A ROBUST LOW BIT RATE QUAD-BAND EXCITATION LSP VOCODER

BY

CHIU KIM MING

A MASTER THESIS
SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT
FOR THE DEGREE OF MASTER OF PHILOSOPHY
IN
THE DEPARTMENT OF ELECTRONIC ENGINEERING
THE CHINESE UNIVERSITY OF HONG KONG

HONG KONG
JUNE, 1994

Acknowledgment

I would like to gratefully acknowledge Dr. P.C. Ching for his guidance and encouragement during the research work. I would also like to thank Mr. M.H. Ko and Mr. Brian Mak for the technical supports and the help on computer programming. In addition, I would like to express my gratitude to Mr. H.C. So and Mr. T. Lee for the advice and suggestions on digital signal processing and speech processing techniques.

Last but not least, I would like to thank Mr. S. Dodson, Mr. W.K.Lai, Mr. H.C. So, Mr. C.M. Pang and Mr. C.Y. Chik for the participation in the listening tests.

Abstract

Speech is one of the most important tools in communications. In recent telecommunication applications, digital mobile telephony plays an essential role for interchanging of information and messages. Due to the increasing demand on wireless communications, the capacity of frequency channels needs to be enlarged to satisfy a huge number of users. One of the methods for increasing the channel capacity in a limited transmission space is to develop data compression techniques for speech coders that can be operated at very low bit rates. The aim of this thesis is to design a Quad-Band Excitation Line Spectral Pair vocoder that is capable to operate at a bit rate as low as 2.2 kb/s.

The first part of the study mainly concentrates on the development of an effective excitation model. Since most short time speech spectra have strong voiced characteristics in the low frequency region whereas unvoiced features are most found in the high frequency area, a Dual-Band Excitation model is developed in which no more than 2 excitation bands are used to represent the excitation spectrum. In order to retain the finer spectral properties of the excitation signals, a Quad-Band Excitation model is next introduced. In this method, a maximum of 4 excitation bands are used in a single speech frame and a fairly good replica of the excitation spectrum can be obtained. When compared with the existing Multiband Excitation model which normally uses up to 12 frequency bands in each frame, Quad-Band Excitation has a lower computational complexity and is more suitable for low bit rate speech encoding.

In the second part of the study, a Quad-Band Excitation Line Spectral Pair vocoder has been developed for low bit rate transmission. This vocoder is based on linear predictive model of speech signals. Line spectral pairs are used to represent the spectral envelope because of its effectiveness in encoding. Different quantization schemes are applied to parameter coding including non-linear quantization and split vector quantization. The vocoder is simulated on computer with a operational bit rate of 2.2 kb/s, and the synthesized speech has been found to be highly intelligible with a reasonably good quality.

Table of contents

Chapter 1 Introduction

Speech is one of the most important tools for human communication, it contains not only information but also messages. Communication through speech is the easiest and the most direct way to understand each other. New technologies make good use of this kind of communication so that interchange of information and messages can be easily done throughout the world. Recent development of speech communication not only covers man-man communication but also extends to man-machine communication with many practical applications including text-to-speech synthesis and other features in multimedia application.

As speech is a natural tool in communication, development of speech encoding techniques acts as a leading role in communication applications, such as telephony and voice mail system. Its application would be much more enormous if it is incorporated in Integrated Services Digital Network (ISDN) for both civil and military use [1]. Particularly, in telecommunication applications such as mobile telephony, due to the dramatic increase in user demand but the space for wireless data transmission is limited, the transmission bandwidth of speech codec need to be as narrow as possible while maintaining the output quality at the same time, so that more channels can be available to the huge amount of users. On speech codec in digital telephony, the transmission bandwidth can be confined by reducing the operating bit rate of the codec so that more channels can be available to satisfy the market. Current studies show that model-based speech codec can be operated as low as 4 kbps with high output quality [2]. Due to the rapid development on signal processor chips, low bit rate speech coding algorithms can be used in hardware implementation even the encoding and decoding algorithms of them are much more complicated. As a result, low bit rate speech coding has attracted continuous worldwide attention with an aim to develop a speech codec that is robust, and can be operated at a suitably low bit rate with reasonable output quality to satisfy future demand in telecommunication systems.

1.1 Speech production

Speech sounds can be classified into three distinct classes according to their mode of excitation, namely the voiced sounds, the unvoiced sounds and the plosive sounds. In human speech production system, voice originates from the vibration of a pair of muscles called the vocal cord. They modulate the air flow from the lungs. To produce voiced sound such as vowel, the vocal cord vibrates quasi-periodically. If unvoiced sound is to be produced, like fricatives, noise-like signal will be given. Those vibrating air flow passes through a volume formed by the vocal tract, the mouth cavity and the nasal duct. The spectral shape of the signal is then modified by the resonant structure of the volume to produce voice. Figure 1.1 shows the human speech production system diagramatically . Different voices come from both the variation of the vocal cord and the transfer function of the resonant path which changes with its volume. By examining time domain speech waveforms, it is noted that voiced sounds contain a quasi-periodic structure whereas the unvoiced sounds are rather noisy [3].

Physically, the production process of voiced sound can be modeled by a lossless tube, which is excited by a source of quasi-periodic glottal pulses. This tube has its resonance modes and are usually called formants. In the frequency domain, voiced sounds are characterized by the existence of several formant peaks, usually 4 to 5, in their power spectra. Since unvoiced sounds are not generalized by vocal cord vibrations, there are no resonance properties observed in their power spectra. Unvoiced sounds are, in general, characterized by a high energy contribution from the high frequency components in the power spectra.

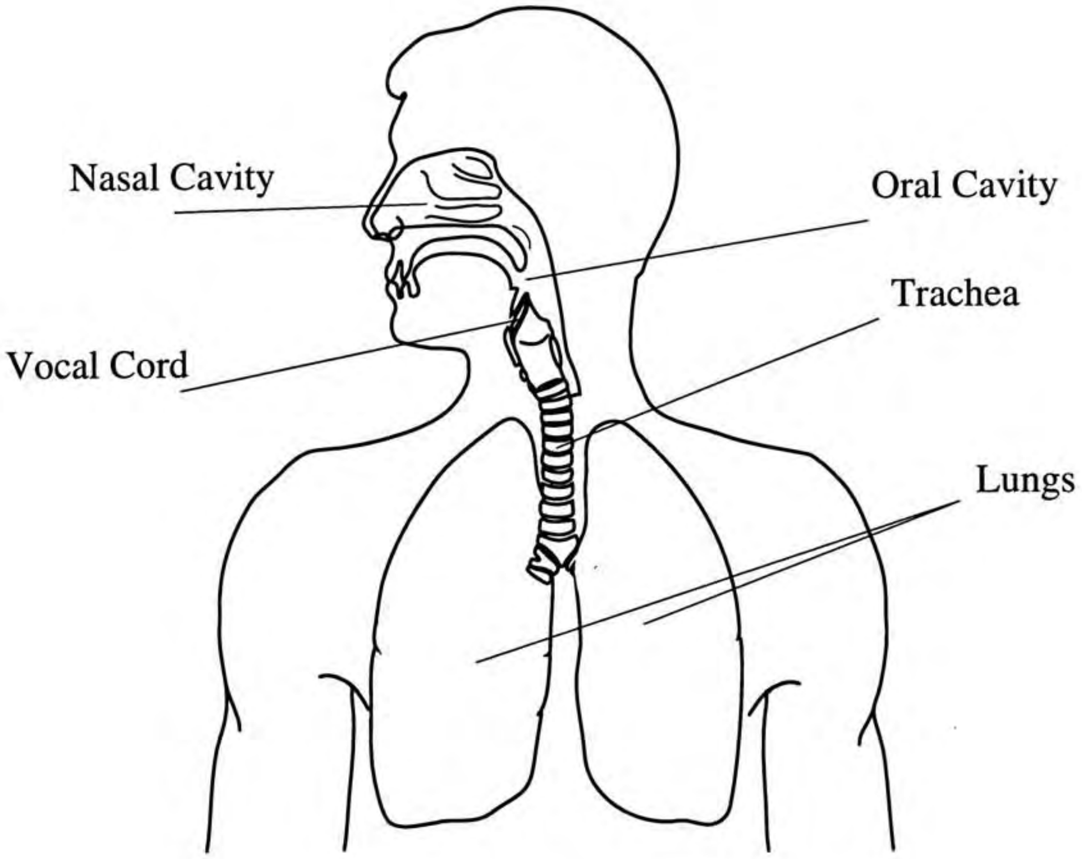


Figure 1.1 Human speech production system

Furthermore, periodicity of the pulses controls the pitch or fundamental frequency of voice tone while temporal variation of pitch determines accent and intonation of uttered words, phrases and sentences. Male speakers usually have a lower fundamental frequency, typically ranges from 100 Hz to 200 Hz, and female speakers, on the other hand, have pitch value as high as 300 Hz [3].

1.2 Low bit rate speech coding

In digital speech coding, the simplest technique is called waveform coding [4]. In this method, the speech waveform is sampled at Nyquist frequency and the sample data sequence is then quantized and transmitted. It can provide high output speech quality and the major error is mostly due to quantization. The operation bit rate is relatively high. Typical bandwidth of speech signal is 4 kHz, results in the minimum possible Nyquist frequency of 8 kHz. If 8 bits are used for coding, as in standardized Pulse Code Modulation (PCM) [5], the resulting bit rate would be 64 kbps. An improved version called Adaptive Differential Pulse Code Modulation (ADPCM) [6] has been introduced to reduce the transmission of redundant information. It requires a bandwidth of 32 kbps and is commonly used in telephone network. In Delta Modulation (DM) [5][7], since only one bit is used in each sample data, the bit rate is essentially equal to the sampling frequency. However, in order to achieve toll quality speech, the sampling rate normally adopted is 16 kilo-samples per second. It can be seen that the algorithms based on waveform coding requires a rather large transmission bandwidth even though it can provide highly intelligible output speech. Consequently, the application of waveform coding is somewhat limited particularly for radio communication in which bandwidth is rather stringent.

In order to further reduce the transmission bandwidth, model-based algorithms are used, in which speech characteristics parameters instead of sample data are encoded and transmitted. The main difference between waveform coder and model-based coder is that the former intends to reconstruct the signal with the exact waveform, whereas the latter tries to retain the characteristics of the signal. Model-based speech coder is introduced by representing the physical speech production system in terms of a mathematical model in which system parameters are used to characterize the model through analysis [1]. These parameters are transmitted to the receiver to regenerate the speech signal. Channel vocoder and formant vocoder [8] are typical examples of model-based vocoders. A more important and widely used model is called Linear Predictive Coding (LPC) [9]-[13]. In this system, the instant speech data is represented by a linear combination of a finite number of past samples. The parameters which characterize the spectral properties of the speech signal are calculated by minimizing the mean-squared error of the system and it can be computed easily. Presently many speech codec are developed based on the LPC model.

The output quality of the synthesized speech using the original LPC model is rather poor due to insufficient interpretation of the excitation signal. Extensive studies have been concentrated, in the past decade, on the expression of the excitation signal to improve the quality of the synthesized speech. In Code Excited Linear Predictive Coding (CELP) [14], a number of excitation patterns are stored in a code-book. The code index which denotes the suitable excitation is found by closed loop optimization, and needs to be sent to the transmitter. CELP can be operated at 4 kbps with communication quality [1]. An alternative method is called Vector Sum Excited Linear Predictive Coding (VSELP) [15]. The excitation, in this case, is a linear combination of a number of basis vectors and optimization is performed during speech analysis. Multipulse Excited Linear Predictive Coding (MPELPC) [16] is yet another method but has proved to be more efficient and effective. In this approach, the location and the magnitude of a certain number of pulses are found in each speech segment to construct an optimum excitation signal. Toll quality can be produced when it is operated at 8 kbps [1].

Another type of model-based vocoder which is different from LPC is called the Multiband Excitation (MBE) vocoder [17]. In contrast to the widely used speech production model in which each speech frame is either classified as voiced or unvoiced, MBE vocoder allows a combination of voiced and unvoiced spectrum in a single frame of speech signal, so that it gives a higher flexibility in expressing the excitation spectrum. The first MBE vocoder is designed to operate at 8 kbps and is capable of producing high synthetic speech quality. There are several factors in MBE vocoder which limit its lowest possible bit rate of operation. Firstly, instead of using a single voiced/unvoiced (v/uv) decision for each speech segment, MBE model requires a series of decisions which in turn consumes more bits for transmission. The excessive number of data bits required, of course, depends on the number of frequency bands in the spectrum that need v/uv decision. Secondly, the substantial amount of bits required to encode the parameters of the spectral envelope is also a limiting factor. By using data compression technique, it can be operated at a bit rate as low as 4.8 kbps [18], but a higher computational complexity is expected.

It is noted that most ordinary speech spectra contain voiced portion in the low frequency region and unvoiced portion in the high frequency region. Thus the large number of frequency bands that are commonly employed in conventional MBE vocoder is likely to be redundant. It is possible that a speech spectrum can be divided into two parts: the low frequency region which is basically a voiced band while the remaining is regarded as an unvoiced band. Thus an excitation consists of only two bands may be adequate to model the input spectrum [19]. Of course, if a finer spectral structure of the excitation signals is required, more frequency bands could be used for each speech frame to render good quality.

The main theme of this thesis is to develop a speech codec that can be operated at a fairly low bit rate while maintaining the synthesized speech quality at a highly acceptable level for commercial use. A Quad-Band Excitation model is proposed in which no more than 4 non-overlapping frequency bands with variable bandwidth are used. It retains not only the flexibility on the representation of the excitation spectrum as in MBE model, but also reduces the complexity of the codec system. A Quad-Band Excitation Line Spectral Pair Vocoder has been implemented. In this vocoder, the number of bits in encoding the voiced/unvoiced decision is reduced when compared with the MBE vocoder. Furthermore, the spectral envelope is modeled by a set of LPC coefficients which are then converted into line spectral pairs (LSP) for efficient transmission. This has increased the possibility of the vocoder to operate at a low bit rate for digital telephony. The proposed vocoder is operated at 2.2 kbps and it has been tested using computer simulation and the synthesized speech output is found to be highly intelligible with reasonably good quality.

The organization of this thesis is as following: In Chapter 2, linear prediction of speech signal will be reviewed with different kinds of excitation being used. Some commonly used speech analysis/synthesis methodologies are discussed as well. The proposed Dual-Band and Quad-Band excitation will be introduced in details in Chapter 3. In Chapter 4, the Quad-Band Excitation LSP vocoder will be described and the implementation details of the vocoder are also included. Simulation results of the proposed vocoder are given in Chapter 5. Finally, a conclusion will be provided in Chapter 6.

Chapter 2 Speech analysis & synthesis

The most widely used technique in speech analysis and synthesis is linear prediction of speech signal. In this method, the current speech sample is assumed to be a linear combination of a number of past samples, which can be denoted by an auto-regressive process. The parameters that characterize the speech signal in this model are the pitch value, voiced/unvoiced decision and the filter coefficients which represent the spectral envelope. In this chapter, a general discussion about speech analysis and synthesis based on linear prediction is given. We shall also review different types of excitations such as coded excitation, vector sum excitation, regular pulses and multipulse excitation, that can be incorporated in linear predictive coding. Furthermore, the concept of multiband excitation and the corresponding vocoder will also be introduced.

2.1 Linear prediction of speech signal

In order to represent the human speech production mechanism effectively by using a mathematical model, linear prediction method is widely used to extract the speech parameters accurately [9]-[11]. Recall that the major physical component to produce human speech is the excitation from the vocal cord and the resonator formed by the vocal tract volume together with the oral cavity [3]. Linear predictive coding (LPC) consists of an all-pole time varying digital filter that models the transfer function of the vocal tract. The transfer function of the synthesis filter *H*(*z*) is given by

$$H(z) = \frac{s(z)}{x(z)} = \frac{G}{1+\sum_{k=1}^{N} \alpha_k z^{-k}} \tag{2.1}$$

where G is the gain value, N is the order of the filter and $\{\alpha_i\ ,\ i\ =\ 1\ldots N\}$ are called the LPC coefficients which characterize the frequency response of the filter and equivalently the resonant frequencies or the formant frequencies of the vocal tract. The z-transform of the output speech is denoted by $s(z)$ while $x(z)$ is the z-transform of the input to the filter and is referred to the excitation. In time domain notation, equation (2.1) can be written as

$$s(n) = Gx(n) - \sum_{k=1}^{N} \alpha_k s(n-k) \tag{2.2}$$

It can be seen that the current speech sample is obtained by a linear combination of the past values and the scaled excitation signal. The denominator in equation (2.1) is called the analysis filter $A(z)$, and it can be expressed as

$$A(z) = 1 + \sum_{k=1}^{N} \alpha_k z^{-k} \tag{2.3}$$

When $s(z)$ is fed to the analysis filter, the prediction error, viz. $Gx(n)$, is obtained. For producing voiced sound, $x(z)$ equals to the z-transform of a set of periodic impulses in which the periodicity should be the same as the pitch period of the required synthesized speech. For unvoiced sound, the z-transform of random noise is employed instead. Figure 2.1 shows a simple speech production system both physically and mathematically.

As speech signal is quasi-stationary, the LPC coefficients and the gain value are usually determined in a short-time frame by frame basis and the parameters are found by minimizing the mean-squared prediction error of individual speech waveform segment which is normally 20 - 35 ms long. LPC analysis method is widely used because the coefficients and the gain value can be evaluated efficiently, which will be discussed in more details in the following section.

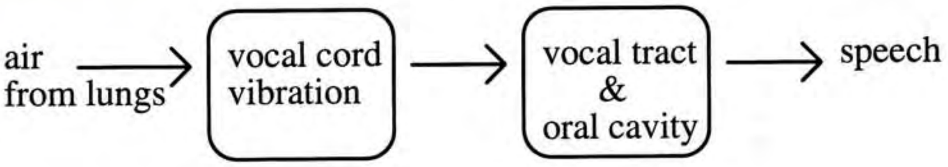


Figure 2.1(a). Human voice production system

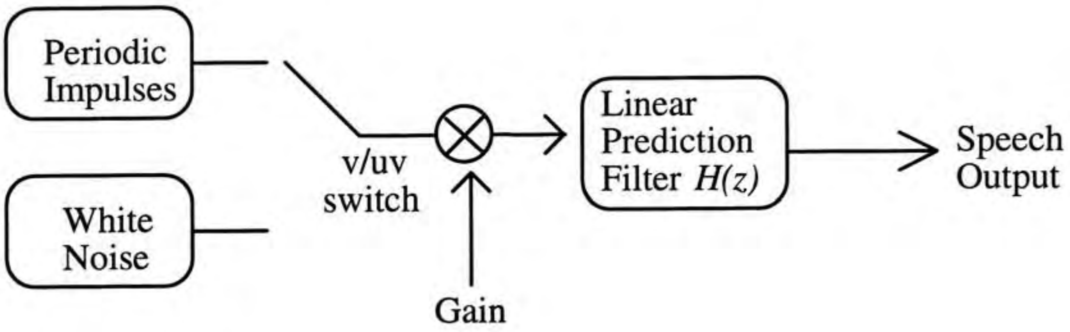


Figure 2.1(b). Linear Predictive Speech Model

2.2 LPC vocoder

Linear predictive analysis is applicable in low bit rate speech coding, and the corresponding codec is called LPC vocoder [20][21]. Since it is a model based vocoder, speech analysis and synthesis is the main constituents of the vocoder. The basic functional block diagram of a LPC vocoder is shown in Figure 2.2. Speech analysis consists of pitch detection and LPC analysis to find the filter coefficients. At the receiver, simple LPC synthesis is performed to produce synthesized speech. It is carried out by feeding a suitable input to the LPC synthesis filter to produce the output speech. Periodic impulse train is used as input to produce voiced sound, with a periodicity equals to the pitch period of the voiced speech. To produce unvoiced sound, white noise sequence is used. The excitation parameters normally include the gain, the pitch value and the voiced/unvoiced decision. The spectral parameters are the set of filter coefficients which are called the LPC coefficients.

2.2.1 Pitch and voiced/unvoiced decision

Pitch period is a very important parameter to be determined in speech analysis. It states the fundamental frequency of the corresponding voiced frame which equivalently shows the 'quasi-periodicity' of the signal. This information is important in producing synthetic voiced sounds since most of the energy in this case is mainly from the contribution of the harmonic frequencies.

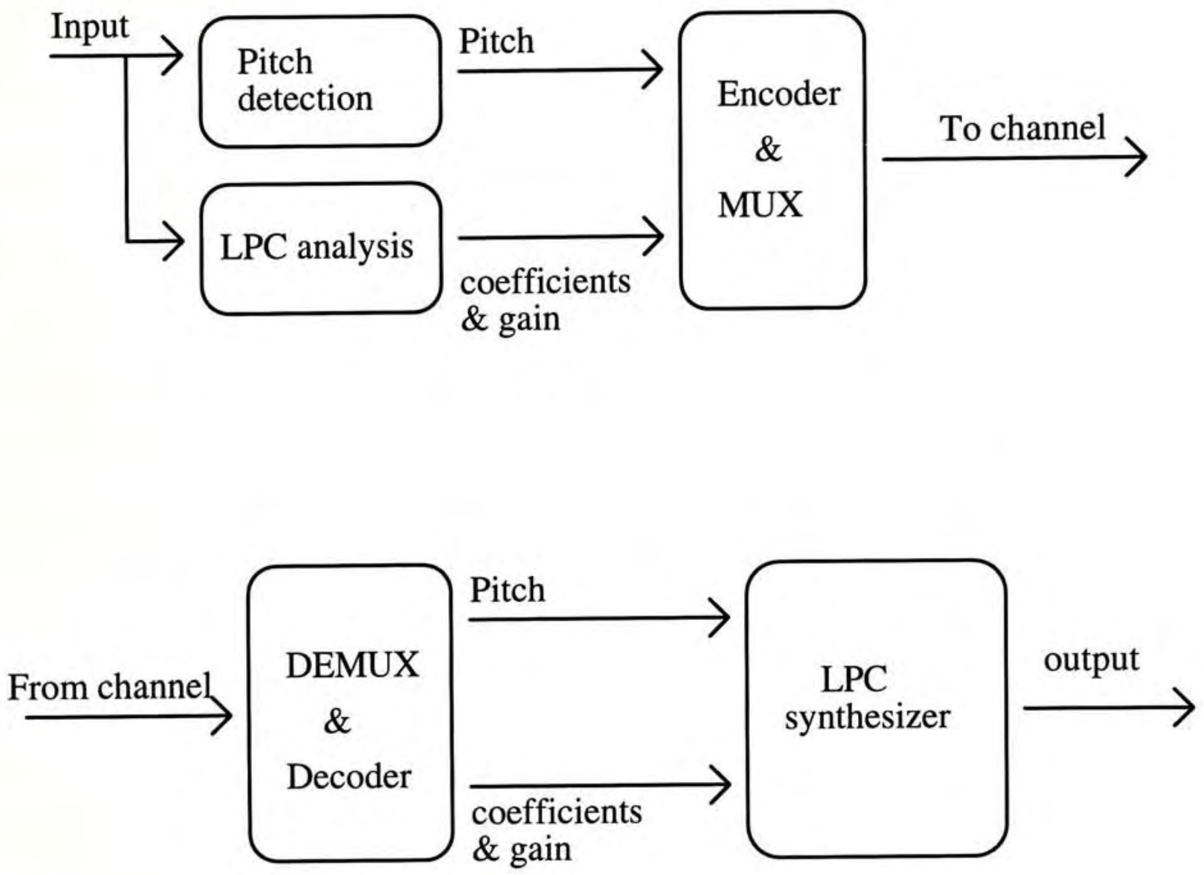


Figure 2.2 LPC vocoder

A number of methods can be used to determine the pitch period [22][23]. The simplest method is by using zero-crossing rate. Pitch period can be estimated by examining the rate in which the waveform cut the axis of zero amplitude. Typically the crossing rate is estimated every 10 ms. However, the result can be affected by the occurrence of dc offset which in turn affects the accuracy of the method.

A more commonly used method to find the periodicity of the speech signal is the auto-correlation technique [24]. The auto-correlation function of a windowed discrete time sequence $s_w(n)$ can be written as

$$corr(d) = \sum_{n=0}^{N_f-d-1} s_w(n) \cdot s_w(n+d) \tag{2.4}$$

where d is the time delay and N_f is the frame size. This function has a maximum value when d equals zero as this is equal to the sum of squares of the sequence. For a periodic signal, the auto-correlation function would be produced with peaks occur when d equals to the multiples of the signal period. From the summation, it is noted that the portion of the signal for computing the correlation reduces as the time delay increases. Thus the peaks in the function becomes smaller as the value of d becomes larger. By searching the first peak of the function, the pitch period of the speech signal can be found.

In order to ensure that the correlation function will emphasize the peaks which occur periodically, the speech signal is center clipped such that only the waveform which is contributed by the fundamental frequency is retained for calculating the correlation function [24]. The process of center clipping is performed according to the following operation

$$s'(n) = \begin{cases} s(n) - C_{upper}, & s(n) > C_{upper} \\ 0, & C_{lower} \le s(n) \le C_{upper} \\ s(n) - C_{lower}, & s(n) < C_{lower} \end{cases} \tag{2.5}$$

where $s'(n)$ is the clipped sequence. C_{upper}, C_{lower} are the upper and lower limit respectively. These limits are normally chosen to be ±30% of the maximum data value of the corresponding frame. As an alternative, three level clipping [25] can also be used and in this case the operation becomes

$$s'(n) = \begin{cases} 1, & s(n) > C_{upper} \\ 0, & C_{lower} \le s(n) \le C_{upper} \\ -1, & s(n) < C_{lower} \end{cases} \tag{2.6}$$

Auto-correlation method requires a substantial amount of multiplication which makes real time implementation difficult, if not possible. A magnitude difference function can be used to solve the problem [26]. It is done by replacing the multiplication in the auto-correlation function in equation (2.4) by a subtraction. In this case the minimum of the function is searched instead of searching the maximum in the previous method.

In ordinary LPC model, it is known that all the vowel sounds are voiced while the fricatives are mostly unvoiced. It is noted that only voiced sounds have detectable pitch value. In other words, the availability of the pitch value can be used to determine whether the corresponding frame is voiced or unvoiced in ordinary LPC vocoder.

2.2.2 Spectral envelope representation

The spectral envelope of a speech frame is modeled by the frequency response of the LPC filter. Each pair of the poles in the analysis filter represents the location of one formant frequency and normally 5 formant frequencies are required. Consequently, the order of the filter is usually set to 10 and the filter coefficients are used to represent the spectral envelope of a speech segment.

The LPC coefficients are found by minimizing the mean square error between the original speech signal and the predicted one [9][11]. The mean-squared error of the *n*th speech frame is

$$E_n = \sum_{m=0}^{N+N_f-1} (s_n(m) + \sum_{k=1}^{N} \alpha_k s_n(m-k))^2 \tag{2.7}$$

where $s_n(m)$ is the *n*th frame of the original signal for $0 \le m \le N_f\text{-}1$. After minimizing the error by setting the partial derivatives of E_n equals to zero with respect to α_i for all i, the following equations can be obtained

$$\sum_{m=0}^{N+N_f-1} s_n(m-i)s_n(m) = \sum_{k=1}^{N} (\alpha_k \cdot \sum_{m=0}^{N+N_f-1} s_n(m-i)s_n(m-k)) \quad 1 \le i \le N \tag{2.8}$$

If the auto-correlation function of $s_n(m)$ is expressed as

$$R_{ss}^n(i,k) = \sum_{m=0}^{N+N_f-1} s_n(m-i)s_n(m-k) \tag{2.9}$$

equation (2.8) can be simplified to

$$R_{ss}^{n}(i,0)=\sum_{k=1}^{N}\alpha_k R_{ss}^{n}(i,k) \quad 1\le i\le N \tag{2.10}$$

The coefficients can be calculated by solving equations (2.10). A number of algorithms are available to solve the problem and Durbin's recursive algorithm [9] is found to be most effective and is commonly used.

2.3. Excitation

The output quality of a linear predictive vocoder depends very much on the accuracy of representing the excitation signal. If the excitation signal is faithfully regenerated, synthesized speech with high output quality can be obtained. In basic LPC model, excitation of either periodic impulse train or random noise is found to be inadequate to represent the realistic input signals and thus the synthesized speech sounds somewhat 'robotic'. There are a number of techniques that are capable of obtaining more accurate representation of the excitation signal [27]. The following sub-sections introduce some typical examples.

2.3.1 Regular pulse excitation (RPE) and Multipulse excitation (MPE)

Speech analysis can be performed under a closed loop optimization scheme. Based on this scheme, the parameters to be determined are initialized and used to produce a synthetic speech which will be compared with the original one. The difference of two speech signals is fed back to improve the determination of the parameters until the error between the synthetic speech and the original speech is minimized. In Regular Pulse Excitation (RPE) model [28], the excitation signal is determined by using the above procedures.

In RPE model, a speech frame is divided into a number of sub-frames, typically each of 5 ms long. An impulse train with a fixed inter-pulse distance is used as excitation for each sub-frame, normally a distance of 5 samples is used. By using the closed loop optimization, the phase of the impulse train and the magnitude of each pulse is optimized by minimizing the error between the generated speech and the original one. A block diagram to illustrate the excitation analysis in RPE model is provided in Figure 2.3 . Generally speaking, RPE model uses extra pulses other than those representing the pitch period when compared with ordinary LPC model, and the optimization of the phase and the pulse magnitudes leads to a finer structure of the excitation signal.

Another excitation model that is similar to the RPE model is the Multipulse excitation (MPE) model [16]. The improvement over RPE model is that the distance between pulses is not fixed. Both the amplitudes and the locations of a certain number of pulses are determined using closed loop optimization. Normally, 6-8 pulses are used in sub-frames of 10 ms. This variation can model the true excitation signal more effectively and thus it is capable to give high output speech quality.

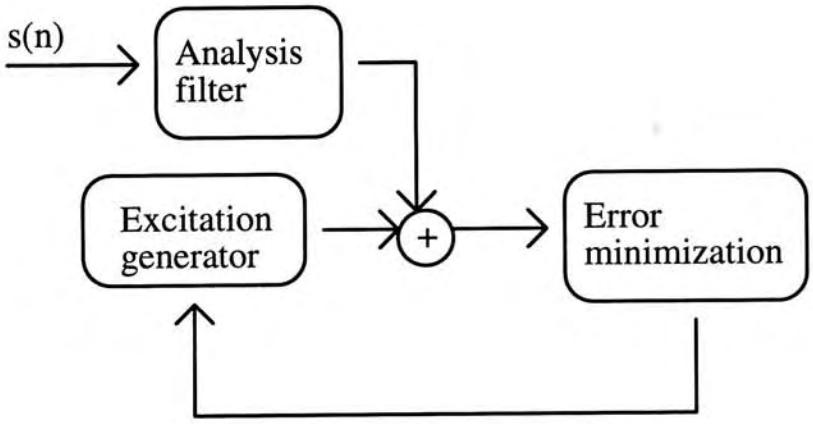


Figure 2.3 Optimization in the RPE model

2.3.2 Coded excitation and vector sum excitation

Another way of finding the optimal excitation in speech analysis is using codebook searching. It is found that [29] the probability density function of the prediction error samples is close to a Gaussian function. Consequently, a random codebook can be applied which consists of Gaussian random numbers which are the possible excitation for a particular speech frame. During speech analysis, synthesized speech is produced by each of the possible excitations and the optimal excitation is searched in the codebook by minimizing the error between the synthesized and the original speech. It is called Coded Excited Linear Prediction (CELP) [14]. In this method, the excitation is expressed by the codeword in the codebook for transmission, and it enhances the possibility of low bit rate speech coding. The procedures of finding the optimal excitation in CELP is shown in the Figure 2.4 . In this figure, it is noted that to generate the synthesized speech, the excitation is fed to two prediction filters separately. The first filter is used to generate the periodicity of the voiced speech and the second one is employed to restore the spectral envelope.

Similar to CELP, Vector Sum Excited Linear Prediction (VSELP) [15] also use codebook searching technique to find the optimal excitation in speech coding. Instead of using a single codebook as in CELP, VSELP consists of 3 codebooks to generate the excitations. The first one is an adaptive codebook which is updated after every sub-frame. The codevectors of the remaining 2 codebooks are essentially a set of linear combinations of 7 basis codevectors. Excitation signals are generated by adding the scaled codevectors from the 3 codebooks. During speech analysis, optimal codevectors are searched sequentially. The optimal codevector in the first codebook is chosen first, and it is used to determine the optimized codevector in the second codebook. The suitable codevector in the third codebook is then found by using the optimal codevectors from the previous codebooks. Figure 2.5 shows the steps of speech analysis in VSELP.

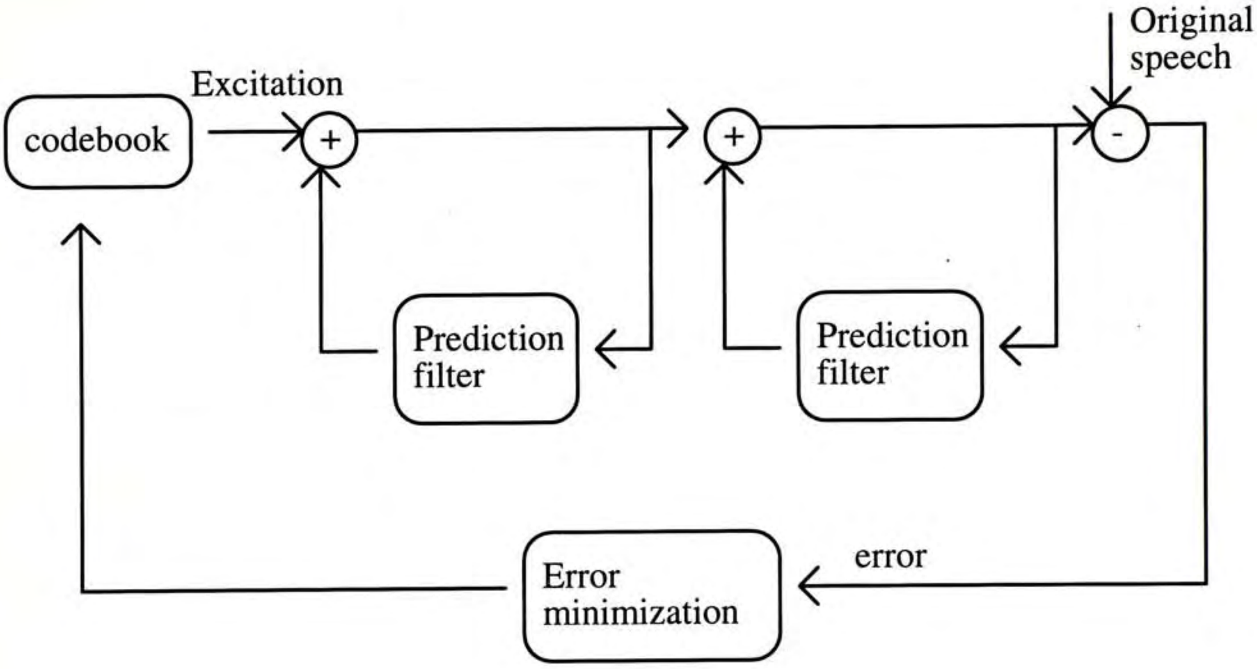


Figure 2.4 Speech analysis in CELP

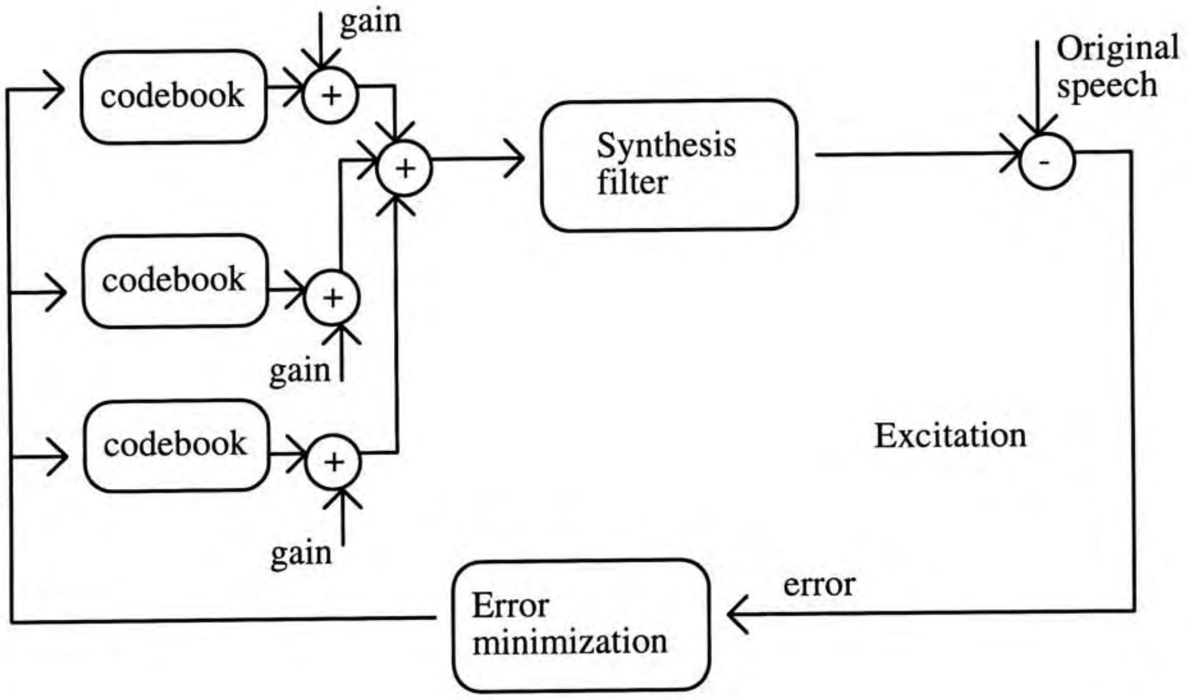


Figure 2.5 Speech analysis in VSELP

2.4 Multiband excitation (MBE)

In order to provide more variation and flexibility in expressing the excitation in conventional LPC, a multiband excitation model is introduced [17]. In this case, the synthetic speech spectrum $S'(\omega)$ equals to a multiplication of the excitation spectrum $E(\omega)$ and the spectral envelope $H(\omega)$, i.e.

$$|S'(\omega)| = |E(\omega)||H(\omega)| \tag{2.11}$$

By examining ordinary speech spectra, it is found that some frequency areas are dominated by the energy of harmonic frequencies, while some regions contain more noise-like energy. Accordingly, the speech spectrum can be divided into a number of frequency bands, and voiced/unvoiced decision is determined in each of them. If the energy of a certain frequency band is mainly contributed by the harmonic frequencies, it is declared as voiced. Otherwise it is declared as unvoiced. An example of MBE model is shown in Figure 2.6 . Ideally, such a determination should be performed for each frequency band with a bandwidth that contains only one harmonic frequency. If the frequency band is fine enough, the spectral envelope according to a particular band can be regarded as a constant M_k such that equation (2.11) can be rewritten as

$$\left|\hat{S}_k(\omega)\right| = M_k\left|E_k(\omega)\right| \tag{2.12}$$

where k is the band index. Thus in MBE model, the parameters required to synthesize speech includes the fundamental frequency, a sequence of voiced/unvoiced decisions and a set of spectral envelope values M_k. The size of these parameters depends on the number of frequency bands used to divide the speech spectrum during analysis.

Practically, a maximum of 12 non-overlapping frequency bands are commonly used. Each band, except the highest band, contains 3 harmonic frequencies [18].

Compared with the conventional speech model, multiband excitation gives a flexibility that both voiced and unvoiced characteristics are allowed in a single speech frame. A finer spectral structure of the excitation signal can then be provided. It is interesting to note that MBE model is different from LPC model since the spectral envelope is now being represented by the parameters M_k instead of the filter coefficients.

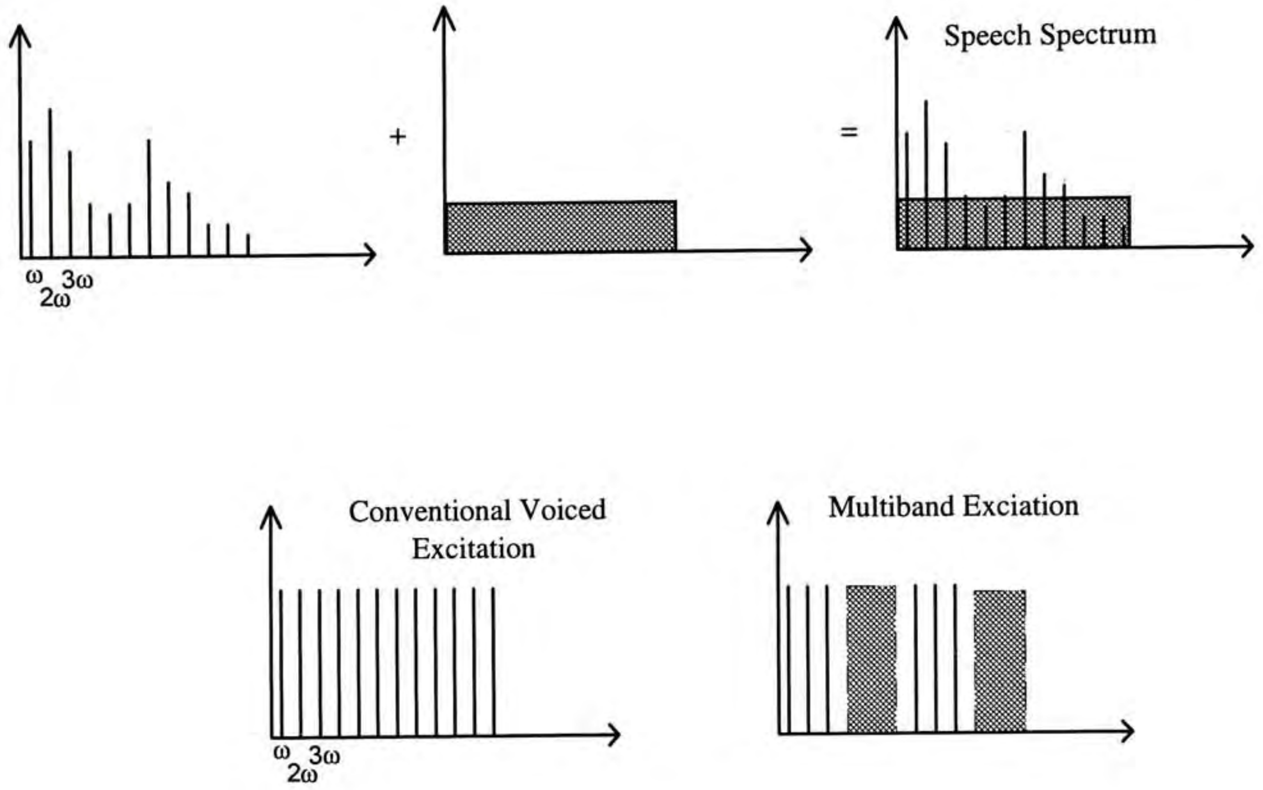


Figure 2.6 Multiband excitation model

2.5 Multiband excitation vocoder

To apply MBE model in speech coding, multiband excitation vocoder have been introduced [17] for producing high quality synthesized speech. As voiced and unvoiced characteristics coexists in a single speech frame, the algorithm of speech analysis in MBE vocoder is different from that of LPC vocoder. First of all, a pitch value needs to be determined. After dividing the particular speech spectrum into a number of frequency bands, voiced/unvoiced decision is performed for each and every frequency band. The number of frequency band being used mainly depends on the fundamental frequency of the speech frame. According to the voiced/unvoiced decision, the spectral envelope parameter in each frequency band is determined. A block diagram summarizing the procedure of speech analysis in MBE vocoder is shown in Figure 2.7 .

To determine whether a speech frame is voiced or unvoiced is just as important as to find the pitch value during speech analysis in MBE vocoder. Voiced/unvoiced decision for each frequency band is found by a method of spectrum comparison. Assuming that the spectrum in the corresponding frequency band is voiced, a periodic spectrum $P(\omega)$ is then used as the excitation spectrum, $E_w(\omega)$, to generate a synthetic signal spectrum. The periodicity of $P(\omega)$ is determined by the fundamental frequency of the current speech frame. It can be obtained by centering a window spectrum on each of the harmonic frequencies to form a periodic spectrum. From [17], the mean squared error (MSE) between these two spectra can be expressed as

$$\varepsilon_k = \frac{1}{2\pi}\int_{a_k}^{b_k}[|S_w(\omega)| - |S'_w(\omega)|]^2 d\omega \tag{2.13}$$

where $S_w(\omega)$ and $S'_w(\omega)$ is the original and the synthetic windowed speech spectrum respectively, and (a_k, b_k) are the boundaries of the kth frequency band. The synthetic speech spectrum, in fact, can be written as a multiplication of the periodic excitation spectrum $P(\omega)$ and the spectral envelope function $H(\omega)$ as in (2.11). The MSE in (2.13) is then given by

$$\varepsilon_k = \frac{1}{2\pi}\int_{a_k}^{b_k}[|S_w(\omega)| - |H(\omega)||P(\omega)|]^2 d\omega \tag{2.14}$$

For a frequency band with a suitably narrow bandwidth, the spectral envelope is assumed to be a constant. Using (2.12), the MSE in the kth frequency band is as follows,

$$\varepsilon_k = \frac{1}{2\pi}\int_{a_k}^{b_k}[|S_w(\omega)| - |M_k||P(\omega)|]^2 d\omega \tag{2.15}$$

where M_k is the spectral magnitude. The parameter M_k can be found by minimizing the MSE and is given by

$$|M_k| = \frac{\int_{a_k}^{b_k}|S_w(\omega)||P(\omega)|d\omega}{\int_{a_k}^{b_k}|P(\omega)|^2 d\omega} \tag{2.16}$$

Once the envelope is found, the MSE for each individual frequency band can then be computed.

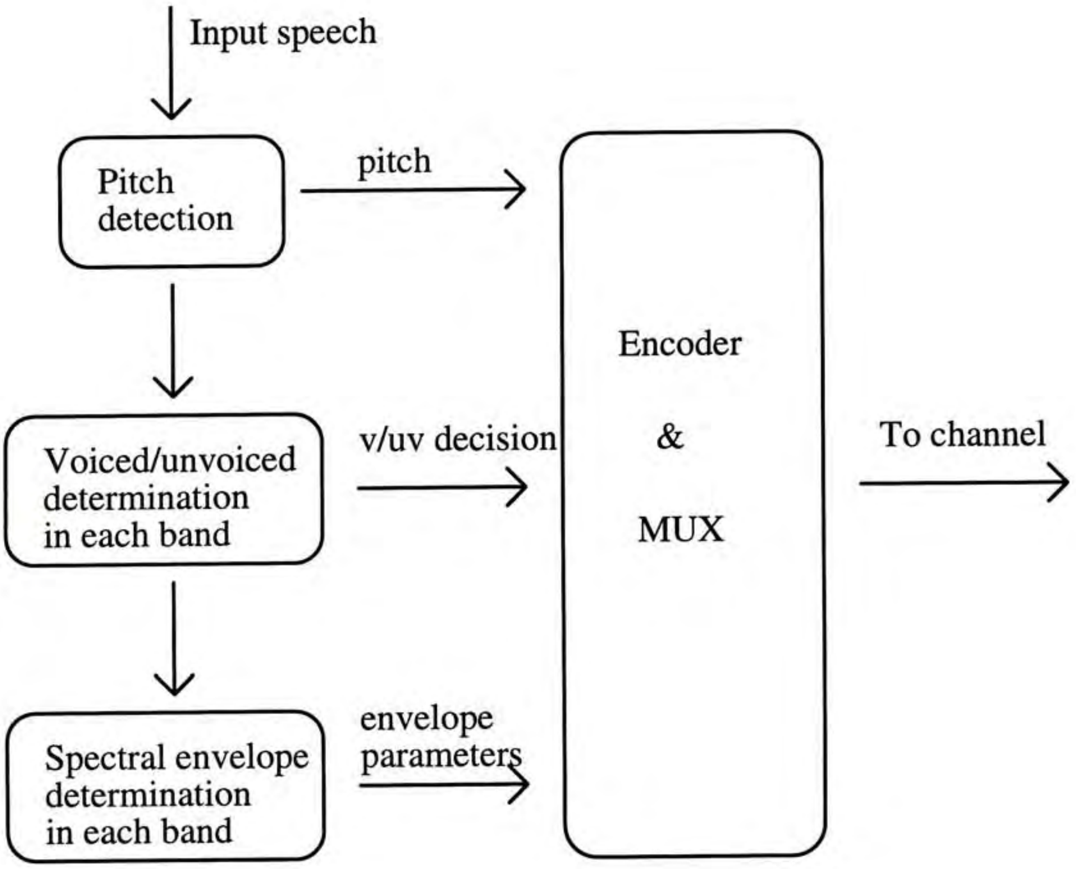


Figure 2.7 Speech analysis in MBE vocoder

To make the voiced/unvoiced decision, the MSE obtained is being normalized which is denoted by ε'_k,

$$\varepsilon'_k = \frac{\varepsilon_k}{\frac{1}{2\pi}\int_{a_k}^{b_k} |S_w(\omega)|^2 \, d\omega} \tag{2.17}$$

This normalized error is then compared with an empirical threshold value, which is set to 0.2 [17]. If the error is smaller than the threshold value, it means that the synthetic spectrum using voiced excitation is close to the original one and a voiced decision is made. Otherwise, the original spectrum is assumed to have some noise-like energy, and the corresponding frequency band is declared to be unvoiced. The procedure is carried out for every frequency band to get a sequence of voiced/unvoiced decision to express the whole spectrum.

In MBE vocoder, since it is not related to LPC model, the spectral envelope parameters are not derived from the filter coefficients. Actually, during voiced/unvoiced decision, the spectral envelope magnitude have already been computed for each frequency band as given by equation (2.16). Therefore the spectral envelope parameters for voiced bands are in fact available during the voiced/unvoiced decision process. The spectral magnitude for unvoiced band, on the other hand, can be found by using a white noise spectrum to replace $E_w(\omega)$ in (2.16). Since the magnitude of the random noise spectrum is a constant, the calculation of the envelope is equal to taking an average of the original spectrum in the unvoiced band.

During speech synthesis, voiced speech and unvoiced speech is generated individually and superimpose together to form the synthesized speech [17]. Voiced sounds are produced in the time domain, whereas unvoiced sounds are generated in the frequency domain. For those harmonic frequencies in voiced bands, a bank of oscillators are used to produce these harmonic frequencies. The amplitudes of the oscillations are determined by the corresponding spectral magnitude parameters of the harmonic frequencies. In addition, the phase between the oscillators are also estimated. The waveforms of the oscillations are then superimposed to form the voiced sounds. To produce unvoiced sounds, an unvoiced envelope is determined initially by using the unvoiced spectral magnitudes, and setting the frequency components that correspond to the voiced bands to zero. An unvoiced spectrum is then regenerated by replacing the spectral envelope of a random noise spectrum with the pre-determined envelope. The unvoiced waveform is then obtained by using inverse Fast Fourier Transform. Finally, synthesized speech is obtained by a superposition of voiced and unvoiced sounds. A block diagram that illustrates the speech synthesis procedures in MBE vocoder is shown in Figure 2.8 .

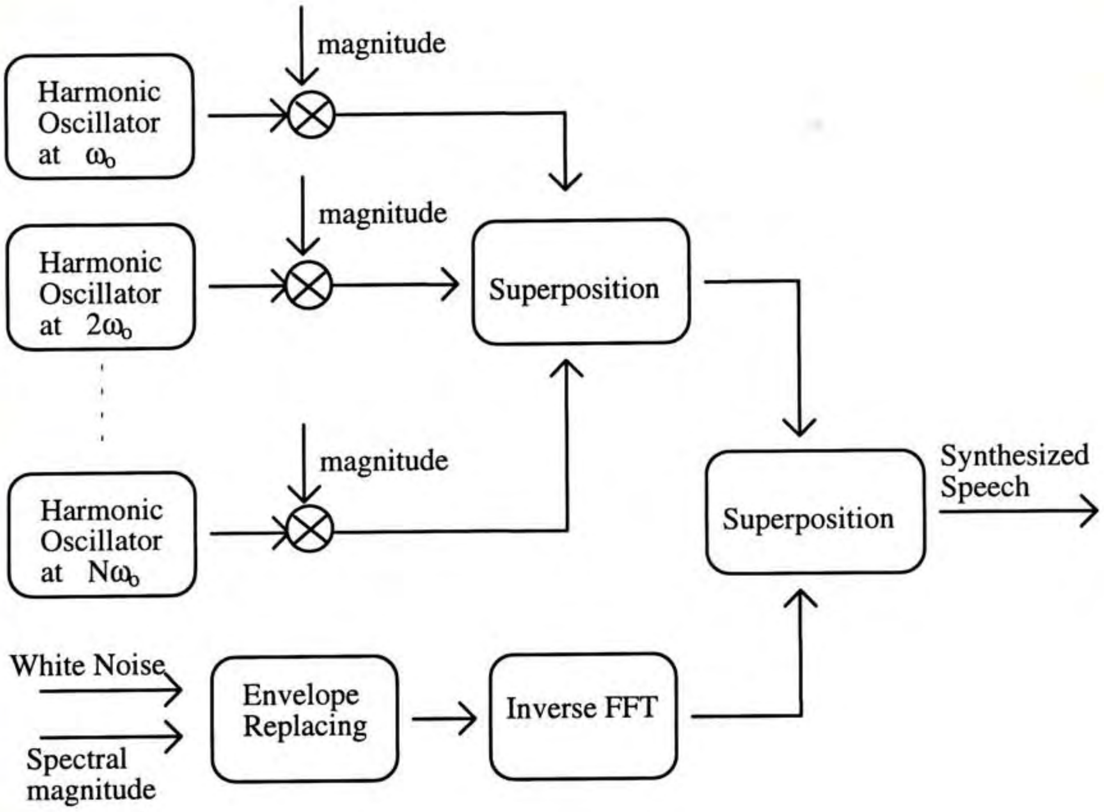


Figure 2.8 Speech synthesis in MBE vocoder

Chapter 3 Dual-band and Quad-band excitation

Multiband Excitation, as discussed earlier, provides a flexibility in modeling the frequency spectrum of an excitation signal by allowing a combination of both voiced and unvoiced characteristics in a single frame of speech. However, it is found that dividing a speech spectrum into as much as 12 bands, as commonly used, not only produces a great deal of redundancy but also involves extensive computation. In the following sections, a simpler but yet effective model called Dual-Band Excitation (DBE) is first introduced. Then, an improved version called Quad-Band Excitation (QBE) model is developed for low bit rate coding which is capable of reproducing good quality synthesized speech signals.

3.1 Dual-band excitation

When examining the spectrum of an ordinary speech segment, it is noted that the low frequency region is usually dominated by the harmonic frequencies. Whereas in the higher frequency area, the spectrum contains mostly noise-like energy. Some typical examples are shown in Figure 3.1 . The speech waveform in part (a) was obtained from a male voice, and that in part (b) was extracted from an utterance uttered by a female. From their frequency spectra which were drawn in log scale as shown in Figure 3.1 (c) and (d), it can be seen that the lower harmonic frequencies have much larger energy and are obviously noticeable by their peaks in the spectral plots. In the upper frequency region, the spectral shape was fairly uniform, and the peaks of the envelope due to the higher harmonic frequencies did not contribute to a substantial amount of energy. This is found to be generally true for speech signals because the speech spectrum often has a roll-off of 20 dB per octave as frequency increases. As a result, the first two formant frequencies are usually more dominant than the higher formants. While, on the other hand, the energy of the higher harmonic frequencies are only somewhat comparable with the other frequency components, and thus in the high frequency region it is more like unvoiced. Hence, it might not be necessary to use as many as 12 voiced/unvoiced decisions within a speech frame. Indeed, a fine spectral structure of the excitation signal can still be obtained by using a fewer number of variable bandwidth frequency bands to partition the speech spectrum accurately, in contrast to the conventional MBE model which uses many narrow frequency bands.

An experiment has been carried out to investigate the statistical distribution of the number of frequency band that is required for satisfactory reproduction of the excitation spectrum for ordinary conversational speech synthesis. In this experiment, two sets of recordings were used. They were recorded from different sources to increase the generality of the speech data. One set of the recordings was obtained from a cassette tape which was tailored made for a comprehensive listening examination. The other set was recorded from TV news broadcasting. Both recordings included male and female voices and the language used was English. Each of the recordings was about 10 minutes long, and was band limited from 100 Hz to 3.4 kHz before digital sampling. These recordings were then digitized at a sampling rate of 8 kHz with 16-bit resolution. The speech data were segmented into frames of 35 ms long by using hamming window, with a time shift of 20 ms. For those frames that possessed a detectable pitch value P in Hz, the spectrum of individual speech frame was divided into K frequency bands for voiced/unvoiced determination. Each of these bands, except the highest band, contained 3 harmonic frequencies only. In this method, a maximum of 12 frequency bands are allowed in a speech frame. The number of harmonic frequencies N_h in a speech frame is calculated from

$$N_h = \left\lfloor \frac{4000}{P} \right\rfloor \tag{3.1}$$

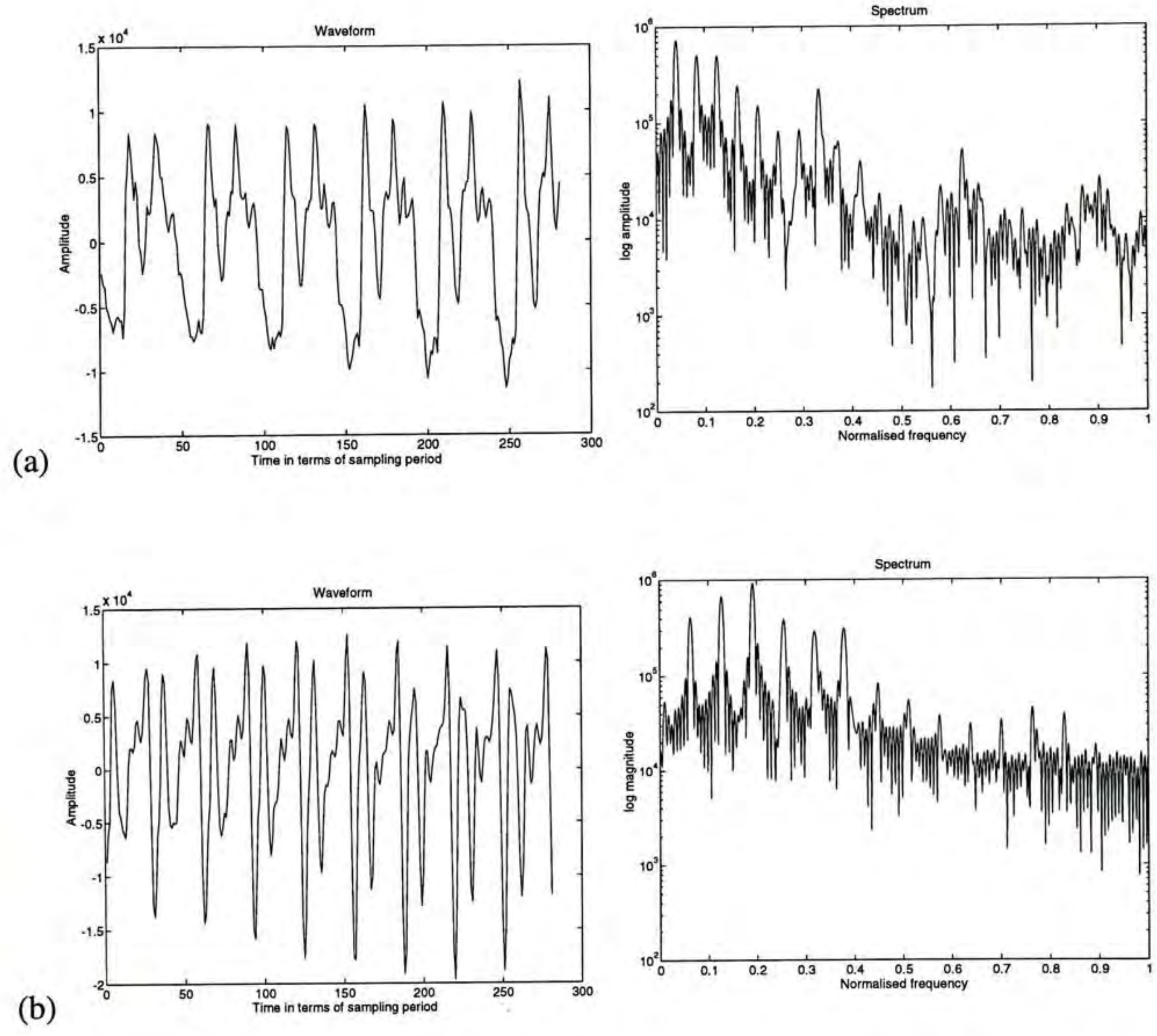


Figure 3.1 Examples of speech waveforms and their corresponding spectra

where $\lfloor x \rfloor$ equals to the largest integer that is less than or equal to x. Then

$$K = \begin{cases} \left\lceil \dfrac{N_h}{3} \right\rceil & \text{if } N_h \leq 36 \\ 12 & \text{if } N_h > 36 \end{cases} \tag{3.2}$$

where $\lceil \text{x} \rceil$ equals to the smallest integer that is larger than or equal to x. For normal conversational speech, the pitch value usually lies within the range of 100-350 Hz. Therefore, K will be bounded between

$$4 \leq K \leq 12 \tag{3.3}$$

Each of these frequency bands can then be described by a band boundary B_i (Hz) which is given by

$$B_i = \begin{cases} \dfrac{(6i+1)P}{2} & 1 \leq i \leq K-1 \\ \dfrac{f_s}{2} & i = K \end{cases} \tag{3.4}$$

where f_s is the sampling frequency. This is in fact similar to the approach in MBE model. However, if consecutive bands have the same v/uv decision, they are allowed to combine to form a larger band, and the band boundary is then adjusted accordingly. Thus individual frequency bands will have different bandwidths. The statistical result of the voiced/unvoiced distribution is given in Table 3.1 .

Number of excitation bands	frequency band distribution	%	
	dc $f_s/2$	1st set	2nd set
1	v	*48.43*	*37.56*
	uv	*0.39*	*0.72*
	sub-total	48.82	38.28
2	v uv	*21.89*	*38.13*
	uv v	*0*	*0.03*
	sub-total	21.89	38.16
3	v uv v	*11.72*	*7.94*
	uv v uv	*0.22*	*0.40*
	sub-total	11.94	8.34
4	v uv v uv	*10.21*	*10.71*
	uv v uv v	*0.03*	*0.03*
	sub-total	10.24	10.74
5	v uv v uv v	*4.10*	*2.43*
	uv v uv v uv	*0.04*	*0.05*
	sub-total	4.14	2.48
6	v uv v uv v uv	*2.24*	*1.68*
	uv v uv v uv v	*0.01*	*0*
	sub-total	2.25	1.68
more than 6		0.72	0.32

v = voiced band uv = unvoiced band

f_s = sampling frequency

Table 3.1 Voiced/unvoiced decision statistics for multiband modeling of speech signals

It is interesting to note from the statistical result that, by considering those segments with a detectable pitch value, about 70% of them contained either 1 or 2 frequency bands, an approximately 10% consisted of 3 bands, while another 10% of them have 4 bands. In addition, there was roughly 6% on average of the speech frames that contained more than 4 frequency bands. It was also found that the lowest band was usually voiced, almost without any exception. Thus using excitations with a maximum of 2 bands can already represent the spectrum of about 70% of the speech frames. This demonstrated that a smaller number of frequency bands is adequate to intimate the excitation spectrum, provided that the frequency bands have a variable bandwidth to cover the entire frequency range.

Based on these statistical results, a Dual-Band Excitation (DBE) model is proposed [19]. In this method, no more than 2 frequency bands are used to express the frequency contents of the excitation signal. Furthermore, v/uv pattern starting with an unvoiced band in the lowest frequency band is not allowed. Accordingly, for those frames that consist of more than 2 frequency bands, all other bands except the last one are declared as voiced since they must end with an unvoiced band. Consequently, in DBE model, only two possible v/uv patterns are allowed: {v} and {v uv}. Therefore, there is at most 12 possible frequency boundaries to identify the voiced band from the unvoiced band. These boundaries are in fact given by B_i , $1 \leq i \leq K$, and $B_K = f_s/2$ (this represent a single voiced band), and are available once the pitch period is computed which does not require any additional computation. For transmission purpose, we need to send the v/uv boundary for each speech segment to the receiver only. Hence, a mere of 4 bits per frame will be sufficient to encode this information. This greatly reduces the number of bits that are needed to encode the v/uv decision.

3.2 Quad-band excitation

Although DBE model only requires a fairly low bit rate for encoding the excitation signal, it does not provide, however, a faithful intimation of the original excitation spectrum. From Table 3.1 , we can see that over 25 % of the speech frames actually have more than 2 voiced/unvoiced bands. In order to retain the detail information of the spectral envelope, we develop a modified multiband approach which allows a combination of 4 voiced/unvoiced bands in the excitation. By so doing, we find that over 90% of the excitation spectra can be well represented. We called this Quad-Band Excitation (QBE) model. Since the voiced/unvoiced pattern that starts with an unvoiced band in the lowest frequency region are generally not allowed, therefore there are only 4 possible band arrangements, namely, {v}, {v uv}, {v uv v} and {v uv v uv}. The operation of the QBE model is similar to the DBE model. For a particular speech frame, we again compute the pitch value P as well as the frequency boundaries B_i according to equation (3.4). Again, the number of frequency bands K depends on P and in any case is not more than 12. Each frequency band is then individually identified as voiced or unvoiced. If adjacent bands fall into the same category, they are combined together to form a larger band. In this method, we allow up to a maximum of 4 voiced/unvoiced bands after grouping, and the boundaries between adjacent bands need to be re-adjusted accordingly. It will be shown in chapter 4 that there are a total of 232 voiced/unvoiced patterns with different band boundaries for this method. If these patterns are stored in a table, only 8 bits per frame will be adequate for encoding this information. When compared with the conventional MBE model that uses 12 bits per frame, over 30% of the number of bits is saved.

If a particular segment of a speech signal is found to have more than 4 v/uv bands, the excessive frequency bands will be declared as unvoiced so that they can be amalgamated with the fourth band to form a larger unvoiced region. This is based on the assumption that the high frequency region in a speech spectrum usually contains noise-like energy and the higher harmonic frequencies are not pre-dominant.

Speech analysis/synthesis using DBE and QBE has been simulated on computer. Figure 3.2 shows the spectrograms of some synthesized speech using conventional MBE, DBE and QBE in linear predictive model. From the spectrograms, it can be seen that the frequency contents of the speech signal can be retained using both DBE and QBE. When compared with the MBE model, the spectrograms of the synthesized outputs all look very similar. Informal listening tests also showed that there was little differentiation among them and all output speech were highly intelligible. Further evaluation and comparison between DBE and QBE will be given in more details in Chapter 5, and it will be demonstrated that QBE can give an overall better performance than DBE.

We have shown that, QBE model can, on one hand, retain the flexibility of expressing excitation spectrum by using both voiced and unvoiced spectrum in a single frame, and, on the other hand, avoid the redundancy of using too many frequency bands as in the MBE model. By using no more than 4 frequency bands with variable bandwidth, it is efficient and effective enough to intimate the fine structure of the excitation spectrum. Although the voiced/unvoiced patterns are limited to 4 possible cases only, it still have sufficient v/uv variations and this simplification shows negligible effects on the quality of the synthesized speech. From the encoding point of view, it provides a possibility for very low bit rate transmission. In addition, the complexity of the system is also reduced particularly in the speech synthesis part. A comparison between these two excitation models is given in Table 3.2.

Based on the above findings, QBE is chosen to be used for the development of a low bit rate vocoder.

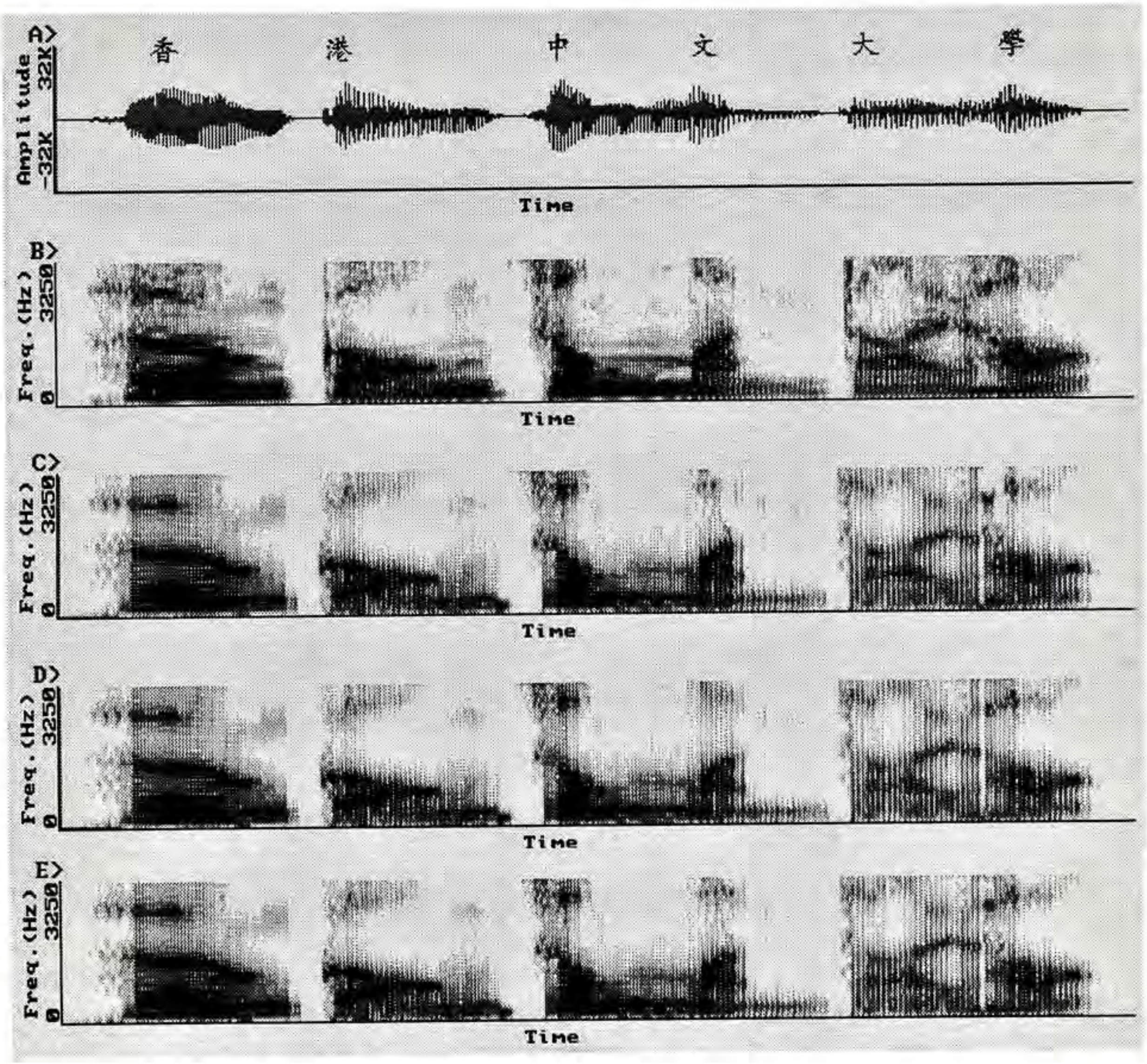


Figure 3.2 (A) The input signal, (B) Spectrogram of (A), (C) Spectrogram of synthesized output using Multiband Excitation, (D) Spectrogram of synthesized output using Dual-Band Excitation, and (E) Spectrogram of synthesized output using Quad-Band Excitation

	Multiband Excitation	**Quad-Band Excitation (Dual-Band Excitation)**
Number of band used per frame	normally 12 bands	no more than 4 (2) bands
Band nature	non-overlapping	non-overlapping
Band width	equal bandwidth	variable bandwidth
Computational complexity	high complexity in speech synthesis	relatively low complexity in speech synthesis
Number of bits required per frame	1 bit per band, normally 12 bits per frame	8 (4) bits
Output quality in conjunction with LPC	comparable quality, both highly intelligible	

Table 3.2 MBE Vs QBE/DBE

3.3 Parameters determination

Excitations in the QBE model are expressed in the frequency domain. The parameters for characterization of speech properties in this method include the pitch value and the v/uv pattern. In the following sections, details of the parameter extraction process will be given.

3.3.1 Pitch determination

The method employed for pitch detection in the QBE model is the auto-correlation technique [24]. Recall that in this method the auto-correlation function of the data frame is first calculated and the pitch period is then found in terms of the number of samples by finding the location of the peak on the correlation function. The pitch detection method being used in the vocoder is summarized in Figure 3.3 . In order to increase the accuracy of the detection method, a pre-processing is carried out for the speech sequence before the auto-correlation function is being computed. As the correlation function is a discrete time function and the corresponding pitch value is denoted by the number of samples within a period, the estimated pitch value suffers a quantization error. In order to reduce this error, the data sequence is linearly interpolated before any calculation such that the time difference between adjacent samples is halved, and the resolution of the pitch value can thus be improved to half a sample. Since only the ac component is being analyzed, the data sequence is processed with a zero mean. In addition, the speech samples are center clipped so that the correlation function can concentrate on showing the peaks on the fundamental period and it can be found more accurately. The process of center clipping is performed according to equation (2.5).

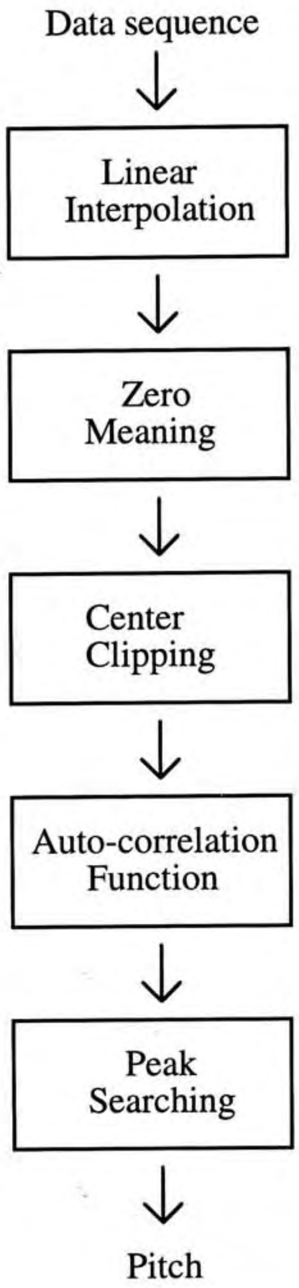


Figure 3.3 Pitch detection algorithm

Once the auto-correlation function in equation (2.4) is calculated, the location of the first peak is searched to determine the pitch value. In order to locate this peak easily, the correlation function is normalized, and any value smaller than 0.25 is reset to zero. The auto-correlation function has the largest value when the delay is zero, which is equivalent to its energy and does not provide information on the pitch value. The peaks usually become gradually smaller as time delay increases. The first peak, which is always the largest except the one at zero delay, will give the pitch period. The range of the pitch value to be considered is 100 - 350 Hz. Any value out of this range is not allowed and the respective frame is regarded as to have no fundamental frequency. A computer C program for the pitch detection algorithm is written for simulation and is included in Appendix A.

3.3.2 Voiced/unvoiced pattern generation

Similar to the MBE model, voiced/unvoiced decision in QBE is performed on each of the pre-divided frequency bands, but a grouping operation is applied to the voiced/unvoiced stream which is then transformed into one of the allowable pattern. A flow chart to summarize the procedures for voiced/unvoiced decision in QBE model is shown in Figure 3.4. The speech segment is first undergone Fast Fourier Transform (FFT) to obtain a discrete frequency spectrum. With the fundamental frequency obtained in the pitch detection process, the spectrum is then divided into a maximum of 12 frequency bands and each band, except the highest band, contains 3 harmonic frequencies. A voiced/unvoiced decision is assigned for each of the frequency bands by using the method of spectrum comparison similar to that being used for the MBE model as discussed earlier [17]. Adjacent bands with the same voiced/unvoiced decision would be combined together to form a larger band, and the band boundary is adjusted accordingly. A maximum of 4 frequency bands are allowed. If the speech frame has more than 4 frequency bands, the excessive high frequency bands are assigned to be unvoiced and they are grouped with the fourth band to form a larger unvoiced band. After frequency band grouping, a quad-band voiced/unvoiced pattern is then generated to represent the structure of the excitation spectrum. It is reminded that voiced/unvoiced decision is performed only for those frames that have a detectable pitch value. For other frames that do not have a fundamental frequency, the whole frequency spectrum would be declared as unvoiced. A computer C program of the voiced/unvoiced detection is included in Appendix B for reference.

As mentioned earlier, during spectrum comparison in voiced/unvoiced decision for a particular frequency band, the normalized difference between the original speech spectrum and the synthetic voiced spectrum is compared with a threshold value. If the error is smaller than the threshold, the corresponding spectrum is declared as voiced. Otherwise an unvoiced decision is assigned. In conventional MBE model, the threshold value is fixed for all frequency bands. In fact, clean speech prefers a high threshold value so that more harmonic frequencies can have contribution in producing synthetic output. However, if it is used in noisy speech, the unwanted voiced energy in the high frequency region would become more dominant [30]. To compromise these trade off, a frequency dependent threshold function is used instead of a constant value throughout the entire frequency range and a simple linear function is used. Based on informal listening tests, an empirical threshold function is chosen which is given in Figure 3.5. The largest threshold value is 0.7 at the dc level, and is linearly decreased to 0.4 at the sampling frequency. This threshold function is chosen such that the lower frequency region of speech spectrum is more likely to be declared as voiced and unvoiced in the higher frequency area. It is noted that during voiced/unvoiced decision for a particular frequency band, an average threshold value is calculated and used before spectral comparison.

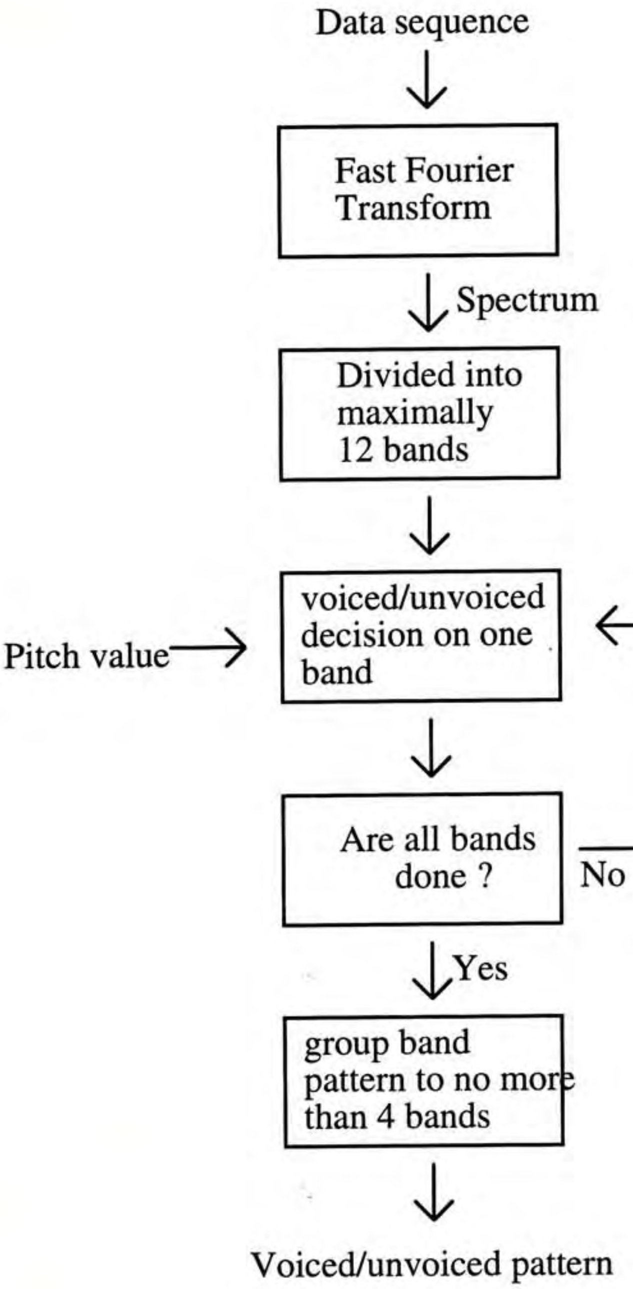

Data sequence
Fast Fourier Transform
Spectrum
Divided into maximally 12 bands
Pitch value
voiced/unvoiced decision on one band
Are all bands done ?
No
Yes
group band pattern to no more than 4 bands
Voiced/unvoiced pattern

Figure 3.4 Voiced/unvoiced decision algorithm

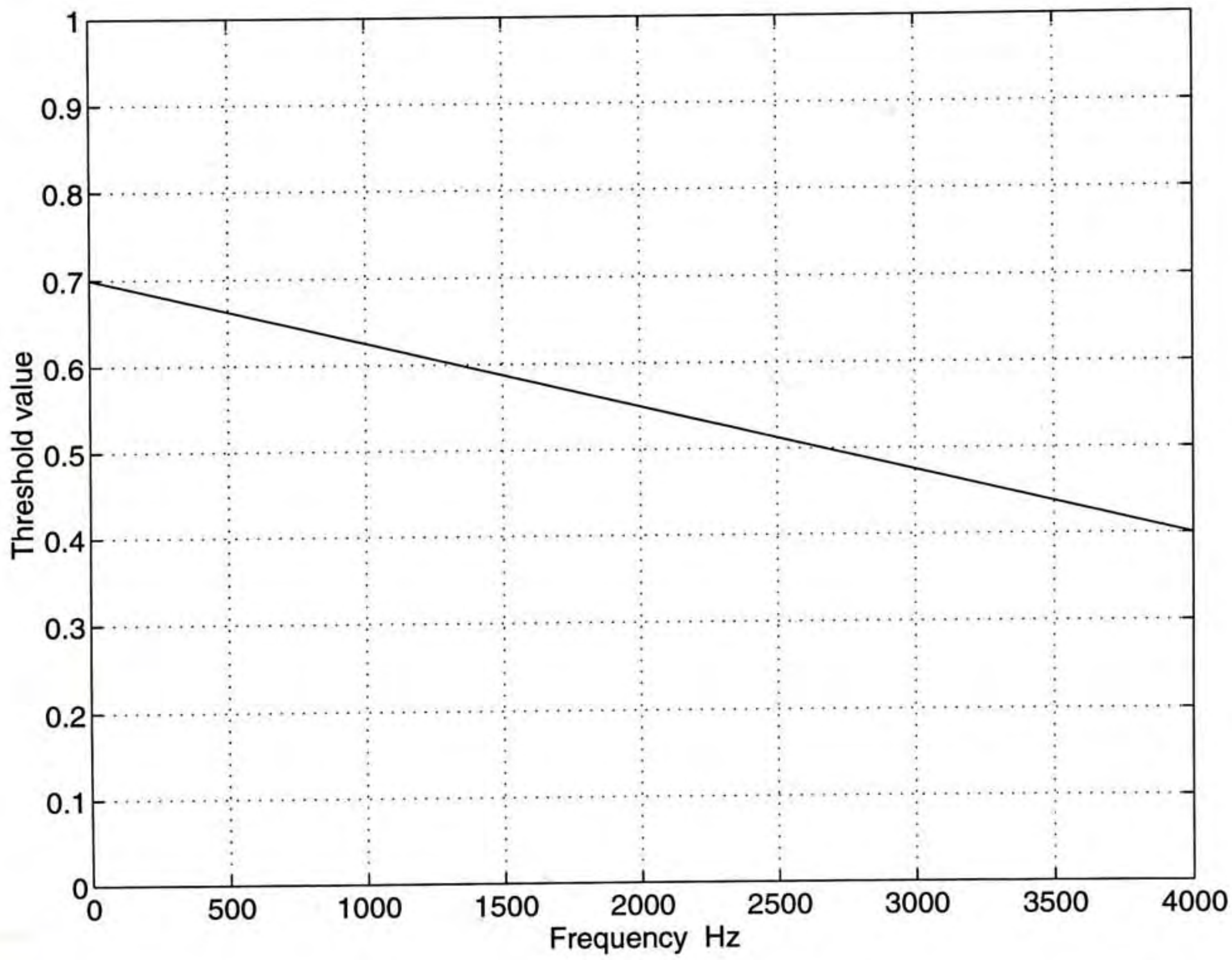


Figure 3.5 Frequency dependent threshold value.

3.4 Excitation generation

Once the pitch value and the v/uv pattern are received at the receiver, the frequency spectrum of the excitation signal for that particular frame can be regenerated. For voiced bands, impulses with equal magnitude are inserted at the harmonic frequencies and for unvoiced bands, white noise spectrum is used instead. The energy ratio between voiced and unvoiced bands are obtained from the linear prediction residual. Once the v/uv pattern is known, the v/uv energy ratio can be calculated from the spectrum of the linear prediction error signal, and the magnitude of the impulses in the voiced bands is then set accordingly. For simplicity, consider a speech frame that contains 1 voiced band and 1 unvoiced band, with a band boundary B_i. The voiced/unvoiced energy ratio would then be

$$r_{v/uv} = \frac{\sum_{n=0}^{B_i N_s/2\pi} S_{error}^2(n)}{\sum_{B_i N_s/2\pi}^{N_s} S_{error}^2(n)} \tag{3.5}$$

where $S_{error}(n)$, $0 < n < 2N_s$, is the $2N_s$ point Fast Fourier Transform of the prediction error signal. By setting the magnitude of all frequency components in the regenerated unvoiced spectrum to a constant M_{noise}, the impulse magnitude, which is the same for all impulses, that is being used to regenerate the excitation voiced spectrum is given by

$$M_{impulse} = \sqrt{\frac{r_{v/uv} \cdot \sum_{B_i N_s/2\pi}^{N} M_{noise}^2}{\lfloor B_i / P \rfloor}} \tag{3.6}$$

where P is the fundamental frequency of the speech frame. The regenerated excitation spectrum EX(n) will become

$$EX(n)=\begin{cases} M_{impulse}\sum_{m}\delta(n-\frac{mPN_s}{2\pi}) & 0\le n\le\frac{B_iN_s}{2\pi} \\ 1 & \frac{B_iN_s}{2\pi}<n\le N_s \end{cases} \qquad (3.7)$$

Subsequently, the excitation signal is obtained by applying Inverse Fast Fourier Transform (IFFT) to the regenerated spectrum. The procedures in generating the excitation signal is summarized in Figure 3.6 .

In order to avoid an abrupt change in the pitch period due to concatenation of consecutive frames, signal segments are chosen from the IFFT excitation sequences such that the pitch period is consistent with the current frame. An example is given in Figure 3.7 . It is shown in the figure that the first two data pulses in the excitation sequence of the (N+1)th frame are disregarded and the excitation sequence is chosen starting from the third data pulse. As a result, the pitch period within the frame boundary is retained. Otherwise, a higher frequency component may occurred at the junction and it would lead to irrevocable error during the process of speech synthesis.

After generating the excitation signal, it is used as an input to the linear prediction synthesis filter to produce synthetic speech at the receiver of the vocoder.

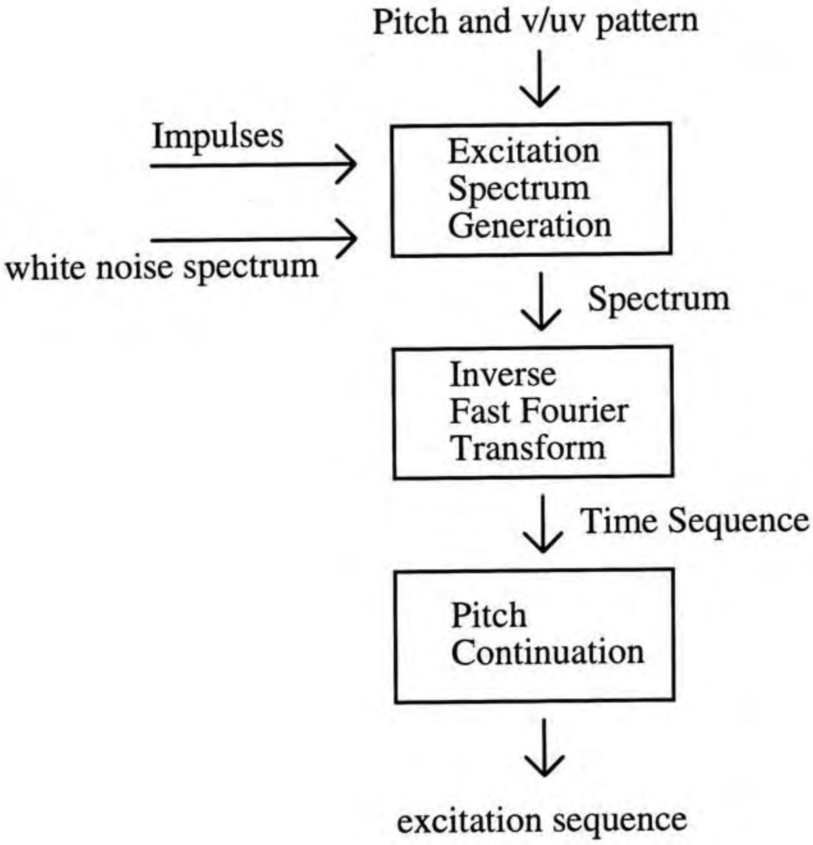


Figure 3.6 Generation of the excitation signal

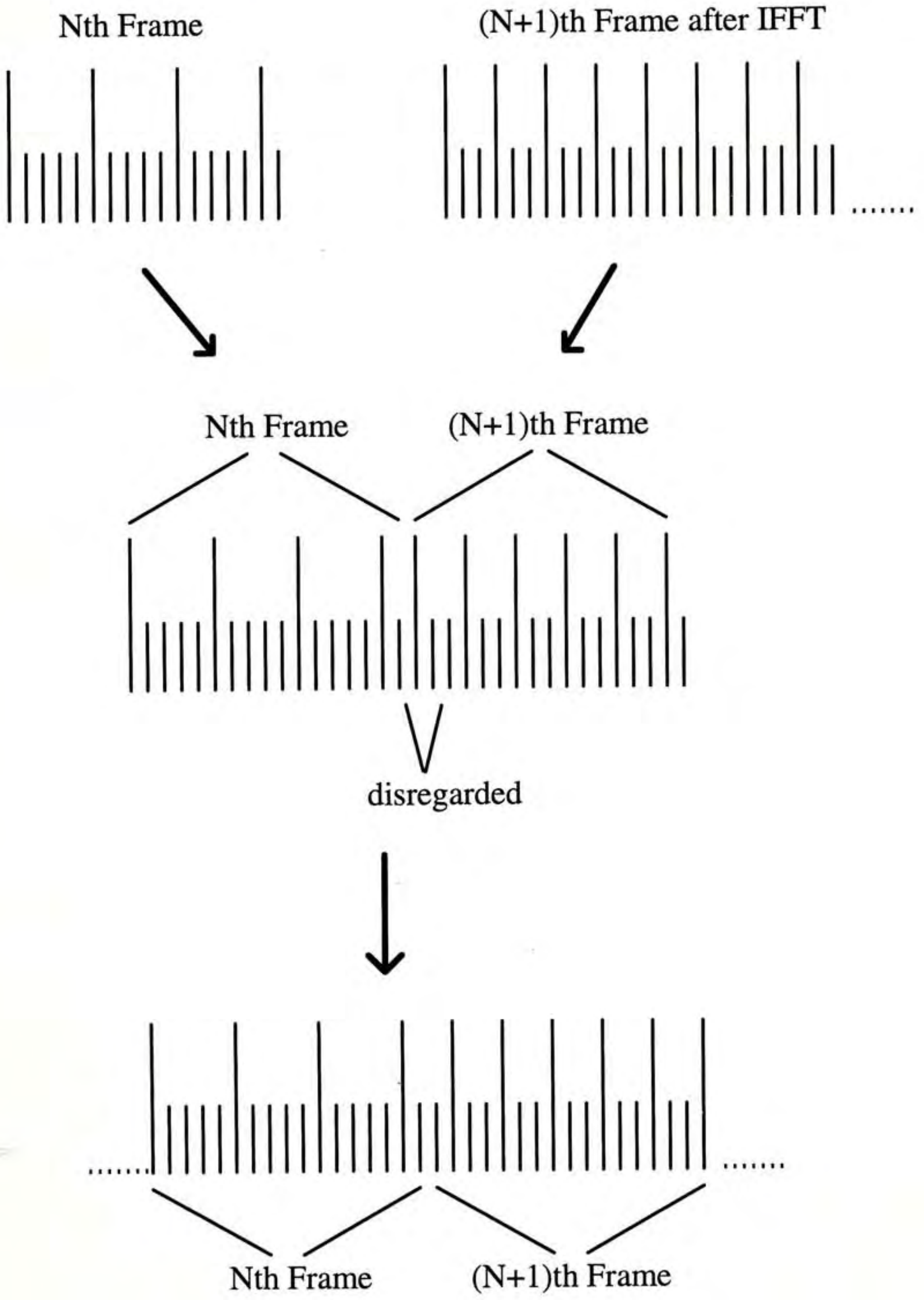


Figure 3.7 Pitch matching in excitation signal generation

Chapter 4 A low bit rate Quad-Band Excitation Line Spectral Pair Vocoder

Based on the proposed QBE model, a Quad-Band Excitation Line Spectral Pair (QBELSP) vocoder has been implemented. It can be operated at a lower bit rate with less complexity as compared with the MBE vocoder. Details of the vocoder will be discussed in the following sections, including its architecture and the encoding/decoding algorithms of the speech parameters.

4.1 Architecture of QBELSP vocoder

A vocoder using QBE has been developed for low bit rate transmission. In Chapter 2, we have mentioned that the spectral envelope of the speech signal is represented by a set of spectral magnitudes M_k in MBE vocoder, which requires a large number of bits for coding. In order to minimize the operating bit rate of the proposed vocoder, QBE is applied to linear predictive coding so that the spectral envelope can be expressed in terms of LPC coefficients for efficient coding. The parameters need to be sent by the transmitter then include the pitch value, v/uv pattern index, a set of LPC coefficients and the gain value. For efficient and robust transmission purpose, the LPC coefficients are converted into line spectral pairs (LSP) before quantization, which will be elaborated in more details later on.

The speech input, band limited from 100 Hz to 3.4 kHz, of the vocoder is sampled at 8 kHz with 16-bit resolution. The data sequence is then divided into segments of 35 ms long with 15 ms overlapping for analysis by multiplying the sequence with a Hamming window. For speech synthesis, a rectangular window of 20 ms is used for each frame without overlapping. The windowing arrangement in the vocoder is given in Figure 4.1 .

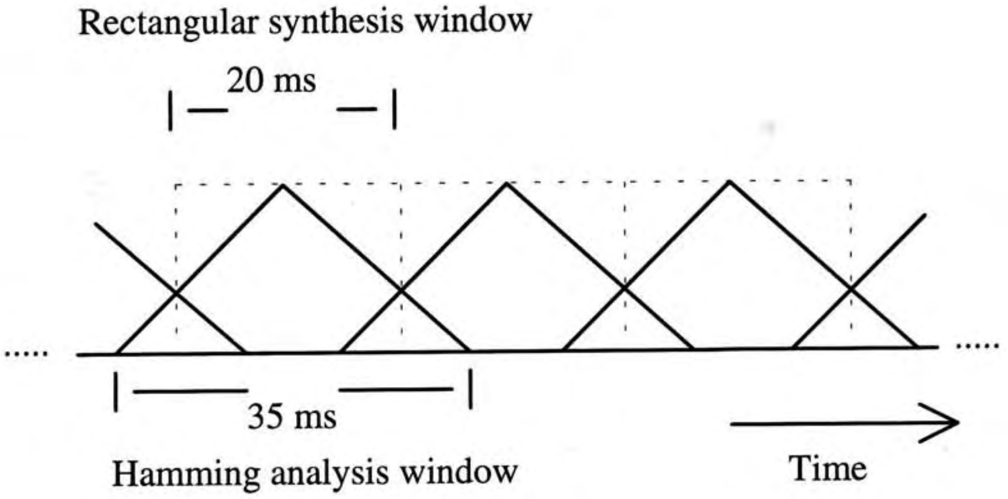


Figure 4.1 Windowing arrangement for speech analysis & synthesis in Quad-Band Excitation LSP vocoder

A block diagram illustrating the structure of the transmitter of the vocoder is shown in Figure 4.2. It involves a number of processes for speech analysis: pitch detection, voiced/unvoiced decision, linear prediction analysis, LPC to LSP conversion and parameter encoding. First of all, the pitch value is determined by using auto-correlation technique with center clipping. The permissible range for pitch value is between 100 Hz to 350 Hz. Any pitch value out of this range would be ignored and the respective speech frame is considered to have no fundamental frequency. The estimated pitch value is used in the voiced/unvoiced decision process to generate the voiced spectrum for spectral comparison. The spectral envelope of the speech data is represented by the LPC coefficients. The order of the predictive filter for analysis is 10. The coefficients are computed by using the well known Durbin's recursive method [9], and they are then converted into line spectral pairs (LSP) [31] before quantization. A gain value is also estimated during the LPC analysis procedure to control the loudness of the output speech.

Different quantization schemes are used for different speech parameters. The pitch value and the v/uv pattern are encoded by the method of table lookup, and the gain value is encoded by using non-linear quantization. For encoding the LSP parameters, vector quantization technique is employed. The feature vector is split into 2 sub-vectors so that the memory required for storing the codebook and the computation needed for codebook searching can be reduced. The number of bits assigned to each of the parameters are summarized in Table 4.1 . The transmission bit rate of the vocoder is operated at 2.2 kbps. The coding procedures of these parameters are described thoroughly in the following sections of this chapter.

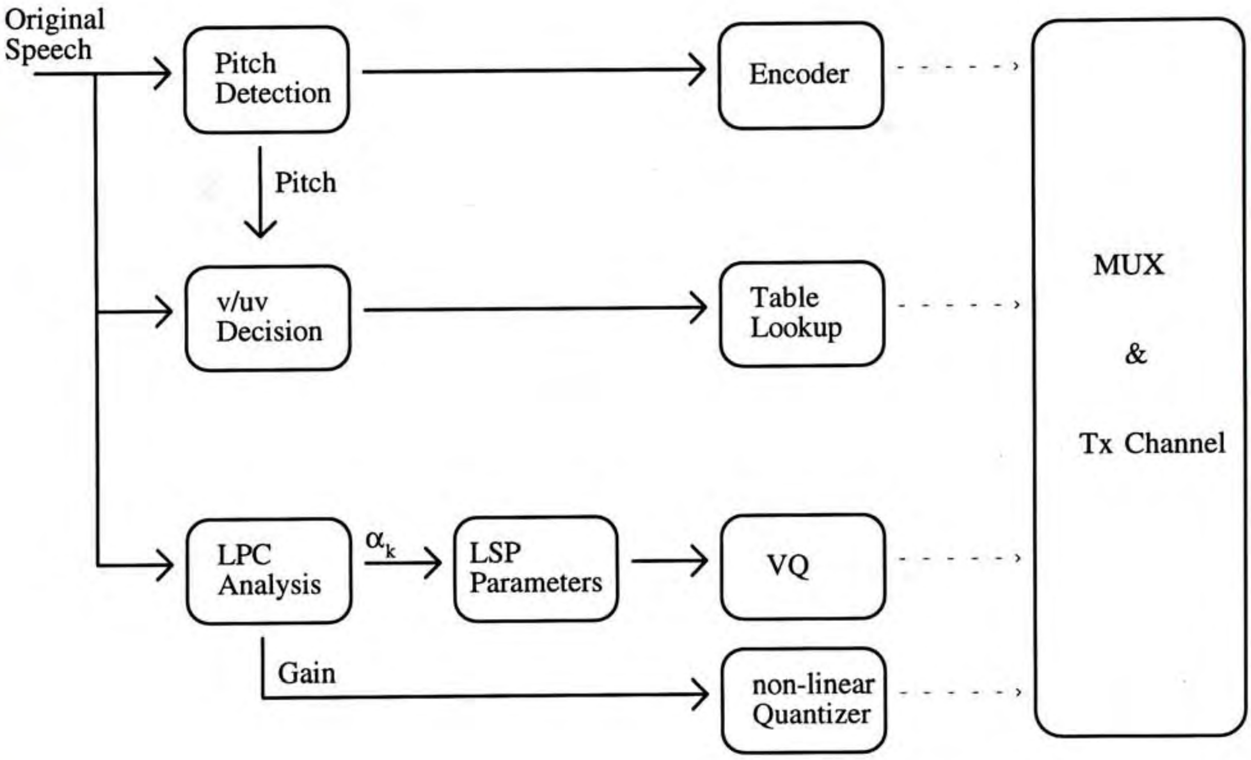


Figure 4.2 Transmitter of the QBELSP vocoder

Parameter	**Number of bits per frame**	**Quantization scheme**
Pitch	7	table-look-up
v/uv pattern	8	table-look-up
LSP frequencies	24	split vector quantization
Gain	5	non-linear quantization

Table 4.1 Bit assignment in QBELSP vocoder

At the receiver, speech output is synthesized by reversing the procedures in the transmitter. After parameter decoding, the pitch value and the v/uv pattern are used to regenerate the excitation spectrum by putting periodic pulses in the voiced bands with the periodicity equals to the fundamental frequency of the corresponding speech frame. While in the unvoiced bands, white noise spectrum is used instead. The voiced/unvoiced energy ratio can be approximated in the following way. After receiving locations of the voiced/unvoiced band boundaries, bandwidth of voiced bands and unvoiced bands are calculated. The voiced/unvoiced energy ratio can be approximated by

$$\text{v / uv energy ratio} = \frac{\text{Total bandwidth of voiced bands}}{\text{Total bandwidth of unvoiced bands}} \tag{4.1}$$

After regenerating the excitation spectrum, the corresponding time-domain excitation signal can be obtained by using inverse Fast Fourier Transform. At the same time, the LPC coefficients can be derived from the decoded LSP parameters. Subsequently, the excitation sequence is fed to the LPC synthesis filter to produce the synthesized speech output. Figure 4.3 shows the functional block diagram of the receiver.

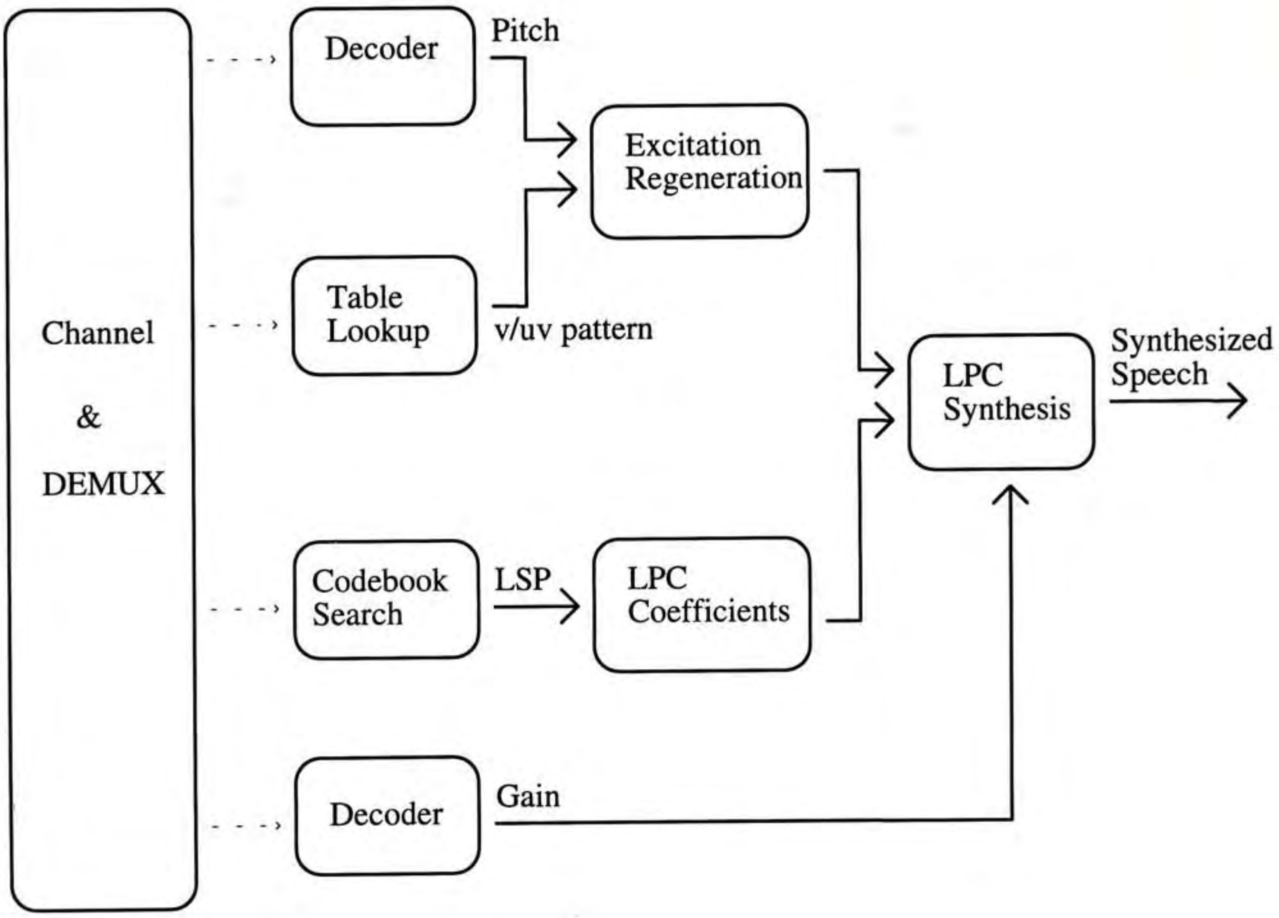


Figure 4.3 Functional block diagram of the receiver of QBELSP vocoder

4.2 Coding of excitation parameters

In the vocoder, parameters for representing the excitation signal is the pitch value and the v/uv pattern. The pitch can be encoded by using simple quantization scheme, while for v/uv pattern, only the locations of the frequency boundaries have to be denoted. In this section, their encoding/decoding algorithm for these parameters are described.

4.2.1 Coding of pitch value

In the auto-correlation method which is used for pitch detection in the vocoder, the estimated pitch period is expressed in terms of the number of data samples. Since linear interpolation is applied to the speech samples in order to reduce quantization error, the final pitch period can have a resolution of up to half a sample. In the vocoder, the permissible fundamental frequency range is approximately 100-350 Hz, which is equivalent to 80 - 23 samples at the sampling rate of 8 kHz. The available set of pitch values in terms of the number of samples will be

$$\{ 23, 23.5, 24, 24.5, \dots , 79, 79.5, 80 \}$$

There are totally 115 pitch values, and the coding process is simple and apparently 7 bits are sufficient. The codeword representing the pitch value is shown in Table 4.2 . At the receiver of the vocoder, the pitch value can be retrieved by a simple searching procedure.

Codeword	Pitch	
	Number of samples	Frequency (Hz)
000 0001	23	348
000 0010	23.5	340
000 0011	24	333
. . .	. . .	. . .
111 0011	80	100

Table 4.2 Codeword representation of pitch period

4.2.2 Coding of voiced/unvoiced pattern

As described previously, only 4 possible voiced/unvoiced band patterns are allowed in QBE, which are {v}, {v uv}, {v uv v} and {v uv v uv} respectively. Since we have initially divided the excitation spectrum into K, $4 \leq K \leq 12$, sub-bands depending on the fundamental frequency P, therefore there might be up to 11 sub-band boundaries B_i, $1 \leq i \leq K\text{-}1$. After grouping to at most 4 voiced/unvoiced bands, we have to redefine these band locations from the sub-band boundaries B_i. However, since B_i can be derived implicitly from the pitch period, we only need to encode the particular voiced/unvoiced pattern by identifying the band boundaries in terms of the sub-band index. In fact, there are totally 232 possible v/uv boundary patterns and Table 4.3 illustrates how they can be encoded effectively. For simplicity, these patterns are stored in a table, and the index corresponding to a particular pattern can be determined by a table lookup operation. As a result, 8 bits are adequate to represent all the v/uv pattern of the excitation spectrum.

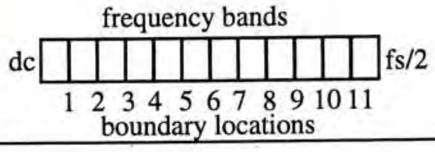
v/uv patterns dc $f_s/2$	possible v/uv boundaries  frequency bands dc fs/2 1 2 3 4 5 6 7 8 9 10 11 boundary locations	Number of possible cases	8-bit Codeword
v	$f_s/2$	1	0000 0000
v uv	{1},{2},{3},....,{11}	11	0000 0001
			:
			0000 1011
v uv v	{1,2},{1,3},{1,4},....,{1,11}		0000 1100
	{2,3},{2,4},....,{2,11}		:
	{3,4},....,{3,11}		:
	:		:
	{10,11}	55	1000 0100
v uv v uv	{1,2,3},{1,2,4},{1,2,5},....,{1,2,11}		1000 0011
	{1,3,4},{1,3,5},....,{1,3,11}		:
	:		:
	{1,10,11}		:
			:
	{2,3,4},{2,3,5},{2,3,6},....,{2,3,11}		:
	{2,4,5},{2,4,6},....,{2,4,11}		:
	:		:
	{2,10,11}		:
			:
	.		:
	.		:
	{9,10,11}	165	1110 0111
	Total	232	

Table 4.3 Encoding of v/uv band patterns in Quad-Band Excitation

4.3 Spectral Envelope Estimation and Coding

As the vocoder is based on linear prediction of speech signal, the spectral envelope is represented by 10 LPC coefficients and the gain value. The LPC coefficients are transformed into line spectral pairs for efficient encoding. A number of quantization scheme for coding the line spectral pairs have been examined and are explained in the following sections.

4.3.1 Spectral Envelope & the gain value

During the voiced/unvoiced decision in MBE vocoder, the spectral envelope parameter M_k is evaluated which represent the envelope of the kth frequency band. However, for QBE, using M_k is basically not practical. It is because only a maximum of 4 frequency bands are employed to cover the whole spectrum from dc to half the sampling frequency. The bandwidth for a particular band would not be narrow enough to assume a flat spectral shape. Therefore, the method of linear prediction of speech signal is adopted in the proposed vocoder and LPC coefficients are used to represent the spectral envelope. The LPC coefficients are obtained by the Durbin's recursive method [9] and a subroutine of this method is included in Appendix C.

For QBE, the gain value of the LPC synthesis filter can be calculated as follows. From equation (2.2),

$$s(n) = Gx(n) - \sum_{k=1}^{N} \alpha_k s(n-k) \tag{4.2}$$

By squaring both sides and taking expectation,

$$E[G^2x^2(n)] = E[(s(n)+\sum_{k=1}^{N}\alpha_k s(n-k))^2]$$

$$G^2E[x^2(n)] = E[s^2(n)+2s(n)\sum_{k=1}^{N}\alpha_k s(n-k)+(\sum_{k=1}^{N}\alpha_k s(n-k))^2]$$

$$G^2E[x^2(n)] = E[s^2(n)]+2E[\sum_{k=1}^{N}\alpha_k s(n)s(n-k)]+E[\sum_{k=1}^{N}\sum_{k=1}^{N}\alpha_i\alpha_j s(n-i)s(n-j)] \quad (4.3)$$

If the auto-correlation function of *x(n)* and *s(n)* are denoted by $R_{xx}(\ )$ and $R_{ss}(\ )$ respectively, equation (4.3) can be rewritten as

$$G^2R_{xx}(0) = R_{ss}(0)+2\sum_{k=1}^{N}\alpha_k R_{ss}(k)+\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j R_{ss}(i-j)$$

$$G = \sqrt{\frac{1}{R_{xx}(0)}[R_{ss}(0)+2\sum_{k=1}^{N}\alpha_k R_{ss}(k)+\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j R_{ss}(i-j)]} \qquad (4.4)$$

By assuming the synthesized signal *s(n)* is the same as the original one, the gain value can then be found accordingly.

4.3.2 Line Spectral Pair (LSP)

LPC coefficients are not normally encoded for transmission because of their large dynamic range. A number of methods are available to represent the LPC coefficients in other formats for encoding, such as log area ratio (LAR) [10] and partial correlation coefficients (PARCOR) [32]. In QBELSP vocoder, the LPC coefficients are converted into line spectral pairs (LSP) frequencies [31][33], of which there exists a much narrower dynamic range which is suitable for quantization, especially the intraframe frequency differences. From LPC analysis, the analysis filter transfer function, order *N*, is given by

$$A(z) = 1 + \sum_{k=1}^{N} \alpha_k z^{-k}$$

The filter can be reconstructed into one symmetric and one asymmetric filter by the following equation

$$\begin{aligned} P(z) &= A(z) - z^{-(N+1)} A(z^{-1}) \\ Q(z) &= A(z) + z^{-(N+1)} A(z^{-1}) \end{aligned} \tag{4.5}$$

where *P*(*z*) is an asymmetric filter and *Q*(*z*) is a symmetric filter, and they are called the LSP polynomials [32]. Both of them are of order (*N*+1). Since *P*(*z*) and *Q*(*z*) has the root of z = -1 and z = +1 respectively, there are *N* more roots in each polynomial and they exist in conjugate pairs. Let the roots of *P*(*z*) and *Q*(*z*) be ω_i and β_i respectively for $1 \le i \le N/2$. These roots have the following characteristics

1. All ω_i and β_i lies on the unit circle for $1 \le i \le N/2$
2. The roots of *P*(*z*) and *Q*(*z*) exist such that

$$\omega_1 < \beta_1 < \omega_2 < \beta_2 < \ \dots\ < \omega_{N/2} < \beta_{N/2} \tag{4.6}$$

3. *A*(*z*) would maintain the minimum phase property if the above two criteria are both satisfied.

By using the first characteristics, $\{\omega_i\}$ and $\{\beta_i\}$ can be found by substituting $z = e^{j\omega}$ and setting the polynomials to be zero. These frequencies are called the line spectral frequencies.

Line spectral frequencies are closely related to the linear prediction coefficients, and both of them are used to represent the location of the formant frequencies of the speech frame. It can be shown that [34] the LSP frequencies are at the locations very close to the formant frequencies, and normally 2 LSP frequencies are used to characterize one formant frequency. It is important to note that the spectral sensitivities of the LSP frequencies are localized. It means that, for example, if the LSP frequency that characterize a certain formant frequency is varied, only the spectral envelope around the corresponding formant frequency would be changed and the rest would remain unchanged. Consequently, it is beneficial for encoding since the spectral distortion due to the quantization error to one of the LSP frequencies only distort the envelope in a limited frequency range.

There are a number of methods for computing the LSP frequencies [35][36]. One simpler method is the computation of *P*(*z*) and *Q*(*z*) with the help of Chebyshev Polynomials [37]. In this method, since -1 and +1 are one of the roots of *P*(*z*) and *Q*(*z*) respectively, after considering the other roots to be on the unit circle, the polynomials in (4.5) can be written as

$$\begin{aligned} P(e^{j\omega}) &= (1-e^{-j\omega})e^{-j\omega N}P'(\omega) \\ Q(e^{j\omega}) &= (1+e^{-j\omega})e^{-j\omega N}Q'(\omega) \end{aligned} \tag{4.7}$$

where

$$\begin{aligned} P'(\omega) &= 2\cos N\omega + 2p_1'\cos(N-1)\omega + \ldots + 2p_{N-1}'\cos\omega + p_N' \\ Q'(\omega) &= 2\cos N\omega + 2q_1'\cos(N-1)\omega + \ldots + 2q_{N-1}'\cos\omega + q_N' \end{aligned} \tag{4.8}$$

and the coefficients in polynomials (4.8) are functions of LPC coefficients. The computation for the LSP frequencies that represent the spectral envelope is then equal to the computation of the roots of the polynomials (4.8). By using the

Chebyshev polynomials, the polynomials can be calculated more effectively. Consider the mapping

$$x = \cos \omega$$

then

$$\cos k\omega = T_k(x) \tag{4.9}$$

where $T_k(x)$ is the kth order of Chebyshev Polynomial with initial conditions $T_0(x) = 1$ and $T_1(x) = 1$, and polynomials in (4.8) can be written as

$$\begin{aligned} P'(x) &= 2T_N(x) + 2p_1'T_{N-1}(x) + \ldots + 2p_{N-1}'T_1(x) + p_N' \\ Q'(x) &= 2T_N(x) + 2q_1'T_{N-1}(x) + \ldots + 2q_{N-1}'T_1(x) + q_N' \end{aligned} \tag{4.10}$$

which can be generalized into

$$Y(x) = \sum_{k=0}^{N} c_k T_k(x) \tag{4.11}$$

By using a recurrence relationship

$$r_k(x) = 2xr_{k+1}(x) - r_{k+2}(x) + c_k \tag{4.12}$$

and the recursion of Chebyshev Polynomials,

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \tag{4.13}$$

equation (4.11) can be rewritten into

$$Y(x) = \sum_{k=0}^{N} [r_k(x) - 2xr_{k+1}(x) + r_{k+2}(x)] T_k(x)$$

$$Y(x) = \frac{r_0(x) - r_2(x) + c_0}{2} \tag{4.14}$$

As a result, the polynomials in (4.8) can be computed more efficiently by using (4.14), and the roots can be found by using numerical methods. The line spectral frequencies can be obtained by the following mapping

$$\omega_i, \beta_i = \cos^{-1} x_i \tag{4.15}$$

where $\{x_i\}$ are the roots of $Y(x)$.

Since the polynomials in (4.10) are of an order N and the roots occur in conjugate pairs, $N/2$ LSP frequencies will be solved in each of the polynomials. As a result, a set of LPC coefficients of size N can be transformed into a set of LSP frequencies which is also of size N. The procedures for determining the LSP frequencies has been written as a computer C program subroutine and is included in Appendix D for reference.

The conversion of the LSP frequencies back to LPC coefficients can be carried out much more directly. A second order polynomial can be obtained from each conjugate pair of the LSP frequencies

$$\begin{aligned} &1 - 2\cos\omega_i z^{-1} + z^{-2} \\ &1 - 2\cos\beta_i z^{-1} + z^{-2} \end{aligned} \tag{4.16}$$

and equation $P(z)$ and $Q(z)$ can be computed by direct multiplication of the second order section and the first order polynomial by the root of unity, i.e.

$$P(z)=(1-z^{-1})\prod_{i=1}^{N/2}(1-2\cos\omega_i+z^{-2}) \tag{4.17}$$

$$Q(z)=(1+z^{-1})\prod_{i=1}^{N/2}(1-2\cos\beta_i+z^{-2}) \tag{4.18}$$

The LPC analysis filter $A(z)$ would be obtained easily as

$$A(z)=\frac{P(z)+Q(z)}{2} \tag{4.19}$$

4.3.3 Coding of LSP frequencies

Due to the narrow dynamic range of the LSP frequencies, efficient encoding using less bits is possible. Different encoding schemes have been investigated, including linear and non-linear quantization of the intraframe LSP frequency differences and split vector quantization on LSP frequencies. The performance of the encoding techniques is evaluated by the calculation of a spectral distortion measure, which is defined as [38]

$$\sqrt{\frac{1}{\pi}\int_0^{\pi}[10\log P_1(\omega)-10\log P_2(\omega)]^2\,d\omega} \tag{4.20}$$

where $P_1(\omega)$ and $P_2(\omega)$ are the frequency power spectra before and after quantization respectively.

Linear quantization on intraframe LSP differences

It is known [33] that the LSP frequencies are bounded by

$$0 < \omega_1 < \beta_1 < \omega_2 < \beta_2 < \ \dots \ < \omega_5 < \beta_5 < \pi \tag{4.21}$$

and each of the LSP frequencies have very narrow dynamic range, especially the intraframe frequency differences [39][40]. When they are normalized to the sampling frequency, it is found that each of the frequency difference is bounded between 0 and 0.15 . Uniform quantization have been used to encode the frequency differences by dividing the range from 0 to 0.15 into levels with a uniform step size, and equal number of bits were used to encode each difference. This quantization scheme was examined by applying to 2900 speech frames which were recorded from TV news broadcasting. Spectral distortion measure was carried out for each frame and finally an average distortion measure was obtained. Different number of bits per frame have been allocated to encode the frequency differences with an equal number of bits for each difference. The performance of the quantizer is plotted on Figure 4.4 . It was found that if 1 dB spectral distortion is required, almost 40 bits are needed for encoding 10 frequency differences. In other words, an average of 4 bits is necessary to encode each individual frequency difference.

Non-linear quantization on intraframe LSP differences

It has been shown that [39] the distribution of each of the LSP frequency differences, in fact, is not uniform within the dynamic range. Thus another approach has been proposed which is to apply non-linear quantization for LSP frequency difference encoding [39]. In this method, non-even step sizes are used which are evaluated using the concept of dynamic programming [41]. In general, one quantization level is optimized initially and then it is used to optimize the second quantization level. The procedure repeats until all the quantization levels are optimized. In other words, dynamic programming is used to achieve a global optimization by an accumulation of the local optimizations. During the quantizer design, a set of speech frames which contains both male and female voice was used for quantization level optimization.

Similar to the case of linear quantization, the performance of the non-linear quantizer was evaluated by allocating different number of bits per frame for quantization. Again, the number of bits assigned for each difference was the same. The testing of the encoding scheme is the same as in linear quantization by using the spectral distortion measure. The average spectral distortion is also depicted in Figure 4.4 for the ease of comparison. It is noted that the performance of the quantizer is better than that of using linear quantization. When compared with linear quantization scheme, almost 5 bits are saved per frame to obtain an average distortion of 1 dB.

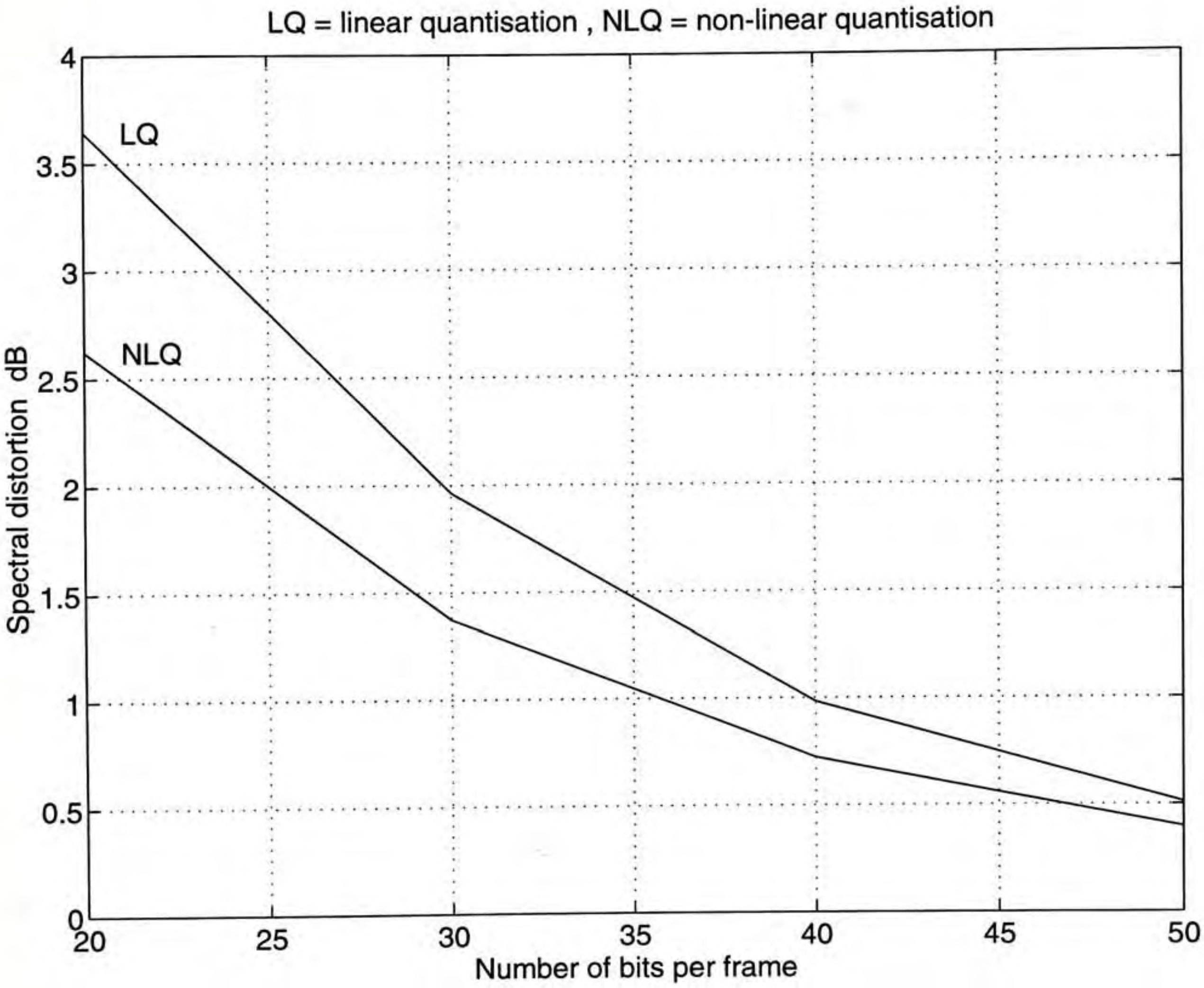


Figure 4.4 Average spectral distortion for quantizing LSP frequencies

Vector quantization on LSP frequencies

Vector quantization (VQ) has been widely used in speech coding [42][43], especially in the quantization of spectral parameters for speech signals. It allows a further reduction on spectral distortion when compared with scalar quantization at any given bit rate, but in the expense of a high computation load. The LPC parameters can be vector-quantized effectively in line spectral frequency domain [44]. In addition, due to the characteristics of localized spectral sensitivities of the LSP frequencies, they can be split into 2 or more sub-vectors for quantization [38] so as to reduce the memory size and the computational complexity. In the design of our QBELSP vocoder, split vector quantization is applied to enable low bit rate transmission.

When designing a VQ codebook, the most important factor affecting the performance of the quantizer is the choice of a suitable distance measure. For LSP frequencies, a weighted Euclidean distance measure [38] is used as follows,

$$d(\mathbf{f},\hat{\mathbf{f}}) = \sum_{k=1}^{10}[(P(f_k))^r(f_k - \hat{f}_k)]^2 \tag{4.22}$$

where $\mathbf{f} = \{f_1, f_2, \ldots, f_{10}\}$ and $\hat{\mathbf{f}} = \{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_{10}\}$ are the original and the quantized LSP frequency vector respectively, $P(f_k)$ is the power spectrum of the signal at frequency f_k, and r is an empirical constant which is set to 0.15 . By introducing a weighting factor depending on the power spectrum, it gives more weighting to the dominant formant frequencies and a less weighting to the formant frequencies that have lower energy. In addition, since human ear has a more sensitive frequency response at the low frequency region, equation (4.20) is further modified into

$$d(\mathbf{f},\hat{\mathbf{f}})=\sum_{k=1}^{10}[c_k(P(f_k))^r(f_k-\hat{f}_k)]^2 \tag{4.23}$$

where $$c_k=\begin{cases}1.0 & 1\le k\le 8\\ 0.8 & k=9\\ 0.4 & k=10\end{cases}$$

The LBG algorithm [45] is employed to design the VQ codebook. In this method, a codebook for M-bit quantizer is trained through the training of a k-bit codebook where k is varied from 1, 2, ... till k = M. The details of the algorithm is as follows,

1. From the set of training vectors, one centroid $\mathbf{C}_1$ is found initially by determining the mean vector of the data. Set the number of quantization level i = 1 .
2. For each centroid $\mathbf{C}_j$ for j = 1,..., i , split the centroid into 2 candidates $\mathbf{C}_j + \mathbf{x}$ and $\mathbf{C}_j - \mathbf{x}$, where x is a fixed vector. Set i = 2i. It is an initial guess by the procedure of splitting.
3. The training vectors are partitioned into i set such that each vector set has its own centroid, It is done by using the method of minimization of the distance measure between each training vector and the centroids. An average distortion is then calculated after partitioning which is the average distance measure between the training vectors and their centroids.
4. If the percentage reduction on the average distortion compared with the last iteration is within a tolerable value, the training of the i-level quantizer completes. Otherwise, finding, the new centroid in each partition by calculation of the mean vector and the procedures repeats from (3).
5. If the training of the i-level quantizer completes, repeats the procedures from (2) until the number of quantization levels equals to 2^M.

During codebook training, 67200 LSP frequency training vectors were used, which were taken by many utterances uttered by many different speakers. Another set of 2500 vectors were used to test the performance of the quantizer.

A number of trials have been performed for quantization of the LSP frequencies, including using different splitting of the feature vector and using different codebook size. The spectral distortion measure in (4.20) is employed to evaluate the performance of the quantizers and an average value is obtained from the distortion of the testing frames of speech. When splitting the LSP frequency parameters into 2 sub-vectors, each contains 5 frequencies, it is found that after consideration of the computation complexity, 24 bits per frame is suitable to quantize the LSP frequencies, and it is noted that equal number of bits assigned to each of the vectors provides the best performance. The experimental results are summarized in Table 4.4 .

Since normally 2 or 3 LSP frequencies are used to characterize one formant frequency and in general the first two formant frequencies are more dominant, splitting of the LSP frequencies into 3 sub-vectors have also been tested. The lowest 3 frequencies are grouped into one vector, and then the next 3 frequencies into another vector, while the highest four frequencies are represented by the third vector. The performance of this quantizer is shown in Table 4.5 .

Number of bits for the 1st codebook	Number of bits for the 2nd codebook	Average spectral distortion (dB)
10	10	1.92
12	12	1.53
11	13	1.56
13	11	1.54

Table 4.4 Average distortion measure using split-2 VQ on LSP frequencies

Number of bits in the 1st codebook (low frequencies)	Number of bits in the 2nd codebook	Number of bits in the 3rd codebook (high frequencies)	Average distortion (dB)
8	8	8	1.60 (24 bits)
10	10	4	2.14 (24 bits)
10	8	6	1.88 (24 bits)
12	8	4	2.24 (24 bits)
10	10	6	1.75 (26 bits)
8	8	10	1.31 (26 bits)
12	8	8	1.54 (28 bits)
10	10	8	1.40 (28 bits)
10	10	10	1.08 (30 bits)

Table 4.5 Average distortion using split-3 VQ on LSP frequencies

From the results, it is found that when using split VQ for LSP frequencies, the performance of using 3 sub-vectors is not as good as using 2 sub-vectors, and an even allocation of the number of bits to each parameter for a particular bit rate produces the best result. After taking into consideration of the computational complexity as well, it is concluded that 24 bits should be used in split-2 VQ for quantizing LSP frequencies.

In order to make the quantizer robust, the codevectors should be reordered such that the Hamming distance of the neighboring codevectors has a close value. However, it is found [38] that the ordering formed by the splitting method as initial guess of the codevectors in LBG algorithm already contains the robust characteristics. In other words, the vector quantizers used in the vocoder already provides a good robustness.

4.3.4 Coding of gain value

The gain value of the LPC synthesis filter in the vocoder is calculated using equation (4.4), which is

$$G=\sqrt{\frac{1}{R_{xx}(0)}[R_{ss}(0)+2\sum_{k=1}^{N}\alpha_k R_{ss}(k)+\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j R_{ss}(i-j)]} \qquad (4.24)$$

The distribution of the gain value has been examined by using the speech file with 2900 frames, and it was found that the gain value was bounded between 0 and 600, most of them had the value below 100. The distribution is shown in Figure 4.5.

As a result, a non-linear quantization scheme can be used to quantize the gain value more effectively. The quantization levels are determined by using the method of dynamic programming, which is similar to the non-linear quantizer design for encoding the LSP frequency differences. In QBELSP vocoder, 5 bits are used to encode the gain value.

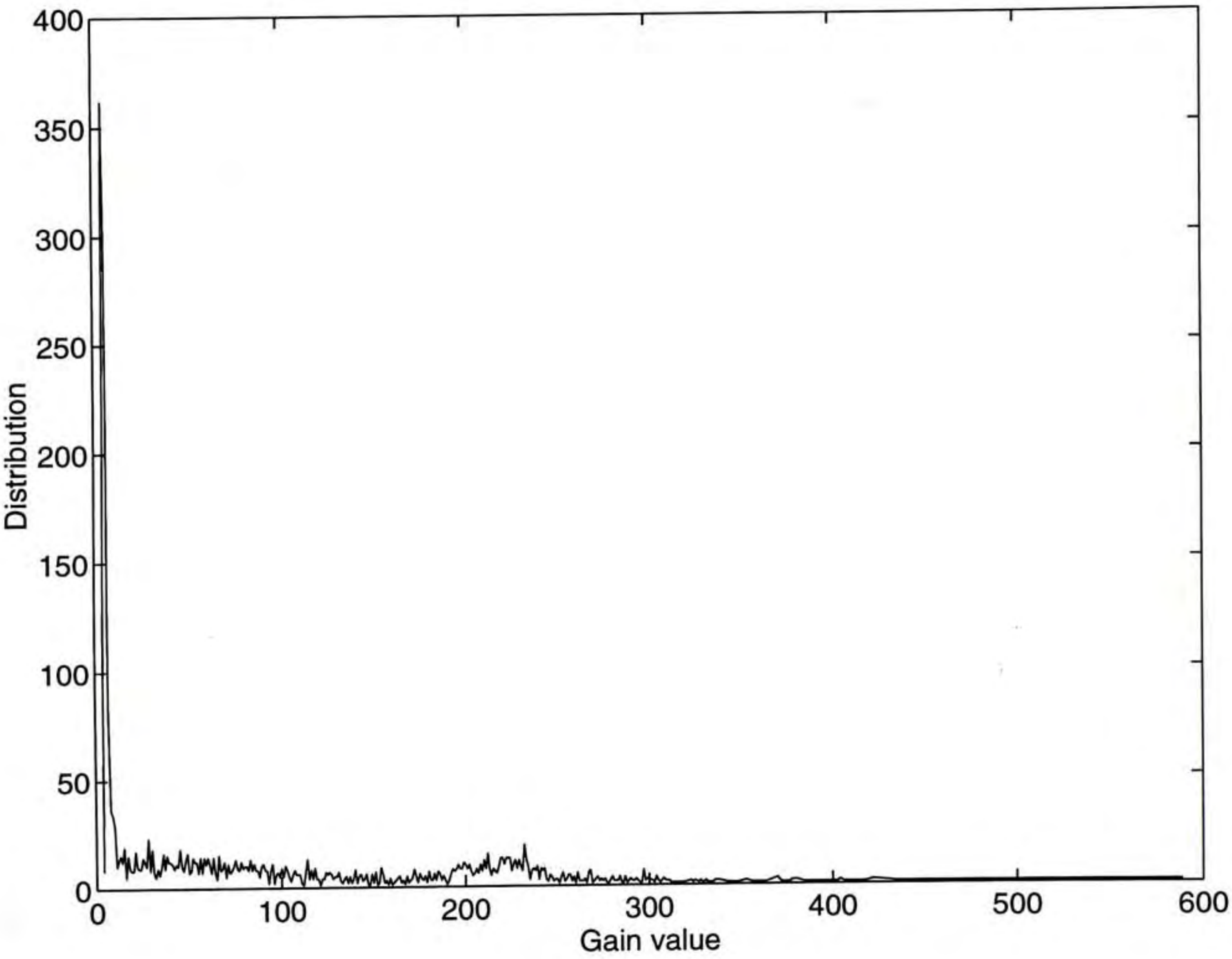


Figure 4.5 Distribution of gain value

Chapter 5 Performance evaluation

The performance of the Dual-Band and Quad-Band Excitation LSP vocoder have been studied by using computer simulation on a DEC workstation. Different speech samples have been used as inputs to the vocoder and the synthesized speech from the receivers have been analyzed and evaluated. In addition, subjective listening tests have been taken to examine the quality and the intelligibility of the synthesized speech. The details of the evaluation will be given in the following sections.

During the simulation, both Cantonese and English languages spoken by different speakers were used as input speech. They were band-limited before digitization by using decade filters to avoid the problem of aliasing. The input speech signals were sampled at a sampling rate of 8 kHz with 16-bit resolution (15 data bit and 1 sign bit). In the vocoder, the size of the analysis Hamming window was 35 ms, with a time shift of 20 ms. In speech synthesis, rectangular non-overlapping window with a window length of 20 ms was employed. Both the output speech that were synthesized by using unquantized and quantized parameters have been examined. The QBELSP vocoder was operated at 2.2 kbps. When DBE is used, the operating bit rate was reduced to 2 kbps.

5.1 Spectral analysis

During analysis, spectrograms of the synthesized speech were computed and examined. A spectrogram is a 3-D spectral plot with the 3rd dimension denoting the energy axis. Four speech sentences uttered by different speakers have been used as input to the vocoder during the analysis,

Sample 1	English short sentence of a male speaker,

Sample 2 Cantonese short sentence of a male speaker,

Sample 3 English short sentence of a female speaker, and

Sample 4 Cantonese short sentence of a female speaker.

These speech were mainly extracted from TV news broadcasting. The input samples and the synthesized speeches were fed to a computer and the corresponding spectrograms were calculated. The spectrograms of the output speech from the QBELSP vocoder are shown in Figures 5.1 - 5.4, and those of the outputs from the DBELSP vocoder are shown in Figures 5.5 - 5.8. In each of the figures, part (A) and (B) show the input speech waveform and its spectrogram respectively. Part (C) is the spectrogram of the synthesized speech using unquantized parameters and part (D) is the spectrogram of the output speech synthesized by using quantized parameters.

It was found from the spectrograms that both DBE and QBE LSP vocoder could produce speech in which the frequency contents were very close to that in the original speech. Due to the effective spectral representation by using LPC coefficients, the formant frequencies were successfully retained at the synthesized speech. It is also noted that the vertical striations occurred in the voiced portions had similar periodicity when compared with that in the original spectrogram. These striations related to the quasi-periodicity of the voiced speech. In other words, the pitch value was accurately determined during speech analysis. In the frequency domain, the frequency components that contain more energy could be preserved. However, some details were lost in the lower energy region. This discrepancy could be seen at the high frequency region of the second last syllable in Figure 5.2 and 5.6 . It was due to the use of unvoiced band in the high frequency region instead of voiced band during the v/uv decision simplification. The unvoiced excitation possessed a random spectrum that essentially destroy the details in the corresponding frequency region of the synthesized speech. By examining the synthesized speech using unquantized and quantized parameters, the difference between their spectrograms was found to be insignificant. In other words, the quantization errors of the speech parameters did not produce serious effects in the synthesized speech.

When the spectrograms of the synthesized speech using DBE and QBE were further investigated, it was noted that they were very similar and the difference was not obvious. When compared with QBE, the details in the high frequency region were not exact when DBE was applied. This can be noticed at the fourth syllable in Figure 5.2 and 5.6 . Another example can be found when comparing the end of the fourth utterance in Figure 5.1 and 5.5 .

In order to further evaluate the difference in the vocoder performance when using these two different types of excitation, spectral distortion measure between the input and the synthesized speech by using each of the excitation schemes has been calculated. The results were compared with that of using MBE model. The spectral distortion measure is defined as the root mean squared error between the log power spectrum of the input speech $P_1(\omega)$ and that of the synthesized speech $P_2(\omega)$ [38], which is given by

$$\text{Spectral distortion} = \sqrt{\frac{1}{\pi}\int_0^{\pi}[10\log P_1(\omega) - 10\log P_2(\omega)]^2\, d\omega} \qquad (5.1)$$

The distortion measure was calculated for each speech frame and an average value was calculated. The results are listed in Table 5.1 .

From Table 5.1 , the average spectral distortion of the synthesized speech was found to be about 8 dB when DBE or QBE was used. In addition, all the three excitation schemes gave similar performance. It showed that DBE and QBE could be used as an alternative to MBE with insignificant distortion. Furthermore, the operating bit rate of the vocoder could also be greatly reduced. Generally, the distortion using QBE was found to be slightly lower than that using DBE. This was expected because more voiced/unvoiced decision patterns were simplified when DBE was used and it would induce more distortion. Consequently, we conclude that QBE can produce a better performance than DBE, even though the improvement might not be very obvious in many instances.

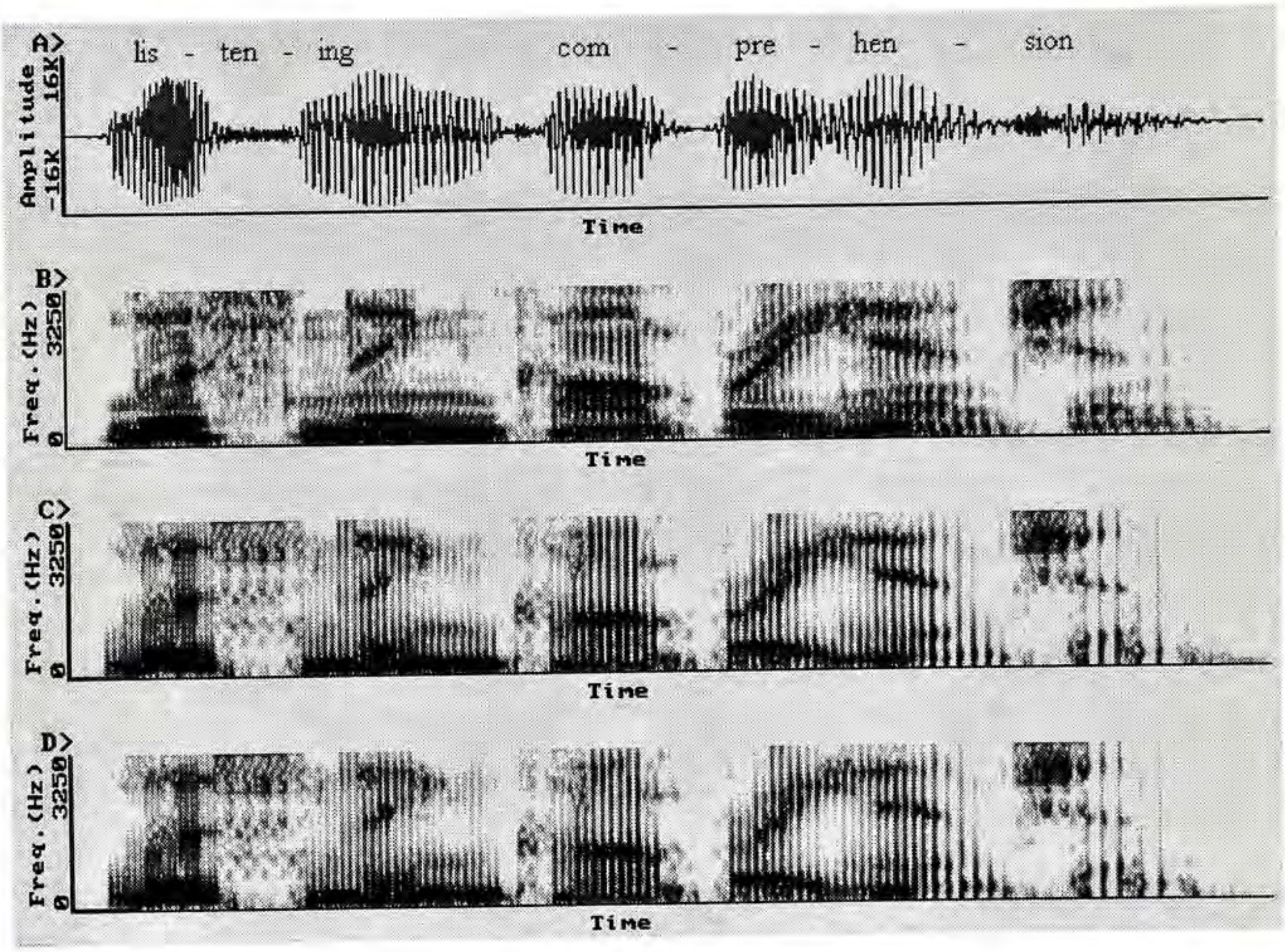


Figure 5.1 Spectrograms of synthesized speech using QBE
A) Input waveform (sample 1), B) Spectrogram of (A),
C) Spectrogram of output using uncoded parameters,
D) Spectrogram of output using coded parameters

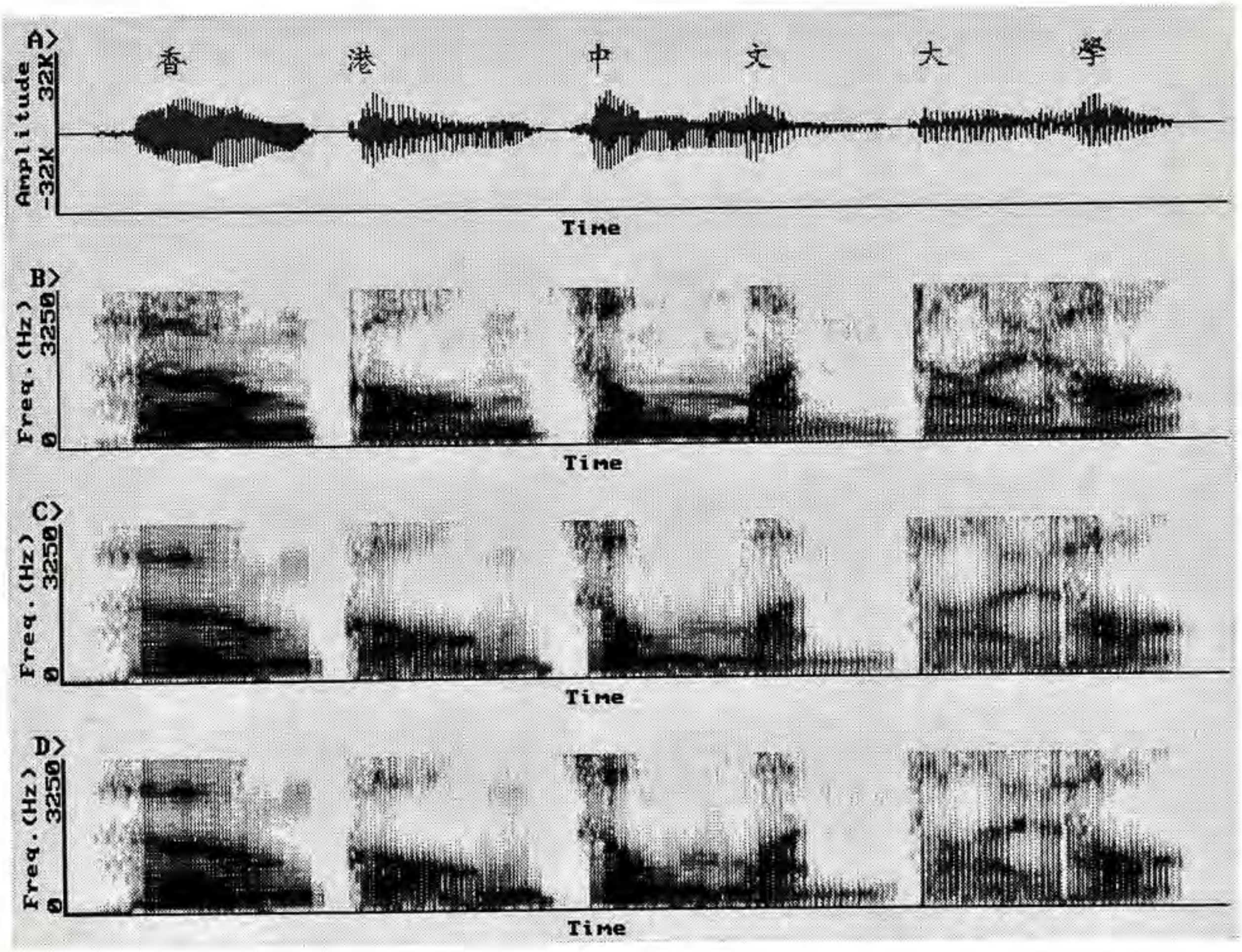


Figure 5.2 Spectrograms of synthesized speech using QBE
A) Input waveform (sample 2), B) Spectrogram of (A),
C) Spectrogram of output using uncoded parameters,
D) Spectrogram of output using coded parameters

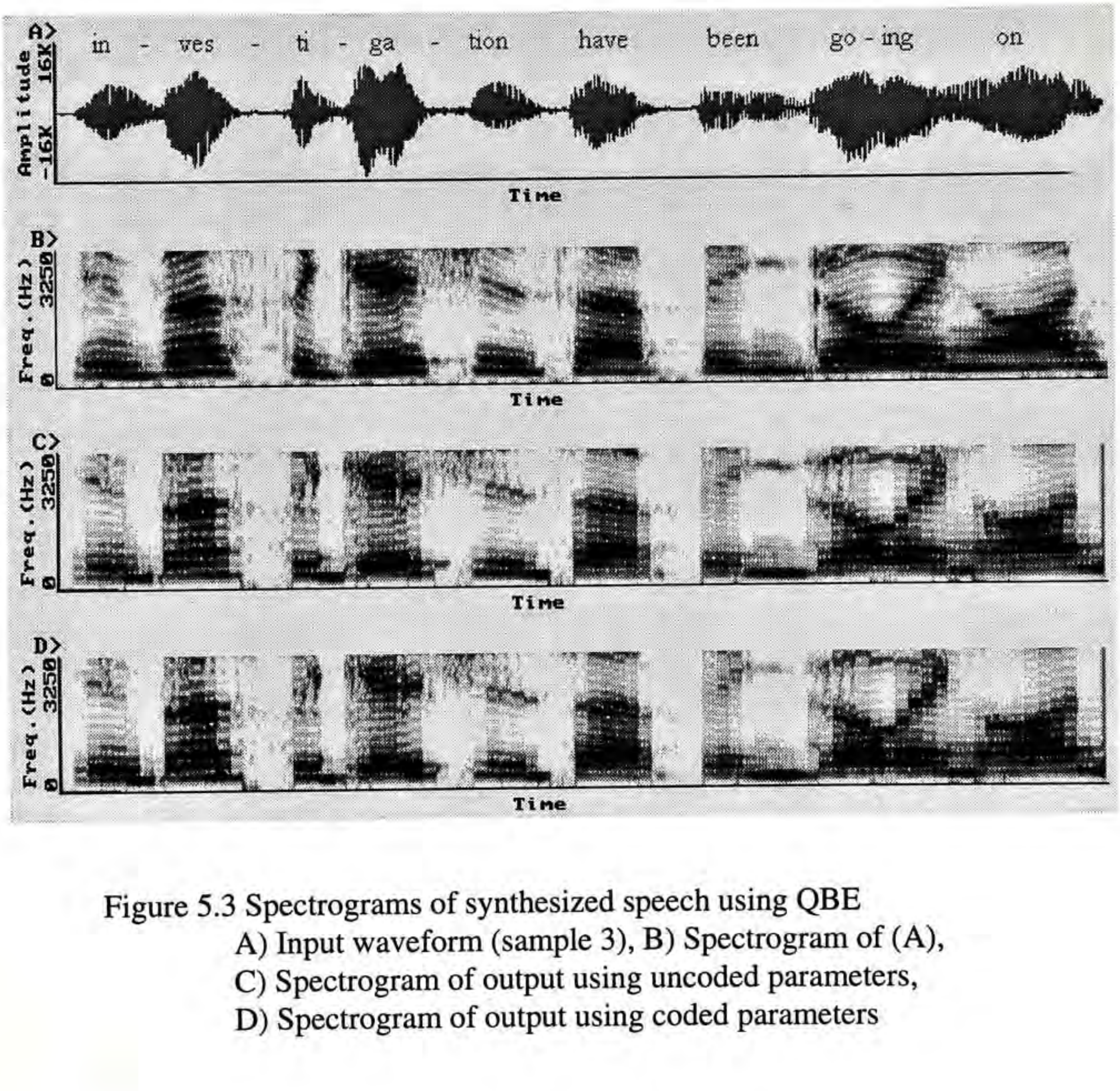


Figure 5.3 Spectrograms of synthesized speech using QBE
A) Input waveform (sample 3), B) Spectrogram of (A),
C) Spectrogram of output using uncoded parameters,
D) Spectrogram of output using coded parameters

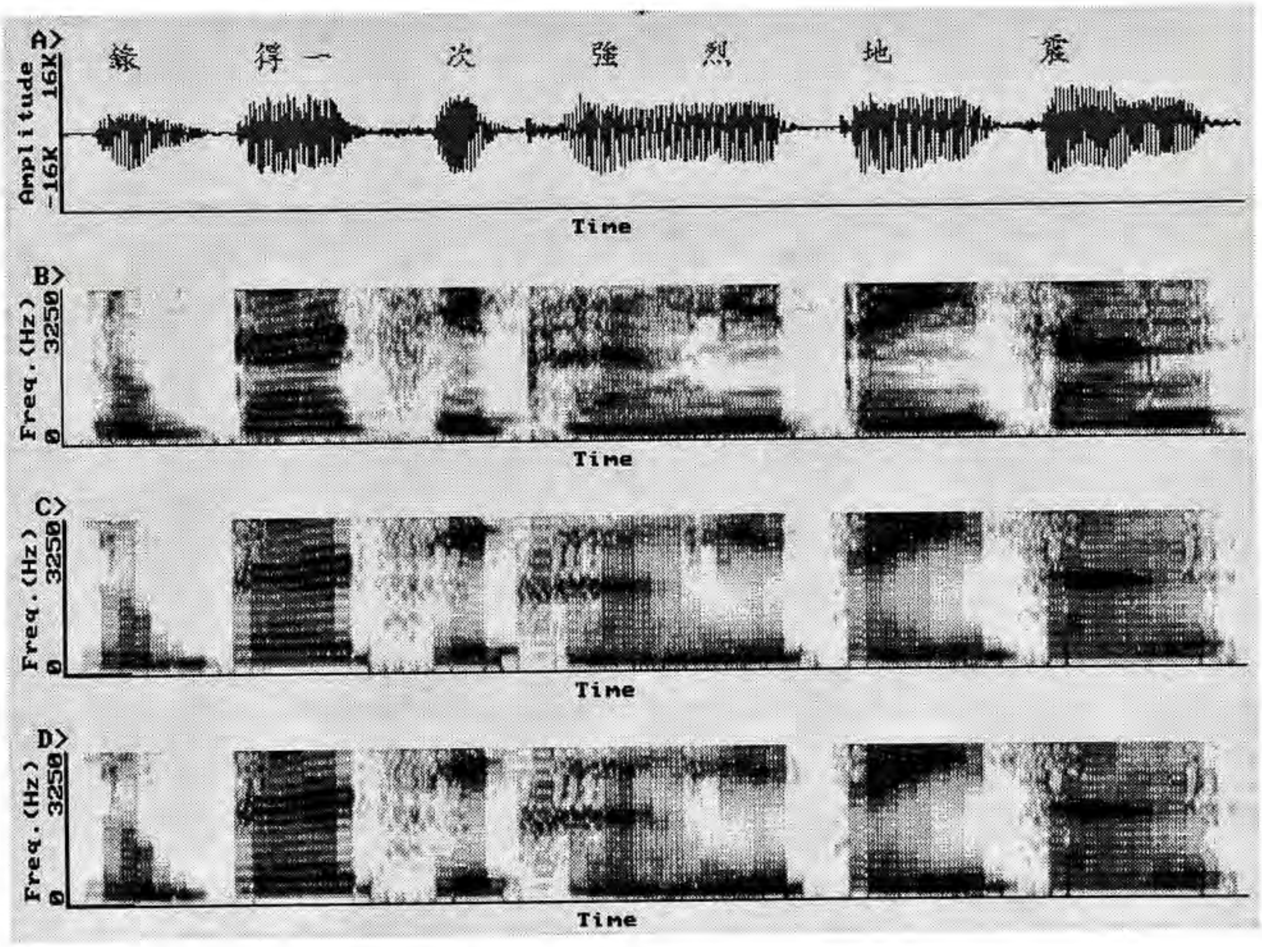


Figure 5.4 Spectrograms of synthesized speech using QBE
A) Input waveform (sample 4), B) Spectrogram of (A),
C) Spectrogram of output using uncoded parameters,
D) Spectrogram of output using coded parameters

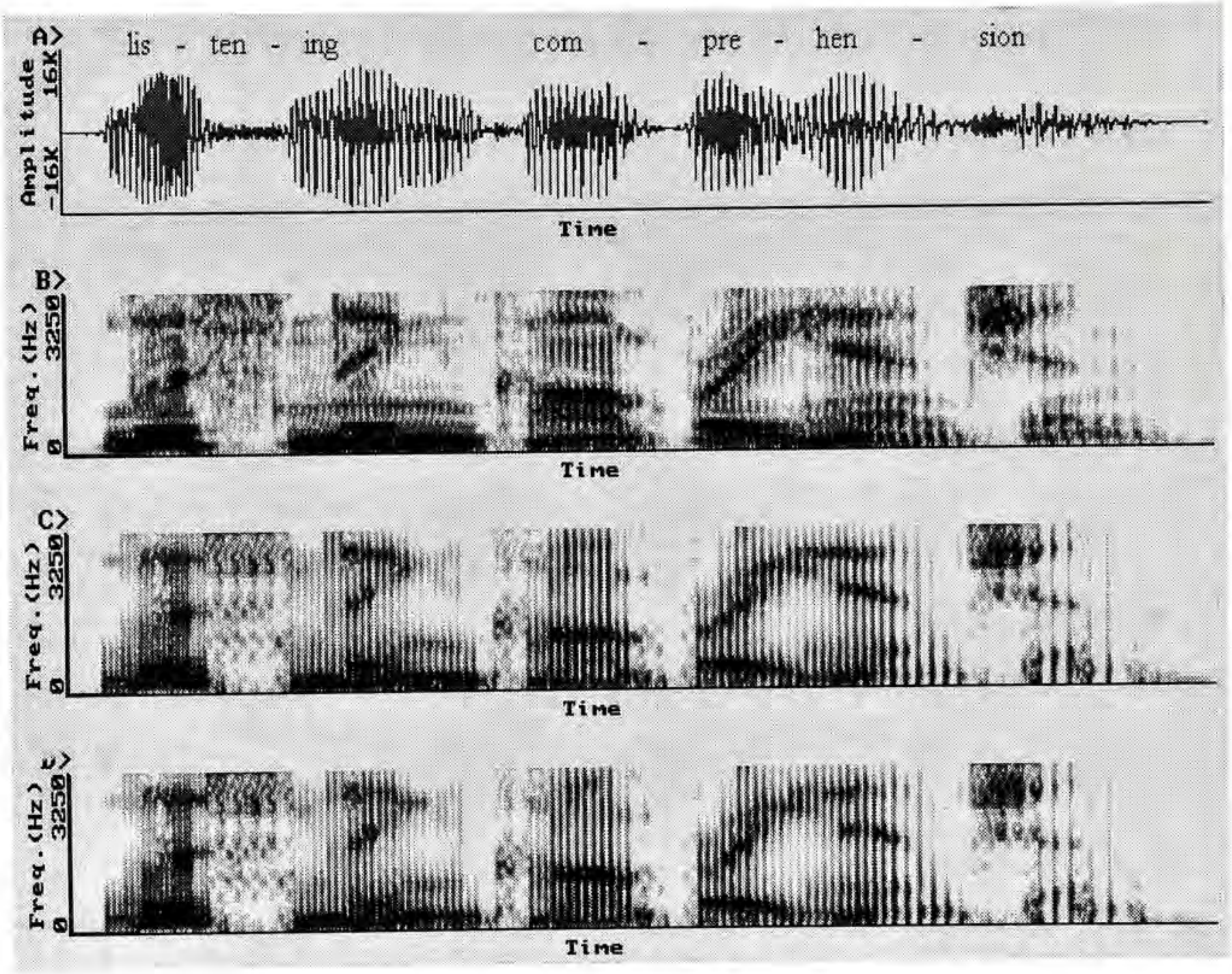


Figure 5.5 Spectrograms of synthesized speech using DBE
A) Input waveform (sample 1), B) Spectrogram of (A),
C) Spectrogram of output using uncoded parameters,
D) Spectrogram of output using coded parameters

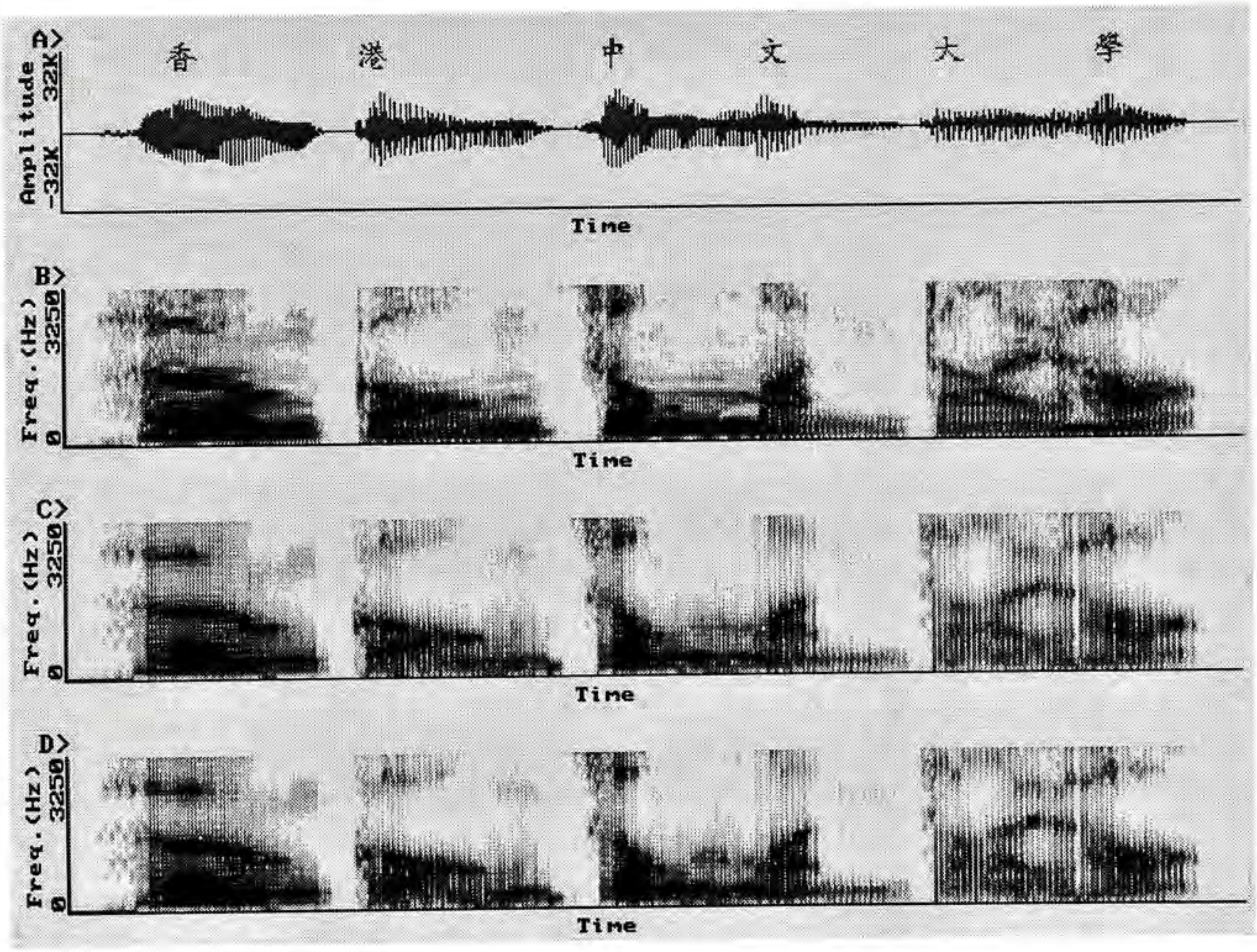


Figure 5.6 Spectrograms of synthesized speech using DBE
A) Input waveform (sample 2), B) Spectrogram of (A),
C) Spectrogram of output using uncoded parameters,
D) Spectrogram of output using coded parameters

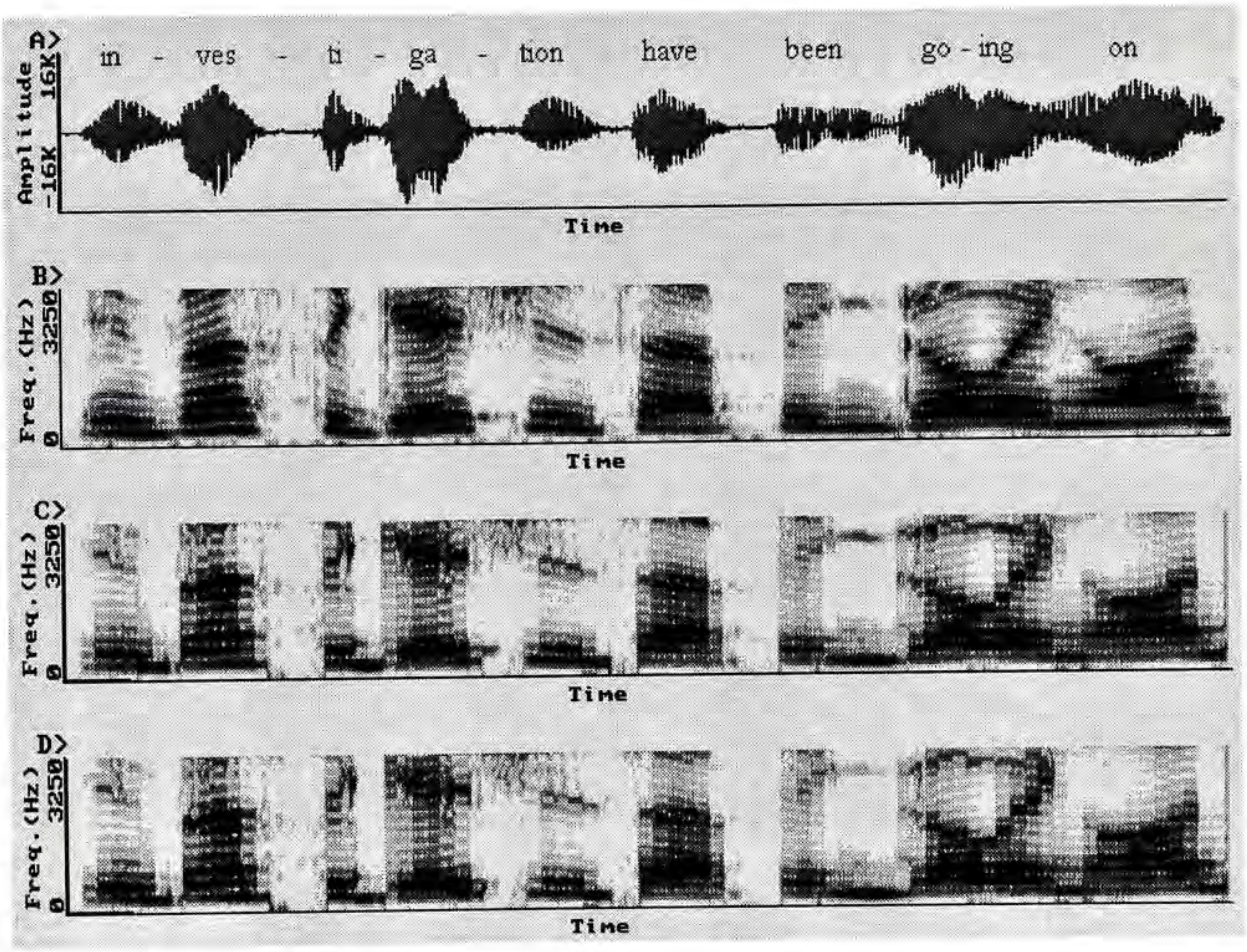


Figure 5.7 Spectrograms of synthesized speech using DBE
A) Input waveform (sample 3), B) Spectrogram of (A),
C) Spectrogram of output using uncoded parameters,
D) Spectrogram of output using coded parameters

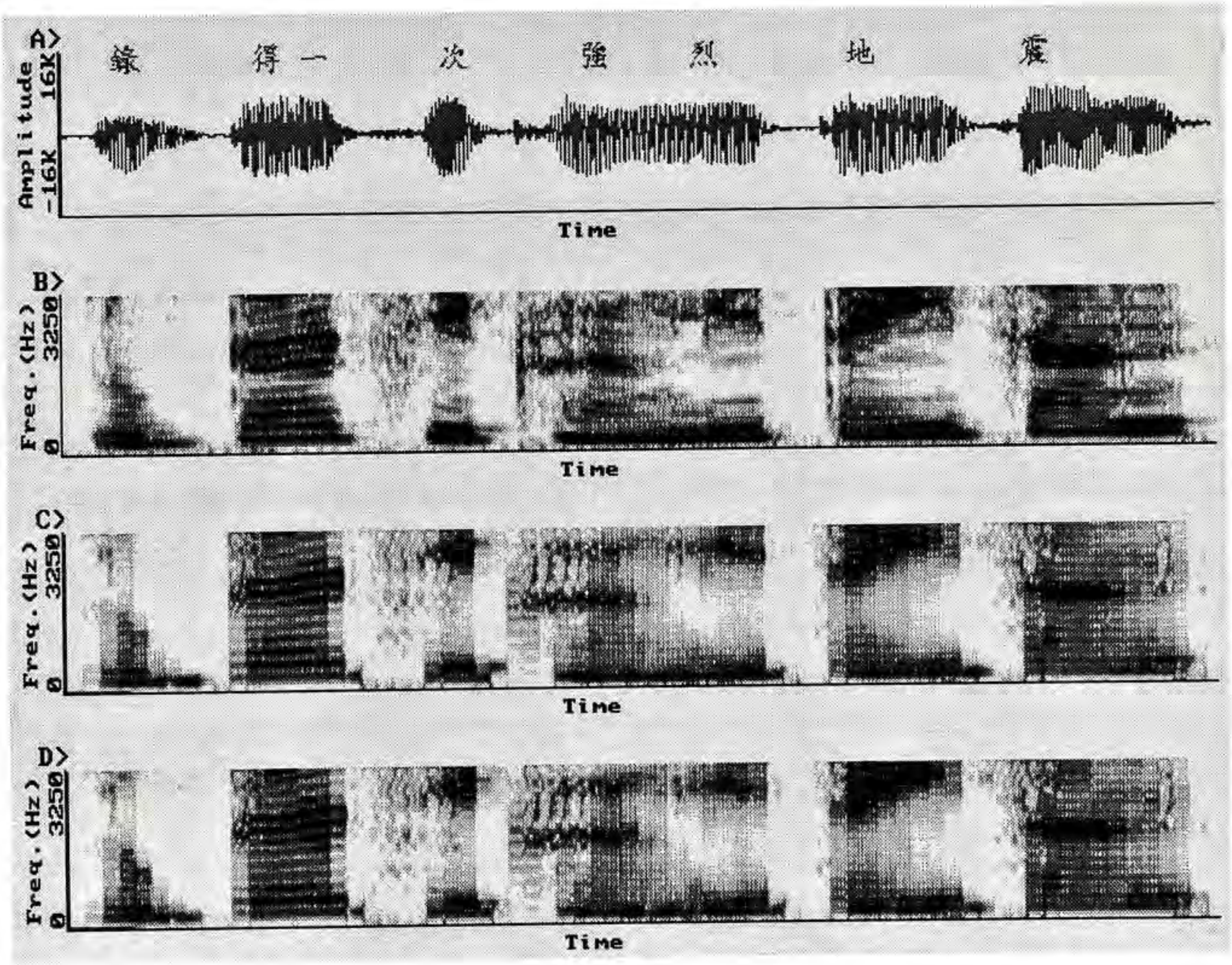


Figure 5.8 Spectrograms of synthesized speech using DBE
A) Input waveform (sample 4), B) Spectrogram of (A),
C) Spectrogram of output using uncoded parameters,
D) Spectrogram of output using coded parameters

Testing speech	**Average spectral distortion (dB) of synthesized speech of LSP vocoder using**		
	MBE	**DBE**	**QBE**
Cantonese spoken by a male speaker	7.5621	7.6043	7.6029
English spoken by a male speaker	7.9485	8.0222	8.0000
Cantonese spoken by a female speaker	7.6799	7.6888	7.6860
English spoken by a female speaker	7.7066	7.7239	7.7664

MBE = Multiband Excitation, DBE = Dual-Band Excitation,

QBE = Quad-Band Excitation

Table 5.1 Spectral distortion measure on synthesized speech using different types of excitations in the LSP vocoder.

5.2 Subjective listening test

Both DBE and QBE LSP vocoder can produce clear synthesized speech, even though the quality is slightly degraded. The degradation is due to the existence of some abnormal vibration on the voiced sounds, such that the smoothness of the output speech is slightly affected. This effect is more noticeable in female speech than in male speech. However, the output speech quality of the vocoder is a lot better when the speaker talks slowly. In addition, most synthesized speech was found to be highly intelligible. When noisy speech is input to the vocoder, the synthesized speech still maintains a reasonable quality, and the intelligibility of it can be kept at a high level. After comparing the output speech by using DBE and QBE with that using MBE, it is noted that the difference in quality is almost indistinguishable.

In order to evaluate the synthesized speech more subjectively, two listening tests were employed. The Mean Opinion Score (MOS) [46] was used to test the quality of the synthesized speech. In addition, the Diagnostic Rhyme Test (DRT) [47] was employed to evaluate the intelligibility of the output speech. In each of the tests, five listeners were participated and the scores were made subjectively. The tests results and discussions are given in the following sections.

5.2.1 Mean Opinion Score (MOS)

The Mean Opinion Score [46] is one of the methods for testing subjective speech quality. It is a kind of absolute category rating (ACR) [38] as no comparison is made between the listening samples and the scores are made depending on the quality of the most recently heard sample only. There are five markings in the test: excellent (5 marks), good (4 marks), fair (3 marks), poor (2 marks) and bad (1 mark). During the test, a total of 10 short sentences were used as inputs to the vocoder. These sentences includes both Cantonese (a dialect of Chinese) and English which are spoken by different speakers of both sexes. Synthesized speech of using 3 different excitations (MBE, DBE & QBE) were produced for each sentence and being tested. After the test, average scores were computed and recorded. The results of the MOS test are listed in Table 5.2 .

It was noted from the results that most synthesized speech had an average MOS score of higher than 3. In other words, most listeners realized that the quality of the speech were better than 'fair', although it is slightly degraded when compared with 'good' input speech. The output speech has a similar MOS score irrespective to whether it is generated by using unquantized or quantized parameters. This indicated that the quantization error of speech parameters only produced minor effect on the synthesized speech quality. In addition, it is found that all three different excitations could synthesize output speech with similar quality. Both DBE and QBE has comparable performance when compared with the conventional MBE model. In general, a slightly higher score can be obtained when QBE, instead of DBE, is used in the vocoder.

Testing speech		MOS result
Original speech		4.5
Synthesized speech using unquantized parameters	MBE	3.38
	DBE	3.19
	QBE	3.25
Synthesized speech using quantized parameters	MBE	3.38
	DBE	3.00
	QBE	3.12

MBE = Multiband Excitation, DBE = Dual-Band Excitation,

QBE = Quad-Band Excitation

Table 5.2 Results of mean opinion score

5.2.2 Diagnostic Rhyme Test (DRT)

The Diagnostic Rhyme Test [47] is an useful method for testing the intelligibility of speech signals. In this test, a set of synthesized single-syllable words were used to examine the response of the vocoder to the consonants. Many words may have the same intonation but the starting consonants can be different such as 'key', 'tea' and 'see'. The listeners have to determine the correct leading consonants of the synthesized words. In order to eliminate the factor of guessing, two possible answers were given to each synthesized word. During the test, two sets of English words were used as input to the vocoder. One set was clean recordings and the other set was a superposition of the clean recordings with a noisy background. The signal-to-noise ratio of the noisy words was kept at about 10 dB. The set of words that used in the test [3] are shown in Appendix E. The DRT score was evaluated by the following calculation

$$\text{DRT score} = \frac{\text{Number of correct answer}}{\text{Number of tested words}} \times 100\% \qquad (5.2)$$

If no mistake is found from the answers, a full mark of 100 will be given. Again, the test was performed on the synthesized words by using three different excitation models. At the end of the test, the DRT scores from different listeners were obtained and the average value of them was computed. The test results are listed in Table 5.3 .

Most testing results on different excitations gave a DRT score of higher than 75 out of 100. This implied that the synthesized words were highly intelligible. It was also noted that the intelligibility of the word generated by using unquantized parameters were found to be highly comparable to that of the original words. However, the intelligibility of the noisy words were slightly lower than that of the clean words, but the results were still acceptable. When comparing the performance of each excitation model in the LSP vocoder, it can be seen that both DBE and QBE had similar capability to synthesize intelligible speech. Moreover, QBE and MBE in fact had very close performance, which indicated that QBE can be used as an alternative to MBE in the LSP vocoder.

Conclusively, DBE and QBE LSP vocoders can produce clean synthesized speech with nearly 'good' quality, even though the smoothness of the speech is slightly affected mainly due to inadequate articulatory representation. Most frequency contents of the input can be retained on the synthesized speech, particularly the formant frequencies can be accurately located. The quantization errors from the parameter quantizers in the vocoder introduce only marginal effect on synthesized speech. In addition, the output speech were found to be highly intelligible. When compared with DBELSP vocoder, QBELSP vocoder offers a better overall performance in the listening tests.

Testing speech		**DRT score**	
		Clean	**Noisy**
Original words		97.5	90.9
Synthesized words using unquantized parameters	MBE	91.1	78.8
	DBE	89.4	75.8
	QBE	90.2	78.2
Synthesized words using quantized parameters	MBE	90.9	77.8
	DBE	82.6	68.8
	QBE	84.8	71.1

MBE = Multiband Excitation, DBE = Dual-Band Excitation,

QBE = Quad-Band Excitation

Table 5.3 Results of Diagnostic Rhyme Test

Chapter 6 Conclusions and Discussions

Multiband excitation model allows a combination of voiced and unvoiced frequency bands in a single speech frame which thus provides a flexibility in the representation of speech excitation spectra. Each harmonic frequency in a speech spectrum can now be examined individually and determine whether energy is contributed mainly by the harmonic frequency or it is given within a broad frequency range. In other words, a fine structure of excitation spectrum can be obtained by dividing the entire speech spectrum into many frequency bands for analysis, in which the number of frequency bands depends on the number of harmonic frequencies in the spectrum.

By examining a large number of speech segments, most spectra of ordinary conversational speech signals have strong harmonic frequencies in the low frequency region, and the high frequency regions in the spectra are rather noisy. Statistically, it has been shown that more than 70% of the speech frames have a characteristics that their spectra can be divided into maximally 2 frequency bands. Their speech spectra consist of either one single voiced band, or 2 frequency bands with different bandwidth. The low frequency band is usually voiced whereas the high frequency band is unvoiced. As a result, a Dual-Band Excitation (DBE) method is proposed in low bit rate speech coding, since the number of bits used to describe voiced/unvoiced decisions is greatly reduced when compared with the MBE model. In most practical MBE vocoders, a total of 12 bits are needed to represent the voiced/unvoiced decisions in each frame for transmission, but now only one-third of them are used when DBE is applied. In order to obtain a finer spectral structure of excitation signal, the method is modified such that no more than four frequency bands are allowed in each speech frame. From the statistical results, it can be seen that if a maximum of four frequency bands with variable bandwidth are used, more than 90% of the excitation spectra can be well expressed. This kind of excitation, named as Quad-

Band Excitation (QBE), requires 8 bits for encoding the voiced/unvoiced decision patterns. It is interesting to note that only four extra bits per frame are needed to get a better spectral intimation of the excitation signal when compared with DBE.

As an application of DBE and QBE in low bit rate speech coding, vocoders of using these two kinds of excitations have been developed and evaluated. The proposed vocoders are based on linear prediction method of speech signals. In order to encode the spectral envelope more effectively, the LPC coefficients are converted into line spectral pairs (LSP) frequencies before quantization. It is because the LSP frequencies, in particular the intraframe frequency differences, have a narrow dynamic range so that the number of bits used for quantization can be reduced. The performance of a number of quantization methods for encoding the spectral information have been evaluated, including linear & non-linear quantization of the intraframe LSP frequency differences, and vector quantization of LSP frequencies. The performance of non-linear quantization of the intraframe LSP frequency differences is found to be better than the method of linear quantization. It is also noted that 30 bits are sufficient to encode the frequency differences with an acceptable output quality through informal listening tests. In addition, split vector quantization on LSP frequencies can be used to provide similar performance, and the number of bits required can be further reduced even though the computational complexity would be increased. Simulation results showed that splitting the LSP frequency vector into two sub-vectors have a lower spectral distortion after quantization than that from the quantization on the entire LSP feature vector at a given bit rate. In addition, split-2 vector quantization has a lower quantization distortion when compared with using split-3 vector quantization. Furthermore, splitting of the LSP feature vector into two sub-vectors with equal vector size produces the best performance. After compromising the trade off between the output speech quality and the computational complexity, 24-bit split-2 vector quantization is decided to be used to encode the LSP frequencies, and 12 bits are used on each codebook. As a result, the operating bit rate of the DBELSP vocoder is 2 kbps and that of the QBELSP vocoder is 2.2 kbps.

The performance of the DBE and QBE LSP vocoder have been studied thoroughly. Both vocoders can produce synthesized speech that most of the frequency contents, in particular the formant frequencies, are successfully maintained. In addition, the synthesized speech is found to be highly intelligible, with a nearly 'good' output quality. The performances of both vocoders are similar, but the listening test results show that QBE is a little more superior to DBE. From the statistical results of the voiced/unvoiced distribution, QBE has an ability to represent almost all excitation spectra by occupying only 4 extra bits per frame compared with DBE. As a consequence, a QBELSP vocoder is finalized for low bit rate transmission.

Further developments in this context can be carried out in order to improve the synthesized speech quality of the vocoder. Firstly, since the pitch value is a very important parameter in QBE to determine the voiced/unvoiced pattern, a more accurate pitch detection algorithm should be applied. The existing pitch detection algorithm in the vocoder has a resolution of half a sampling period, and a pitch value with less quantization error would be more preferable. Secondly, the algorithm of generating spectrum of excitation signal in speech synthesis also limits the performance of the vocoder. In this method, impulses with equal magnitude are inserted to the locations of the voiced harmonic frequencies to form voiced excitation bands. However, since the generated spectrum must be a discrete frequency spectrum, the location of the harmonic frequencies will be quantized. It leads to the degradation of the synthesized speech. The current solution is using discrete frequency spectrum with more number of points to reduce the frequency quantization error, but it requires more computation power when the excitation signal is retrieved by applying inverse Fast Fourier Transform. Thirdly, the computation complexity of the vocoder have to be considered so that it is more feasible to be implemented in real time applications. One of the simplification in computation can be done is for the voiced/unvoiced decision. Currently, spectrum comparison is made between the original speech spectrum and a voiced synthetic spectrum. This synthetic spectrum needs to be generated in each frame as the fundamental frequency is different. This can be simplified by using an unvoiced spectrum as reference, such that it is fixed for all frames. Accordingly, the empirical threshold value used for the decision have to be redefined. Besides, the full searching algorithm of finding optimal codevectors in split-2 vector quantization of LSP frequencies is also time consuming, and a simpler searching algorithm is preferred.

Furthermore, hardware implementation of the vocoder can also be carried out in the future work. The vocoder can be implemented by using digital signal processor chips to further investigate the time delay of the vocoder and the robustness when it is used in noisy channels.

In conclusion, there are two major contributions in this thesis to the design of a low bit rate vocoder. The first contribution is the development of DBE and QBE as an alternative to MBE model. It not only retains the flexibility in the representation of the excitation spectrum as in MBE model, but also reduces the system complexity of the vocoder particularly in the process of speech synthesis. The second contribution is that a QBELSP vocoder is implemented that can be operated at a bit rate as low as 2.2 kb/s with acceptable output quality. Since the proposed vocoder is of a first kind, its performance can be used as a baseline for later development.

References

1. A. Nejat Ince, *Digital Speech Processing: Speech Coding, Synthesis and Recognition*, Massachusetts, Kluwer Academic Publishers, 1992

2. B.S. Atal, V. Cuperman and A. Gersho, *Advances In Speech Coding*, Massachusetts, Kluwer Academic Publishers, 1991

3. J.R. Deller Jr., J.G. Proakis and F.H.L. Hansen, *Discrete-Time Processing Of Speech Signals*, New York, Macmillan, 1993

4. N.S. Jayant, *Waveform Quantization and Coding*, IEEE Press, 1976

5. N.S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM and DM Quantizers", *Proc. IEEE*, vol. 62, pp. 611-632, May 1974

6. P. Cummiskey, N.S. Jayant and J.L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech", *Bell System Tech. J.*, vol. 32, no. 7, pp. 1105-1118, September 1973.

7. H.R. Schindler, "Delta Modulation," *IEEE Spectrum*, vol. 7, pp. 69-78, October 1970

8. M.R. Schroeder, "Vocoders: Analysis and Synthesis of Speech", *Proc. IEEE*, vol. 54, pp. 720-734, May 1966

9. J. Makhoul, "Linear Prediction: A Tutorial Review", *IEEE Proceedings*, vol. 63, pp. 561-580, April 1975.

10. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ; Prentice Hall, 1978

11. L.D. Markel and A.H. Gray Jr., *Linear Prediction of Speech* , New York, Springer Verlag, 1976

12. B.S. Atal and S.L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. Soc. Am.*, vol. 50, pp. 637-655, 1971

13. J. Marhoul, "Spectral Linear Prediction: Properties and Applications", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 23, no. 3, pp. 283-296, June 1975

14. M.R. Schroeder and B.S. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech At Very Low Bit Rates", *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 937-940 , March 1985

15. I.A. Gerson and M.A. Jasiuk, "Vector Sum Excited Linear Prediction Speech Coding at 8 kbps", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, vol. 1, pp. 461-464, April 1990

16. S. Sharad and B.S. Atal, "Amplitude Optimization and Pitch Prediction in Multipulse coders", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, no. 3, pp. 317-327, March 1989

17. D.W. Griffin and J.S. Lim, "Multiband Excitation Vocoder", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223-1235, August 1988

18. J.C. Hardwick and J.S. Lim, "A 4.8KBPS Multiband Excitation Speech Coder", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, vol.-1, pp. 374-377, April 1988

19. K.M. Chiu and P.C. Ching, "A Dual-Band Excitation LSP Codec For Very Low Bit Rate Transmission", *Int. Symposium on Speech, Image Processing and Neural Networks*, vol. 2., pp. 479-482, April, 1994

20. L.D. Markel and A.H. Gray Jr., "A Linear Prediction Vocoder Simulation Based Upon the Autocorrelation Method," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 22, no. 2, pp. 124-134, April 1974

21. M.R. Sambur, "An Efficient Linear Prediction Vocoder," *Bell Syst. Tech. J.*, vol. 54, no. 10, pp. 1693-1723, December 1975

22. L.R. Rabiner, M.J. Cheng, A.E. Rosenberg and C.A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 399-418, October 1976

23. B. Gold and L.R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," *J. Acoust. Soc. Am.*, vol. 46, no. 2, pp. 442-448, August 1969

24. L.R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 24-33, February 1977

25. J.J. Dubnowski, R.W. Schafer and L.R. Rabiner, "Real-Time Digital Hardware Pitch Detector", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 1, pp. 2-8, February 1976

26. M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg and H.J. Manley, "Average Magnitude Difference Function Pitch Extractor", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 22, pp. 353-362, October 1974

27. R.C. Rose and T.P. Barnwell, "Design and Performance Of An Analysis-By-Synthesis Class of Predictive Speech Coders", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 38, no. 9, pp. 1489-1503, September 1990

28. P. Kroon, E.F. Deprettere and R.J. Sluyter, "Regular-Pulse Excitation - A Novel Approach To Effective And Efficient Multipulse Coding of Speech", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, no. 5, pp. 1054-1063, October 1986

29. B.S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Communications*, vol. 30, pp. 600-614, April 1982

30. J.C. Hardwick, "A 4.8Kbps Multi-Band Excitation Speech Coder", MPhil. thesis, M.I.T., Cambridge, MA. 1988

31. F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals", *J. Acoust. Soc. Am.*, vol. 57, pp. 535(A), 1975

32. N. Sugamura and F. Itakura, "Speech Analysis and Synthesis Methods Developed At ECL In NTT: From LPC To LSP", *Speech Communication*, vol. 5, pp. 199-215, 1986

33. F.K. Soong and B.H. Juang, "Line Spectral Pair (LSP) And Speech Data Compression", *Int. Conf. on Acoustics, Speech and Signal Proc.*, vol. 1, pp. 1.10.1-1.10.4, 1984

34. K.K. Paliwal and B.S. Atal, "Efficient Vector Quantization Of LPC Parameters At 24 Bits/Frame", *Int. Conf. on Acoustics, Speech and Signal Proc.*, vol. 1, pp. 661-664, May 1991.

35. C.F. Chan, "Computation Of LSP Parameters From Reflection Coefficients", *Electronic letters*, vol. 27, no. 19, pp. 1773-1774, September 1991.

36. S. Saoudi, J.M. Boucher and A.L. Guyader, "A New Efficient Algorithm To Compute The LSP Parameters For Speech Coding", *Signal Processing*, vol. 28, pp. 201-212 1992

37. P. Kabal and P. Ramachandran, "The Computation Of Line Spectral Frequencies Using Chebyshev Polynomials", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, no. 6, pp. 1419-1426, December 1986

38. B.S. Atal, V. Cuperman and A. Gersho, *Speech and Audio Coding For Wireless And Network Applications*, Massachusetts, Kluwer Academic Publishers, 1993

39. F.K. Soong, "Optimal Quantization of LSP Parameters", *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp. 15-24, January 1993

40. C.C. Kuo, F.R. Jean and H.C. Wang, "Low Bit Rate Quantization Of LSP Parameters Using Two-Dimensional Differential Coding", *Int. Conf. on Acoustics, Speech and Signal Proc.*, vol. 1, pp. 97-100, March 1992

41. N. Sugamura and N. Farvardin, "Quantizer Design in LSP Speech Analysis-Synthesis", *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 432-440, February 1988

42. J. Makhoul and H. Gish, "Vector Quantization in Speech Coding", *IEEE Proceedings*, vol. 73, no. 11, pp. 1551-1587, November 1985

43. A. Buzo, A.H. Gray Jr., R.M. Gray, and J.D. Markel, "Speech Coding Bases Upon Vector Quantization", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 5, pp. 562-574, October 1980

44. J.S. Collura and T.E. Tremain, "Vector Quantizer Design For The Coding Of LSF Parameters", *Int. Conf. on Acoustics, Speech and Signal Proc.*, vol. 2, pp. 29-32, April 1993

45. Y. Linde, A. Buzo and R.M. Gray, "An Algorithm for Vector Quantizer Design", *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84-95, January 1980

46. J.R. Rosenberger, "Quality Assessment for Speech Coding", *Telecommunications Journal*, vol. 55, no. 12, pp. 820-825, 1988

47. W.D. Voiers, "Evaluating Processed Speech Using the Diagnostic Rhyme Test", *Speech Technology*, pp. 30-39, January/February 1983

Appendix A   Subroutine of pitch detection

```
float   Pitch_detection (short int *data, frame_index) {

        int     frame_index,
                delay,
                start_pt,
                i,
                m;

        float   interpolated[561],
                clipped[561],
                correlation[561],
                high_level,
                low_level,
                mean,
                max,
                MAX,
                Pitch;


        /* ** linear interpolation ** */

        for (i = 0; i < 281; i++)
                interpolated[2 * i] = data[frame_index + i];

        for (i = 1; i < 561; i += 2)
                interpolated[i] = (interpolated[i - 1] + interpolated[i + 1]) / (float) 2.0;


        /* ** make it zero-mean ** */

        mean = 0;
        for (i = 0; i < 561; i++)
                mean += interpolated[i];
        mean /= (float) 561;
        for (i = 0; i < 561; i++)
                interpolated[i] = interpolated[i] - mean;


        /* ** finding local maximum ** */

        max = interpolated[0];
        for (i = 0; i < 561; i++) {
                if (interpolated[i] > max)
                max = interpolated[i];
}


/* ** center clipping ** */

high_level = floor ((double) max * (double) 0.3);
low_level = high_level * -(double) 1.0;
for (i = 0; i < 561; i++) {
        if ((float) interpolated[i] > high_level)
                clipped[i] = interpolated[i] - high_level;
        else
        if ((float) interpolated[i] < low_level)
                clipped[i] = interpolated[i] - low_level;
        else
                clipped[i] = 0;
}


/* ** calculating normalized correlation function ** */

correlation[0] = 0;
for (m = 0; m <= 560; m++)
        correlation[0] += clipped[m] * clipped[m];

if (correlation[0] != 0) {

        for (delay = 1; delay < 341; delay++) {
                correlation[delay] = 0;
                for (m = 0; m <= 560 - delay; m++)
                        correlation[delay] += clipped[m] * clipped[m + delay];
                correlation[delay] /= correlation[0];
                if (correlation[delay] < 0.25)
                        correlation[delay] = 0;
        }

        correlation[0] = (float) 1.0;


 /* ** pitch searching ** */

        start_pt = 0;
        do {
                start_pt++;
        } while (correlation[start_pt] > 0 && start_pt != 340);

        if (start_pt != 340) {
                Pitch = 0;
                MAX = 0;
```

```
            for (delay = start_pt; delay < 341; delay++) {
                if (correlation[delay] > MAX) {
                    MAX = correlation[delay];
                    Pitch = delay;
                }
            }
        }
        else
            Pitch = 0;

        Pitch /= (float) 2.0;

        if (Pitch < 23 || Pitch > 150)
            Pitch = 0;

    }
    else
        Pitch = 0;

    return (Pitch);

}
```

Appendix B Subroutine of voiced/unvoiced decision

```
void    v_uv_decision (float Pitch, float number_of_band,
                       float number_of_harmonics, COMPLEX *dft_data) {

        extern int     vuv[13];

        int            k,
                       i,
                       m,
                       number_boundary;


        float          fundamental,
                       twpi;
                       a,
                       b,
                       numerator,
                       denominator,
                       dummy1,
                       dummy2,
                       error,
                       theta;

        COMPLEX        syn_spec;

        void           synthetic_spectrum(int, float, COMPLEX);

        float          threshold(int, int);



        twopi = 2*3.1426;
        fundamental = twopi/Pitch;

        /* ** v/uv determination on each band ** */

        for (k = number_of_band; k >= 1; k--) {

                a = (3*k-2.5)*fundamental*512/twopi;
                if (k != number_band)
                        b = (3*k+0.5)*fundamental*512/twopi;
                else
                        b = (number_of_harmonic+0.5)*fundamental*512/twopi;
                numerator = 0;
        denominator = 0;
        for (m=ceil((double)a); m <= (ceil((double)b)-1); m++) {

/* ** calculate synthetic spectrum ** */

                synthetic_spectrum (m, fundamental, syn_spec);


                dummy1 = dft_data[m].real*dft_data[m].real +
                                dft_data[m].imag*dft_data[m].imag;
                dummy1 = sqrt((double)dummy1);

                dummy2 = syn_spec.real*syn_spec.real +
                                syn_spec.imag*syn_spec.imag;
                dummy2 = sqrt((double)dummy2);

                numerator += ((dummy1 - dummy2)*(dummy1 - dummy2));

                denominator += dft_data[m].real*dft_data[m].real +
                                dft_data[m].imag*dft_data[m].imag;

        }
        error = numerator / denominator;


/* ** calculating threshold value ** */

        theta = threshold(ceil((double)a),ceil((double)b)-1);


/* ** make v/uv decision ** */

        if (error < (double) theta)
                vuv[k] = 1;     /* ** 1 = voiced ** */
        else
                vuv[k] = 0;     /* ** 0 = unvoiced ** */

 }


 /* ** v/uv band pattern grouping ** */

 if (vuv[1] == 1) {

        k = 1;
        number_of_boundary = 0;

        do {
                k++;
```

```
                        if (vuv[k] != vuv[k-1])
                                number_of_boundary++;
                } while ( k < number_of_band && number_of_boundary < 3);

                if (number_boundary == 3) {
                        if (k != number_band) {
                                i = vuv[k];
                        for (;k<=number_band;k++)
                                vuv[k] = i;
                        }
                }

        }
        else {
                for (k=1; k<= number_band; k++)
                        vuv[k] = 0;
        }

}


void    synthetic_spectrum (int x, float y, COMPLEX *output)
{
        int                     i,
                                L,
                                k;

        float                   a,
                                b,
                                twopi,
                                out,
                                AL,
                                inter,
                                upper,
                                lower;

        COMPLEX                 p, q;

        static int              pre_L = 0;

        static float            pre_AL;

        extern COMPLEX          dft_window[16384];


        twopi = 2*3.1416;
        L = -1;
        do {
                L++;
        b = 512 * ((float) L + 0.5) * y / (twopi);
} while ((double) x >= ceil ((double) b));
a = 512 * ((float) L - 0.5) * y / (twopi);

if (L != pre_L) {
        upper = 0;
        lower = 0;
        for (i=ceil((double)a); (double) i<=(ceil((double)b)-1); i++) {
                p.real = dftrdata[i].real;
                p.imag = dftrdata[i].imag;

                k = (int)(32.0*(float)i - 16384.0*(float)l*(float)y/twopi + 0.5);
                if ( k >= 0) {
                        q.real = dft_window[k].real;
                        q.imag = dft_window[k].imag;
                }
                else {
                        q.real = dft_window[k+16384].real;
                        q.imag = dft_window[k+16384].imag;
                }

                inter = sqrt((double)(p.real*p.real+p.imag*p.imag));
                upper += inter*sqrt((double)(q.real*q.real+q.imag*q.imag));
                lower += q.real * q.real + q.imag * q.imag;
        }
        AL = upper / lower;
}
else
        AL = pre_AL;

if (l != 0)
        k = (int)(32.0*(float)x - 16384.0*(float)l*(float)y/twopi + 0.5);
else
        k = 32*x;

if ( k >=0 ) {
        q.real = dft_window[k].real;
        q.imag = dft_window[k].imag;
}
else {
        q.real = dft_window[k+16384].real;
        q.imag = dft_window[k+16384].imag;
}

output.real = AL * q.real;
output.imag = AL * q.imag;


pre_l = l;
```

```
        pre_AL = AL;

}


float   threshold(double x_1, double x_2)
{
        float   y_1,
                y_2,
                out;


        y_1 = -(float)0.001172*(float)x_1 + (float)0.7;
        y_2 = -(float)0.001172*(float)x_2 + (float)0.7;

        out = (y_1 + y_2)/(float)2.0;
        return(out);
}
```

Appendix C Subroutine of LPC coefficients calculation using Durbin's recursive method

```
void    Durbins_method (float *windowed_data) {

        extern float    LPC[56];

        int             i,
                        j,
                        k,
                        m,
                        index1,
                        index2;

        float           correlation[],
                        E[11],
                        K[11],
                        dummy;


        /* ** calculate correlation function ** */

        for (k = 0; k <= 10; k++) {
                correlation[k] = 0;
                for (m = 0; m <= 280 - k; m++)
                        correlation[k] += (windowed_data[m] * windowed_data[m + k]);
        }


        /* ** normalize correlation function ** */

        dummy = correlation[0];
        for (k=0; k<=10; k++)
                correlation[k] = correlation[k] / dummy;


        /* ** calculate LPC coefficients ** */

        E[0] = correlation[0];
        for (i = 1; i <= 10; i++) {
                if (i != 1) {
                        if (i == 2)
                index1 = 0;
            else
                index1 += (i-2);
            dummy = 0;
            for (j = 1; j <= i-1; j++)
                dummy += (LPC[index1 + j] * correlation[i - j]);
        }
        else
            dummy = 0;

        K[i] = (correlation[i] - dummy) / E[i-1];

        if (i == 1)
            index2 = 0;
        else
            index2 += i - 1;
        LPC[index2 + i] = K[i];

        if (i != 1) {
            for (j = 1; j <= i-1; j++)
                LPC[index2 + j] = LPC[index1 + j] -
                                  (K[i]*LPC[index1 + i - j]);
        }
        E[i] = ((float)1.0 - (K[i] * K[i])) * E[i - 1];
    }

}
```

Appendix D Subroutine of LSP calculation using Chebyshev Polynomials

```
void    vq_cheby(float *lpc)
{

        extern double  lsp_freq[11];

        int            i,
                       j,
                       index;

        double         sym[12],
                       assym[12],
                       G_sym[6],
                       G_assym[6],
                       limit,
                       n,
                       x,
                       y1,
                       y2,
                       lsp_freq[11],
                       alpha[11],
                       sym_coef[6],
                       assym_coef[6],
                       interval,
                       increment,
                       twopi;

         double        cheby(double, int, double*, double*),
                       cal_b(int, double, double*);



         interval = (double)0.001;
         increment = (double)0.000125;
         twopi = ((double)(2.0*3.141592654));


         for (i=1;i<=10;i++)
                 alpha[i] = -(double)1.0*(double)lpc[i];
```

```
/* ** determine the coefficients of the symmetric &assymetric polynomial ** */

j = 10;
for (i = 1; i <= 5; i++) {
        sym[i] = alpha[i] + alpha[j];
        sym[j] = sym[i];
        assym[i] = alpha[i] - alpha[j];
        assym[j] = -(double)1.0*assym[i];
        j--;
}
sym[0] = (double)1.0;
sym[11] = sym[0];
assym[0] = (double)1.0;
assym[11] = -(double)1.0*assym[0];

G_sym[0] = (double)1.0;
G_assym[0] = (double)1.0;
for (i = 1; i <= 5; i++) {
        G_sym[i] = sym[i] - G_sym[i - 1];
        G_assym[i] = assym[i] + G_assym[i - 1];
}

j = 0;
for (i = 5; i >= 1; i--) {
        sym_coef[i] = (double)2.0*G_sym[j];
        assym_coef[i] = (double)2.0*G_assym[j];
        j++;
}
sym_coef[0] = G_sym[5];
assym_coef[0] = G_assym[5];


/* ** find LSP by searching the roots of the polynomial ** */

index = 1;
for (n = (double)1.0; n > -(double)1.0; n -= interval) {
        y1 = cheby(n, index, sym_coef, assym_coef);
        y2 = cheby(n-interval, index, sym_coef, assym_coef);

        if ( (y1>0 && y2<0) || (y1<0 && y2>0) ) {
                limit = n - interval;
                for (x = n; x > limit; x -= increment) {
                        y1 = cheby(x, index, sym_coef, assym_coef);
                        y2 = cheby((x-increment), index, sym_coef, assym_coef);

                if ( (y1>0 && y2<0) || (y1<0 && y2>0) ) {
                        lsp_freq[index] = (double)2.0*x - increment;
                        lsp_freq[index] /= (double)2.0;
                        lsp_freq[index] = acos((double)lsp_freq[index]);
```

```
                                index++;
                                x = limit - (double)1.0;
                        }
                }
        }

        if (index > 10)
                n = -(double)2.0;

        }

}


double cheby(double x, int index, double *sym_coef, double *assym_coef) {

        double          out,
                        b0,
                        b2;


        if (index % 2 != 0) {
                b0 = cal_b(0, x, sym_coef);
                b2 = cal_b(2, x, sym_coef);
                out = (b0 - b2 + sym_coef[0])/(double)2.0;
        }
        else {
                b0 = cal_b(0, x, assym_coef);
                b2 = cal_b(2, x, assym_coef);
                out = (b0 - b2 + assym_coef[0]) / (double)2.0;
        }

        return(out);
}


double cal_b(int k, double x, double *coef) {

        double          out;

        if (k <= 5)
                out = (double)2.0*x*cal_b((k+1), x, coef) - cal_b((k+2), x, coef) +
                        coef[k];
        else
                out = 0;

        return(out);
}
```

Appendix E Single syllable word pairs for Diagnostic Rhyme Test

vee	bee	daunt	taunt	dip	nip
jest	guest	gnaw	daw	thick	tick
mad	bad	shaw	chaw	neck	deck
jab	gab	bong	dong	pence	fence
than	dan	got	dot	shad	chad
sing	thing	thong	tong	sank	thank
gill	dill	sole	thole	foo	pooh
chair	care	meat	beat	goose	juice
yen	wren	weed	reed	moon	noon
gaff	calf	yield	wield	poop	coop
nab	dab	gin	chin	bowl	dole
shag	sag	mitt	bit	got	jot
moot	boot	jilt	gilt	fop	hop
moan	bone	bid	did	news	dues
joe	go	hit	fip	choose	shoes
ghost	boast	mend	bend	pool	tool
moss	boss	met	net	coat	goat
jaws	gauze	keg	peg	note	dote
yawl	wall	bank	dank	show	so
jock	chock	bat	gat	taught	caught
mom	bomb	bean	peen	pond	bond
wad	rod	need	deed	knock	dock
von	bon	sheet	cheet	pot	tot
dune	tune	cheep	keep	pent	tent
chew	choo	peak	teak		
you	rue	key	tea		