefficients $\{\beta_j\}$), are determined in a closed-loop fashion (analysis-by-synthesis loop) and updated every subframe.

The weighted synthesis filters in the upper and lower branches of Fig. 4.3 are updated at the beginning of every subframe, and their response is kept fixed for the subframe duration. The upper filter memory is updated by passing the last subframe of residual samples while the lower filter memory is updated by filtering the optimally selected LP excitation $\nu_{\mathbf{opt}}$.

Due to the all-pole nature of the weighted synthesis filter, a recursive formulation is used:

$$\bar{s}_w(n) \;=\; \nu(n) \;+\; \sum_{k=1}^{p} a_k \gamma^k \bar{s}_w(n-k), \tag{4.3}$$

where p is the linear prediction order and $0 \leq n \leq L-1$. However, a direct convolution with the weighted synthesis filter impulse response $\{h'(n)\}$ is more suitable for a clear derivation of the synthesis parameters optimization scheme. The filtering operation takes then the form:

$$\bar{s}_w(n) \;=\; \sum_{k=-\infty}^{\infty} \nu(k) h'(n-k). \tag{4.4}$$

The convolution operation has the same outcome as the recursive formulation only in the case where the Zero-State Response (ZSR) of the weighted synthesis filter (zero initial memory) is sought. The different initial conditions at the subframe boundaries with the filters being time-varying lead to different filtering responses for the two formulations. Nevertheless, as it will be seen in the coming derivations, the effect of initial conditions can be grouped into one term that does not interfere in the minimization loop, namely the Zero-Input Response (ZIR), obtained by letting the weighted synthesis filter ring for one subframe duration. After subtracting the ZIR from the reference waveform, zero initial conditions for the weighted synthesis filter can be set, and either the convolution or the recursive computation can be used.

4.2.2 Synthesis Parameters Optimization

Coding of the parameters in the CELP algorithm is done on a subframe basis. As can be seen from Fig. 4.3, the weighted error samples, $\epsilon(n)$, are obtained as the difference between the original weighted speech and the reconstructed weighted speech, $s_w(n) - \bar{s}_w(n)$, for one subframe of length L $(0 \leq n \leq L-1)$. The optimal parameters are obtained by minimizing the mean squared weighted error:

$$\varepsilon \;=\; \sum_{n=0}^{L-1} \epsilon(n)^2. \tag{4.5}$$

The optimal LTP delay and the codebook index are selected by performing an exhaustive search over all allowable pairs (d, i), with the corresponding computed gain G and LTP coefficients minimizing the error of Eq. (4.5).

The spectral deemphasis of the synthesis filter coefficients $\{a_k\}$ by the bandwidth expansion factor γ leads to a quicker attenuation of the weighted synthesis filter impulse response $\{h'(n)\}$. For a value of γ around 0.8 and a sampling frequency of 8 kHz, the quasi totality of the impulse response energy is in the first 20 samples. The impulse response of the filter $\frac{1}{1-F(z/\gamma)}$ can thus be approximated by a finite impulse response (FIR) filter, $h'(0).h'(1),\ldots,h'(R-1)$, with R being smaller than L, the subframe length. The weighted reconstructed speech can therefore be obtained by passing the LP excitation through this FIR approximation,

$$\bar{s}_w(n) \;=\; \sum_{k=n-R+1}^{n} \nu(k)h'(n-k). \tag{4.6}$$

Some of the terms in the above filtering operation depend on past LP excitation samples. Such terms can be grouped separately to yield:

$$\bar{s}_w(n) \;=\; \sum_{k=n-R+1}^{-1} \nu(k)h'(n-k) \;+\; \sum_{k=0}^{n} \nu(k)h'(n-k), \tag{4.7}$$

where the first summation corresponds to the initial boundary conditions while the second gives the ZSR of the weighted synthesis filter. Since the initial boundary conditions depend only on past LP excitation samples, they are identical to the ZIR of the weighted synthesis filter, and they do not interfere in the synthesis parameters selection for the current subframe. They can thus be subtracted from the original weighted speech samples to yield a new reference waveform,

$$\hat{s}_w(n) \;=\; s_w(n) \;-\; \sum_{k=n-R+1}^{-1} \nu(k)h'(n-k), \tag{4.8}$$

and the weighted error samples can be written as

$$\epsilon(n) \;=\; \hat{s}_w(n) \;-\; \sum_{k=0}^{n} \nu(k)h'(n-k). \tag{4.9}$$

The LP excitation samples $\nu(n)$ can be written as the recursive filtering output of the pitch synthesis filter $\frac{1}{1-P(z)}$, when fed by codebook excitation vector samples $e^i(n)$ scaled by G:

$$\nu(n) \;=\; Ge^i(n) \;+\; \sum_{j=-M}^{M} \beta_j \nu(n-d-j). \tag{4.10}$$

Substituting for $\nu(n)$ in Eq. (4.9), the expression for the weighted error samples becomes

$$\epsilon(n) \;=\; \tilde{s}_w(n) \;-\; G\tilde{e}^i(n) \;-\; \sum_{j=-M}^{M} \beta_j \tilde{\nu}^{d+j}(n), \tag{4.11}$$

where $\tilde{e}^i(n)$ is the ZSR of the weighted synthesis filter for the input $e^i(n)$, and $\tilde{\nu}^d(n)$ is the filtered version of $\nu(n-d)$, respectively obtained by:

$$\begin{aligned} \tilde{e}^i(n) \;&=\; \sum_{k=0}^{n} e^i(k) h'(n-k) \\ \tilde{\nu}^d(n) \;&=\; \sum_{k=0}^{n} \nu(k-d) h'(n-k). \end{aligned} \tag{4.12}$$

In order to solve for the optimal values of G and the LTP coefficients $\{\beta_j\}$, either the autocorrelation or the covariance methods can be applied to the mean squared error criterion ε of Eq. (4.5) [55]. A system of $(2M+2)$ equations results, written in matrix form as $\mathbf{\Phi c} = \phi$. The autocorrelation matrix $\mathbf{\Phi}$ is formed by the correlations of all the vectors obtained in Eq. (4.12):

$$\mathbf{\Phi} \;=\; \sum_{n=0}^{L-1} \mathbf{t_{(n)}} \mathbf{t_{(n)}}^T, \tag{4.13}$$

with $\mathbf{t_{(n)}}$ defined to be

$$\mathbf{t_{(n)}} \;=\; \begin{bmatrix} \tilde{e}^i(n) \\ \tilde{\nu}^{d-M}(n) \\ \tilde{\nu}^{d-M+1}(n) \\ \vdots \\ \tilde{\nu}^{d+M}(n) \end{bmatrix}. \tag{4.14}$$

The vector of parameters to be optimized is defined to be

$$\mathbf{c} \;=\; \begin{bmatrix} G \\ \beta_{-M} \\ \beta_{-M+1} \\ \vdots \\ \beta_M \end{bmatrix}, \tag{4.15}$$

and the cross-correlation vector φ is found to be

$$\varphi = \begin{bmatrix} \sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{e}^i(n) \\ \sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{\nu}^{d-M}(n) \\ \sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{\nu}^{d-M+1}(n) \\ \vdots \\ \sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{\nu}^{d+M}(n) \end{bmatrix}. \tag{4.16}$$

From the expressions in Eq. (4.12), it is clear that if the minimum LTP delay d is constrained to be greater than the subframe length L, the filtered LTP contribution $\tilde{\nu}^d(n)$ depends only on past LP excitation samples, i.e. $\nu(n)$ for $n < 0$. At the beginning of the current subframe, this information is already available from the past subframe, and thus autocorrelation matrix $\mathbf{\Phi}$ as well as the cross-correlation vector ϕ are ready known quantities. Finding the optimal set of LTP coefficients and codebook gain amounts therefore to solving the above linear system of equations.

The following section proposes a joint and a sequential optimization schemes for the determination of the optimal synthesis parameters in the case of a one-tap long term predictor. It will be shown that combining these two schemes yields the best compromise between computation complexity and quality.

4.2.3 Optimization for a One-Tap Predictor

LTP Minimum Delay Greater than Subframe Length

It was previously seen that the alternative representation for a one-tap LTP was an adaptive codebook of delayed LP excitation vectors, scaled by the LTP coefficient β. The current LP excitation sample can therefore be written as the sum of the fixed codebook excitation and the adaptive codebook excitation:

$$\nu(n) = Ge^i(n) + \beta\nu(n-d). \tag{4.17}$$

The optimization equations, $\mathbf{\Phi c} = \phi$, take in this case the form:

$$\begin{bmatrix} \sum_{n=0}^{L-1} \tilde{e}^i(n)^2 & \sum_{n=0}^{L-1} \tilde{e}^i(n)\tilde{\nu}^d(n) \\ \sum_{n=0}^{L-1} \tilde{\nu}^d(n)\tilde{e}^i(n) & \sum_{n=0}^{L-1} \tilde{\nu}^d(n)^2 \end{bmatrix} \begin{bmatrix} G \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{e}^i(n) \\ \sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{\nu}^d(n) \end{bmatrix}. \tag{4.18}$$

With the minimum lag d being larger than the subframe length L, the above system matrix and the right hand side cross-correlation vector can be computed at the start of the current subframe. The first strategy consists in jointly optimizing the parameters to obtain the LP excitation of Eq. (4.17). This consists in solving the linear system of Eqs. (4.18) for the optimal gains (G, β) for every pair of codebook index and LTP delay (i, d) chosen each from a predetermined dictionary. It was reported in [55] that CELP coders using this joint optimization scheme registered SNR increases up to 3 dB when compared to coding schemes where the gains and indices are optimized at the analysis stage (open-loop on the original speech). However, even with moderate size codebooks, such as 512 entries for the fixed excitation codebook and 128 allowable LTP delay values, the computational complexity is quite high. The sequential optimization scheme proposed next helps reducing the complexity at the expense of a minor decrease in the objective performance measure values.

The two components of the LP excitation of Eq. (4.17) are sequentially optimized in this alternative strategy. The periodic component $\beta\nu(n-d)$ is considered first by discarding the fixed codebook contribution (setting G to zero). The corresponding optimal LTP coefficient β_{opt} is found from Eq. (4.18) to be:

$$\beta_{opt} = \frac{\sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{\nu}^d(n)}{\sum_{n=0}^{L-1} \tilde{\nu}^d(n)^2}. \tag{4.19}$$

With this optimal value of the LTP coefficient, the error criterion to be minimized becomes:

$$\varepsilon = \sum_{n=0}^{L-1} \tilde{s}_w(n)^2 - \frac{(\sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{\nu}^d(n))^2}{\sum_{n=0}^{L-1} \tilde{\nu}^d(n)^2}. \tag{4.20}$$

The optimal LTP delay is found by maximizing the second term on the right hand side of Eq. (4.20) with respect to d. This is accomplished by an evaluation of the term for all permissible lags (vectors in the adaptive codebook) and the selection of the delay that yields the largest value. Once the LTP parameters are determined, their values are reported in Eqs. (4.18), and the optimal codebook gain G is computed for each fixed codebook entry $\mathbf{e}^{\mathbf{i}}$. The pair (i, G) that yields the minimum mean squared weighted error ε is chosen to represent the codebook contribution to the LP excitation.

In comparing both strategies for the synthesis parameters optimization, it was found that the degradations in segSNR could reach 2 dB when the sequential scheme is employed. The reconstructed speech quality was however very similar for both versions, with the exception of isolated distortions that could be heard at unvoiced/voiced transitions for the latter version. This is most likely due to the decoupling of the LTP coefficient β and the codebook gain G, resulting in large values of β trying to track the sudden periodic structure. Sudden bursts of LTP coefficient values can be avoided if scalar quantized values of β are considered in the evaluation of the minimum mean squared error. The quantization of the synthesis parameters will however be addressed in the next chapter, where highly performing quantizers well suited for bit rate reductions are used.

The synthesis parameters optimization experiments are conducted on 4 ms speech subframes, corresponding to 32 samples per subframe at a speech sampling rate of 8 kHz. At many occasions the pitch lag for female speakers falls well below 30 samples (the pitch frequency exceeds sometimes 300 Hz). With the minimum LTP delay constrained to be greater than 32 samples, the only manner to capture such pitch lags is at doubled or tripled pitch cycle values. However, as the number of pitch doubling increases, the speech harmonic structure is degraded and wavered sounds become audible. Additional measures should be taken to allow the LTP delay to fall below the subframe length, as will be seen in the next section.

Pitch Recycling

For LTP delays d smaller than the subframe length L, the autocorrelation matrix $\mathbf{\Phi}$ of Eq. (4.13) and the cross-correlation vector of Eq. (4.16) will depend on the

LP excitation samples $\nu(n)$ for $n > 0$, which in turn can only be obtained with the knowledge of the synthesis parameters that are being optimized. The set of equations to be solved becomes nonlinear and not conveniently implementable in practice. With a jointly optimized solution being impractical, the sequential approach remains the only alternative for the pitch predictor to "recycle" the current LP excitation output. Indeed, considering once again the codebook gain G to be zero, with the LTP delay d not falling below half the subframe length $L/2$, the resulting LP excitation is now:

$$\nu(n) = \begin{cases} \nu_1(n) = \beta\nu(n-d) & 0 \le n < d \\ \nu_2(n) = \beta^2\nu(n-2d) & d \le n < L \end{cases} . \tag{4.21}$$

The weighted error samples $\epsilon_w(n)$ have now to be split into two terms, one for the d newly computed LP excitation samples, and the other for the remaining part of the subframe including the recycled outputs. Thus, for $0 \le n < d$:

$$\epsilon_{w1}(n) = \tilde{s}_w(n) - \sum_{k=0}^{n} \nu_1(k)h'(n-k), \tag{4.22}$$

and for $d \le n < L$, one more term should be considered:

$$\epsilon_{w2}(n) = \tilde{s}_w(n) - \sum_{k=0}^{d-1} \nu_1(k)h'(n-k) - \sum_{k=d}^{n} \nu_2(k)h'(n-k). \tag{4.23}$$

The total mean squared error is the sum of the squares of the above contributions, given by:

$$\varepsilon = \sum_{n=0}^{d-1} \epsilon_{w1}(n)^2 + \sum_{n=d}^{L-1} \epsilon_{w2}(n)^2. \tag{4.24}$$

Expanding Eq. (4.24) and replacing with the expressions of the LP excitation in Eq. (4.21), the error to be minimized becomes:

$$\begin{aligned} \varepsilon = \sum_{n=0}^{L-1} \tilde{s}_w(n)^2 & - 2\beta \sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{\nu}_1^d(n) \\ & + \beta^2 \Big(\sum_{n=0}^{L-1} [\tilde{\nu}_1^d(n)]^2 - 2\sum_{n=d}^{L-1} \tilde{s}_w(n)\tilde{\nu}_2^{2d}(n) \Big) \\ & + 2\beta^3 \sum_{n=d}^{L-1} \tilde{\nu}_1^d(n)\tilde{\nu}_2^{2d}(n) \\ & + \beta^4 \sum_{n=d}^{L-1} [\tilde{\nu}_2^{2d}(n)]^2. \end{aligned} \tag{4.25}$$

The optimal LTP parameters are found by selecting among the real roots of the cubic equation $\partial\varepsilon/\partial\beta = 0$, at a given LTP delay d, the root value β that yields the minimum mean squared error. This procedure is repeated for all permissible delay values less than the subframe length, and the pair (d, β) that results in the minimum value for ε is considered to be optimal. The excitation codebook parameters are then found employing the standard analysis-by-synthesis search procedure. It is worth noting, however, that the computational burden involved in solving for the roots of the cubic can be avoided if the LTP coefficient is scalar quantized. Indeed the quantization values of β can be successively tried in Eq. (4.25) along with the delays, with the pair that minimizes the error energy selected for transmission. Other schemes based on periodic continuation of the LP excitation instead of recycling have been evaluated [55], but the amplitude of the successive pitch pulses in the subframe could not be adapted (such as the scaling by β^2 in $\nu_2(n)$). As a result, LTP coefficient values greater than unity had a degrading impact on the reconstructed speech signal.

Sequential-Lag/Joint-Gains Optimization Scheme

Previous work [55] has reported that constraining the pitch predictor to operate at a delay corresponding to multiples of the speech subframe fundamental period, instead of allowing that delay to fall below the subframe length, resulted in spurious energy peaks in the reconstructed speech spectrum. Perceptually these peaks corresponded to sudden noise bursts in voiced regions. The LTP delay range considered in the coding scheme of this thesis starts at 20 samples (400 Hz pitch frequency) and ends at 147 samples (54.42 Hz pitch frequency), covering the whole pitch range of 8 kHz sampled speech. A 7-bit adaptive codebook is therefore necessary to represent the one-tap pitch predictor if only integer LTP delay values are considered. As the subframe duration in the coding algorithm is of 4 ms ($L = 32$ samples), it is clear that some of the transmitted LTP delay values will likely be smaller than the subframe size.

In the adopted synthesis parameters optimization scheme, the sequential approach is considered first. The pitch recycling technique described earlier can in this manner be applied for the lag search between 20 and 31 samples. Setting thus the codebook gain G to zero, an optimal pair (d_{opt}, β_{opt}) is found by an exhaustive search along the LTP delay range. However, the value of β_{opt} is discarded while the selected LTP delay

Optimization technique	Average PG (dB)	SNR (dB)	segSNR (dB)
Sequential	7.84	14.66	12.02
SL/JG	7.42	16.72	14.46

Table 4.1: Performance of synthesis parameters optimization schemes. The objective measures are given for female coded speech in a CELP coder with unquantized parameters. The prediction gain (PG) is averaged over the total number of update intervals.

d_{opt} is transmitted. Now in order to perform the optimization of the LTP coefficient β and the codebook parameters (i, G) jointly, the periodic contribution to the LP excitation is redefined for $d_{opt} < L$ to be:

$$\nu_{new}(n) = \begin{cases} \nu(n - d_{opt}) & 0 \leq n < d_{opt} \\ \nu(n - 2d_{opt}) & d_{opt} \leq n < L. \end{cases} \tag{4.26}$$

With the LP excitation periodic component being now formed by the periodic extension of a pitch cycle, the system of Eqs. (4.18) is rendered linear, and one can solve for β and G for every trial codebook excitation vector $\mathbf{e}^{\mathbf{i}}$. The combination that minimizes the mean squared error ε is transmitted along with d_{opt}.

The performance of the sequential-lag/joint-gains (SL/JG) optimization scheme can truly be assessed when compared to the sequential optimization scheme introduced earlier. Both LTP predictors are implemented in a basic CELP coding environment, with the only quantized parameters being the codebook excitation index and the LTP delay. Table 4.1 gives a general idea on the reconstructed speech quality when the LTP coefficient is either determined before the codebook gain or jointly computed with this latter. Although the average prediction gain is slightly lower for the SL/JG scheme, the reconstructed speech quality is much better perceptually, confirming the 2 dB SNR and segSNR difference between the sequential and the sequential/joint approaches. The lower prediction gain can in fact be attributed to portions of the speech signal where the LTP delay drops below the subframe length. The LTP coefficient in such cases is optimized for the periodic extension of one pitch

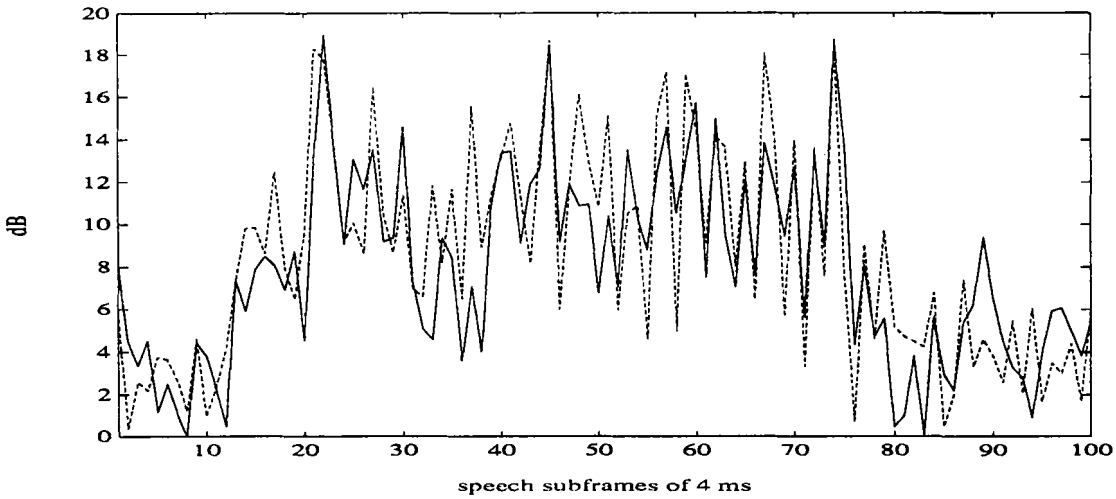
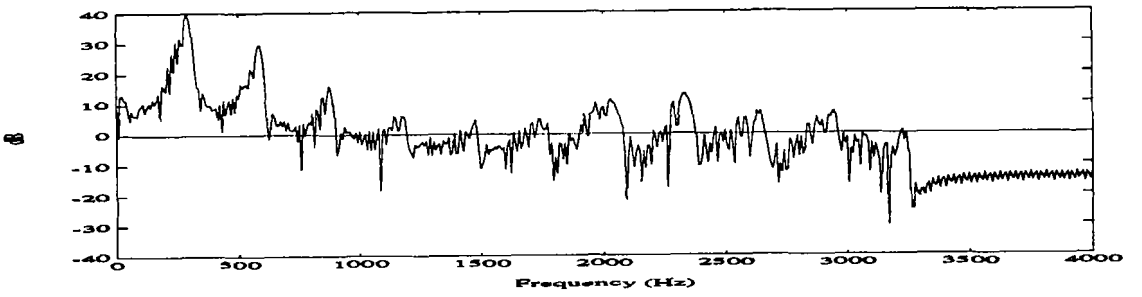


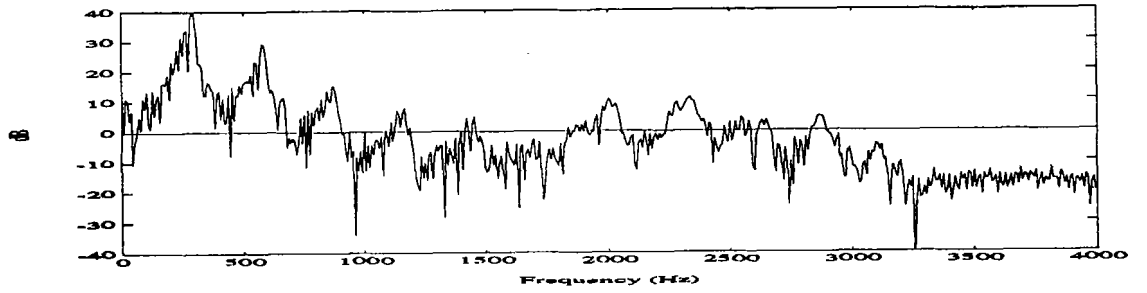
Figure 4.4: Segmental prediction gains (4 ms speech segments) for 400 ms of female speech when the LTP coefficient is computed in the sequential method (solid line), and in the SL/JG technique (dashed line).

cycle, and lacks thus the fullness of the pitch recycling procedure. The improved performance of the SL/JG is illustrated in Fig. 4.4, where the segmental prediction gain, taken over 4 ms subframes, is given for 0.4 s of speech. It is clear from the voiced regions (subframes with high prediction gain) in Fig. 4.4 that the SL/JG optimization scheme outperforms the sequential approach. The segmental prediction gain difference can easily reach 6 to 7 dB. These regions correspond however to LTP delays greater than the subframe length.

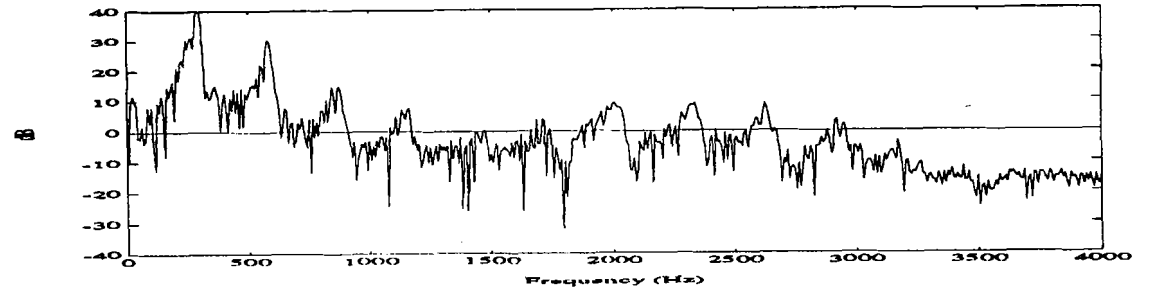
Finally, to complete the performance evaluation of the implemented scheme, it is worth investigating its impact on speech segments where the pitch period drops below the set subframe length of 4 ms. Fig. 4.5 exhibits the energy spectra for an original segment of female speech, its reconstructed version with sequential parameters optimization and with the SL/JG technique. The LTP delay in this speech segment hovers around 29 samples (pitch frequency around 275 Hz), for ten subframes of 32 samples each. The excessive number of sharp spectral dips around the 250 Hz harmonic and its multiples is apparent for the SL/JG method speech spectrum of Fig. 4.5 (c). The sequential optimization technique displays a smoother energy spectral envelope at the lower frequencies. It does also in the limit better represent the orig-



(a)



(b)



(c)

Figure 4.5: Energy spectrum of 10 ms of (a) original female speech, (b) reconstructed speech with sequential parameters optimization, (c) reconstructed speech with optimization based on the SL/JG scheme. The pitch period follows a smooth evolution from 28 samples to 31 samples in this segment.

inal speech spectrum. On the other hand, clearly better results are obtained for the SL/JG technique in the case of LTP delays greater 32 samples, as those displayed in Fig. 4.6. The pitch frequency for the now considered segment is around 222 Hz (a period of 36 samples). The excessive energy at the second harmonic of the pitch frequency (around 450 Hz) is apparent between the peaks of the spectrum corresponding to the sequential approach (Fig. 4.6 (b)). The low frequency region is much better recovered in the case of the sequential/joint technique, as the dips below 0 dB are faithfully represented at the harmonics of the pitch frequency. This translates perceptually into the removal of noise bursts or clicks in the reconstructed speech. It is worth also noting the better spectral representation at the high frequency end for the adopted scheme.

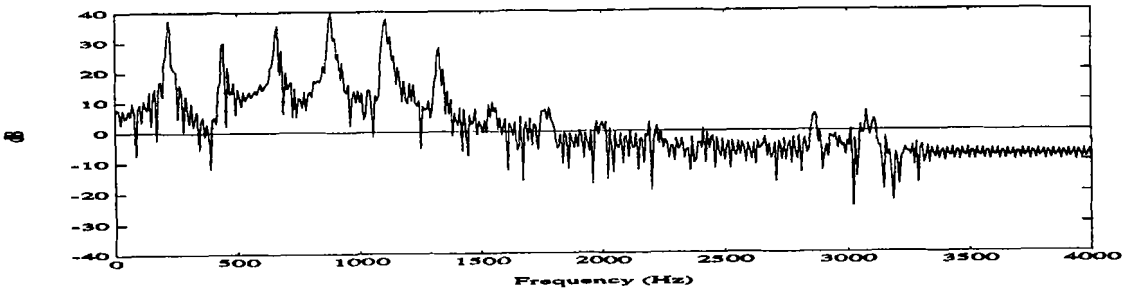
The LTP delay is selected in the closed-loop optimization procedure by matching the **past** reconstructed speech subframe to the current original speech subframe. This delay can therefore yield inaccurate pitch periods and the LTP coefficient will be underestimated. By reducing the quantization error of the LTP delay, finer LTP coefficient estimates can be obtained. In narrowband speech coding, the delay resolution is limited to the sampling rate of 8 kHz. Increasing the resolution of the delay while keeping a sufficiently accurate quantization of the LTP coefficient results in a significant enhancement of pitch prediction. Fractional LTP delays obtained either by multiple-tap predictors or by interpolation of the speech signal are the object of the next section.

4.3 Increased Delay Resolution Pitch Predictors

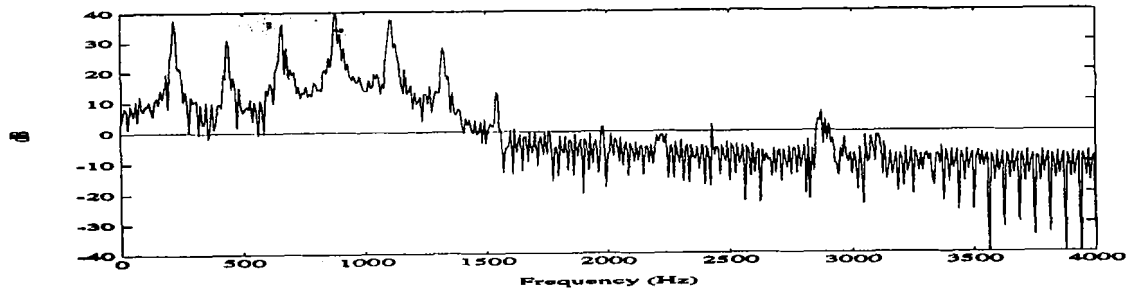
4.3.1 Three-Tap Predictors

The higher performance of three-tap long term predictors when compared to single tap pitch predictors can be accounted for the LTP coefficients dependence on frequency and the variability in frequency of the harmonics. The LP excitation obtained from a three-tap pitch synthesis filter can be written as:

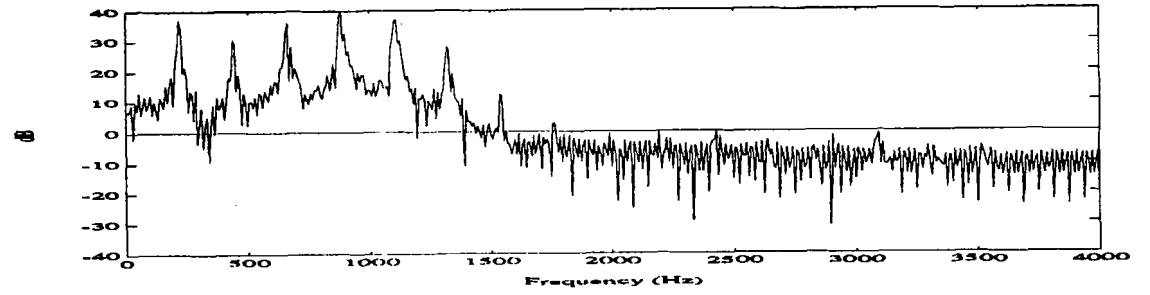
$$\nu(n) = Ge^{i}(n) + \sum_{k=-1}^{1} \beta_k \nu(n-d-k). \tag{4.27}$$



(a)



(b)



(c)

Figure 4.6: Energy spectrum of 40 ms of (a) original female speech, (b) reconstructed speech with sequential parameters optimization, (c) reconstructed speech with optimization based on the SL/JG scheme. The pitch period follows a smooth evolution from 36 samples to 37 samples in this segment.

With the center delay being d, the poles of the pitch synthesis filter are obtained by solving the polynomial $z^d - \beta_{-1}z^{-1} - \beta_0 - \beta_1 z$. A study of the three-tap LTP for very small coefficient variations gives a good general idea on the frequency dependence of the LTP parameters. For this purpose, the LTP end coefficients are split into odd and even normalized contributions b_e and b_o, and the polynomial to be solved is expressed as:

$$z^d - \beta_{-1}z^{-1} - \beta_0 - \beta_1 z \;=\; z^d - ((b_e + b_o)|\beta_0|^{1/d}z^{-1} + \beta_0 + (b_e - b_o)|\beta_0|^{-1/d}z). \quad (4.28)$$

If the even and odd contributions of the taps are set to zero, the single-tap LTP results, with the poles z_k located at $z_k = |\beta_0|^{1/d}\, e^{j2\pi k/d}$. The resonant frequencies (harmonics) are clearly evenly distributed around a circle.

The even contribution of the coefficients is now analyzed by setting the odd contribution b_o to zero. For a very small variation of b_e, the new pole location becomes z'_k, given by:

$$z'_k \;=\; |\beta_0 \;+\; 2b_e \cos(2\pi k/d)|^{1/d}\; e^{j2\pi k/d}, \quad (4.29)$$

as $|z_k - z'_k| \ll 1$. The b_e component of the LTP coefficients contributes to a radial movement of the poles, where for positive b_e, the poles will move outward at low frequencies and inward at high frequencies in the z-plane. The envelope of the pole locations has thus low-pass characteristics (consequently high-pass characteristics for negative b_e). Instability may result when the pole location is near the unit circle (β_0 near 1). The work in [4] has demonstrated the tendency of b_e towards positive values, thus emphasizing the low-pass nature of the three-tap pitch predictors. As long term prediction becomes more and more inaccurate at high frequencies (Figs. 4.5 and 4.6), the low-pass characteristics provide a compensation by allowing high gains at low frequencies and preventing error increases at higher frequencies.

The small odd contribution of the coefficients ($b_e = 0$) for very small pole displacements, $|z_k - z'_k| \ll 1$, is responsible for the tangential movements of the poles in the z-plane. The new pole location is now:

$$z'_k \;=\; |\beta_0|^{1/d} \left[1 \;-\; 2j\frac{b_o}{\beta_0}\sin(2\pi k/d)\right]\; e^{j2\pi k/d}. \quad (4.30)$$

For nonzero values of b_o, the even distribution of the resonant frequencies is lost as well as the linearity of the LTP phase. The small variations of the LTP coefficients

odd contribution do not thus simply represent a fractional adjustment of the LTP center delay d. The phase of the filter transfer function is adjusted only for the dominating regions of the spectrum, corresponding in the time domain to delay refinement for particular frequency bands. A large number of LTP coefficients are required to achieve subsample resolution of the delay, but the scarce bit resources render this approach harmful to the coding quality. An alternative approach to multiple-tap pitch predictors is proposed next, resulting in only a slight bit rate increase from single-tap prediction.

4.3.2 Fractional Delay Pitch Predictors

The prediction gain of long term predictors is usually dependent on the rate of update of the parameters, the prediction order and the amount of periodicity in incoming signals. In addition, it was shown in [25] that increasing the speech sampling frequency, f_s, increases the average prediction gain. As it was previously mentioned, higher order (multiple-tap) predictors yield higher prediction gains, as the LTP coefficients enable for certain frequency bands inter-sample interpolation. However, 2 to 3 bits are needed on average for each LTP coefficient to be quantized [3], making multiple-tap predictors a very expensive choice as update rates get close to 200 updates/second. The employed scheme is a variation of the single-tap predictor where the LTP delay d is allowed to have arbitrary temporal resolution. With much lower bit allocation requirements, this scheme is equivalent in performance to three-tap predictors. The achievement of such higher time resolutions is formally described next, followed by a performance evaluation of the proposed fractional delay pitch predictor.

Subsample Resolution of Prediction Lags

In the one-tap long term predictor of Section 4.2.3, the LTP delay was represented by an integer number of samples, d, obtained at the sampling frequency f_s (8 kHz). The increased temporal resolution of the delay is now achieved by expressing this latter as an integer number of samples d plus a fraction of a sample l/D, where l and D are integers and $l = 0, 1, \ldots, D-1$. The optimal fractional delay can therefore be obtained by shifting forward the past LP excitation by the noninteger delay l/D before performing the closed-loop search. A very efficient method used to perform

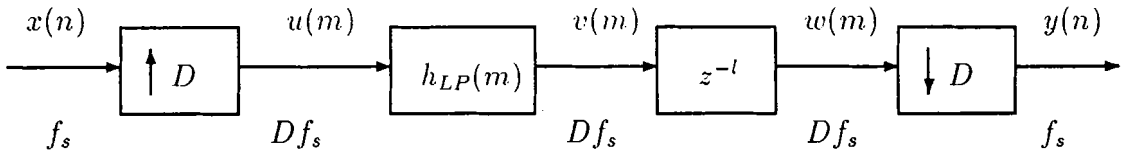


Figure 4.7: A basic structure for achieving a fractional delay of l/D samples.

shifting of discrete signals by fractional delays is *polyphase filtering*. The polyphase filters structure [56] is described in Appendix A, with some of the properties listed.

Assuming that the signal $y(n)$ is an advanced (or delayed) version of the input signal $x(n)$ by a fraction of a sample l/D, this corresponds in the Fourier domain to a linear phase shift:

$$Y(e^{jw}) = e^{jwl/D}X(e^{jw}). \tag{4.31}$$

The ideal system to achieve this operation is seen from Eq. (4.31) to be an all-pass filter with a linear phase $\Phi(w) = lw/D$. It is shown in Appendix A that an FIR polyphase filter approximates the characteristics of the desired system, and thus FIR polyphase filters will be the basis of the fractional delay practical implementations. It is important to realize that a fractional delay l/D at the sampling frequency f_s corresponds to an integer delay l at the higher sampling rate Df_s. Fig. 4.7 displays the various stages of the phase shift procedure. The sampling frequency is at first increased by a factor of D by inserting $(D-1)$ zeros between successive samples of $x(n)$, yielding in the frequency domain the relationship:

$$U(e^{jw}) = X(e^{jwD}). \tag{4.32}$$

The resulting output is then passed through an ideal low-pass filter $h_{LP}(m)$ with cutoff frequency $f_s/2$ (at the sampling rate Df_s) in order to remove the mirror images of $u(m)$. The interpolated version, $v(m)$, of the input signal $x(n)$ results:

$$V(e^{jw}) = H_{LP}(e^{jw})X(e^{jwD}). \tag{4.33}$$

The interpolated signal is next advanced by l samples at the higher sampling rate:

$$W(e^{jw}) = H_{LP}(e^{jw})X(e^{jwD})e^{jwl}, \tag{4.34}$$

and then downsampled again to the original sampling frequency. With the assumption of an ideal low-pass filter, the images of $u(m)$ are sufficiently attenuated to be neglected (no aliasing components), and the output is finally obtained as:

$$Y(e^{jw}) = \frac{1}{D} H_{LP}(e^{jw/D}) e^{jwl/D} X(e^{jw}). \tag{4.35}$$

The low-pass filter $H_{LP}(e^{jw})$ can be approximated by a FIR filter $h(0), h(1), \ldots, h(N-1)$ with exactly linear phase. The delay at the higher sampling frequency will then be $(N-1)/2$. In order to keep an integer filtering delay at the low sampling rate, N is chosen such that the delay is a mutiple of D:

$$\frac{N-1}{2} = ID, \tag{4.36}$$

and I becomes thus the delay at the sampling rate f_s. Furthermore, if the magnitude response of $h(m)$ is equal to D in the passband, the output can be written as:

$$Y(e^{jw}) = e^{jwI} e^{jwl/D} X(e^{jw}). \tag{4.37}$$

The structure in Fig. 4.7 with the FIR filter $h(m)$ becomes hence an approximation to an all-pass network with a fixed integer delay of I samples and a variable fractional delay of l/D samples.

Polyphase filters $p_l(k)$ are used to realize the sampling rate increase and the interpolation (*cf* Appendix A). They are obtained from the coefficients of the FIR interpolation filter as:

$$p_l(k) = h(kD - l) \quad 0 \leq l \leq D-1. \tag{4.38}$$

With the filter $h(m)$ being causal ($h(m) = 0$ for $m < 0$), the first coefficient $p_l(0)$ for nonzero l will be zero. Moreover, each one of the D polyphase filters will have q coefficients ($0 \leq k \leq q-1$), with q given by:

$$q = 2I + 1. \tag{4.39}$$

It is worth noting that if one wanted to implement phase delays instead of advances, the expression for the polyphase filter coefficients would be:

$$p_l(k) = h(kD + l) \quad 0 \leq l \leq D-1. \tag{4.40}$$

The signal $y(n)$ corresponding to a shifted version of $x(n)$ by l/D can now be obtained by the convolution with the $l-th$ polyphase filter, given by:

$$y(n) = \sum_{k=0}^{q-1} p_l(k)x(n-k). \tag{4.41}$$

The FIR approximation to the ideal low-pass filter should be accurate enough to sufficiently attenuate the aliasing components in the downsampling process. The cutoff frequency of the stopband should thus be $f_s/2$, and the stopband ripples sufficiently small. The $\sin(x)/x$ interpolation function weighted with a Hamming window is used in this work as a Nyquist filter, with a cutoff frequency of 4 kHz. The 0-th polyphase filter corresponding to the shifting by the integer delay I operation will have in this case coefficients $p_0(0) = 1$ and $p_0(k) = 0$ for $k > 0$, greatly simplifying the convolution of Eq. (4.11). It is worth mentioning that other interpolator design methods, such as by minimizing the mean-squared interpolation error [57], yield equally performant interpolation filters. In fact, it is even pointed out by [25] that such filters have the same performance of the $\sin(x)/x$ prototype interpolator at lower fixed sample fitering delays I, which reduces the effective number of impulse response samples required for an accurate approximation of the ideal interpolator.

With the arbitrary fractional delay shifting scheme being now set, the expression for the single-tap long term predictor with LTP delay $d + l/D$ becomes:

$$P(z) = 1 - \beta \sum_{k=0}^{q-1} p_l(k) z^{-(d-I+k)}. \tag{4.42}$$

If the fixed delay I value is guaranteed to be less than the minimal LTP integer delay d, the filter will be causal and the polyphase filtered past LP excitation will only need to be shifted backward by I samples. Shifting of the past LP excitation is performed for all allowable fractional delays l/D before the closed-loop optimization procedure. It is also essential not to forget at reconstruction time to shift by the optimally selected fractional delay while adding the periodic structure to the selected codebook excitation.

Performance Evaluation of the Proposed Fractional Predictor

A fractional delay pitch predictor with temporal resolution up to 1/6 of a sample was found to considerably improve the perceptual quality of the reconstructed speech. As

D	Lower delay	Upper delay	Number of entries
4	20	$24\ ^{3/4}$	20
6	25	$69\ ^{5/6}$	270
4	70	$99\ ^{3/4}$	120
3	100	$126\ ^{2/3}$	81
1	127	147	21

Table 4.2: LTP delay distribution for a fractional delay predictor with 9-bit lag quantization.

will be found shortly, prediction gain measures confirm the subjective results. At an operating coding rate of 8 kb/s, it was concluded after various reallocations that not more than 9 bits per 4 ms subframe are needed to be assigned to LTP delay quantization. The delay distribution was limited to the range of the pitch period in male and female speech, namely from 20 samples (2.5 ms) to 147 samples (18.375 ms). Lags corresponding to the most frequent fundamental periods for both male and female speakers had their resolution increased up to 1/6 of a sample. This increased resolution range, mainly applied to the shorter delays, helped compensating for the more perceivable distortion produced by the CELP algorithm in coded female speech and not in coded male speech. Other less frequent intervals, such as high pitched and grave sounds were resolved to 1/4 or 1/3 of a sample, while the very long delays that rarely occur were left at integer sample resolution. The respective ranges of the 512 LTP delay values are given in Table 4.2. With the chosen values for D in Table 4.2, the fixed delay I of the FIR linear phase interpolator must be set to a relatively high value in order to provide a good approximation to the ideal low-pass filter while still preserving the causality of the pitch predictor ($I < \min\ d$). The value $I = 16$ proved to be a good compromise between quality (respectively 97, 129 and 193 coefficients for the FIR interpolators by 3, 4 and 6), and real-time implementability of polyphase filters (33 coefficients each).

The basis of performance comparison between the integer delay single-tap predictor and the adopted fractional delay LTP scheme is the average prediction gain measure. The prediction gain values (in dB) were averaged on a collection of male

Delay resolution	Average Prediction Gain (dB)	
	female	male
Integer	7.16	5.71
Fractional	8.75	6.64

Table 4.3: Average pitch prediction gains for single-tap long term predictors with integer and fractional delay temporal resolution ($I = 16$).

and female speech sentences, with all speech segments yielding gains below 1.5 dB excluded as they usually either represent silence on nonperiodic sounds. The long term predictors are, once again, evaluated in a basic CELP coder environment, with an LTP update rate of 4 ms and LPC parameters computed every 20 ms. The codebook gain and the LTP coefficient are left unquantized. Table 4.3 summarizes the experimental results. The increase in performance for the fractional delay single-tap LTP is comparable to prediction gain values recorded with a 3-tap pitch predictor with integer delays [25, 34]. If 2 bits/coefficient are needed on average to encode the multiple-tap LTP coefficients, a saving of about 3 bits is realized as only $\log_2(D)$ bits are required to encode the fractional sample delays in excess. Both pitch prediction schemes were successively implemented as part of a 10-bit codebook CELP coding algorithm, with the parameters left unquantized. At an LTP parameter update rate of 4 ms, the reconstructed speech resulted in about 1.5 dB higher SNR values in the case of fractional LTP delays.

4.4 Conclusion

The performance of linear predictive coders such as CELP is strongly related to the prediction gain of the pitch predictor. In these coders, multiple-tap and fractional delay predictors are more efficient than single-tap long term predictors, as they allow a smoother evolution of the pitch-cycle waveform. For the 3-tap predictor, this enhancement was the result of the low-pass characteristics of the frequency-dependent coefficients envelope along with the capability of moving the poles tangentially. As a consequence, the LTP delay could be refined for high energy spectral regions, with

the coefficients allowing interpolation between samples. However, the high bit rate requirement of multiple-tap pitch predictors impedes their use in medium and low bit rate coding schemes. In addition, stability checking procedures are more complex for such predictors which often become unstable, affected by quantization errors on the LTP coefficients. For a single-tap LTP, stability is guaranteed be keeping the LTP coefficient below unity. Exploiting the simplicity of single-tap predictors, fractional delay pitch predictors perform equivalently or better than multiple-tap predictors in a full coder. They are characterized by one LTP coefficient and an increased time resolution for the delay, yielding LTP delays expressed as an integer number of samples plus a fraction of a sample. The distortion in CELP coders is usually more audible for female speech than for male speech. Noninteger delay pitch predictors allow the enhancement of performance for female speakers by increasing the time resolution for the shorter delays. Moreover, a small number of bits is needed to quantize the fractional LTP delays which results in significant bit rate savings as no extra coefficients need to be quantized. This allows the refinement of other components in the coding scheme (such as increasing the excitation codebook size). The advantage of using fractional delay pitch predictors in the analysis-by-synthesis loop of the CELP algorithm is therefore the elimination of matching errors due to limited time resolution. This increases the significance of the pitch prediction scheme in the error matching procedure and deemphasizes the predominant role of the excitation codebook, leading to more flexibility in encoding the excitation vectors.

Chapter 5

Toll-Quality Speech Coding at 8 kb/s

5.1 Introduction

High speech coding quality at a rate of 8 kb/s has been achieved by two coding schemes based on the CELP algorithm; the Low-Delay 8 kb/s CELP coder (LD-CELP) [7], unique candidate for CCITT standardization, and the Vector Sum Excited Linear Prediction (VSELP) [29] coder selected by the Telecommunications Industry Association (TIA) as the standard for North American digital cellular telephone systems. Both coding schemes registered a similar performance, perceptually equivalent to around a 3.95 on a MOS scale. However, the low coding delay in the first scheme and the bit costly LPC parameters scalar quantization method in the latter scheme have limited the finer quantization of the long term predictor parameters. It is not to say, on the other hand, that a reallocation of the quantization bits among the optimized encoding blocks described in the previous chapters would result in toll-quality reconstructed speech. Going over the 4.0 barrier in mean opinion score at this given bit rate is not related anymore to bit economics. With the human listener being the ultimate judge of quality, the best attainable coding performance is reached by reducing the perceptual objectionable distortion to the lowest level possible. For this purpose, perceptual speech enhancement techniques such as harmonic noise weighting and postfiltering will be made an integrant part of the coding scheme. Moreover, if a

coder can ideally afford to postpone the transmission of the optimized and quantized parameters until after several speech subframes have been coded, one can then select the successive parameters by minimizing an accumulated minimum mean squared error criterion over those subframes. Such an improvement technique is reminiscent of trellis coding, and will be called in this work delayed decision coding. The performance improvements obtained with this method prove indeed the "relative" suboptimality of synthesis parameters that are selected on a subframe-by-subframe basis in the basic CELP algorithm. The major problem remains however the added computational complexity.

Nevertheless, all the speech enhancement techniques added to the 8 kb/s coding scheme require first of all an economical quantization scheme for the synthesis parameters and then a careful integration of every component in the overall coding scheme. Quantization of the long term predictor delay and the LPC parameters has already been discussed. A very efficient vector quantization method for jointly quantizing the codebook gain and the pitch predictor coefficient will be fully detailed. With all the coder building blocks being now reoptimized, the bit allocation policy will be set first followed by an overview on the structure and operation of the full 8 kb/s CELP coder. Each performance improvement technique will then be discussed and appropriately incorporated in the redesigned coding scheme.

The difficulty with the perceptual improvement techniques used to deemphasize the coding distortion is the lowering of the objective measures that results from the even poorer sample-to-sample match between the original and the coded speech signals. Measures such as SNR and segSNR loose their quality rating significance and will only be used in the case of the operation of the coder with unquantized parameters, or to set lower bounds for acceptable coding degradations. The real assessement of toll-quality coding will come from informal comparison tests between the enhanced 8 kb/s CELP coder and a 7-bit log PCM speech coder.

5.2 Coder Structure

Many of the CELP coder building blocks have been detailed in the previous chapters. Fig. 5.1 is a block diagram of the speech encoder, which, like in all analysis-by-

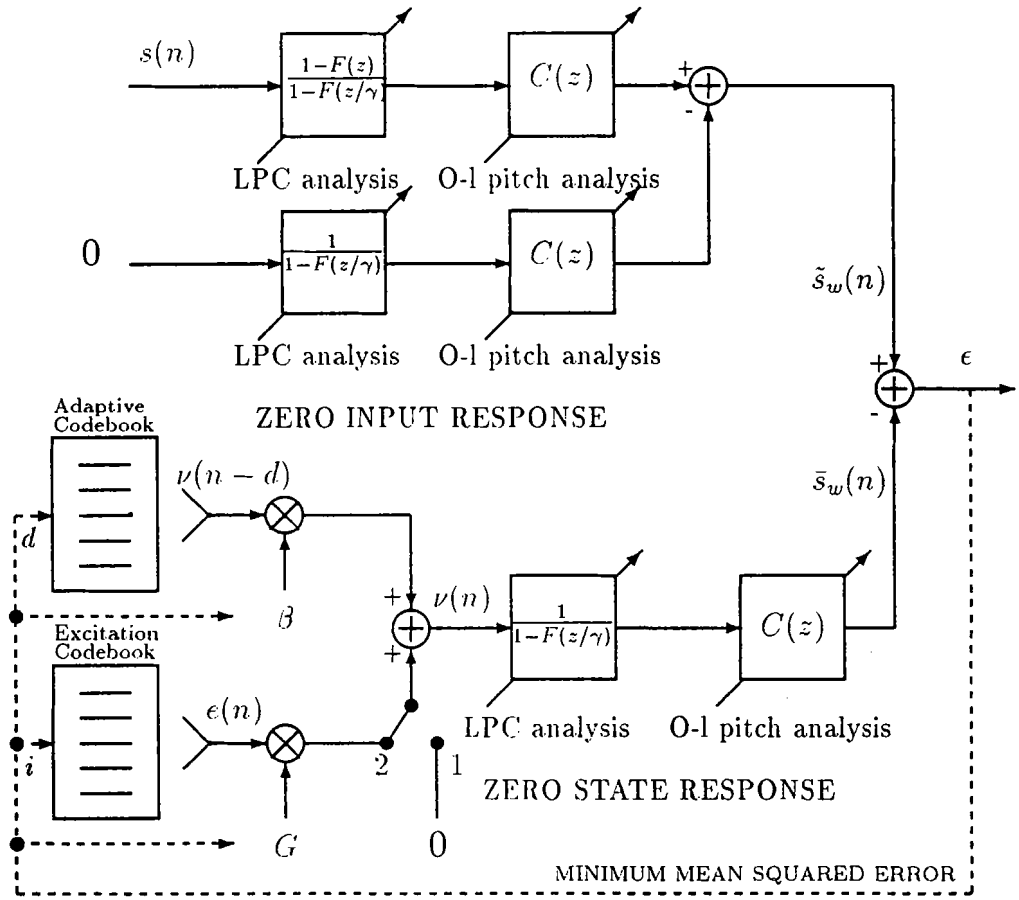


Figure 5.1: Block diagram of the enhanced CELP speech coder.

synthesis coders, contains an embedded decoder. The optimization of the synthesis parameters is based on the minimization of a weighted mean squared error criterion. As can be seen from Fig. 5.1, the weighting consists of a spectral noise weighting filter, $\frac{1-F(z)}{1-F(z/\gamma)}$, whose task is to relocate the coding distortions to high energy regions of the spectrum where they are less audible, and a harmonic noise weighting filter, $C(z)$, used to enhance the periodic structure of the speech signal. Both weighting filters are incorporated into the analysis-by-synthesis loop. The value of γ is fixed at 0.8 and the spectral noise weighting filter is updated from the LPC analysis stage. The harmonic noise weighting filter is, on the other hand, updated at the subframe level by an open-loop (O-1) pitch analysis. The latter weighting scheme will be discussed in greater detail in a subsequent section.

The frame length is 20 ms, corresponding to 160 samples at a speech sampling rate of 8 kHz. The speech frame is further subdivided into five subframes of 32 samples each (4 ms). The analysis frame also consists of 160 samples, but, as was previously detailed, encompasses the four last subframes of the current frame being coded plus one subframe of the next speech frame to be coded. The speech spectral envelope is described by 10 LPC coefficients, vector quantized in the LSF domain using the 24 bits/frame split VQ scheme decribed in Chapter 3.

The linear prediction excitation (LP excitation), $\nu(n)$, is composed of a periodic contribution, $\beta\nu(n-d)$, which consists essentially of a subframe of past samples of this excitation amplitude scaled by β (single-tap pitch filter (*cf* Chapter 4)), and a stochastic excitation vector of 32 samples $\epsilon(n)$, scaled by the codebook gain G. Those synthesis parameters are updated on a subframe-by-subframe basis according to the sequential/joint approach described in Section 4.2.2. The method is illustrated in Fig. 5.1. By setting the switch to position 1, the excitation codebook contribution to the LP excitation is cancelled, and the LTP delay d is determined in these conditions. Once the delay d (adaptive codebook contribution) is found, the switch is reset to position 2 and the remaining parameters, namely the excitation codebook index i, the codebook gain G, and the LTP coefficient β are jointly optimized by minimizing the weighted mean squared error between the original and the reconstructed speech versions.

Once all the parameters for a given subframe are optimized and quantized, the adaptive codebook and the filters state are updated by computing the optimal LP excitation and by passing it through the weighted synthesis filter and the harmonic noise weighting filter in the lower branch.

An efficient way of reducing the computational complexity without affecting the speech quality is to consider the effect of the Zero Input Response (ZIR) of the weighted synthesis filter and the cascaded Harmonic Noise Weighting (HNW) filter outside the analysis-by-synthesis loop. At the beginning of every speech subframe to be coded, the ZIR is obtained by letting the cascaded filters ring for the duration of 32 samples, then subtracted from the weighted original speech subframe to yield a new reference waveform $\tilde{s}_w$. The state of these filters is then reset to zero and the Zero State Response (ZSR) will determine what synthesis parameters are best suited

Parameter	Bits/Subframe	Bits/Frame
LPC coefficients		24
Frame energy		5
Codebook index	10	50
LTP delay	9	45
Gains (G, β)	7	35
UNUSED		1
TOTAL		160

Table 5.1: Bit allocations for the 8 kb/s CELP coder.

to match the reference waveform.

Table 5.1 gives the bit allocations for the coder operating at a rate of 8 kb/s. The adaptive codebook consists of 512 entries for an LTP delay range 20–147 samples. As was specified in the previous chapter, the LTP delay resolution is increased up to 1/6 of a sample in certain critical ranges and up to 1/3 or 1/4 in others. Adaptive codebook entries for LTP delays smaller than the subframe length are formed by periodic a extension. The excitation codebook contains 512 stochastically generated (iid Gaussian) excitation vectors and their negative counterparts.

The remaining bits are spent on gain quantization, namely the LTP coefficient β and the codebook gain G. An efficient vector quantization scheme for the gains has been developed and used in the VSELP coder [29]. Based on this model, a 7-bit gains vector quantizer is employed in this work. However, it requires for proper operation an acceptable estimate of the current speech frame energy, encoded by 5 bits/frame.

5.3 Gains Quantization

Upon joint optimization of the excitation codebook index and the gains, the optimal LP excitation can be written as:

$$v(n)_{opt} = Gc^{l}(n) + \beta v(n-d), \tag{5.1}$$

where I is the excitation codebook index selected for transmission. A scalar quantization of G and β would probably yield good results, but not enough bits are available for the desired resolution of the quantizers. Moreover, the correlation that exists between the excitation components is totally neglected when separate quantizers are used. Vector quantization offers both bit rate reduction capabilities and consideration of the interaction between the LTP coefficient and the codebook gain. The distortion criterion will be the perceptually weighted mean squared error between the original and reconstructed speech subframes, with all parameters being quantized. Denoting by $\bar{\beta}$ and $\bar{G}$ the quantized versions of β and G, the error sample, according to Fig. 5.1, is:

$$\epsilon(n) \;=\; \tilde{s}_w(n) \;-\; \bar{\beta}\tilde{\nu}^d(n) \;-\; \bar{G}\tilde{e}^I(n), \tag{5.2}$$

with $\tilde{\nu}^d(n)$ and $\tilde{e}^I(n)$ being, as defined in Chapter 4, the weighted synthesis filtered and harmonic noise weighted sequences $\nu(n-d)$ and $e^I(n)$. The mean squared error is then expressed as:

$$\varepsilon \;=\; \sum_{n=0}^{L-1} \epsilon(n)^2. \tag{5.3}$$

In order to make this expression resemble a distortion criterion for vector quantizing the gains, it can be rewritten as:

$$\varepsilon \;=\; R_{ss} + \bar{\beta}^2 R_{aa} + \bar{G}^2 R_{ee} - 2\bar{\beta}R_{sa} - 2\bar{G}R_{se} + 2\bar{\beta}\bar{G}R_{ae}, \tag{5.4}$$

with the precomputed values being:

$$\begin{aligned}
R_{ss} &= \sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{s}_w(n) \\
R_{aa} &= \sum_{n=0}^{L-1} \tilde{\nu}^d(n)\tilde{\nu}^d(n) \\
R_{ee} &= \sum_{n=0}^{L-1} \tilde{e}^I(n)\tilde{e}^I(n) \\
R_{sa} &= \sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{\nu}^d(n) \\
R_{se} &= \sum_{n=0}^{L-1} \tilde{s}_w(n)\tilde{e}^I(n) \\
R_{ae} &= \sum_{n=0}^{L-1} \tilde{\nu}^d(n)\tilde{e}^I(n).
\end{aligned} \tag{5.5}$$

5.3.1 Vector Quantization Scheme

A direct vector quantization of the gains values is usually avoided due to the occasional spurious behaviour of the LTP coefficient. Indeed, the optimal value for β can occasionally get very large, at unvoiced/voiced segment transitions for example. A better behaved pair of parameters is employed as a basis for vector quantization: $P0$, the *normalized approximate energy contribution of the optimal adaptive codebook vector* scaled by β^2, and GS, an *energy offset* for refining the normalized energy contribution estimates.

The normalization of the energy contributions of the adaptive codebook and the excitation codebook entries is obtained after dividing by an estimation of the speech residual energy, RS, for the current subframe. Using the set of p reflection coefficients $\{k_i\}$ corresponding to the interpolated subframe predictor coefficients, the speech residual energy can be obtained by [8]:

$$RS = L\bar{R}_s(0)\prod_{i=1}^{p}(1-k_i^2), \tag{5.6}$$

where L is the subframe length and $\bar{R}_s(0)$ is the quantized current speech subframe per sample energy estimate. This subframe energy estimate is in fact obtained from interpolating the quantized per sample energy estimates of the past and the present analysis speech frames, $\bar{R}_{past}(0)$ and $\bar{R}_{current}(0)$. The interpolation scheme for the subframe energy estimates corresponds to the one employed for the LPC parameters, given in Section 3.5. The quantization of the per sample energy for the current frame is based on a 32-level uniform quantizer in the log domain, with a bin width of 2 dB. Appendix B details the steps of the frame energy quantization process and outlines the interpolation scheme that yields $\bar{R}_s(0)$.

The expressions for $P0$ and GS can now be formally defined. Given that $R_x(0)$ and $R_x(1)$ correspond respectively to the energy contributions of the optimally selected adaptive codebook and excitation codebook vectors, they can be expressed as:

$$R_x(0) = \begin{cases} \sum_{n=0}^{L-1} v(n-d)^2 & \text{if } d \geq L \\ \sum_{n=0}^{d-1} v(n-d)^2 + \sum_{n=d}^{L-1} v(n-2d)^2 & \text{otherwise} \end{cases}, \tag{5.7}$$

$$R_x(1) = \sum_{n=0}^{L-1} e^I(n)^2. \tag{5.8}$$

With GS considered to be the correction factor for the energy estimates, the optimal adaptive codeword normalized energy contribution is obtained by:

$$P0 = \frac{\beta^2 R_x(0)}{RS\ GS}, \tag{5.9}$$

where $0 \leq P0 \leq 1$. The optimal excitation codebook vector normalized energy contribution is obtained in a similar fashion. Ideally, if all the subframe residual energy is accounted for $P0$, the excitation codebook vector energy contribution should vanish. Thus this latter contribution is linearly equivalent to $1 - P0$, and can be expressed as:

$$1 - P0 = \frac{G^2 R_x(1)}{RS\ GS}. \tag{5.10}$$

From Eqs. (5.9) and (5.10), one can solve for $\bar{\beta}$ and $\bar{G}$ once the best $(P0, GS)$ pair has been selected from the vector quantization codebook according to the criterion of Eq. (5.4). Precomputing some of the factors in the mean squared error criterion increases the efficiency of the codebook search for the optimal quantized gains. By defining:

$$\begin{aligned} a &= 2R_{sa}\sqrt{RS/R_x(0)} \\ b &= 2R_{s\epsilon}\sqrt{RS/R_x(1)} \\ d &= 2R_{a\epsilon}RS/\sqrt{R_x(0)\ R_x(1)} \\ e &= R_{aa}RS/R_x(0) \\ f &= R_{\epsilon\epsilon}RS/R_x(1), \end{aligned} \tag{5.11}$$

the distortion criterion of (5.4) can be expressed in terms of $P0$ and GS as:

$$\begin{aligned} \varepsilon = R_{ss} - a\sqrt{GS\ P0} - b\sqrt{GS\ (1-P0)} + d\ GS\sqrt{P0\ (1-P0)} \\ + e\ GS\ P0 + f\ GS\ (1-P0)\ . \end{aligned} \tag{5.12}$$

By successively trying the 128 codebook entries, the $(P0, GS)$ pair that minimizes the total weighted error is transmitted to the decoder, where the gains are recovered as follows:

$$\begin{aligned} \beta &= \sqrt{\frac{RS\ GS\ P0}{R_x(0)}} \\ G &= \sqrt{\frac{RS\ GS\ (1-P0)}{R_x(1)}}. \end{aligned} \tag{5.13}$$

It is clear from Eq. (5.13) that G and β assume only positive values. While the sign problem for G is taken care of by the structure of the excitation codebook where each stochastic excitation vector has a negative counterpart, β is not allowed to fall below zero even if its optimal unquantized value is negative. However, it has been experimentally determined in [55] that negative LTP coefficients occur only in unvoiced speech segments, and consequently, the role of the pitch predictor is insignificant in such cases. Therefore, whenever the cross-correlation, R_{sa}, between the weighted reference signal and the synthesized adaptive codebook contribution is negative, the pitch predictor is deactivated by setting the quantized value of β to zero. In this case, the mean squared error simplifies to:

$$\varepsilon \;=\; R_{ss} - b\,\sqrt{GS\;(1-P0)} + f\;GS\;(1-P0), \tag{5.14}$$

and the quantized value for G can still be determined by selecting the codebook pair $(P0, GS)$ that minimizes the above error. It is also important to mention that this expression for the error is also valid for initial subframes to be coded (i.e. when the adaptive codebook is entirely populated by zeros).

5.3.2 Discussion

The codebook of 128 $P0 - GS$ vectors is designed using the standard LBG training algorithm described in Chapter 3, and the training is based on a large speech database equally distributed between male and female speakers. Fig. 5.2 shows the distribution of the gains codebook vectors, where $P0$ is displayed versus $10\log(GS)$. By factoring out the average residual subframe energy, the gains can be quantized equally well at all signal levels. The dynamic range problem is hence solved by quantizing the speech average energy once per frame. The behaviour of the quantized GS and $P0$ parameters is illustrated in Fig. 5.2. As the adaptive codebook energy contribution increases yielding $P0$ values near unity (voiced segments), the corresponding GS values become less significant (around 1). However, as the pitch prediction role diminishes with decreasing $P0$ values, the corresponding GS values quickly tend towards zero, further attenuating the gains. Such a behaviour helps in dealing with situations such as unvoiced/voiced segment transitions where sudden energy increases are regulated by low GS values. The LTP coefficient β can occasionally get very large in similar

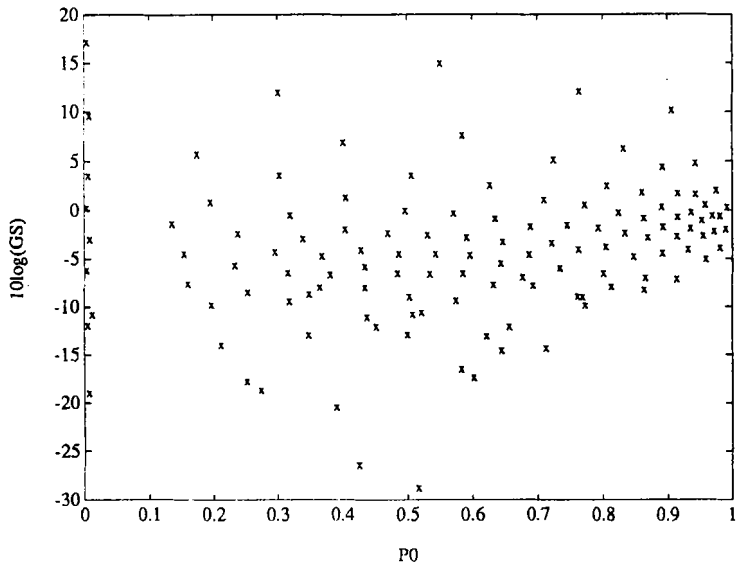


Figure 5.2: Gains codebook vectors represented as P0 vs GS in dB.

situations, whereas the $P0$ range is always limited between zero and unity. The GS and $P0$ parameters are thus much more suitable for vector quantization than are the codebook and LTP gains. Finally, as long as the average frame energy is properly transmitted to the decoder, the reconstructed speech energy will not exceed the desired energy specified by the range of GS in Fig. 5.2, and thus sudden bursts are avoided. Results on the performance of the implemented gain vector quantization scheme are reported at the end of this thesis.

5.4 Perceptual Enhancement of Coded Speech

As limitations are imposed on the operating bit rate of the CELP coder, maintaining good coding quality becomes a much more involved task. The loss of accuracy in the waveform matching approach should then be compensated by emphasizing the perceptually significant features of the speech signal. It was already explained how exploiting the masking property of the human auditory system has led the spectral noise weighting to improve the matching process in CELP. Other quality enhancement techniques have been employed with postfiltering [58] and pitch prefiltering [29] being the most common.

Accentuating the coded speech periodic structure has also drawn the attention of many researchers, as some interharmonic distortions were audible in voiced speech segments. The proposed periodicity enhancement technique in this work derives from the spectral noise weighting approach, since it exploits in this case the noise masking potential of the harmonic structure in the speech signal. As it will be shortly seen, this *Harmonic Noise Weighting* (HNW) technique incurs no cost in the bit allocations of the coder since it is included in the analysis-by-synthesis loop of the encoder for the sole purpose of improving the perceptual weighting of the error matching criterion [34]. Furthermore, in view of the close relationship between the long term predictor delay and coefficient and the HNW parameters, the complexity of the synthesis parameters optimization technique can be greatly reduced by performing a limited LTP parameters search around the selected HNW parameters. The minor objective quality measurement degradations that follow are not perceptible.

On the decoder side, the reconstructed speech quality is enhanced by adaptive postfiltering. An efficient postfilter, originally introduced in [31] to improve the performance of the CCITT standard 16 kb/s LD-CELP coder, is implemented in the coding scheme of this thesis. A full description of the postfilter adaptation process is also given in this section.

5.4.1 Harmonic Noise Weighting

Reducing the presence of noise between harmonics was in the earlier versions of the CELP coder accomplished at the decoder side by pitch postfiltering [32]. More recently, an attempt to enhance periodicity at the synthesis parameters optimization stage was carried in [33]. It consisted in reducing the contribution of the excitation codebook vectors in voiced segments, as they were considered to be undesired noisy components. The codebook gain in those circumstances was set to values below the optimally calculated ones. The improvement of the subjective coding quality with this technique (constrained excitation approach) addresses the limitations of the common weighted error criterion used in the CELP algorithm.

Weighting Structure

By exploiting the noise masking potential of the speech signal harmonic structure, the perceptual accuracy of the CELP error criterion can be increased within the analysis-by-synthesis procedure. This is accomplished by cascading the spectral noise weighting filter $W(z)$ with a harmonic noise weighting filter (HNW filter) $C(z)$ as illustrated in Fig. 5.1. The combination of both weighting schemes leads to a significant quality enhancement over the usual spectrally weighted error criterion. The form of the harmonic noise weighting filter is similar to that of the long term predictor, given by:

$$C(z) = 1 - \varepsilon_p \sum_{i=-M}^{M} c_i z^{(-D+i)} \tag{5.15}$$

with ε_p a parameter set between zero and unity to specify the amount of weighting to be applied. The HNW filter delay D and multiple taps $\{c_i\}$ are determined from an **open-loop pitch analysis** on the spectrally weighted input speech. However as the number of taps increases the spectral envelope of the HNW filter looses its flatness (*cf* Section 4.3.1), and may degrade the weighting performance. A 3-tap HNW filter was found to be a good compromise between complexity and performance [34]. The proposed harmonic noise weighting scheme in this work uses a single tap HNW filter, with the delay D resolution increased up to 1/6 of a sample over the whole pitch period range considered (20 to 147 samples), obviating the need for multiple taps.

Complexity Reduction

The incorporation of harmonic noise weighting in the closed-loop synthesis parameters optimization affects the computational complexity at two stages: the long term predictor (LTP) delay determination and the joint selection of the codebook index and gains. It was experimentally concluded in [34] that spectral noise weighting was sufficient during the LTP delay search, and subsequently, harmonic noise filtering is only necessary during the joint optimization stage. Furthermore, the HNW filter parameters found at the outcome of the open-loop analysis on the spectrally weighted input speech, especially the fractional delay D, can be employed to reduce the complexity of the adaptive codebook search. The method is similar to a hybrid open-loop/closed-loop search, where the open-loop stage determines a list of candi-

date lags to be evaluated in the closed-loop search.

From the spectrally weighted input speech $s_{sw}(n)$ the correlation arrays $C_0(d)$ and normalization arrays $G_0(d)$ are computed first for all integer delays d in the lag range considered according to:

$$\begin{aligned} C_0(d) &= \sum_{n=0}^{L-1} s_{sw}(n)s_{sw}(n-d), \\ G_0(d) &= \sum_{n=0}^{L-1} s_{sw}(n-d)^2. \end{aligned} \tag{5.16}$$

The optimal one-tap predictor delay J is then found by maximizing the normalized correlation,

$$\frac{C_0(d)}{\sqrt{G_0(d)}} \tag{5.17}$$

over the lag range. Once the submultiples of J are checked an optimal integer lag determined, the resolution of the delays is increased to 1/6 of a sample by polyphase filtering the arrays of Eq. (5.16) (*cf* Appendix A). Fractional delays are then classified as surviving candidates by selecting around the optimal integer lag those lags that maximize the interpolated normalized correlation function with their associated pitch prediction gain exceeding a certain threshold. The threshold can be for example a percentage (75 % in this work) of the prediction gain obtained for J. At this stage, the value of the HNW filter delay D is chosen to be the smallest surviving lag (integer or fractional), and the corresponding filter coefficient is computed. Additional surviving lags can also be determined from doubling and tripling D at the higher resolution and the open-loop search stage is concluded by rearranging all the surviving lags in decreasing prediction gain order. The closed-loop adaptive codebook search procedure is in turn performed around the best few surviving lags. The amount of computation is therefore reduced with no loss in quality as the estimated open-loop surviving lags are highly correlated with the immediate past reconstructed speech pitch cycles for voiced frames. However, the relationship between open-loop and closed-loop long term correlations is not as obvious for unvoiced frames. HNW filtering will be turned off for such frames (with prediction gain values less than a set threshold) as it adversely affects the coded speech quality, and the adaptive codebook search is conducted on the whole delay range.

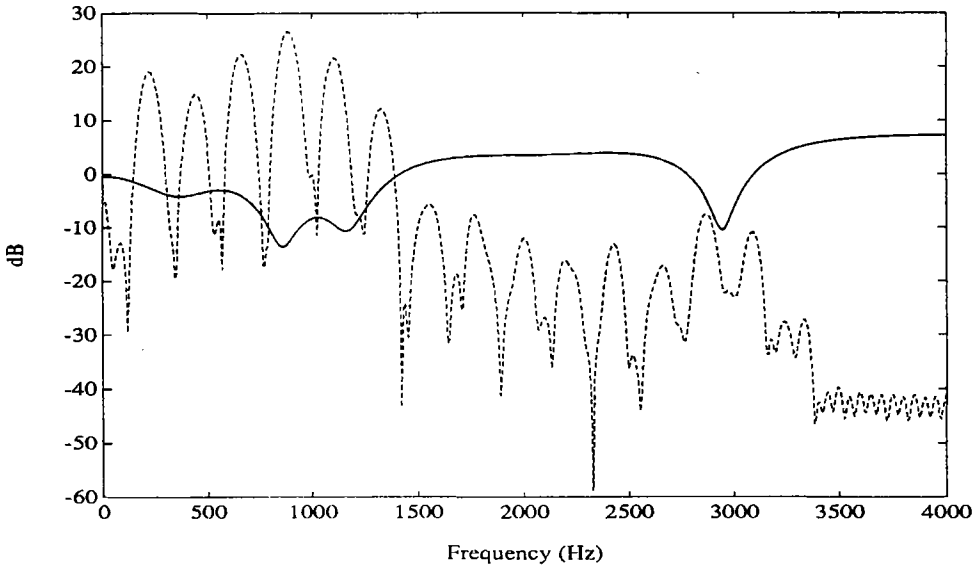
Discussion

Improved perceptual speech quality results from incorporating harmonic noise weighting in the analysis-by-synthesis loop. The HNW filter is updated at the subframe level by an open-loop pitch analysis (every 4 ms) with a fixed harmonic noise weighting parameter $\varepsilon_p = 0.3$. Two surviving LTP delay values are kept at the outcome of the open-loop analysis. The adaptive codebook search is then performed for all fractional lags within one sample of the top two selected delays. Increasing the resolution of the HNW filter delay was achieved by using 33-tap polyphase filters. Fig. 5.3 displays the energy spectrum of a 20 ms segment of input speech with superimposed the spectral noise weighting scheme frequency response in one case and the combined harmonic and spectral weighting cascade frequency response in the other. As can be seen, the spectral weighting envelope is preserved with high energy spectrum portions less emphasized than lower energy portions, thus masking coding distortions. However, the error is assigned more weight at the spectral dips resulting from the harmonic structure of the spectrum, emphasizing in this case the interharmonic quantization noise.

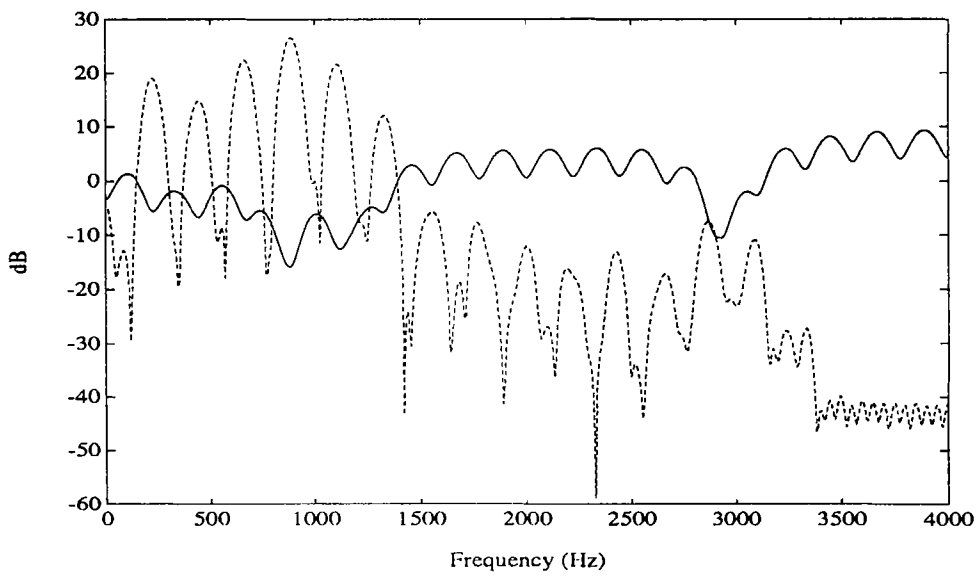
The net effect of harmonic noise weighting can be depicted in Fig. 5.4. A comparison between an original female speech voiced segment and the corresponding coded versions reveals that the introduction of harmonic noise weighting contributes to improving the envelope of the time waveform. By accentuating the periodicity of the reconstructed speech segment, HNW filtering helps also by attenuating the impact of sudden pitch period variations and increased noisy behaviour of the speech waveform. This can be viewed as reducing the role the stochastic codebook excitation plays in the speech synthesis process, with most of the speech reconstruction attributed to the periodic contribution.

5.4.2 Adaptive Postfiltering

Enhancement of the reconstructed speech takes place at the very last stage of the decoding process by the means of postfiltering. Postfilters are usually pole/zero filters, with their coefficients either kept fixed or adapted with the LPC parameters. However, while postfilters contribute to coding quality improvement in the case of a single encoding, they can also be the cause of drastic performance degradations in tandeming

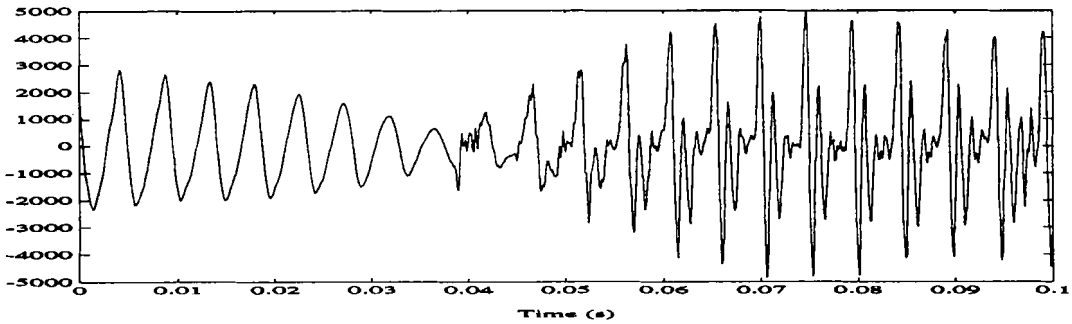


(a)

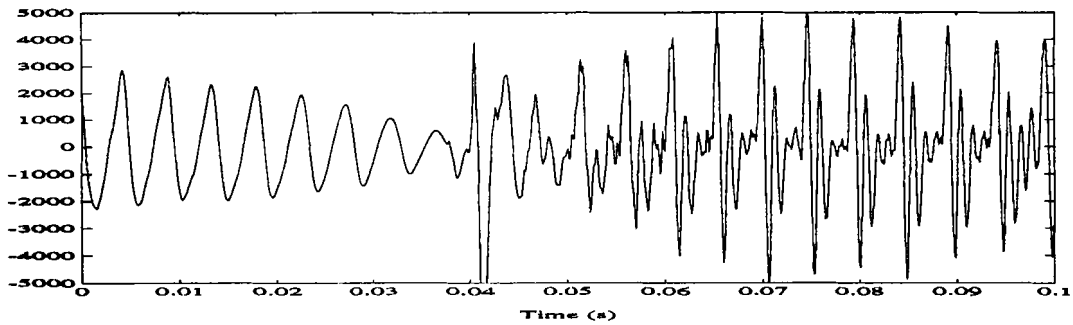


(b)

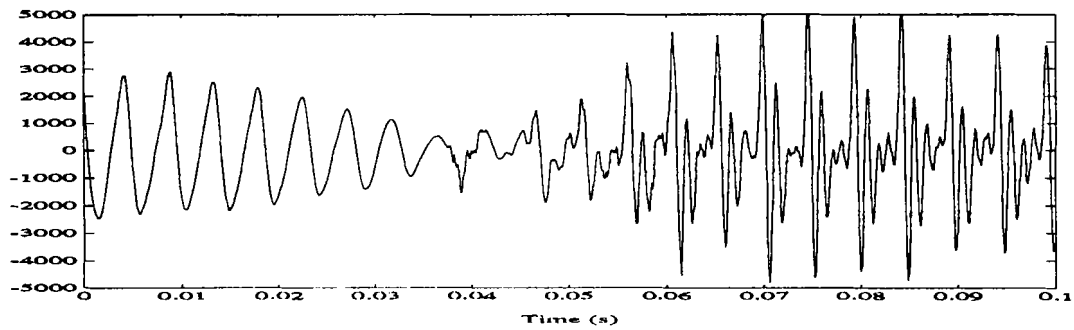
Figure 5.3: Energy spectrum of a voiced speech segment (dashed) with the (a) spectral noise weighting frequency response (solid) and the (b) combined spectral and harmonic noise weighting frequency response (solid) superimposed.



(a)



(b)



(c)

Figure 5.4: 100 ms of a voiced segment of (a) female speech along with its corresponding coded version with (b) only spectral noise weighting and (c) combined spectral and harmonic noise weighting.

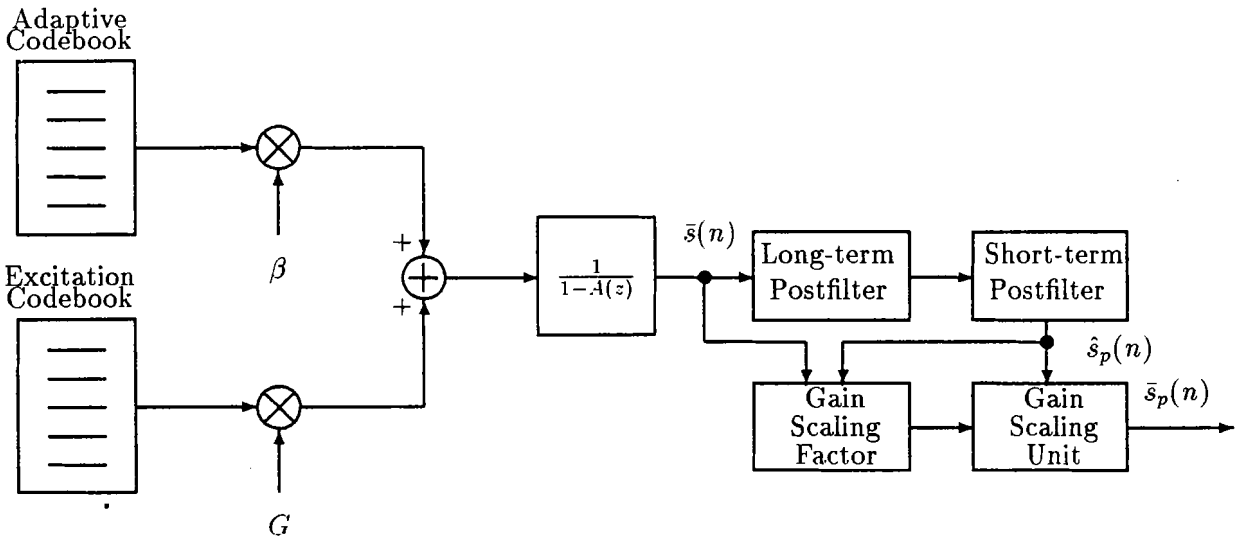


Figure 5.5: CELP decoder with adaptive postfiltering.

situations. Nevertheless, the work in [31] has shown that if the postfilter is tuned for every encoding stage, its effect at the early stages can be controlled, resulting in an overall beneficial performance. It is therefore crucial to adapt the postfilter to the spectral characteristics of the speech segment being coded, and to guarantee flexibility through some tunable parameters. The postfiltering scheme adopted in this work is based on the improved model introduced to the 16 kb/s LD-CELP CCITT standard. Fig. 5.5 represents the decoding portion of the 8 kb/s CELP coding scheme, with the postfilter components added.

Long-term postfiltering is carried out by the single tap FIR filter:

$$H_l(z) = g_l(1 + b\, z^{-d}), \tag{5.18}$$

with d being the optimal fractional LTP delay (up to 1/6-th of a sample resolution). The scaling factor g_l is dependent on b:

$$g_l = \frac{1}{1+b}, \tag{5.19}$$

and the long-term postfilter coefficient is defined as a function of the optimal LTP

coefficient β:

$$b = \begin{cases} 0 & \text{if } \beta < 0.6 \\ \lambda\beta & \text{if } 0.6 \leq \beta \leq 1 \\ \lambda & \text{if } \beta > 1 \end{cases} \tag{5.20}$$

The amount of long term postfiltering is controlled by the tunable parameter λ.

The short-term postfilter has the form:

$$H_s(z) = \frac{1 - \sum_{i=1}^{10} \bar{b}_i z^{-i}}{1 - \sum_{i=1}^{10} \bar{a}_i z^{-i}} \left[1 + \mu z^{-1}\right] \tag{5.21}$$

where

$$\begin{aligned} \bar{b}_i &= a_i \, \gamma_1^i \\ \bar{a}_i &= a_i \, \gamma_2^i \\ \mu &= \gamma_3 k_1. \end{aligned} \tag{5.22}$$

The short-term postfilter parameters are adapted every subframe in accordance to the set of interpolated LPC parameters $\{a_k\}$ with k_1 being the corresponding first reflection coefficient. An appropriate choice of the bandwidth expansion factors γ_1 and γ_2 will yield a spectral weighting scheme that enhances the reconstructed speech quality, while the tunable parameter γ_3 controls the first order low-pass filter $[1+\mu z^{-1}]$ appended to the postfiltering scheme in order to increase the coded speech brightness.

To ensure unity power gain between the input $\bar{s}(n)$ and the output $\hat{s}_p(n)$ of the postfilters, a gain scale factor is computed and used to scale the postfiltered reconstructed speech. It is obtained as:

$$\delta = \sqrt{\frac{\sum_{n=0}^{L-1} \hat{s}_p(n)^2}{\sum_{n=0}^{L-1} \bar{s}(n)^2}}. \tag{5.23}$$

However, before being used, this scale factor is passed through a first order low-pass filter yielding:

$$\delta'_{(k)} = 0.9875 \, \delta'_{(k-1)} + 0.0125 \, \delta, \tag{5.24}$$

where k refers to the time index of the current subframe. The postfiltered speech is then multiplied by $\delta'_{(k)}$, resulting in the decoded speech $\bar{s}_p$. The scaling factor

computation method allows the gain values to gradually adapt to energy increases and drops.

With tunable parameter values $\gamma_1 = 0.65$, $\gamma_2 = 0.75$, $\gamma_3 = 0.15$, and $\lambda = 0.65$, the introduction of the postfilter resulted in sharper perceptual speech quality, despite a loss of about 1.3 dB in SNR.

5.5 Delayed-Decision Coding

The essence of analysis-by-synthesis based coders such as CELP coders is speech coding on a blockwise basis. The parameters that are quantized and transmitted at every subframe are optimized for the current input speech subframe and in fact take into account the effects of the previously transmitted parameters as the filter states are updated at the beginning of the subframe. However, by allowing slightly suboptimal parameters to be selected for a given subframe, the choice of optimal parameters for the following subframe with the now suboptimal initial conditions can possibly yield a smaller average mean squared error when evaluated for both subframes. Departing from this line of thought, transmission of the subframe synthesis parameters can be delayed until the end of the speech frame, where for each one of the five subframes, the optimally selected parameters are kept along with a number of other surviving suboptimal parameters. The procedure takes then the form of a trellis coding scheme performed on the subframe level. Unfortunately this scheme is not practically implementable without a pruning operation at every stage in order to keep the number of surviving paths reasonable. Otherwise, even with a small number of starting paths such as 4, the number of alternatives at the 5-th subframe would be 1024.

Fig. 5.6 illustrates the delayed-decision coding scheme that the CELP coder follows. For every starting point of a subframe synthesis stage, one suboptimal LTP delay is kept along with the optimal delay. By suboptimal delay it is meant the one that yields the next to the lowest mean squared error obtained for the optimal delay value. For each surviving LTP delay d, the excitation codebook index and the gains are jointly optimized, and the two synthesis parameter sets that yield the smallest subframe mean squared error are kept, resulting in further branching in the tree. Therefore, at the end of this procedure for the first subframe, four paths are con-

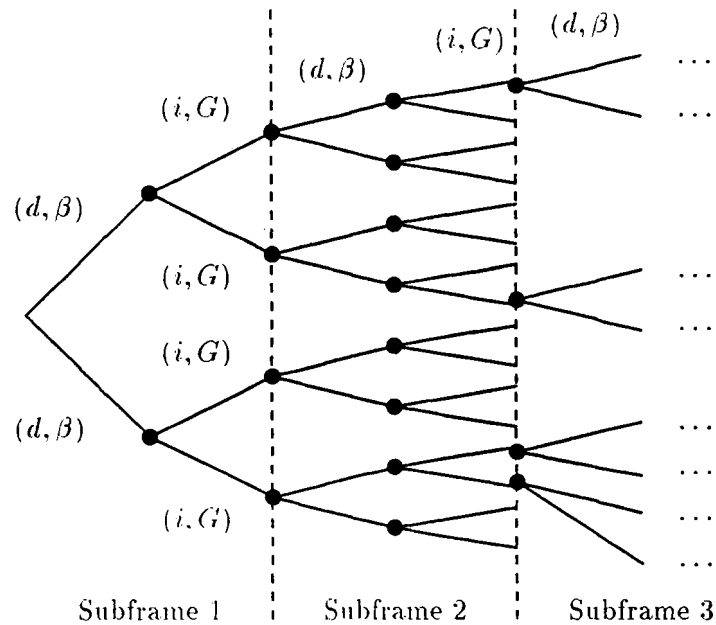


Figure 5.6: Delayed-decision coding tree (four surviving paths are kept at every stage).

sidered to be potential candidates for speech reconstruction. The selection process is repeated in the second subframe for each one of the four starting points, yielding in all 16 paths to be considered at the start of the third subframe. However, by bearing in mind that different filter memories and adaptive codebooks are associated with different paths, it is easily seen how computational complexity quickly rises. The maximum number of surviving paths is therefore limited to four, with the rest discarded. The cost associated with each path is the total mean squared error accumulated over the previous subframe stages in the tree. At every stage, the four paths yielding the lowest cost form the starting points for the next stage. At the end of the fifth subframe, the parameters of the path yielding the minimal accumulated mean squared error are transmitted.

Improvements up to 2 dB were recorded in both SNR and segSNR values when subframe parameters transmission was delayed until the fifth subframe. Statistics revealed that suboptimal parameter values were chosen about 89% of the time for the first three subframes of the decision tree, allowing, on a longer span of time, better

Speech	SNR (dB)	
	Unquantized	Quantized
Female	17.1	15.6
Male	13.8	12.7

Table 5.2: SNR average values for male and female coded speech.

coded speech perceptual quality. Indeed, the few clicks heard in unvoiced/voiced transition regions were attenuated with the introduction of delayed-decision coding.

5.6 Coding Scheme Performance

The reconstructed speech quality of a 7-bit log PCM coder was employed as the criterion for the subjective evaluation of the performance of the enhanced 8 kb/s CELP coder. Without the adaptive harmonic weighting and the delayed-decision coding techniques, the coded speech quality came very close to the reference quality but was still inferior in some particular transition regions (such as unvoiced/voiced, vowel/stop consonant). The delayed-decision method greatly contributed to the speech quality improvement at those transitions while harmonic noise weighting resulted in clearer voiced speech segments. The net result was a coded speech quality comparable to that of a 7-bit log PCM coder, and even superior for voiced regions. Objective SNR measures were also recorded for a collection of both male and female sentences. Although their actual value do not constitute a good quality evaluation criterion, they were used to evaluate the performance of the different parameter quantizers. Table 5.2 summarizes the obtained results for two versions of male and female speech sentences: one with unquantized CELP synthesis parameters and one with a fully quantized coder.

The training of the CELP excitation codebook was intentionally avoided in order to keep the coding scheme as much speech-context free as possible. Minor improvements were however obtained with the inclusion of a set of single-pulse excitation vectors in the codebook. By monitoring the codebook optimal index selection, it was found that single-pulse excitation indices were usually chosen at the onset of voiced

regions, ensuring thus a faster adaptation to the input speech pitch periods.

5.7 Conclusion

With all the features of an enhanced 8 kb/s CELP speech coder now added to the overall coding scheme, quality assessement is in order. Objective measures become insignificant at this point and the only way to carry out the performance evaluation is through a comparison with another well-established coding scheme. The G.721 CCITT standard is chosen for this purpose.

As the bit allocation resources became more limited, efficient and economical parameter quantization schemes turned out to be a necessity. With most of the available bits already used by the excitation codebook index, the LTP delay and the LPC parameters, only 8 bits/subframe remained available for the codebook and LTP gains quantization. A virtually transparent quantization could thus only be obtained by vector quantization. However, the erratic behaviour of the gains does not allow one to properly exploit the existing inter-correlation between them. An ingenious way to achieve the vector quantization was to use instead the per sample energy contributions of the excitation and adaptive optimal code vectors for transmission, along with an estimation of the frame overall energy. The gains could then be retrieved from those normalized energies. The subjectively evaluated reconstructed speech quality justifies the sufficiency of a 7-bit codebook energy contribution vector quantizer along with a 32 level uniformely quantized frame energy.

The coding structure sets the physical lower bound that the coding distortion can attain. Improving the coding quality after that becomes a matter of remodeling slightly the reconstructed speech structure to exploit the limitations of the human auditory system. Postfiltering is one way of enhancing the perceptual quality of the coder. Both a long term and a short term postfilters are implemented in this work, resulting in brighter speech quality.

Harmonic noise weighting, although not directly processing the reconstructed speech, contributes greatly to enhancing the speech periodicity for voiced segments by exploiting the masking capabilities of the spectrum harmonics. Spectral noise weighting was proven to lead to a more perceptually appropriate CELP weighted er-

ror criterion, but the improvements obtained with the introduction of harmonic noise weighting in the analysis-by-synthesis loop demonstrate that there is still potential for more perceptually valid distortion criteria.

Finally, by minimizing the spectral and harmonic noise weighted error criterion over a longer interval, the selected synthesis parameters, although suboptimal for a given subframe duration, yield better matched (to reference) reconstructed speech. The concept of trellis coding is adapted to yield delayed-decision coding scheme where the accumulated mean squared error is minimized over one frame of speech (5 subframes) before transmitting the parameters of the individual subframes. The complexity related to the frequent filter state updates along the search tree is the major drawback of this enhancement technique, but the substantial perceptual improvement over the subframe based optimization method renders the implementation worthwhile.

Chapter 6

Conclusion

The work in this thesis carried out investigations on the possibilities of achieving toll-quality speech coding at an operating rate of 8 kb/s. After standardizing the LD-CELP 16 kb/s coder, The CCITT has issued a set of requirements and recommendations for their next target, namely low-delay high-quality coding at 8 kb/s. The only existing potential candidate for standardization is the 8 kb/s LD-CELP coder proposed in [7]. However, due to the one-way coding delay constraint of 10 ms, the mean opinion score for this coding scheme did not exceed the 3.95 mark. Another successful version of high-quality coding at 8 kb/s is the VSELP [29] which was selected by the TIA (Telecommunications Industry Association) as the standard for use in North American digital cellular telephone systems. This coder, widely employed now for its robustness to channel errors and very good coding quality, was also unsuccessful in crossing the 4.0 mark (toll-quality indicator) on the MOS scale. By relaxing some of the constraints imposed on the two previous coding schemes, toll-quality reconstructed speech was indeed obtained at a coding rate of 8 kb/s in this work.

In view of the superiority in bit rate reduction capabilities of analysis-by-synthesis linear prediction based coders while maintaining high reconstructed speech quality, it was only logical that the implemented scheme relied on the Code Excited Linear Prediction (CELP) coding algorithm. For the chosen coding delay of 20 ms, it turned out that a good practical implementation for a corresponding analysis frame of 160 samples was an 10-th order formant predictor cascaded with a long term predictor

equivalent to a 3-tap predictor in performance. The covariance prediction method yields higher objective results (more than 1 dB in overall SNR) than those of the autocorrelation method when both procedures were tried out in turn in a full coder. However, the occasional unstable behaviour of the covariance scheme bends the choice toward the latter prediction method for which synthesis filter stability was guaranteed.

Perceptually smoother LPC parameters transition from one analysis frame to another were also obtained when a small amount of bandwidth expansion was provided to the formant filter coefficients before quantization. To this end, a binomial window of effective bandwidth of 80 Hz was applied to the autocorrelation coefficients before solving for the short term predictor coefficients.

As fewer bits per frame become available for LPC parameters quantization at medium rates, scalar quantization could not possibly yield a performance suitable for a toll-quality speech coder. Transparent quantization of the LPC parameters is a necessary condition for achieving toll-quality, and only vector quantizers are capable of yielding spectral distortions less than 1 dB at such low bit rates. Complexity problems however quickly arise with vector quantization as the codebook size grows. Chapter 3 proposed a vector quantization scheme that circumvents complexity by adopting a product-codebook model. The Line Spectral Frequencies (LSF) representation of LPC parameters was found to be an attractive form of parameterization due to the close relationship between the LSF properties and speech spectral characteristics.

A perceptually weighted Euclidean LSF distance measure was chosen to be the quantization distortion criterion. This weighting scheme takes into account the spectral hearing sensitivity and the speech spectrum related LSF properties to emphasize the more perceptually significant lower frequency regions. A 24 bits/frame split vector quantizer (split VQ) was constructed by creating one 4096-entry quantization codebook for the first four LSF's and another 4096-entry codebook for the remaining six LSF's. The codebooks were trained according to the LBG algorithm. Splitting the LSF parameter vector for quantization corresponds in essence to splitting the speech spectrum into lower energy and higher energy bands. The weighted Euclidean LSF distance, when used as a spectral distortion measure, resulted in average spectral distortion values around 1 dB. Two extra bits would have been necessary to achieve the same transparent quantization performance using the split VQ scheme with a simple

LSF Euclidean distance measure.

Optimization of the CELP synthesis stage parameters was detailed in Chapter 4. The usual closed-loop approach was used for determining the long term predictor parameters assuming that the excitation codebook does not provide any contribution to the reconstructed speech. However, once the optimal predictor delay is selected, the pitch predictor coefficient(s) was jointly optimized along with the excitation codebook index and gain. This sequential lag/joint gains optimization procedure increased substantially the perceptual quality of the coded speech when compared to the sequential optimization technique (pitch parameters followed by excitation codebook parameters). The subjective results agreed with the objective measures increases associated with the joint optimization technique, recording up to 2 dB increases in prediction gain and SNR values.

The quality of the reconstructed speech was further enhanced by allowing subsample resolution of the long term predictor delay. The fractional delays were resolved to 1/6 of a sample in critical pitch lag ranges, such as the female average pitch period range (not fully exploited by the basic CELP coding algorithm), and to 1/3 or 1/4 of a sample for other less sensitive regions. A very efficient interpolation procedure consisting of polyphase filtering rendered the operation of increasing the delay resolution computationally affordable for practical purposes. Hence, for a single-tap fractional delay pitch predictor, up to 1 dB SNR improvements were obtained with a small noticeable increase in perceptual quality, a performance comparable to that of a three-tap pitch predictor.

With few bits remaining for quantizing the excitation codebook and the long term predictor gains, vector quantization was found to be the only alternative for high-quality coding needs. It is however well-known that the gains and especially the pitch coefficient do not lend themselves well to vector quantization due to their occasional erratic behaviour. The correlation that exists between the periodic and stochastic components of the linear prediction excitation was rather exploited by vector quantizing the per sample energy contributions of the formant synthesized adaptive codebook entry and the excitation codebook entry. The quantized gains could then be recovered from these entities and from a uniformly quantized average frame energy. A 7-bit gains vector quantizer achieved very satisfactory results by

allowing only minor degradations in objective quality measures and slight perceptual distortions.

After optimizing the various stages of the CELP encoding process, the different techniques and quantization schemes were assembled to form a preliminary version of the 8 kb/s toll-quality coder. The reconstructed speech quality was however still not entirely convincing upon comparison with the output of a 7-bit log PCM coder. Since finer quantization was not anymore physically possible, the coding quality could only be improved by enhancing the perceptual features of speech signals. Spectral noise weighting of the CELP mean squared error between the original and the reconstructed speech has been until now the most popular way of exploiting the spectral masking properties of the human auditory system. On the same baseline, the implemented work in this thesis showed that the periodicity of voiced speech segments could be greatly enhanced by further weighting the mean squared error between the harmonics of the speech spectrum. The incorporation of the harmonic noise weighting technique in the analysis-by-synthesis loop increased the accuracy of the CELP error criterion, as the masking properties of the spectral harmonic regions were better exploited. Finally, on the decoder side, the reconstructed speech quality was also enhanced by adaptive short-term and long-term postfiltering. A brighter speech was the net perceptual result.

The last developement stage in the coding scheme addressed the limitations of confining the optimization of the CELP parameters to a speech subframe duration. By allowing suboptimal parameter values to be quantized at a given subframe, the consequently optimized parameters for the following subframe turned out to yield in more than 80 % of the cases a lower mean squared error than that resulting from independent parameters optimization for the two consecutive subframes. Those results have led to the elaboration of a delayed-decision coding scheme conceptually similar to trellis coding principles. An accumulated minimum mean squared error cost was assigned to every path in a delayed-decision coding tree where a maximum number of allowable paths were kept at every subframe stage. At the last subframe stage, the path with the minimum accumulated mean squared error had its quantized parameters transmitted for the total of five subframes in one frame. Substantial perceptual improvements in the coding quality resulted from this scheme, quantitatively equiva-

lent to over 2 dB increases in SNR measures. The major drawback in this scheme is however the increased computational complexity issue. One could also worry about the effect of channel errors propagation along the delayed-decision coding tree stages. Nevertheless, since the work in this thesis was only at the experimental level, complexity reduction was not the major target and the coder performance was evaluated in error-free channel conditions.

Informal comparison listening tests between the completed CELP coding scheme and a 7-bit log PCM coder revealed that the quality of the two reconstructed speech versions was perceptually equivalent. Moreover, clearer CELP coded speech resulted in some voiced regions, due to the periodicity enhancement techniques employed.

The CCITT specifications for standardizing the 8 kb/s coder require a one-way coding delay less than 10 ms. Investigations in this work have been carried out to lower the adopted 20 ms coding delay. The speech quality suffered slightly from reducing this delay to 16 ms. and toll-quality was lost. As it was mentioned previously, the Low-Delay CELP coder operating at 8 kb/s [7] has characteristics that are the closest to the CCITT specifications, but does not achieve yet toll-quality coding. The quality enhancement that resulted from the combined harmonic and spectral noise weighting scheme and especially from the delayed-decision coding technique at no extra bit rate penalties is a very encouraging step toward future research in achieving toll-quality coding at medium bit rates. Starting from the Low-Delay 8 kb/s CELP coder, perceptual enhancement techniques should be able to increase the coded speech quality, and eventually reach the performance of a 7-bit log PCM coder with the application of delayed-decision coding. The latter improvement method can however quickly increase the computational complexity of low-delay coding applications, as the parameter update rate becomes much more frequent (shorter subframes) in addition to the LPC parameters being backward adapted. Such characteristics are reflected in an increased number of stages in the delayed-decision tree as well as a separate LPC analysis for every alternative (path) at a given stage in the tree. Procedures to bring down the complexity of delayed-decision coding in a backward adaptive LPC analysis coding environment might be the solution for attaining toll-quality when low coding delay constraints are imposed.

Appendix A

An efficient polyphase network for the common 1-to-D digital interpolator is derived in this appendix, followed by a brief overview on some of the properties of the polyphase filters used in the structure.

A block diagram for a sampling rate increase by D is given in Fig. A.1. The sampling rate expander inserts $L-1$ zero valued samples between each pair of samples of $x(n)$ to yield the signal $w(n)$:

$$w(m) = \begin{cases} x(\frac{m}{D}), & m = 0, \pm D, \pm 2D, \ldots \\ 0, & \text{otherwise} \end{cases} . \tag{A.1}$$

The spectrum of $w(m)$ will contain the baseband frequencies of interest ($-\pi/D$ *to* π/D) plus images of the baseband centered at harmonics of the original sampling frequency $\pm 2\pi/D, \pm 4\pi/D, \ldots$. The baseband signal is recovered by passing $w(m)$ through an ideal digital low-pass filter $h_{LP}(m)$. In the frequency domain, the ideal filter response $H_{LP}(e^{jw})$ is known to be:

$$H_{LP}(e^{jw}) = \begin{cases} D, & |w| \leq \frac{\pi}{D} \\ 0, & \text{otherwise} \end{cases} , \tag{A.2}$$

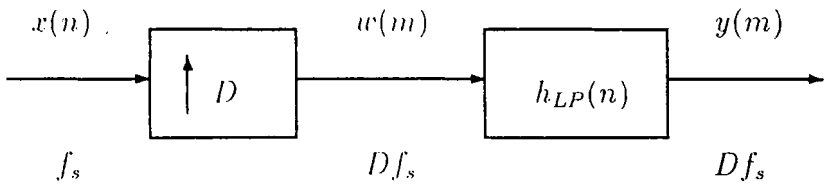


Figure A.1: Block diagram for interpolation by an integer factor D.

and the interpolation output signal $y(m)$ will be:

$$Y(e^{jw}) = \begin{cases} D\ X(e^{jwD}), & |w| \leq \frac{\pi}{D} \\ 0, & \text{otherwise} \end{cases} . \tag{A.3}$$

The output signal $y(m)$ can be expressed as the convolution of the input signal with the impulse response of the ideal low-pass filter $h_{LP}(m)$, written as

$$\begin{aligned} y(m) &= \sum_{k=-\infty}^{\infty} h_{LP}(m-k)\ x(\lfloor \frac{k}{D} \rfloor) \\ &= \sum_{r=-\infty}^{\infty} h_{LP}(m-rD)\ x(r). \end{aligned} \tag{A.4}$$

By introducing the change of variable

$$r = \lfloor \frac{m}{D} \rfloor - n \tag{A.5}$$

where $\lfloor \alpha \rfloor$ is the least integer less than or equal to α, Eq. (A.4) becomes:

$$y(m) = \sum_{n=-\infty}^{\infty} h_{LP}(m - \lfloor \frac{m}{D} \rfloor D + nD)\ x(\lfloor \frac{m}{D} \rfloor - n). \tag{A.6}$$

With the *modulo* notation $m \oplus D$ being more compact for $m - \lfloor \frac{m}{D} \rfloor$, the output $y(m)$ is finally expressed as:

$$y(m) = \sum_{n=-\infty}^{\infty} h_{LP}(nD + m \oplus D)\ x(\lfloor \frac{m}{D} \rfloor - n). \tag{A.7}$$

The coefficients of the low-pass filter impulse response in (A.7) can be denoted by $g_m(n)$, where

$$g_m(n) = h_{LP}(nD + m \oplus D), \tag{A.8}$$

for all m and n. The set of coefficients $\{g_m(n)\}$ can be seen as a periodically time varying filter with period D. $y(D)$ is thus generated using the same set of coefficients $\{g_0(n)\}$ as for $y(0)$, $y(D+1)$, like $y(1)$, uses $\{g_1(n)\}$, and so on. On the other hand, the input signal $x(n)$ increases by one sample for every D output samples. In general the output samples $y(rD), y(rD+1), \ldots, y(rD+D-1)$ are obtained from the input samples $x(r-n)$. The signal $x(n)$ is thus updated at the low sampling rate f_s, while the output $y(m)$ is evaluated at the high sampling rate Df_s.

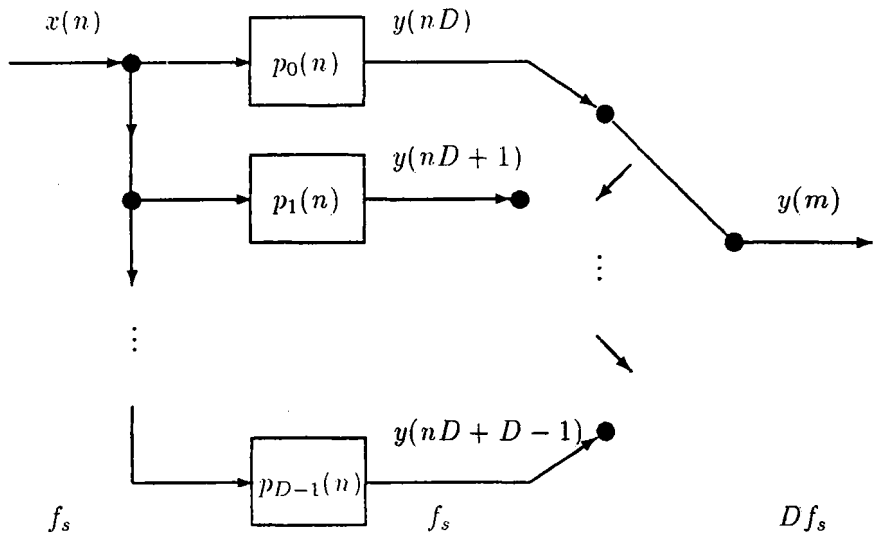


Figure A.2: Commutator model for a $1-to-D$ interpolator.

The ideal low-pass filter impulse response $h_{LP}(m)$ can be partitioned into D filter subsets operating at the low sampling rate. These subsets are D separate linear time-invariant filters, $p_0(n), p_1(n), \ldots, p_{D-1}(n)$, known as *polyphase filters*. The $k-th$ polyphase filter is given by:

$$p_k(n) = g_k(n), \tag{A.9}$$

for $0 \leq k \leq D-1$ and all n. With the help of (A.8), the expression for the polyphase filters becomes

$$p_k(n) = h_{LP}(nD+k) \quad k = 0, 1, 2, \ldots, D-1 \tag{A.10}$$

for all n. For each new input sample $x(n)$, D samples $y(nD+k)$ will thus be generated as the output of the D successive polyphase filters.

With the polyphase filtering structure now introduced, the $1-to-D$ interpolator can be efficiently represented by the counterclockwise commutator model shown in Fig. A.2. The filtering in the polyphase interpolation network is performed at the low sampling rate. For each input sample $x(n)$, the commutator sweeps through the D polyphase paths to get D output samples of $y(m)$.

Taking a closer look at the definition of the polyphase filters in (A.10), it is seen that they correspond to decimated versions (by a factor of D) of the low-pass filter

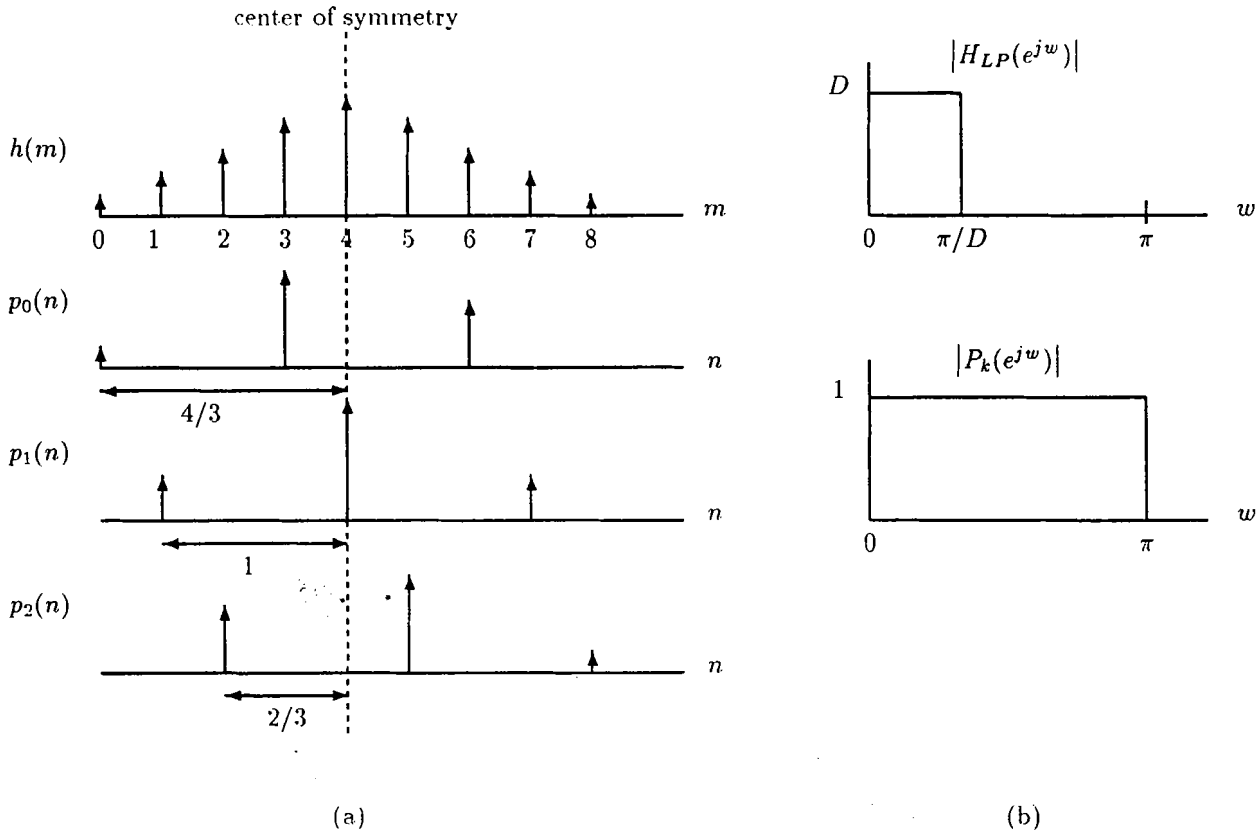


Figure A.3: Polyphase filters properties: (a) fractional sample phase shifts and (b) all-pass frequency response.

$h_{LP}(m)$. This ideal low-pass filter is very often approximated by linear phase FIR filter $h(n)$. The corresponding polyphase filters will naturally also be finite. Fig. A.3 (a) examplifies the decimation process for an interpolation factor $D = 3$ and a 9-tap FIR interpolator. The FIR interpolator is shown to be symmetric about $m = 4$, and thus having a flat delay of 4 samples. The points of symmetry of the envelopes of $p_0(n)$, $p_1(n)$, and $p_2(n)$ are respectively at 4/3 of a sample, one sample, and 2/3 of a sample. Different phase shifts are thus associated with the different FIR polyphase filters, and hence justifying the origin of the terminology. Generally, if the FIR low-pass filter approximation is of length N, the polyphase filters will be of length N/D. Choosing N to be a multiple of D will yield polyphase filters of the same length.

Fig. A.3 (b) shows the scaling of the polyphase filters frequency response $P_k(e^{jw})$ from the range $0 \leq w \leq \pi/D$ corresponding to the ideal low-pass filter response to the range $0 \leq w \leq \pi$, due to the decimation process. It can hence be concluded that FIR polyphase filters approximate ideal *all-pass linear phase* filter characteristics, with each value of k corresponding to a different phase shift.

Appendix B

This appendix briefly details the computation and the uniform quantization scheme of the frame energy. It also illustrates the interpolation procedure used to obtain the subframe energy estimates necessary for the recovery of the codebook gains.

As shown in Fig. B.1, the computation of the frame energy is based on one analysis frame rather than on one frame to be encoded in order to preserve continuity in the subframe energy estimates. Assuming the order of prediction to be p and the analysis frame length to be N_A, the frame energy of the input speech $s(n)$ is given by:

$$R(0) = \frac{\Phi(0,0) + \Phi(p,p)}{2(N_A - p)} \tag{B.1}$$

where

$$\Phi(i,k) = \sum_{n=p}^{N_A} s(n-i)s(n-k). \tag{B.2}$$

The energy normalized by $R_{max} = s_{max}(n)^2$ in expressed in the log domain as:

$$R_{dB} = 10\log(R(0)/R_{max}). \tag{B.3}$$

The implemented 5-bit uniform quantizer has 2 dB width bins uniformly distributed along the log-energy range. The transmitted quantization index I is hence determined according to the following equations:

$$I = \begin{cases} 0 & \text{if } R_{dB} < -72 \\ 1\ldots31 & \text{s.t. } |I - (R_{dB}+66)/2| \text{ is minimal} \end{cases}. \tag{B.4}$$

From this transmitted index, the energy on the decoder side is recovered by:

$$\tilde{R}(0) = \begin{cases} R_{max}\ 10^{(2I-66)/10} & I \neq 0 \\ 0 & I = 0 \end{cases}. \tag{B.5}$$

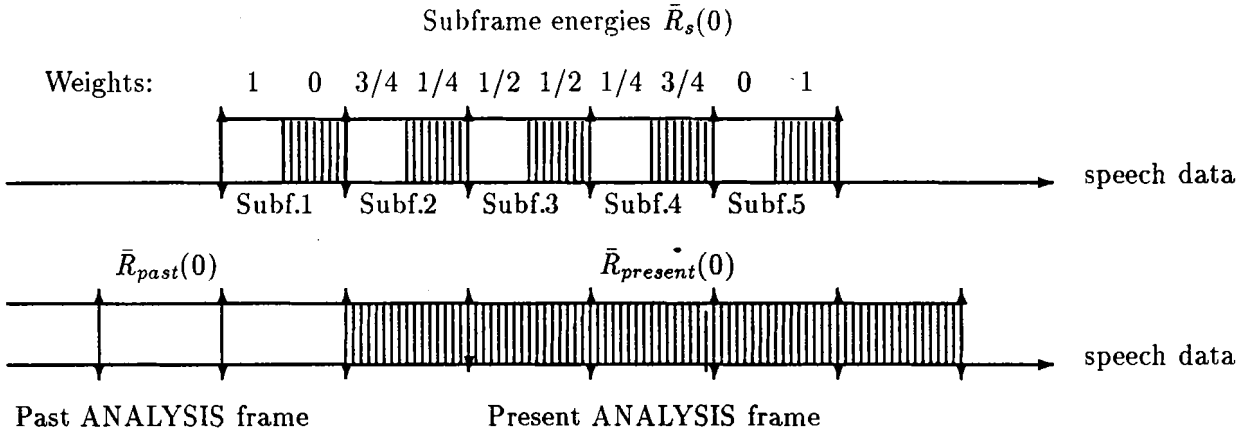


Figure B.1: Interpolation scheme for the subframe energy estimates.

The subframe energy estimation $\bar{R}_s(0)$ is based on a direct interpolation of the past analysis frame quantized energy $\bar{R}_{past}(0)$ and the present analysis frame quantized energy $\bar{R}_{present}(0)$. In other terms, it is obtained as a weighted combination of the quantized frame energies:

$$\bar{R}_s(0) \;=\; w_i \bar{R}_{past}(0) \;+\; (1 - w_i)\bar{R}_{present}(0), \tag{B.6}$$

with the weighting scheme $\{w_i\}$ illustrated in Fig. B.1.

References

[1] R.E. Ziemer, W.H. Tranten and D.R. Fannin, *Signals and Systems: Continuous and Discrete*, 2-nd Edition, Mcmillan Publishing Company, NY (1990).

[2] A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ (1975).

[3] B.S. Atal and M.S. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria," *IEEE Trans. Acoust. Speech Signal Proc.*, **ASSP-27**(3), pp. 247–254 (1979).

[4] W.B. Kleijn, "Analysis-by-Synthesis Speech Coding Based on Relaxed Waveform-Matching Constraints," *Doctorate Thesis*, Delft University of Technology, (1991).

[5] C.E. Shannon, "Communication in the Presence of Noise," *Proc. I.R.E* **37**(3), pp. 10–21 (1949).

[6] N.S. Jayant, "High-Quality Coding of Telephone Speech and Wideband Audio," *IEEE Communication Magazine*, pp. 10–20 (Jan. 1990).

[7] J-H. Chen, "An 8 kb/s Low-Delay CELP Speech Coder," *GLOBECOM '91*, pp. 1894–1897 (1991).

[8] D. O'Shaughnessy, *Speech Communication*, Addison-Wesley, (1990). Englewood Cliffs, NJ (1975).

[9] P. Strobach, *Linear Prediction Theory*, Springer-Verlag, NY (1990).

[10] B.S. Atal, "Predictive Coding of Speech Signals at Low Bit Rates," *IEEE Trans. Comm.*, **COM-30**(4), pp. 600–614 (1982).

[11] B. Townshend, "Nonlinear Prediction of Speech," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Toronto*, pp. 425–428 (1991).

[12] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE* **63**(4), pp. 561–580 (1975).

[13] W. F. LeBlanc and V. Cuperman, "Sequential Optimization of CELP Speech Coding at 4 kb/s," *Abstracts IEEE Workshop on Speech Coding for Telecommunications, Whistler*, pp. 105–106 (1991).

[14] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ (1979).

[15] J. Makhoul, "Stable and Efficient Lattice Methods for Linear Prediction," *IEEE Trans. Acoust. Speech Signal Proc.*, **ASSP-25**, pp. 423–428 (1977).

[16] A. Cumani, "On a Covariance Lattice Algorithm for Linear Prediction," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Paris*, pp. 651–654 (1982).

[17] J. Makhoul and L.K. Cosell, "Adaptive Lattice Analysis of Speech," *IEEE Trans. Speech Signal Proc.* **ASSP-29**(3), pp. 654–659 (1981).

[18] A. Papoulis, *Probability, Random Variables and Stochastic Processes, 2nd ed.*, McGraw-Hill, New-York (1984).

[19] J-H. Chen, "High-Quality 16 kb/s Speech Coding with a one-way delay less than 2 ms," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Albuquerque*, (1990).

[20] M. Foodeei and P. Kabal, "Backward Adaptive Prediction: High-Order Predictors and Formant-Pitch Configurations," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Toronto*, pp. 2105–2109 (1991).

[21] P. Kabal and R.P. Ramachandran, "Joint Optimization of Linear Predictors in Speech Coders," *IEEE Trans. Acoust. Speech Signal Proc.*, **ASSP-37**, pp. 642–650 (1989).

[22] T.P. Barnwell, III, "Recursive Windowong For Generating Auto-Correlation Coefficients for LPC Analysis," *IEEE Trans. Acoust. Speech Signal Proc.*, **ASSP-29**(5), pp. 1062–1066 (1981).

[23] S. Singhal and B.S. Atal, "Improving Performance of Multipulse LPC Coders at Low Bit Rates," *Proc. Int. Conf. Acoust. Speech and Sign. Process., San Diego*, pp. 1.3.1–1.3.4 (1984).

[24] U. Kipper, H. Reininger, and D. Wolf, "Improved CELP Coding Using Adaptive Excitation Codebooks," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Toronto*, pp. 237–240 (1991).

[25] P. Kroon and B.S. Atal, "On the Use of Pitch Predictors with High Temporal Resolution," *IEEE Trans. Acoust. Speech Signal Proc.* **ASSP-39**(3), pp. 733–736 (1991).

[26] B. Berouti, H. Garten, P. Kabal, and P. Mermelstein, "Efficient Computation and Encoding of the Multipulse Excitation for LPC," *Proc. Int. Conf. Acoust. Speech and Sign. Process., San Diego*, pp. 10.1.1–10.1.4 (1984).

[27] E.F. Deprette and P. Kroon, "Regular Excitation Reduction for Effective and Efficient LP-Coding of Speech," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Tampa*, pp. 965–968 (1985).

[28] B.S. Atal and M. R. Schroeder, "Stochastic Coding of Speech at Very Low Bit Rates," *Proc. Int. Conf. Comm., Amsterdam*, pp. 1610–1613 (1984).

[29] I.A. Gerson and M. A. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8 kbps," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Albuquerque*, pp. 461–464 (1990).

[30] J.P. Campbell, V.C. Welch, and T.E. Tremain, *CELP Documentation Version 3.2*, U.S. DoD, Fort Mead, MD (1990).

[31] J-H. Chen, N. Jayant, and R.V. Cox, "Improving The Performance of The 16 kb/s LD-CELP Speech Coder," *Proc. Int. Conf. Acoust. Speech and Sign. Process., San Fransisco*, pp. I-69–I-72 (1992).

[32] 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).

[33] Y. Shoham, "Constrained Stochastic Excitation Coding of Speech at 4.8 kb/s," *Advances in Speech Coding*, pp. 339-348, Kluwer Academic Publishers (1991).

[34] I.A. Gerson and M. A. Jasiuk, "Techniques for Improving The Performance of CELP Type Speech Coders," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Toronto*, pp. 205–208 (1991).

[35] M.J. Sabin and R.M. Gray, "Product Code Vector Quantizers for Waveform and Voice Coding," *IEEE Trans. Acoust. Speech Signal Proc.*, **ASSP-32**(3), pp. 474–488 (1984).

[36] G. Davidson and A. Gersho, "Complexity Reduction Methods for Vector Excitation Coding," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Tokyo*, pp. 3055–3058 (1986).

[37] G. Davidson and A. Gersho, "Multiple-Stage Vector Excitation Coding of Speech Waveforms," *Proc. Int. Conf. Acoust. Speech and Sign. Process., New-York*, pp. 163–166 (1988).

[38] I.M. Trancoso and B.S. Atal, " Efficient Procedures for Finding the Optimum Innovation in Stochastic Coders," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Tokyo*, pp. 2379–2382 (1986).

[39] J-P. Adoul and C. Lamblin, "A Comparison of Some Algebraic Structures for CELP Coding of Speech," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Dallas*, pp. 1953–1956 (1987).

[40] T. Shoham, "Cascaded-Likelihood Vector Coding of The LPC Information," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Glasgow*, pp. 160–163 (1989).

[41] 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., Toronto*, pp. 661–664 (1991).

[42] A.H. Gray and J.D. Markel, "Quantization and Bit Allocation in Speech Processing," *IEEE Trans. Acoust. Speech Signal Proc.*, **ASSP-24**, pp. 459–473 (1976).

[43] F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients," *J. Acoust. Soc. Am.* **57 Supplement(1)**, pp. S.35 (1975).

[44] G.S. Kang and L.S. Fransen, "Application of Line Spectrum Pairs to Low Bit Rates Speech Encoders," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Tampa*, pp. 7.3.1–7.3.4 (1985).

[45] F.K. Soong and B.H. Juang, "Line Spectrum Pairs (LSP) and Speech Data Compression," *Proc. Int. Conf. Acoust. Speech and Sign. Process., San Diego*, pp. 1.10.1–1.10.4 (1984).

[46] P. Kabal and R.P. Ramachandran, "The Computation of Line Spectral Frequencies Using Chebyshev Polynomials," *IEEE Trans. Acoust. Speech Signal Proc.*, **ASSP-34**(3), pp. 1419–1426 (1986).

[47] F.K. Soong and M.M. Sondhi, "A Frequency-Weighted Itakura Spectral Distortion Measure and Its Application to Speech Recognition in Noise," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Dallas*, pp. 625–628 (1987).

[48] G.S. Kang and L.S. Fransen, "Low Bit Rates Speech Encoders Based on Line Spectrum Frequencies (LSF's)," *Naval Research Laboratory Report 8857*, (1984).

[49] J. Grass and P. Kabal, "Quantization of Predictor Coefficients in Speech Coding," *Rapport Technique de L'INRS-Télécommunications no. 91-01*, (1991).

[50] B.H. Juang, D.Y. Gray and A.H. Gray Jr., "Distortion Performance of Vector Quantization for LPC Voice Coding," *IEEE Trans. Acoust. Speech Signal Proc.*, **ASSP-30**, pp. 294–303 (1982).

[51] T. Moriya and M. Honda , "Speech Coder Using Phase Equalization and Vector Quantization," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Tokyo*, pp. 1701–1704 (1986).

[52] Y. Linde, A. Buzo and R.M. Gray, "An Algorithm for Vector Quantization Design," *IEEE Trans. on Comm.*, **COM-30**, pp. 84–95 (1980).

[53] H-Y. Su and P. Mermelstein, "Improving The Speech Quality of Cellular Mobile Systems Under Heavy Fading," *Proc. Int. Conf. Acoust. Speech and Sign. Process., San Fransisco*, pp. II-121–II-124 (1992).

[54] B.S. Atal, R.V. Cox, and P. Kroon, "Spectral Quantization and Interpolation for CELP coders," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Glasgow*, pp. 69–72 (1989).

[55] J-L. Moncet and P. Kabal, "Codeword Selection for CELP Coders," *Proc. Int. Conf. Acoust. Speech and Sign. Process., New-York*, pp. 147–150 (1988).

[56] R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ (1983).

[57] T. Parks and D. Kolba, "Interpolation Minimizing Maximum Normalized Error for Bandlimited Signals," *IEEE Trans. Acoust. Speech Signal Proc.*, **ASSP-26**(4), pp. 381–384 (1978).

[58] J.H. Chen and A. Gersho, "Real-Time Vector APC Speech Coding at 4800 bps with Adaptive Postfiltering," *Proc. Int. Conf. Acoust. Speech and Sign. Process., Dallas*, pp. 2185–2188 (1987).