NOVEL PITCH DETECTION ALGORITHM WITH APPLICATION TO SPEECH CODING

A Thesis

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of masters

Master of Science
in
The Department of Electrical Engineering

by

Vijay B. Kura

B.Tech., Jawaharlal Institute of Technological University, 2000

December 2003

TABLE OF CONTENTS

Acknowledgements

First of all I want to express my gratitude to my advisor Dr. Dimitri Charalampidis for his invaluable guidance during the entire period of this work. I'm very obliged for his suggestions, ideas and concepts without which this work wouldn't have been what it is now.

I also would like to thank my committee members Dr. Vesselin Jilkov, Dr. Jing Ma and Dr. Terry Remier for their suggestions and insightful comments

Finally I would like to thank my parents and my family for their continuous and unconditional love and support**.**

List of Illustrations

List of figures:

List of Tables:

Glossary of Abbreviations

PCS	Personal Communication Systems
VOIP	Voice Over Internet Protocol
DSVD	Digital Simultaneous Voice and Data
NB	Narrowband
WB	Wideband
PSTN	Public Switched Telephone Networks
ASIC	Application Specific Integrated Circuits
FEC	Forward Error Correction
MOS	Mean Square Score
PCM	Pulse Code Modulation
ADPCM	Adaptive Pulse Code Modulation
DAM	Diagnostic Acceptability Measure
DRT	Diagnostic Rhyme Test
AbS	Analysis by Synthesis
STP	Short-time Predictor
LTP	Long-time Predictor
CELP	Code Excited Linear Predictive coding
RPELPC	Regular Pulse Excitation LPC
LPC	Linear Predictive Coder

RELP	Residual Excitation Coding
RPE	Regular Pulse Excitation coding
SELP	Self Excitation Coding
MELP	Mixed Excitation Linear Predictive Coding
MBE	Multi-Band Excitation Coder
SNR	Signal-to-Noise ration
MSE	Mean Square Error
ARMA	Auto Regressive Moving Average
ACR	Auto-Correlation
AMDF	Average Magnitude Difference Function
PDA	Pitch Detection Algorithms
GCI	Glottal Closure Instant
MLE	Maximum Likelihood Estimation
MAWT	Multi-feature, Autocorrelation (ACR) and Wavelet Technique
STE	Short-time Energy
ZCR	Zero-Crossing Rate

ABSTRACT

This thesis introduces a novel method for accurate pitch detection and speech segmentation, named Multi-feature, Autocorrelation (ACR) and Wavelet Technique (MAWT). MAWT uses feature extraction, and ACR applied on Linear Predictive Coding (LPC) residuals, with a wavelet-based refinement step. MAWT opens the way for a unique approach to modeling: although speech is divided into segments, the success of voicing decisions is not crucial. Experiments demonstrate the superiority of MAWT in pitch period detection accuracy over existing methods, and illustrate its advantages for speech segmentation. These advantages are more pronounced for gain-varying and transitional speech, and under noisy conditions.

Chapter 1: Introduction

1.1 Motivation for Speech Coding

Speech communication is arguably the single most important interface between humans, and it is now becoming an increasingly important interface between human and machine. As such, speech represents a central component of digital communication and constitutes a major driver of telecommunications technology. With the increasing demand for telecommunication services (e.g., long distance, digital cellular, mobile satellite, aeronautical services), speech coding has become a fundamental element of digital communications. Emerging applications in rapidly developing digital telecommunication networks require low bit, reliable, high quality speech coders. The need to save bandwidth in both wireless and line networks, and the need to conserve memory in voice storage systems are two of the many reasons for the very high activity in speech coding research and development. New commercial applications of low-rate speech coders include wireless Personal Communication Systems (PCS) and voice-related computer applications (e.g., message storage, speech and audio over internet, interactive multimedia terminals). In recent years, speech coding has been facilitated by rapid advancement in digital signal processing and in the capabilities of digital signal processors. A strong incentive for research in speech coding is provided by a shift of the relative costs involved in handling voice communication in telecommunication systems. On the one hand, there is an increased demand

for larger capacity of the telecommunication networks. Nevertheless, the rapid advancement in the efficiency of digital signal processors and digital signal processing techniques has stimulated the development of speech coding algorithms. These trends are likely to continue, and speech compression most certainly will remain an area of central importance as a key element in reducing the cost of operation of voice communication systems.

1.2 Importance of Pitch Estimation in Speech Coding

The motivation for speech coding is to reduce the cost of operation of voice communication that involves development of various efficient coding algorithms and relating areas. One of the important areas of speech coding is pitch estimation. There is a significant number of speech coding algorithms, which are broadly classified into four categories, namely, phonetics, waveform, hybrid and voice vocoders. A detail explanation of these coders is presented in the [Chapter 3]. Phonetics vocoders are more related to the acoustic characteristics of speech signals, whose investigation is beyond the scope of this thesis. Second form coders are waveform coders, which are based on a simple sampling and amplitude quantization process. These coders include 16-bit PCM [19], companded 8-bit PCM [19], and ADPCM [18]. Since the only concept behind these coder types is amplitude quantization, the compression rate of speech signals is limited to very large numbers. Even the most recently standardized waveform coders require a minimum of 16 kbits/sec. However, the main objective of the current speech coders is to reduce the minimum compression rate to 1 - 4 kbits/sec or even lower. With the increasing demand for further compression (low bit rate coding), and increasing number of different

applications, simple amplitude quantization is not an efficient process for transmission of speech signals.

In contrast to waveform coders, vocoders consider the details in the nature of human speech. In their principles, there is no attempt to match the exact shape of the signal waveform. Vocoders generally consist of an analyzer and synthesizer. The analyzer attempts to estimate and then transmit the model parameters that represent the original signal. Speech is synthesized using these parameters to produce an often crude and synthetic constructed speech signal. These types of algorithms are called perceptual quality coders. A very familiar and traditional speech vocoder is LPC-10e. In this type of coders, speech signals are synthesized with an excitation that consists of a periodic pulse train or white noise. The complete quality of the synthesized speech signal depends on the excitation signal. The excitation is simply a train of narrow pulses. Two consecutive pulses are placed apart by a time difference equal to the pitch period. Therefore, the quality of the synthesized speech signal highly depends on the accurate estimation of the pitch period.

Even the most recently developed algorithms, such as MPLPC [35], RPELPC [36], CELP [26], require the correct estimation of the LP coefficients, LTP coefficients and excitation. The basis for estimating these parameters is the fundamental pitch period. Incorrect estimation of the fundamental period harms the estimation of the LTP coefficients, and consequently, the residual. This in turn causes an incorrect selection of excitation, therefore the final speech quality.

From the above discussion, it is evident that the fundamental pitch period estimation is the deciding factor in the final quality of speech signal. In general, whether speech quality is toll-

quality, communication quality, professional quality or synthetic quality, it all depends on the correct estimation of fundamental pitch.

1.4 Thesis Contribution

The following section describes the contribution of the thesis

The new proposed pitch estimation algorithm based on Gabor filters, and an efficient implementation of the auto-correlation method is presented in the chapter 6. This technique is named as Multi-feature, Autocorrelation (ACR) and Wavelet Technique (MAWT). The algorithm has moderate advantages over all the traditional and most recently PDA's for speech signals with or without noise. The accuracy of pitch estimation for fast pitch changing signals, for low energy speech signals, and for transition is improved. The algorithm is threshold insensitive and independent of frame length. Other contributions of the thesis are the study and comparison of various speech coding algorithms, and the complete implementation of LPC vocoder and MBE vocoder. In addition to the above, the reason why pitch estimation of speech signal plays vital role in the final quality of synthesized signal is highlighted.

1.5 Thesis Outline

Chapter 2 presents a thorough description of the basic speech production mechanism and provides the background information for understanding the major contribution of this thesis. It also describes the main characteristics of speech, and basic speech analysis methods. Chapter 3

provides a description of the main speech coder categories and their principles and concepts. It also includes the complete description of the components involved in a generalized model. This description gives the basis of the importance of the pitch estimation in speech coding. Chapter 4 discusses the difficulties in pitch estimation, and includes the complete methodology of some traditional and some recently developed pitch detection algorithms. Finally, chapter 5 presents a novel pitch detection algorithm, named MAWT. A comparison between the methods presented in chapter 4 and the new pitch detection algorithm presented in chapter 5 is also included. Results are provided in the end of the chapter 5. Finally, chapter 7 closes with some concluding remarks.

Chapter 2: Human Speech Production and Perception

This chapter provides an introductory description of the principles related to speech production and perception of speech. First, human speech production is described from the basic acoustic point of view, and second, through the introduction of the factors influencing the fundamental pitch period. Finally, a spectral analysis of speech production is presented, and different fundamental pitch estimation methods are briefly discussed.

2.1 Human Speech Production

Speech signals are composed of a sequence of sounds. These sounds and the transition between them serve as a symbolic representation of information. The arrangement of sounds (symbols) is governed by the rules of the language. The study of the rules and classification of speech is called phonetics. The purpose of processing speech signals is to enhance and extract information, which is helpful in providing as much knowledge as possible about the signal's structure i.e., about the way in which information is encoded in the signal.

2.1.1 The mechanism of speech production

Human speech production requires three elements – a power source, a sound source and sound modifiers. This is the basis of the source-filter theory of speech production. The power source in normal speech results from a compression action of the lung muscles. The sound source, during the voiced and unvoiced speech, results from the vibrations of the vocal folds and turbulent flow past narrow constriction respectively. The sound modifiers are the articulators, which change the shape and therefore the frequency characteristics of the acoustic cavities through which the sound passes.

The main anatomy of the human speech production mechanism is depicted in figure 2.1 and an ideal block diagram of the functional mechanism is illustrated in the figure 2.2. The three main controls of the speech production are –lungs (power source), the position of the vocal folds (sound source) and the shape of the vocal tract (sound modifiers).

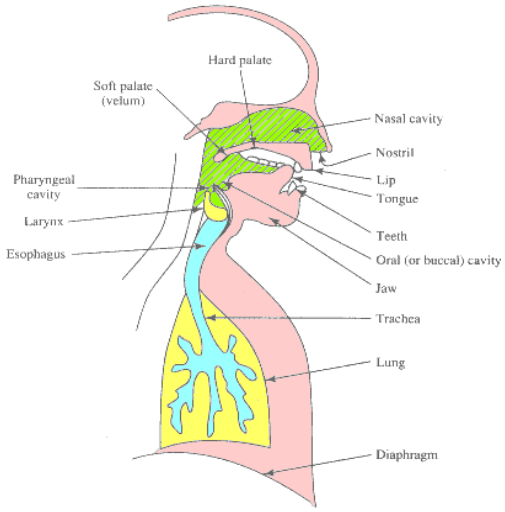


Figure 2.1 Schematic view of human speech production mechanism.

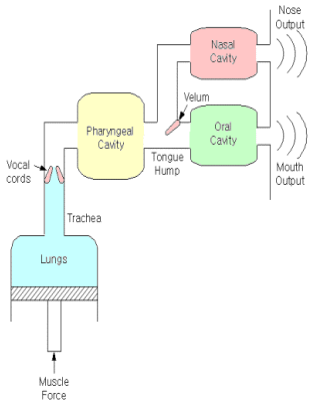


Figure2. 2 Block diagram of human spe production.

2.1.2 Sub-glottal system

This system is composed of the lungs, and the vocal folds. This sub-glottal system serves as a power source and sound source for the production of speech. Speech is simply an acoustic wave radiated from the sub-glottal system, when the air is expelled from the lungs and the resulting flow of air is perturbed by a constriction somewhere in the vocal tract. Speech sounds can be classified into three distinct classes according to their mode of excitation:

- *Voiced* sounds are produced by forcing air through the glottis with tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, there by producing quasi-periodic pulses of air which excite the vocal tract.
- *Unvoiced* or *fricative* sounds are generated by forming a constriction at some point in the vocal tract (usually toward mouth end) and forcing air through the constriction at high enough velocity to produce turbulence.
- *Plosive* sounds result from making a complete closure (again usually toward the front of the vocal tract), building up pressure behind the closure and abruptly releasing it.

2.1.3 Vocal tract

The vocal tract is also termed as *sound modifiers*, and it is depicted in figure 2.2. It is formed by the oral together with the nasal cavities. The shape of the vocal tract (but not the nasal cavity) can be altered during speech production that changes its acoustics properties. The velum can be raised or lowered to shut off or couple the nasal cavity, and then the shape of the vocal tract tube. As the sound is generated, the frequency spectrum is shaped by the shape of the vocal tract tube. Each voiced speech segmented is characterized by a series of peaks in the vocal tract

frequency response curve known as formants. Depending upon the shape of the vocal tract tube, the first three formant frequencies for men, women and children are given in table 2.1.

Parameter	*Men*	*Women*	*Children*
F1 range	270 Hz – 730 Hz	300 Hz – 800 Hz	370 Hz – 1030 Hz
F2 range	850 Hz – 2300 Hz	900 Hz – 2800 Hz	1050 Hz – 3200 Hz
F3 range	1700 Hz – 3000 Hz	1950 Hz – 3300 Hz	2150 Hz – 3700 Hz
f_0 mean	120 Hz	225 Hz	265 Hz

Table 2.1 Typical first three formant frequency ranges f_0 man and ranges of conversational speech of mean, women and children (formant values from [4]).

The average frequency domain effects in speech production are summarized in figure 2.3. Human speech has an approximate 6 dB per octave roll-off with increasing frequency.

The sound source and the sound modifiers make the following contributions to this spectrum. For voiced speech, all harmonics are present with an average spectral roll-off of 12 dB per octave, while for voiceless speech the spectrum is flat. Radiation of the acoustic pressure waveform via the lips gives a +6dB per octave tilt with increasing frequency. This results in an average overall spectral variation with increasing frequency during voiceless speech. Since the amplitude of voiced speech is usually significantly greater than that of voiceless speech, the average spectral shape of speech tends to be close to –6 dB per octave.

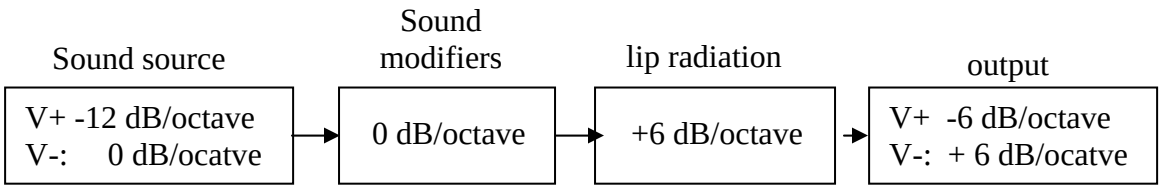


Fig 2.3. Average spectral trends of the sound source, sound modifiers and lip radiation during voiced (V+) and voiceless (V-) speech.

2.2. Factors Influencing the Fundamental frequency

This section explains some of the physiological factors, as well as other factors that influence pitch.

(a) **Body size**

The most obvious influence on pitch that comes to mind is the size of the sound-producing apparatus; we can observe from the instruments of the orchestra that smaller objects tend to make higher-pitched sounds, and larger ones produce lower-pitched sounds. Therefore, it is logical to assume that small people would make high sounds, and large people would make low sounds. And this assumption is borne out by the facts, at least to an extent. Baby cries have a fundamental frequency (referred to as f_0) of around 500 Hz. Child speech ranges from 250-400 Hz, adult females tend to speak at around 200 Hz on average and adult males around 125 Hz. Thus, the body size is one of the factors related to f_0. On the other hand, we know that big opera singers don't always make low sounds; there are very large sopranos, and some rather short, slender basses. So, body weight and height is not a sole determining factor.

(b) **Laryngeal size**

Perhaps, a factor more relevant to the voice source is the size of the larynx. Men, on average, have a larynx about 40% taller and longer (measured along the axis of the vocal folds) than women, as seen in figure 2.4. Nevertheless, this does not completely explain the difference between male and female fundamental frequency f_0; there is a size difference inside the larynx, which fully explains the difference in f_0.

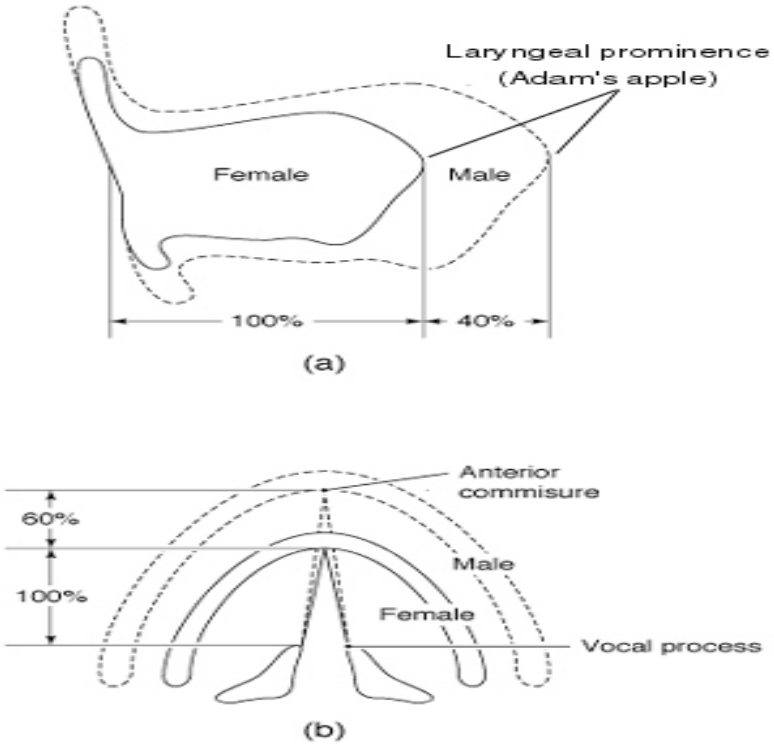


Figure 2.4 (a) Laryngeal shape of female and male speaker (b) Relative sizes of the laryngeal

(c) Vocal fold length

If it is assumed that the vocal folds are 'ideal strings' with uniform properties, their fundamental frequency f_0 is governed by Equation 2.1.

$$F_0 = \frac{1}{2L}\sqrt{\frac{\sigma}{\rho}} \qquad 2.1$$

where

L: Length of vocal folds
σ: Longitudinal stress
ρ: Tissue density

The key variable here is the *length* of the vocal folds part that is actually in vibration, which we call *effective vocal fold length*. If this quantity is examined for men and women, it is

found that men have a 60% longer effective fold length than women, on average, which accounts for the general f_0 difference observed between the sexes.

Briefly, some other factors that influence the fundamental pitch period are (1) the difference between languages (2) the specifics of different applications, (3) the emotional state of the person and (4) the environmental conditions under which speech is produced.

2.3. Speech analysis

One of the important characteristics of a speech waveform is the time-varying nature of the content of the speech pressure. Determination of the time-varying parameters of speech is a key area of analysis required in speech research. Another key area is classification of speech waveform segments into voiced or voiceless (mixed excitation is usually considered voiced). As mentioned previously, in the case where speech is voiced, the most important parameter is the fundamental frequency value f_0.

This section introduces these two areas of analysis and discusses the principles and limitation involved. First, the fundamental frequency f_0 analysis is considered, followed by the spectral analysis method of dynamic speech signals.

2.3.1 Fundamental frequency estimation

The pitch of a sound depends on how our hearing system functions and is based on a subjective judgment by a human listener on a scale from low to high. Therefore, such a psychoacoustic measurement cannot currently be made algorithmically without the involvement of a human listener. The f_0 measurement of the vocal fold vibration is an objective measure, which can be utilized algorithmically. Therefore, the term fundamental frequency estimation is to

be preferred to the term pitch extraction commonly used in the literature. One reason why estimation rather than extraction is adopted, is that although changes in pitch are perceived when f_0 is varied, small changes in pitch can also be perceived when the intensity (loudness) or the sound's spectral content (timbre) is varied when f_0 is kept constant.

The choice of an f_0 measurement technique should be made with direct reference to the particular demands of the intended application in terms of the expected speaker population to be analyzed (adult or child, male or female, pathological or non-pathological), the likely competition from acoustic back ground or foreground noise (others working in the same room, external noises, domestic noises, machine noises, classroom, clinic, children), the material to be analyzed (read speech, conversation speech, shouting, sustained vowel, singing), the effect of the speaker to analysis system signal transmission path (room acoustics, microphone placement, telephone, pre-amplification) and the measurement errors can be tolerated (f_0 doubling, f_0 halving, f_0 smoothing, f_0 jitter).

The operation of f_0 estimation algorithms can be considered in terms of

- The input pressure waveform (time domain)
- The spectrum of the input signal (frequency domain)
- A combination of time and frequency domains (hybrid domain)
- Direct measurement of larynx activity

Most of the errors associated with f_0 estimation are due to

- The quasi-periodicity of voice speech signals
- Formant movements

- The difficulties in locating accurately the onsets and offsets of voiced segments

A highly comprehensive review is given in Hess [5], some of the methods of estimation, errors involved and importance to speech coding will be discussed in chapter (5)

2.3.2 Spectral analysis

Since the 1940's, the time-varying spectral characteristics of the speech signal can be graphically displayed through the use of the sound spectrograph [32,33]. This device produces a two-dimensional pattern called a spectrogram in which the vertical dimension corresponds to frequency and horizontal dimension to time. The 16 bit gray scale level is used to represent the given spectrogram. Even though the color representation is more visually appealing, it sometimes leads to misleading interpretation of the spectrogram. The darkness of the pattern is proportional to signal energy. Thus, the resonance frequencies of the vocal tract show up as dark bands in the spectrogram. Voiced regions are characterized by a striated appearance due to the periodicity of the time waveform, while the unvoiced intervals more solidly filled in. An example, spectrogram of the utterance of "What do you think about that" of a female speaker (in the Figure 2.5a) is shown in the Figure 2.5b. The spectrogram is labeled corresponds to the labeling of Figure 2.5b, so that the time domain and frequency domain can be correlated.

The time scale and frequency resolution of the spectrograph plays a vital role in representation of speech spectral energy. The most rapid changes in time scale occur during the release stages of plosives, which order is of 5-10ms. For individual representation of the harmonics of male speech, a frequency resolution less than the minimum expected f_0 for males – approximately 50 Hz is required. Consequently, there is a direct trade off to be considered

between frequency and time resolution and this can be controlled by altering the bandwidth of the spectrograph's analysis filter. Usually, this is indicated as wide or narrow based on the relation between the filter's bandwidth and the f_0 of the speech being analyzed.

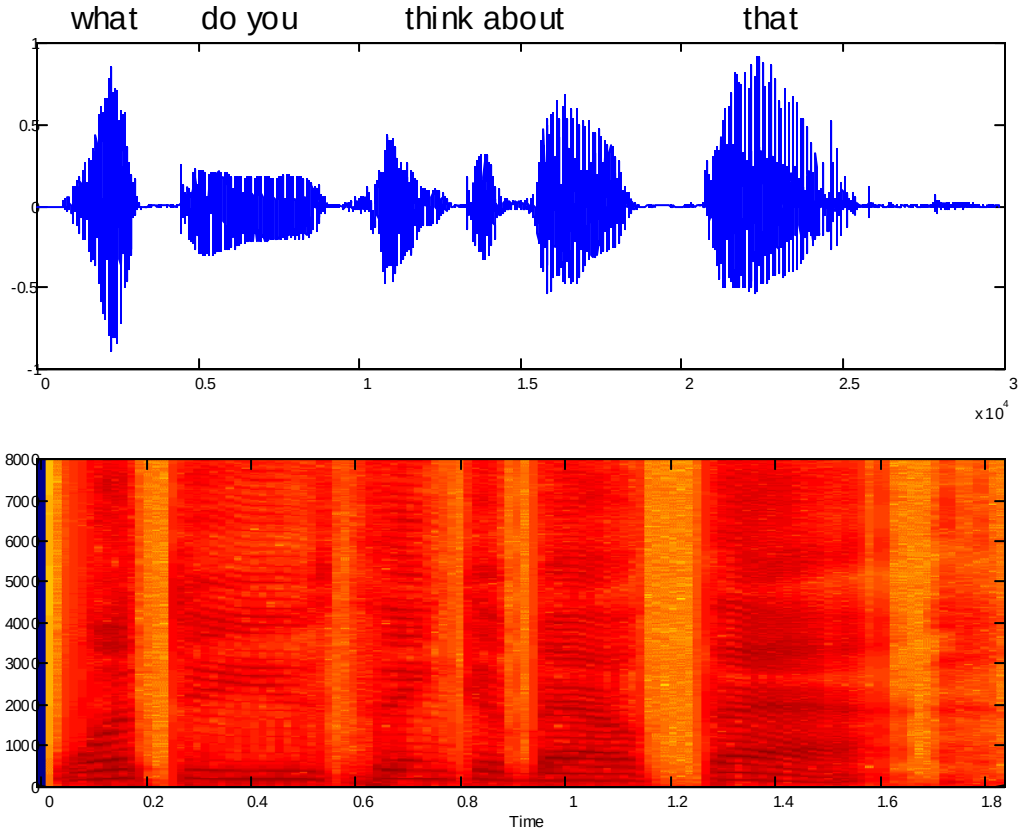


Fig 2.5 Model spectrogram of "what do you think about that" spoken by the healthy adult female.

2.3.3 Wavelet analysis

Another way of applying frequency analysis, considering a broader bandwidth at higher frequencies is wavelet analysis [11]. In this type of analysis, the speech signal is correlated with a set of orthogonal basis functions, which represent the impulse responses of a set of increasing

bandwidth filters. The resulting computation structure is very similar to the tree-structured quadrature filter bank used in speech coding.

In fact, the quadrature-mirror filterbank is form of wavelet transform with the output samples of the filters representing the transform coefficients. Due to the variable bandwidth, which is proportional to frequency, the basis functions are simply rescaled and shifted versions of each other in time.

One of the important characteristics of wavelet transforms, in addition to their variable bandwidth characteristics, is that they are simultaneously localized in time and frequency which allows them to possess, at the same time, the desirable characteristics of good time and frequency resolution.

2.3.4 Cepstral analysis

One of the problems of simple spectral analysis is that the resulting output has elements of both the vocal tract (formants) and its excitation (harmonics). This mixture is often confusing and inappropriate for further analysis, such as speech recognition. Ideally, some method of separating out the effects of the vocal tract and the excitation would be appropriate.
Unfortunately, these two speech aspects are convolved together and they cannot be separated by simple filtering. One speech analysis approach that can help in separating the two elements is the Cepstrum[12]. This finds applications both pitch detection and vocal tract.

The method relies on applying nonlinear operations to map the operation of convolution into a summation. Thus, signals which are convolved together are now signals simply added together. As a result, they can be readily separated, provided they do not overlap in this domain.

This is achieved via two mappings

- Convolution in the time domain is equal to multiplication in the frequency domain
- The sum of the logarithms of two numbers is equal to the logarithm of their product.

Thus, Fourier transforming a signal representing two convolved signals, and then taking the logarithm, results in a transform, which represents the sum of two convolved signals. This additively resulted can then be transformed back to the time domain and processed to separate the signal into excitation and vocal tract

Chapter 3: Speech coders and Classification

The purpose of this chapter is to introduce various speech coder standards, their requirements, and their evolution. A broad categorization and the brief explanation of different categories are presented. Finally, some of speech coders are discussed in detail.

Speech coding is the process of obtaining a compact representation of voice signals for efficient transmission over band-limited wired and wireless media and/or storage. Today, speech coders have become essential components in telecommunications and in the multimedia infrastructure. Commercial systems that rely on efficient speech coding include cellular communication, voice over internet protocol (VOIP), videoconferencing, electronics toys, archiving, and digital simultaneous voice and data (DSVD), as well as numerous PC-based games and multimedia applications.

Speech coding is the art of creating a minimally redundant representation of the speech signal that can be efficiently transmitted or stores in digital media, and decoding signal with the best possible perceptual quality. Like any other continuous-time signal, speech may be represented digitally through the processes of sampling and quantization; speech typically quantized using either 16-bit uniform or 8-bit companded quantization. Like many other signals, however, a sampled speech signal contains a great deal of information between either redundant (nonzero mutual information between successive samples) or perceptually irrelevant (information that is not perceived by human listeners). Most telecommunications are *lossy*,

meaning that the synthesized speech is perceptually similar to the original but may be physically dissimilar.

A speech coder converts a digitized speech signal into a coded representation, which is usually transmitted in frames. A speech decoder receives coded frames and synthesizes reconstructed speech. Speech coders may differ primarily in bit rate (measured in bits per sample or bits per second), complexity (measured in operations in seconds), delay (measured in miiliseconds between recording and playback) and perceptual quality of the synthesized speech. Narrowband (NB) coding refers to speech signals whose bandwidth is less than 4 kHz (8kHz sampling rate), while wideband (WB) coding refers to coding of 7-kHz bandwidth signals (14-16 kHz sampling rate). NB coding is more common than WB coding mainly because of the narrowband nature of the wireless telephone channel (300-3600 Hz). More recently, however, there has been an increased effort in wideband speech coding because of several applications such as videoconferencing.

Section 1 discusses the speech coder requirements and objectives, followed by the section 2, which discusses the broad classification of the speech coders. The rest of the sections 3, 4, and 5, give the detail explanation of above speech coders.

3.1. Algorithm Objectives and requirements

The design and capacity of a particular algorithm often depends upon the target application. Sometimes capacity of the algorithms is bounded by stringent network planning rules, in order to maintain high quality of service and not to degrade the existing service. There are some principle aspects of speech coder, which include:

(i) *Speech quality*: The speech coding consideration is speech quality against the bit rate. The lower the bit rate i.e., the higher the signal compression, the more the quality suffers. How to determine the speech quality is still a matter question. However, other factors that affect the requirements for obtaining the appropriate speech quality are the type of application, the environment and the type of the network technology.

(ii) *Coding delay*: Coding delay includes algorithmic (the buffering of speech for analysis), computational (time taken to process the stored speech samples) and transmission factors. Delay becomes a problem for two reasons. Firstly, speech coders are often interfaced to the PSTN via four to two wire converters or "hybrids". A side effect of using these devices is that a proportion of the output signal from the codec is fed back into the input of the codec. Due to coding delays, this introduces echo. This is extremely disconcerting to the user, who hears one or more echoes of his own voice returned at multiples of 80-120 ms. The second problem with delay is when the coding delay is coupled with long transmission delays such as those encountered with transmission via satellites in geosynchronous orbit (200 ms round trip). In this case, a total delay of over 300 ms may be encountered, making actual conversation difficult. Thus minimization of coding delay is an important research aim.

(iii) *Computational complexity and cost*: Lowering the bit rate while maintaining quality is often achieved at the expense of increased complexity. A complex algorithm requires powerful DSP hardware that is expensive and requires increased power consumption. Until the late 1980's, many speech coding algorithms were not implementable in real time due to the lack of sufficiently powerful real time DSP hardware. The advent of the digital signal processors (DSP) chips and custom application specific integrated circuits (ASIC) chips has lowered considerable power. However, the cost and power consumption is still a major problem in places where

hardware is an important factor. Thus, the search for computationally efficient algorithms is an important research activity to reduce DSP hardware requirements, power consumption, and cost of speech coding hardware.

(iv) *Robust to Channel errors*: For many applications, the quality of speech signals against the channel error, being accomplished by employing the Forward Error Correction (FEC). However, it is important to maintain the acceptable quality for mobile and satellite systems, which suffer from random and burst types of noise. The disadvantage with use of FEC is that extra bandwidth is required and that it is unacceptable for mobile and satellite systems. Thus the robustness to channel error is an important consideration.

(v) *Robust to Background noise*: Most of the low-bit rate speech coders exploit the redundancy in the speech signals. However redundancy is not necessarily the same for other signals such as background noise or single sinusoids. In such cases, the speech coder may distort or corrupt the synthesized speech signals. Another effect is that the signal processing techniques used to extract model parameters may fail when speech corrupted by high levels of background noise is coded. For example, many of the very low rate, synthetic quality vocoders used by the military fail in moving vehicles or helicopters due to the presence of periodic background noise.

(vi) *Tandem connection and transcoding*: As it is the end-to-end speech quality, which is important to the end user, the ability of an algorithm to cope with tandeming with itself or with another coding system is important. Degradations introduced through tandeming are usually cumulative, and if an algorithm is heavily dependent on return characteristics then severe degradation may result. This is a particularly urgent resolved problem with current schemes, which employ post-filtering in the output speech signal [29]. Transcoding into another format, usually PCM, also degrades the quality, and introduces extra cost.

3.2. Speech coding Strategies and Standards

Speech coding schemes are broadly classified into four categories as illustrated in the Figure 3.1. The basic principle of these coders is to analyze the speech signal to remove the redundancies and code the non-redundant parts of the signal in perceptually acceptable manner. In the following sections only three main categories are described. The quality vs bit rate for three main coding methods are shown in Figure 3.3. A summary of the speech coding methods, the bit-rate and mean square score (MOS) ranging from 1 to 5 are listed in the table 3.1 Generally, coding quality with MOS higher than 4 is considered as toll quality, between 3.5 and 4 as communication quality, between 3 and 3.5 as professional quality, and below 3 as synthetic quality [17].

3.3. Waveform coder:

Waveform coders attempt to code the exact shape of the speech signal waveform, without considering in detail the nature of human speech production and speech perception. Waveform coders are the most useful in applications that require the successful coding of both speech and nonspeech signals. In the public switched and telephone network (PSTN), for example signaling tones and switching signals of speech is nearly as important as the successful transmission of speech. The most commonly used waveform coding algorithms are uniform 16-bit PCM, companded 8-bit PCM [19] and ADPCM [18].

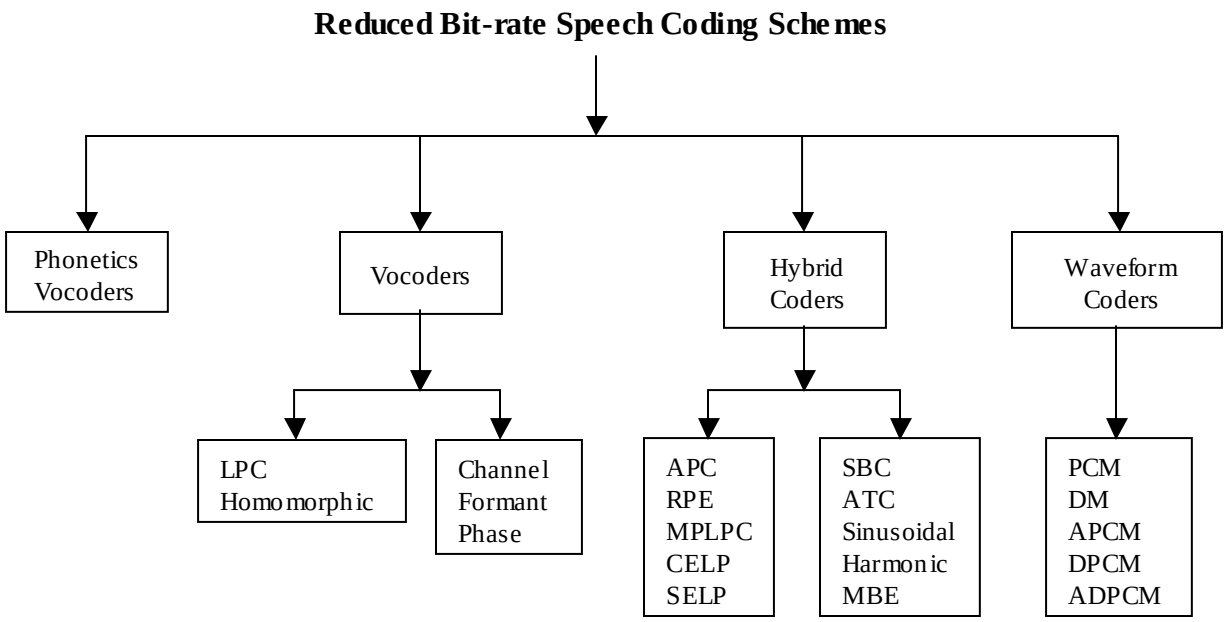


Figure 3.1 Classification of speech coding schemes

Application	Rate (kbps)	MOS	Standard	Algorithm	Year
Landline Telephone	64		ITU-G.711	μ-law or A-law PCM	1972
	32		ITU-G.721	ADPCM	1984
	16-40		ITU-G.726	VBR-ADPCM	1991
	16-40		ITU-G.727	Embedded-ADPCM	1991
Tele conferencing	48-64		ITU-G.722	Split-band ADPCM	1988
	16		ITU-G.728	Low-delay CELP	1992
Digital Cellular	13		GSM-Full rate	LTP_RPE	1989
	12.2		GSM-EFR	ACELP	1995
	7.9		TIA IS-54	VSELP	1991
	6.5		CDMA-TIA IS-96	Qualcomm CELP	1991
	8.0		GSM-Half rate	VSELP	1994
	4.75-12.2		ITU-G.729	CSA-CELP	1995
	1-8		GSM-AMR	ACELP	1998
Multimedia	5.3-6.3		ITU-G723.1	MPLPC, CELP	1996
	2.0-18.2		ISO-MPEG-4	HVXC, CELP	1998
Satellite telephony	4.15		INMARSAT-M	IMBE	1990
	3.6		IMMARSAT Mini-M	AMBE	1995
Secure communications			DDVPC FS1015	LPC-10e	1984
			DDVPC MELP	MELP	1996
			DDVPC FS1016	CELP	1989
			DDVPC CVSD	CVSD	

Table 3.1 representation of speech coding standards

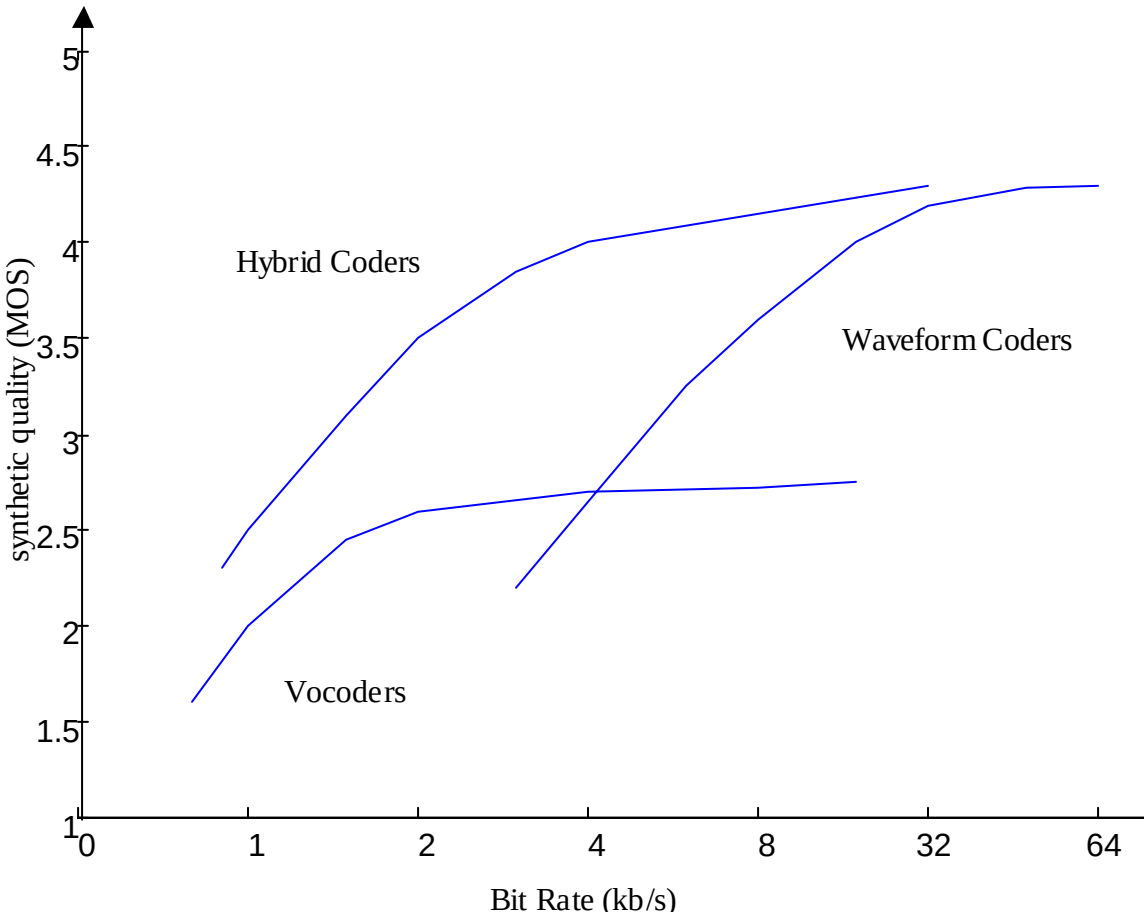


Figure 3.2 Quality comparison of speech coding schemes

3.3.1 Pulse Code Modulation

In pulse code modulation (PCM) coding the speech signal is represented by as series of quantized samples. Since the these are memoryless coding algorithms, each sample of the signal s(n) uses the same number of reconstruction levels k=0……m,…. K. Regardless of the values of previous samples, each sample is estimated from its code word. Thus, it is called as memoryless coding algorithm.

3.3.1a. Uniform PCM:

Uniform PCM is the name given to quantization algorithms in which reconstruction levels are uniformly distributed between S_{max} and S_{min}. The advantage of uniform PCM is that quantization error power is independent of signal power; high-power signals are quantized with the same resolution as low-power signals. Invariant power is considered desirable in many digital audio applications, so 16-bit uniform PCM is standard coding scheme in digital audio.

The error power and SNR of uniform PCM coder vary with bit rate in a simple fashion. Suppose that a signal is quantized using B bits per sample. Then, the quantization step size Δ is

$$\Delta = \frac{S_{\max} - S_{\min}}{2^B - 1} \tag{3.1}$$

Assuming that quantization error are uniformly distributed between Δ/2 and -Δ/2, the quantization error power is

$$10\log_{10} E\left[e^2(n)\right] = 10\log_{10}\frac{\Delta^2}{12} \tag{3.2}$$

$$\approx \text{constant} + 20\log_{10}(S_{\max} - S_{\min}) - 6B \tag{3.3}$$

3.1b Companded PCM

In order for the percentage error to be constant, the quantization levels must be logarithmically spaced. Alternatively, the logarithm of the input can be quantized rather than the input self. This depicted in Fig 3.3, which shows the input amplitudes being compressed by the

logarithm function prior to quantization and being expanded by the exponential function after decoding

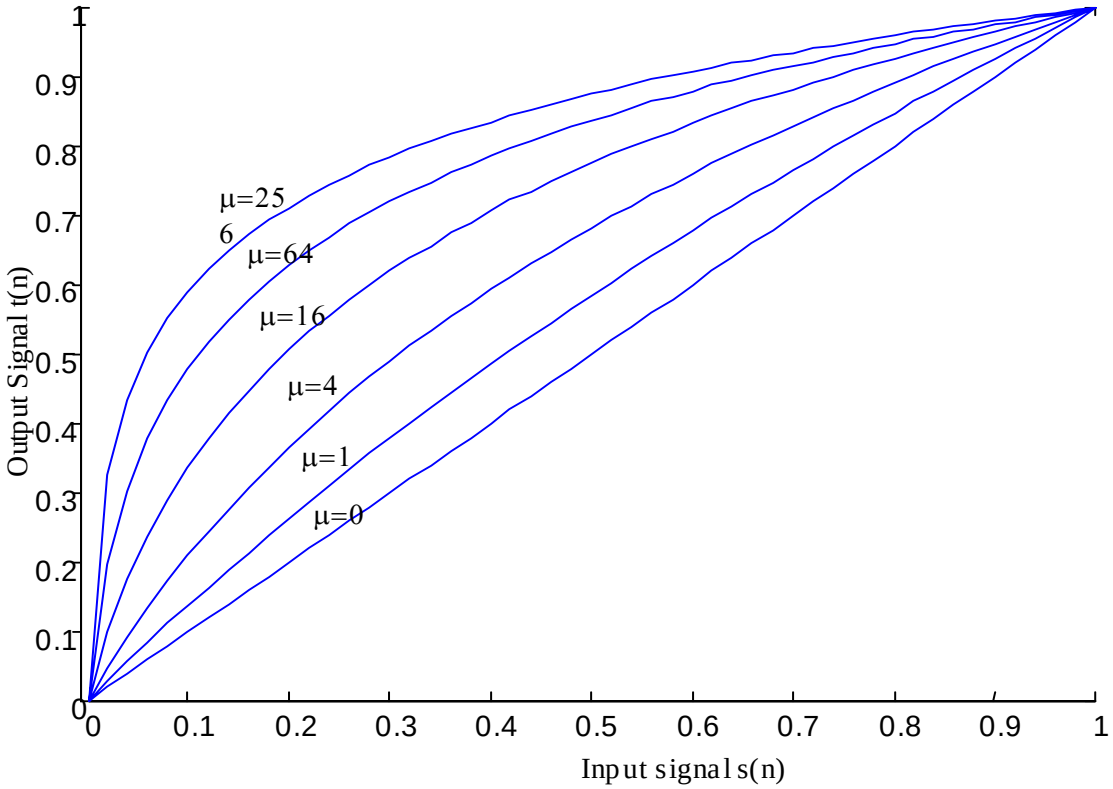


Figure 3.3 μ-law companding function μ=0 ,4,16…256.

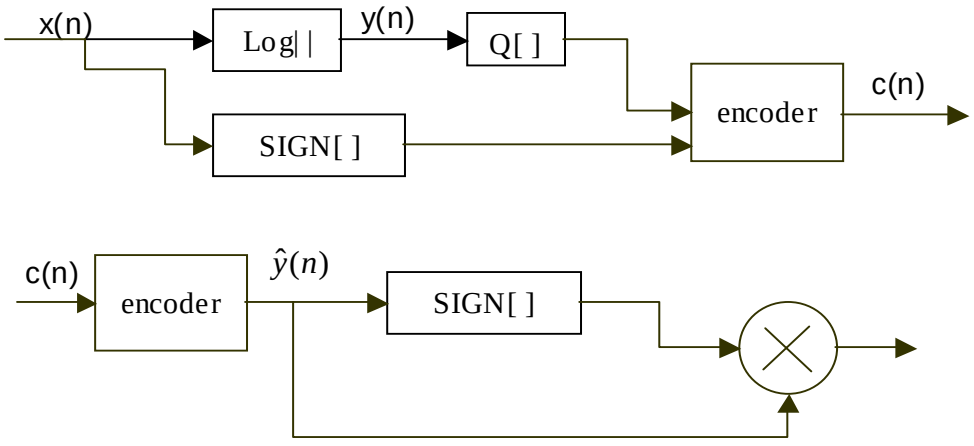


Figure 3.4. Block diagram of a logarithmic encoder-decoder

It can be shown that, if small values of s(n) are more likely than large values, expected error power is minimized by companding function that results in a higher density of reconstruction levels $\hat{x}(k)$ at low signal levels than at high signal levels. A typical example of μ-law companding function [2] (Fig 3.3), which is given by

$$y(n) = F[x(n)] \tag{3.4}$$

$$= x_{\max} \frac{\log\left[1 + \mu \frac{|x(n)|}{x_{\max}}\right]}{\log[1+\mu]} .sign[x(n)] \tag{3.5}$$

Where μ is typically varies between 0 and 256 and determines the amount of nonlinear compression applied.

3.4. Voice vocoders:

In contrast to waveform coders voice vocoders consider the details in the nature of human speech. In their principles, there is no attempt to match the exact shape of the signal waveform. It consists of an analyzer and synthesizer. The analyzer attempts to estimate the model parameters, which represent the original signal, and then transmit them. The speech is synthesized using the parameters to produce an often crude and synthetic constructed speech signal. Since the synthesized signal is either crude or distorted, SNR is not a good measure of the speech quality, hence there is a need of subjective measures such as mean opinion scores (MOS) test, diagnostic rhyme test (DRT) and diagnostic acceptability measure (DAM) [20]. The most current voice

vocoders are 2.4 kbit/sec LPC-10[13], RELP [22], Homomorphic vocoder [23] [24], and the channel and the formant vocoder [25].

The complete description of linear prediction including LPC vocoder and MBE is presented in the chapter 4.

3.5. Hybrid coders:

To overcome the disadvantages of waveform coders and voice vocoders, hybrid coding methods have been developed which incorporate each of the advantages offered by the above schemes. Hybrid coders are broadly classified into two sub-categories:

- Frequency domain hybrid coders:
- Time domain hybrid coders:

3.5.1 Time domain hybrid coders:

These can be classified as analysis by synthesis (AbS) LPC, in which the system parameters are determined by linear prediction and the excitation sequence is determined by a closed loop or open loop optimization. The optimization process determines an excitation sequence, which minimizes a measure of the weighted difference between the input speech and the coded speech. The weighting or filtering function is chosen such that the coder is “optimized” for the human ear. The most commonly used excitation model used for AbS LPC are: the multi pulse, regular pulse excitation, vector or code excitation. Since these methods combines the features of model-based vocoders, by representing the formant and the pitch structure of speech, and the properties of waveform coders, they are called hybrid. The basic

structure of the AbS model and the complete explanation relating the component in the block diagram are presented in the following subsections

3.5.1.1 The Basis LPC Analysis by Synthesis Model:

The basic structure of an AbS model coding system is illustrated in Figure 3.5. It consists by the following three components

(1) Time-Varying filter

(2) Excitation signal

(3) Perceptually based minimization procedure

The model requires frequent updating of the parameters to yield a good match to the original, the analysis procedure of the system is carried out in blocks, i.e., the input speech is partitioned into suitable blocks of samples. The length and update of the analysis block or frame determines the bit rate or capacity of the coding schemes.

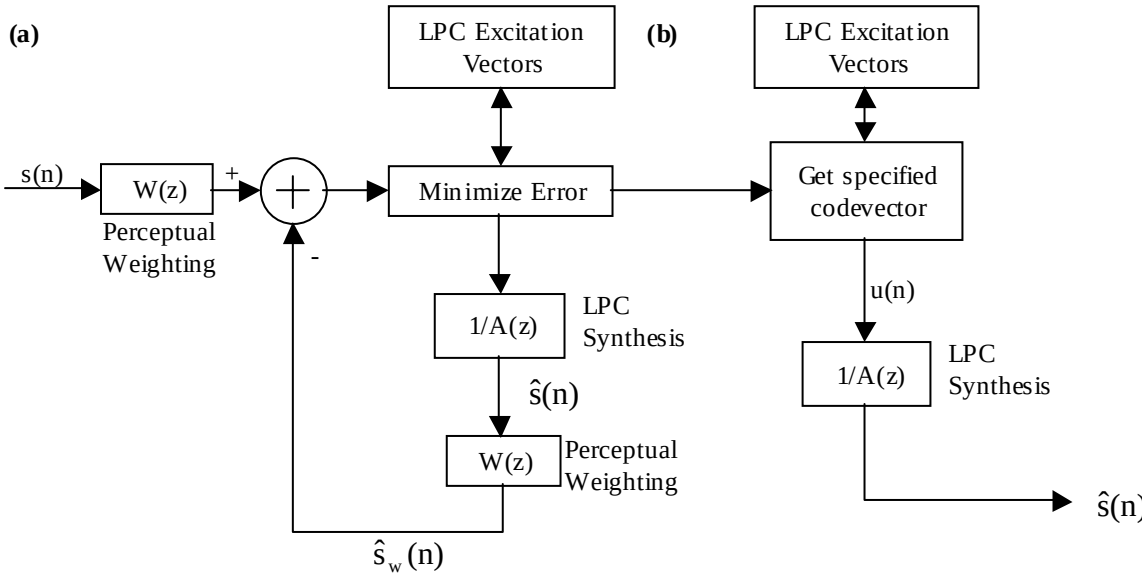


Figure 3.5 General structure of an LPC-AS coder (a) and decoder (b). LPC filter A(z) and perceptual weighting filter W(z) are chosen open-loop, then the excitation vector u(n) is chosen in closed-loop fashion in order to minimize the error metric $|E|^2$.

3.5.1.1(a) Short-Term Prediction filter

In Basic LPC model is also termed as Short-time Predictor (STP), which is illustrated in Figure 3.5. The complete description of LPC model and estimation of the filter coefficients will be discussed in chapter 4. The STP models the short-time correlation in the speech signal (spectral envelope), and has the form given by

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \qquad 3.6$$

where a_i, are the STP (or LPC) coefficients and p is the filter. Most of the zeros in A(z) represents the vocal tract or formant frequencies. Then, the number of LPC coefficients (p) depends on the signal bandwidth. Since each pair of complex-conjugate poles represents one formant frequency and since there is on average, one formant frequency per 1 kHz, p is typically equal to 2BW (in kHz) + (2 to 4). Thus, for a 4 kHz speech signal, a 10^{th}-12^{th} order LPC model would be used.

3.5.1.1(b) Long Term Prediction Filter:

The LTP model the long – term correlation in the speech (fine spectral structure), and has the form given by

$$\frac{1}{P(z)} = \frac{1}{1 - \sum_{i=-I}^{I} b_i z^{-(D+i)}} \qquad 3.7$$

Where D is a pointer to long-term correlation which usually corresponds to the pitch period or its multiples and b_i are the LTP gain coefficients. The process to estimation of parameters is presented in the chapter 4. Again, this filter is time varying and usually has higher adaptation rate than the STP, e.g. every 5-10 ms. The number of filter taps typically form I=0 i.e., 1-tap and I=1,

i.e., 2-tap taps. There is no specific limitation on the order of filters; sometimes the LTP filter is omitted, as in MPLPC.

3.5.1.1(c) Perceptually based minimization procedure

The Abs-LPC coder of Figure 3.4 minimizes the error between the original signal s(n) and the synthesized signal $\hat{s}(n)$ according to a suitable error criterion by varying the excitation signal and the STP and LTP filters. This is achieved via a sequential procedure. First, the time-varying filter parameters are determined, and then the excitation is optimized.

The optimization criterion used for both procedures is the commonly used mean squared error criterion, which is simple and gives an adequate performance. However, at low bit rates, with one or less bit per sample, thus it is very difficult to match the original signal. Consequently, the mean squared error criterion is meaningful but not sufficient. An error criterion, which is near to human perception, is necessary. Although much research on auditory perception is in progress, no satisfactory error criterion has yet emerged. In the meantime, however, a popular but not totally satisfactory method is use of weighting filter in AbS-LPC schemes. The weighting filter is given by

$$W(z) = \frac{A(z)}{A(z/\gamma)} \tag{3.8}$$

$$= \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}} \qquad 0 \le \gamma \le 1 \tag{3.9}$$

A typical plot of its frequency response is shown in Figure 3.6. The factor γ does not alter the center formant frequencies, but just expands the bandwidth of the formants by Δf given by

$$\Delta f = -\frac{f_s}{\pi} \ln \gamma \text{ (Hz)} \tag{3.10}$$

where f_s is the sampling frequency. As can be seen from Figure 3.5, the weighting filter de-emphasis the frequency regions corresponding to the formants as determined by the LPC analysis. By allowing larger distortion in the formant regions, noise that is more subjectively disturbing in the formant nulls can be reduced. The amount of de-emphasis controlled by γ. Most suitable value of γ is usually around 0.8-0.9

3.5.1.1(d) Excitation Signal

The excitation signal is an input to AbS-LPC model and its generation procedure is an important block of the model shown in figure 3.4. This is because excitation signals represent the structure of the residual signal, which is not represented by the time-varying filters (STP and LTP). e.g., speech signals with correlation greater than the LTP delay range, and also the structure that is random in that they cannot be efficiently modeled by deterministic methods. The excitation can be of any form, and can be modeled by the Equation 3.11. A block diagram of an AbS-LPC with different excitation types is shown in Figure 3.7.

$$U_i = g_i X_i \qquad 3.11$$

where U_i is a L-dimensional i^{th} excitaion vector, x_i represents M×L dimensional ‘shape’ vectors and g_i is the M-dimensional gain or scale vector associated with the shape X_i.

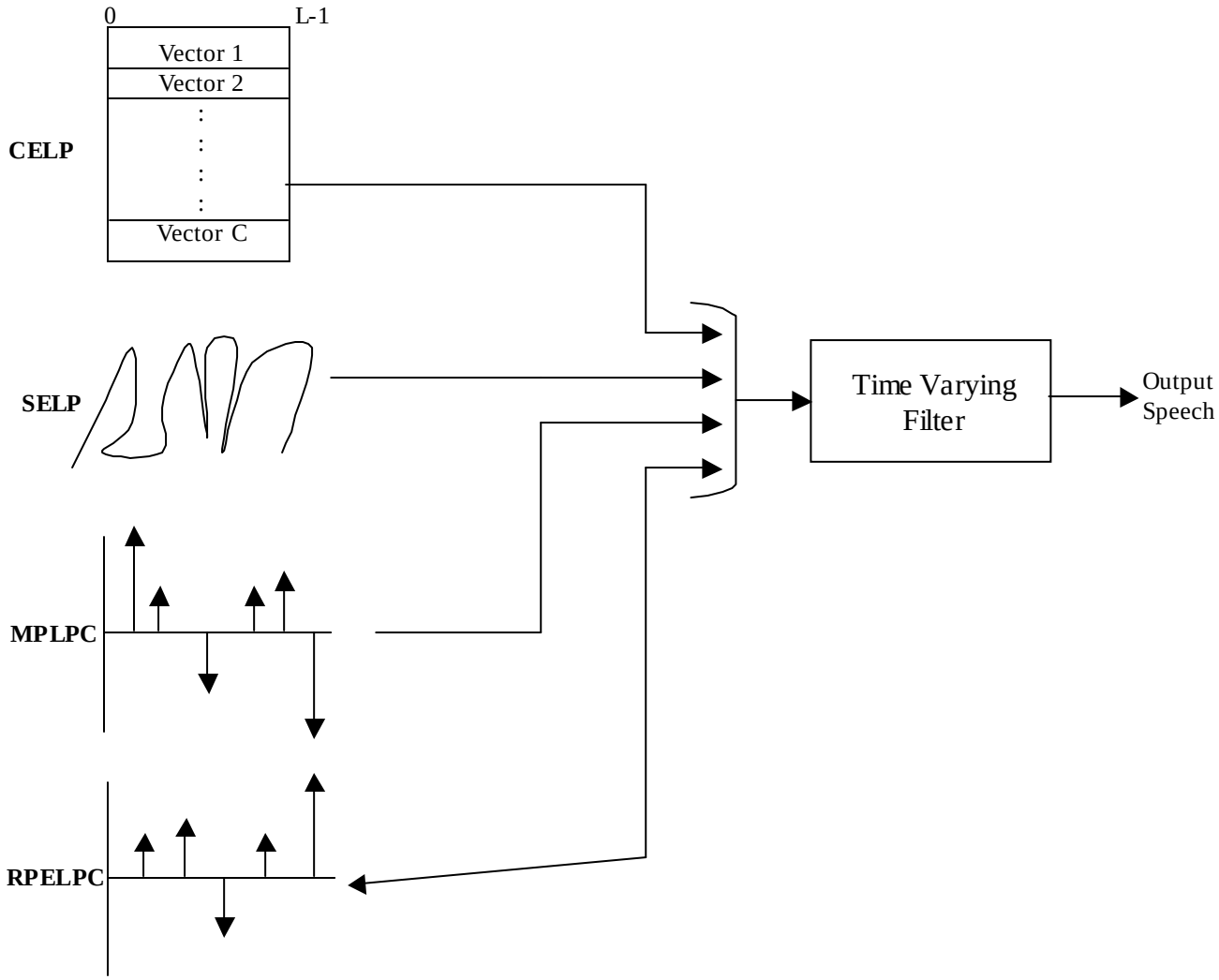


Figure 3.6 Generalised block diagram of AbS-LPC coder with different excitation types

3.5.2.1 Code Excited Linear Predictive coding

In the codebook excitation (CELP) [26], the excitation vector chosen from a set of pre-stored collection of C possible stochastic sequences with an associated scaling or gain vector. Although the scaling vector is usually just a scalar factor, it can be more than one, scaling the excitation vector elements in various parts of the vector accurately. Thus to retain the generality we can divide a code book vector c^k into M equal parts, each of length L/M, and compute the

optimum gains g_{k1}……., g_{kM} for each index k, 1≤k≤c. Thus for codebook excitation, the excitation U can be written as

$$U_i = g_i X_i \; k = 1\ldots\ldots, C \qquad 3.12$$

where $g_k = [g_{k1}^i \ldots\ldots, g_{kM}^i] \; i = 1\ldots\ldots, L/M$

$$X_k = \begin{cases} c_j^k & \text{for} \quad j = 0,1,\ldots, L-1 \\ 0 & \text{otherwise} \end{cases}$$

In the AbS procedure the C possible c sequences are systematically passed through the combined sythesis filter, $H_c(Z)$, and the vector that produces the lowest error is the desired sequence. Since the set of sequences are present at both the encoder and decoder, only index k, to the codebook is required to be transmitted. Therefore, less than 1 bit/sample is possible.

As the codebook is of finite dimension, it must be populated with representative vectors of the excitation to be encoded. In Atal's original proposal, unit variance white Gaussian random numbers were used. This choice of population was reported to give very good results, and was partly due to the fact that pdf of the prediction error samples, produced by inversing filtering the speech through both STP and LTP filters, is very close to Gaussian. Another popular choice of codebook entries is center clipped Gaussian vectors, which both reduce complexity and improve performance.

3.5.2.2 **Multipulse-Excited LPC**

Multipulse coders [27] model the residual, using a series of pulses. The positions and amplitudes of the pulses are chosen to minimize the error between the original and synthesized

speech over the current analysis frame (typically 5ms long). Figure 3.8 illustrates the multipulse analysis loop.

To determine a pulse location and amplitude the excitation generator produces an excitation sequence for each possible pulse location in the analysis frame. These candidate excitations are passed through the synthesis filter, and the MSE between the synthesised and original speech measured. The optimum pulse amplitude is obtained by minimising the MSE at each candidate pulse position. The candidate position and amplitude that minimises the MSE is chosen, and the procedure is repeated for the desired number of pulses.

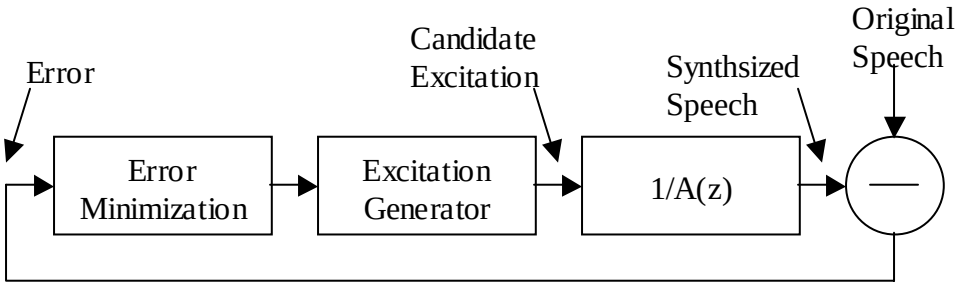


Figure 3.&: Multipulse-Excitation Encoder

This technique is a form of *analysis by synthesis* or *closed-loop* coding, as the canditate excitation signals are synthesised as part of the analysis procedure.

The pulse locations and amplitudes for successive pulses are found iteratively to reduce complexity. After the optimum position and amplitude of pulse *n* has been chosen, the synthesised speech from this pulse is subtracted from the original speech. The result of the subtraction is then used as the original speech for determing pulse *n* +1.

Multipulse-Excitation requires no pitch or voicing detectors, which tends to make it more robust to different speakers and acoustic background noise conditions than the LPC vocoder.

Multipulse coders can produce communications quality speech at bit rates of around 10 kbit/s. Typically around 4-8 pulses per 5ms analysis frame are required for communications quality speech. At bit rates below 10 kbit/s, not enough bits are available for the number of pulses required to produce an adequate excitation signal.

A low complexity development of Multipulse-Excitation is Regular Pulse Excitation (RPE) [28]. This coder only optimizes the position of the first pulse in each analysis frame, the rest are regularly spaced. The amplitudes of each pulse are individually chosen to minimize the MSE in a similar fashion to multipulse coders.

Chapter 4: Linear Prediction of Speech

The purpose of this chapter is to introduce and discuss basic speech coding principles and the complete description of Linear Prediction of Speech vocoder.

4.1. Linear prediction in speech coding

The human speech production process reveals that the generation of each phoneme is characterized basically by two factors: the source excitation and the vocal tract shaping. In order to model speech production we have to model these two factors. To understand the source characteristics, it is assumed that the source and the vocal tract model are independent [1]. The vocal tract model $H(z)$ is excited by a discrete time glottal excitation signal $u(n)$ to produce the speech signal $s(n)$. During unvoiced speech, $u(n)$ is a flat spectrum noise source modeled by a random noise generator. On the other hand, during voiced speech, the excitation uses an estimate of the local pitch period to set an impulse train generator that drives a glottal pulse shaping filter. The speech production process is shown in Fig. 4.1.

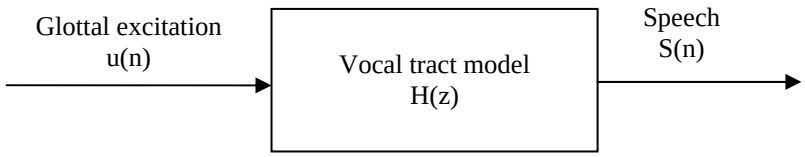


Figure 4.1 Modeling speech production

The most powerful and general linear parametric model used to model the vocal tract is the *autoregressive moving average* (ARMA) model. In this model, a speech signal s(n) is considered to be the output of a system whose input is the excitation signal u(n). The speech sample s(n) is modeled as a linear combination of the past outputs and the present and past inputs [7]. This relation can be expressed in the following difference equation

$$s(n) = \sum_{k=1}^{p} a(k)s(n-k) + G\sum_{l=0}^{q} b(l)u(n-l) \qquad 4.1$$

where G (Gain factor) and a(k) , b(l) (filter coefficients) are system parameters. Since signal s(n) is predictable from the *linear* combinations of past outputs and inputs. Hence the name *linear prediction* is used. The transfer function of the system can be obtained by taking Z-transform on Equation 4.1 with further simplifications:

$$H(z) = \frac{S(z)}{U(z)} = G\frac{1+\sum_{l=1}^{q} b_l z^{-l}}{1+\sum_{k=1}^{p} a_k z^{-k}} \qquad 4.2$$

where $S(z) = \sum_{n=-\infty}^{\infty} s_n z^{-n}$ is the z transform of the s(n) and U(z) is the z transform of u(n). clearly H(z) is a *pole-zero model* or *autoregressive moving average* (ARMA) model. The zeros represent the nasals, while the formants in a vowel spectrum represented by the poles of H(z). There are two special cases of this model.

- When $b_l = 0$, for $1 \le l \le q$, H(z) reduces to an all-pole model, which is also known as an *autoregressive* model.
- When $a_k = 0$, for $1 \le k \le p$, H(z) becomes all-zero model or *moving average* model.

Speech signal is a time varying acoustic pressure wave. For the purpose of analysis and coding, it can be converted to electrical form and sampled. Speech signals are non-stationary; the characteristics of speech evolve over time. As the characteristics vary slowly, speech signals can be approximated as stationary over short periods (in the order of a few tens of milliseconds). With this assumption all-pole model or autoregressive model is widely used for it simplicity and computational efficiency. It models sounds such as vowels well enough. The zeros arise only in nasals and in unvoiced sounds like fricatives. Poles approximately model the voiced sounds. Moreover, it is easy to solve for an all-pole model. In order to solve for a pole-zero model, it is necessary to solve a set of nonlinear equations, but in the case of an all-pole model only a set of linear equations needs to be solved.

The transfer function of an all-pole model is

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad 4.3$$

Actually an all-pole model is a good approximation of the pole-zero model. According to [1], any causal rational system H(z) can be decomposed as

$$H(z) = G' H_{\min}(z) H_{ap}(z) \qquad 4.4$$

where, G' is the gain factor, $H_{\min}(z)$ is the transfer function of a minimum phase filter and $H_{ap}(z)$ is the transfer function of an all-pass filter.

Now, the minimum phase component can be expressed as an all-pole system:

$$H_{\min}(z) = \frac{1}{1 - \sum_{i=1}^{I} a_i z^{-i}} \qquad 4.5$$

where I is theoretically infinite but practically can take a value of a relatively small integer. The all-pass component contributes only to the phase. Therefore, the pole-zero model can be estimated by an all-pole model.

If the gain factor G = 1, then from Equation 4.5 the transfer function becomes

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} \qquad 4.6$$

$$= \frac{1}{A(z)}$$

where the polynomial $\left(1 - \sum_{k=1}^{P} a_k z^{-k}\right)$ is denoted by A(z). The filter coefficients a_k are called the LP (linear prediction) coefficients.

The error signal e(n) is the difference between the input speech and the estimated speech. Thus the following relation holds:

$$e(n) = s(n) - \sum_{k=1}^{P} a_k s(n-k) \qquad 4.7$$

In the z-domain it is equivalent to

$$E(z) = S(z)A(z) \qquad 4.8$$

Now, the whole model can be decomposed into the following parts, the analysis part and the synthesis part (see Figure 4.2)

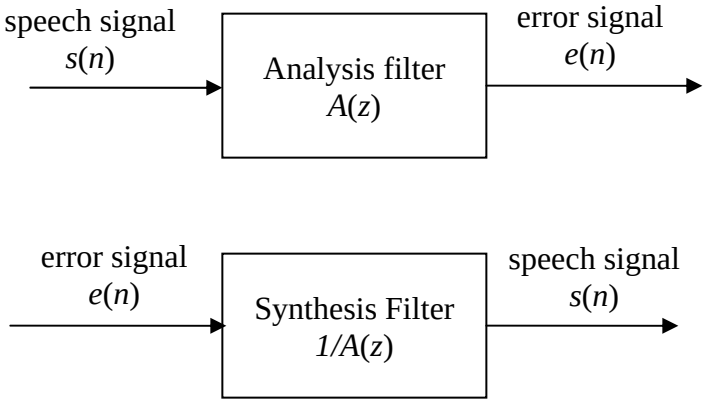


Figure 4.2 LP Analysis and synthesis model

The analysis part analyzes the speech signal and produces the error signal. The synthesis part takes the error signal as an input. The input is filtered by the synthesis filter 1/A(z), and the output is the speech signal. The error signal e(n) is sometimes called residual signal or excitation signal. If the error signal from analysis part is not used in synthesis or if the synthesis filter is not

exactly the inverse of the analysis filter, the synthesized speech signal will not be the same as original signal. To differentiate between the two signals, we use the notation s for the synthesized speech signal.

In general excitation signal for the synthesized filter either error signal or periodic impulse train/white noise input.

4.1.1 Role of Windows

Speech is a time varying signal, and some variations are random. Usually during slow speech, the vocal tract shape and excitation type do not change in 200 ms. But phonemes have an average duration of 80 ms. Most changes occur more frequently than the 200 ms time interval [30]. Signal analysis assumes that the properties of a signal usually change relatively slowly with time. This allows for short-term analysis of a signal. The signal is divided into successive segments, analysis is done on these segments, and some dynamic parameters are extracted. The signal $s(n)$ is multiplied by a fixed length *analysis window* $w(n)$ to extract a particular segment at a time. This is called *windowing*. Choosing the right shape of window is very important, because it allows different samples to be weighted differently. The simplest analysis window is a rectangular window of length N_w:

$$w(n) = \begin{cases} 1, & 0 \le n \le N_w - 1, \\ 0 & otherwise \end{cases} \qquad 4.9$$

A rectangular window has an abrupt discontinuity at the edge in the time domain. As a result there are large side lobes and undesirable ringing effects [8] in the frequency domain representation of the rectangular window. To discard the large oscillations, we should use a

window without abrupt discontinuities in the time domain. This corresponds to low side lobes of the windows in the frequency domain. The Hamming window of Equation 4.10, used in this research, is a tapered window. It is actually a raised cosine function:

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\prod n}{N_w - 1}), & 0 \le n \le N_w - 1, \\ 0, & otherwise \end{cases} \qquad 4.10$$

There are other types of tapered windows, such as the Hanning, Blackman, Kaiser and the Bartlett window. A window can also be hybrid. For example, in GSM 06.90, the analysis window consists of two halves of the Hamming windows with different sizes [31].

4.1.2 LP coefficient computation

There are two widely used methods for estimating LP coefficients. Both methods choose the short-term filter coefficients (LP coefficients) $\{a_k\}$ in such a way that the residual energy (the energy in the error signal) is minimized. The classical least square technique is used for that purpose. In each of the two formulations predictor coefficients are computed by solving a set of p equations with p unknowns. These equations are

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i), \quad 1 \le i \le p \text{ for autocorrelation} \qquad 4.11$$

where $R(i) = \sum_{n=i}^{N_w-1} s_w(n)s_w(n-i)$ s_w is the windowed speech signal $s_w(n) = w(n)s(n)$

$$\sum_{k=1}^{p} a_k \varphi_{ki} = -\varphi_{0i}, \qquad 1 \le i \le p \text{ for covariance method} \qquad 4.12$$

where $\varphi_{ik} = \sum_{n=0}^{N-1} s_w(n-i)s_w(n-k)$

There exits several standard methods to solve the above linear equations eg., the Gauss reduction or elimination method and Crout method. But other methods like square-root or Cholesky decomposition method, needs only half of the number of operations (multiplication or divisions) and half of the storage of the precious methods, because covariance is symmetric and semi-definite. Further reduction in storage and computation is possible in solving the auto-correlation normal Equations 4.11 because of their special form. Equation 4.11 can be expanded in matrix form as

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & R(0) & \cdots & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(3) \end{bmatrix} \qquad 4.13$$

the above $p \times p$ auto-correlation matrix is symmetric and the elements along the diagonal are identical. Levinson derived an elegant recursive procedure for solving this type of equation, later formulated by Robinson. Durbin's procedure can be specified as follows.

$$E_0 = R(0) \qquad 4.13$$

$$k_i = -\left[R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j) \right] \Big/ E_{i-1} \qquad 4.14$$

$$a_i^{(i)} = k_i \qquad 4.15$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \qquad 1 \le j \le i-1 \qquad 4.16$$

$$E_i = (1-k_i^2)E_{i-1} \qquad 4.17$$

Equation form 4.13 to 4.17 are solved recursively for i=1, 2,……, p. Final solution given by

$a_j = a_j^{(p)} \;\; 1 \le j \le p$

4.1.3 Gain computation

It is reasonable to expect that the gain G, could be determined by matching the energy of the signal with energy of the linearly predicted sample. This indeed is true when appropriate assumptions are made abut the excitation signal to the LPC model.

By referring back to Equation 4.7, the excitation signal G.u(n) can be expressed as

$$Gu(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \qquad 4.18$$

where as the prediction error signal given by

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \qquad 4.19$$

In the case where $a_k=\alpha_k$, i.e., the actual predictor coefficients, and those of the model are identical, then

$$e(n) = Gu(n) \qquad 4.20$$

i.e., the input signal is proportional to the error signal with constant of the proportionality is gain constant, G.

Since the Equation 4.20 is approximate, generally it is not possible to solve for G in a reliable way directly from the error signal itself. Instead it is more reasonable to assume that energy in the error signal is equal to the energy in the excitation input.

$$G^2 \sum_{m=0}^{N-1} u^2(m) = \sum_{m=0}^{N-1} e^2(m) = E_n \qquad 4.21$$

Based on the some assumptions, u(n) excitation signal periodic impulse train δ(n), when the signal is voiced, zero mean, unit variance stationary white noise when the signal is unvoiced. The gain G is given by

$$G^2 = R_n(0) - \sum_{k=1}^{p} a_k R_n(k) \qquad 4.22$$

LPC provides an efficient way of coding the vocal tract information. The next stage in coding process is to determine an efficient method of coding the excitation for the filter.

4.2. LPC vocoder:

The synthesizer (decoder) for a simple LPC vocoder is illustrated in Figure 4.3. Speech is synthesized by exciting an LPC synthesis filter with either a periodic (voiced) or white noise (unvoiced) source. The periodic source consists of impulses spaced by the pitch period. Both the periodic and noise sources are scaled by an appropriate gain.

The speech encoder determines the LPC filter coefficients, the pitch, and a single voiced/unvoiced decision for each frame. These parameters are quantized and sent to the decoder. This type of vocoder is capable of sending intelligible speech at bit rates of 2400 bit/s and below.

The main drawback is that the synthesized speech has a mechanical quality, due to the simple excitation model. The LPC vocoder assumes speech to be either voiced or unvoiced. In practice speech often contains both voiced and unvoiced energy, which cannot be adequately modeled by this coder.

The LPC vocoder requires accurate estimation of the excitation model parameters, such as pitch and the voiced/unvoiced decision for each frame. This is a difficult task, which is further complicated when acoustic background noise is present.

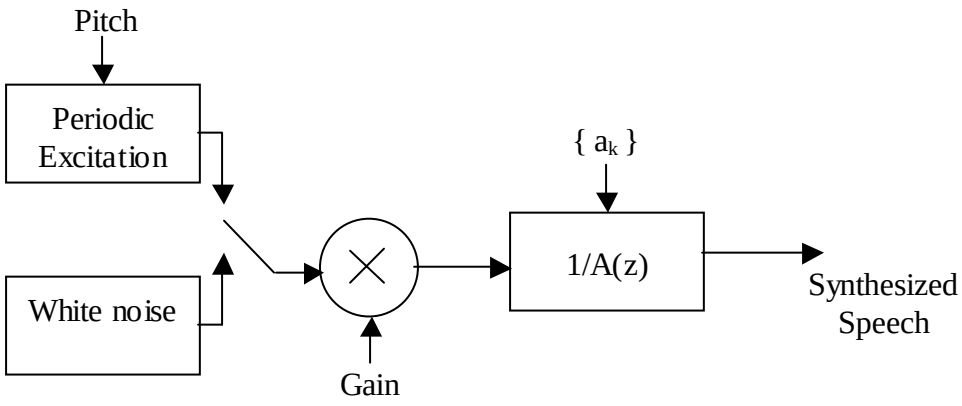


Figure 4.3 LPC Vocoder

Chapter 5: Pitch estimation and Pitch Detection Algorithms

The term “Pitch” corresponds to the name given to the fundamental frequency of a speech signal. This value can be easily seen within the semi-stationary speech waveform signal as the time interval from one peak to the next. In this chapter, applications, importance and difficulties in pitch estimation are discussed. The major classification of pitch estimation methods and traditional and recently developed pitch detection methods are described. Finally, advantages and disadvantages of existing pitch detection algorithms (PDAs) related to speech analysis will be presented.

5.1 Background

5.1.1 Applications of pitch estimation

Speech is used to communicate information from a speaker to a listener. In order to understand how the speech has been produced, many researches have been conducted studies of the human speech production procedure. Since pitch period ("fundamental frequency" or "f_0") provides information in speech that is important for comprehension and understanding, pitch period estimation has become an interesting problem and one of the most important problems in speech processing. There are several immediate applications and systems that require accurate pitch estimation including speaker identification and verification, vocoders, speech analysis, and

aids-to-the handicapped. The period variation over short-time (intrinsic period variation) carries phonemic, linguistic and speaker related information and produces the micromelody (for instance characteristic for plosive consonants). The macromelody (long-time variation) specifies the stress, the difference between a question, an exclamation and a statement and carries also information about the speaker identity. Some of the areas, which require pitch period estimation and its vital role in those areas, are presented in the following subsections.

5.1.2 Importance of pitch estimation in Speech coding:

In various speech coding algorithms, pitch estimation is an important parameter in the final quality of synthesized speech signal. For example, the effect of selecting multiples of correct pitch period, when modeling voiced speech, the spacing between the harmonics (fundamental frequency) is given by $2\pi/\tau$. Instead the second multiple of the pitch period is selected as a fundamental pitch, then the frequency spacing of the harmonics will be $2\pi/2\tau$, i.e., the spectrum will contain twice as many harmonics. This is very annoying for vocoders like LPC10. It produces very rough output speech. On the other hand, if the second sub-multiple is selected as correct pitch selected, then the fundamental frequency will be $2\pi/(\tau/2)$. In this case speech sound thin, e.g., male voice will sound similar to female voice. Since the synthesized quality in formant vocoder is depends on the location of the formats, the similar effect will be observed.

If the pitch change from frame-to-frame is not smooth then the speech produced will have lot of discontinuities. Therefore, not only the correct pitch period estimation, but also the frame-to-frame changes should be smooth.

The advantages of accurate pitch estimation in speech coding are numerous such as estimation of the lag (Γ) in the LTP for MLPC, RELP, CELP.

5.1.3 Difficulties in estimation of pitch estimation

Because of its importance, many solutions to this problem have been proposed. All of the proposed schemes have their limitations. Especially when speech is corrupted by a background noise, several methods yield different performance. A study can compare the efficiency of each scheme associated with various levels of disturbance.

The difficulty in pitch period estimation of speech signals is caused due to several reasons:

(i) The glottal excitation waveform is not a perfect train of periodic pulses. Although finding the period of a perfectly period of periodic waveform is straightforward, measuring the period of a speech waveform, which varies both in period and in the detailed structure of the waveform within a period, can be quite difficult.

(ii) In some instances, the formants of the vocal track can alter significantly the structure of the glottal waveform so that the actual pitch period is difficult to detect [34]. Such interactions generally are most deleterious to pitch detection during rapid movements of the articulators when the formants are also changing rapidly.

(iii) The reliable measurement of pitch is limited by the inherent difficulty in defining the exact beginning and end of pitch period during voiced speech segments.

(iv) Another difficulty in pitch detection is distinguishing between unvoiced speech and low level voiced speech. In many cases, transitions between unvoiced speech

segments and low level voiced speech segments are very subtle and thus are extremely hard to pinpoint

(v) In practical applications, the background ambient noise can also affect the performance of the pitch detector. This is especially serious in mobile communication environments where a high level of noise is present.

In spite of the difficulties in pitch measurement of speech signals, the traditional and recently developed Pitch detection Algorithms (PDAs) are broadly classified into the following two categories:

- Event pitch detectors
- Non-event pitch detectors

5.2 Non-Event pitch detectors

These pitch detectors are referred to as non-event pitch detectors, because they estimate the pitch period by a direct method.

5.2.1 Time domain waveform similarity method

The main principle of this method is based on waveform similarities. The goal is to find the pitch by comparing the similarity between the original signal and its shifted version. If the shifted version has been shifted by a delay equal to the pitch, the two signals should have maximum similarity. The majority of existing traditional PDAs are based on this concept. Among them, the most widely used are the auto-correlation (ACR) method and the average magnitude difference function (AMDF) method.

The most popular method to examine the similarity between two waveforms is

$$E(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}\left[s(n) - s(n+\tau)\right]^2 \qquad 5.1$$

where N is the analysis frame length and τ is the shifted distance. Since the average signal level of a speech is not fixed, the normalized similarity criterion is given by

$$E(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}\left[s(n) - \beta s(n+\tau)\right]^2 \qquad 5.2$$

where β is a scaling factor or pitch gain, controlling the changes in signal level.

5.2.1(a) Auto-correlation PDA:

Under the assumption that the signal is stationary, the error criterion of Equation 5.1 can be written as

$$E(\tau) = \left|R(0) - R(\tau)\right| \qquad 5.3$$

where $R(\tau) = \sum_{n=0}^{N-1} s(n)s(n+\tau)$

The minimization of the estimation error, E(τ), in Equation 5.3 is equivalent to maximizing the auto-correlation R(τ). The variable τ is called lag or delay and the pitch is equal to the value of τ, which results in the maximum R(τ). In actual implementation of auto-correlation PDA involves, determination of auto-correlation coefficients and then the difference between the locations of peaks of auto-correlation is the pitch as shown in the Figure 5.1b. The

advantage of ACPDA is that it is phase insensitive. Hence, it performs well in detecting the pitch of speech, which may suffer some degree of phase distortion.

5.2.1(b) Average Magnitude Difference Function PDA

The AMDF is also a direct similarity criterion, defined as

$$E(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}\left|s(n) - s(n+\tau)\right| \qquad 5.4$$

In contrast to the auto-correlation function, which is a measure of signal agreement, the AMDF measures the disagreement. Consequently, it is referred to as an *anti correlation measure or dissimilarity measure*. Figure 5.1c shows an example of AMDF function for the same speech signal.

The advantage of AMDF is that of computation simplicity, as the subtraction and magnitude computation structures are much faster than multiply-add structures. The AMDF PDA has been used in US government standard LPC10 vocoder [13]. Its advantage has been removed by the introduction of DSPs in mid 1980s, with a hardware multiplier and the pipelined multiply-add instruction. Nevertheless, the fact that the AMDF computation needs much less density of integration it is still an advantage in ASIC implementation. Another advantage is its relatively smaller dynamic range and narrower valleys for stationary signals, which makes pitch tracking make more applicable.

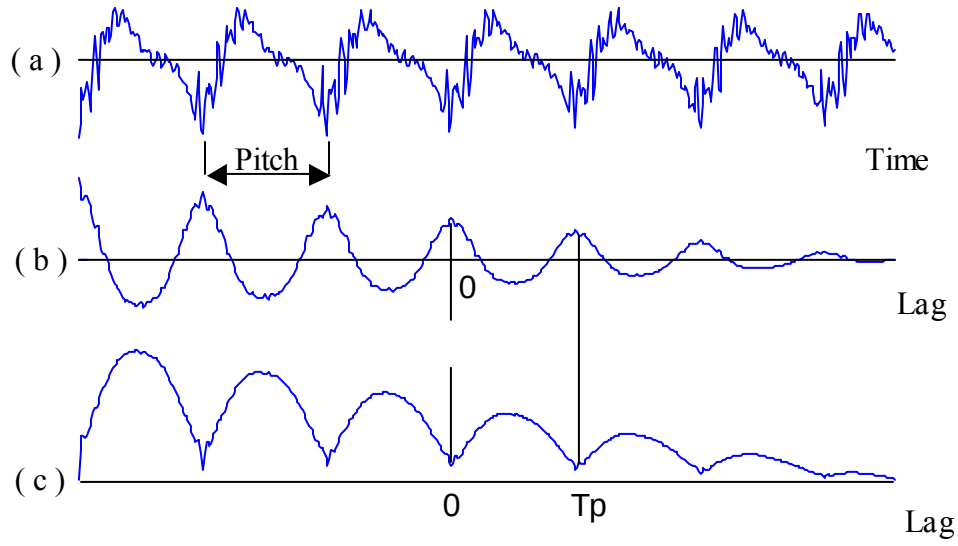


Figure 5.1 (a) Original Speech signal, (b) auto-correlation function and (c) AMDF

5.2.2 Frequency domain spectral similarity methods

Frequency domain PDAs directly operate on the speech spectrum. The main frequency domain feature of a periodic signal is harmonic structure, with distance between harmonics being the reciprocal of the pitch period. The main drawback of the frequency domain methods was their high computational complexity. However, with modern DSP techniques, a real-point transform only needs 0.5 ms in DSP32C implementation. This makes the implementation of the IMBE vocoder with a frequency domain PDA possible for INMARSAT-M land mobile communication standard [14]. In the following section, we briefly explain two frequency domain PDAs.

5.2.2(a) Harmonic Peak detection

An obvious way in determining the pitch in the frequency domain would be the extraction of the spectral peak at the fundamental frequency. This requires the first harmonic to be resent, which in general cannot be expected because of the front end filtering. A more

practical method is to detect the harmonic peaks and then measure the fundamental frequency (pitch frequency) as either the common divisor of these harmonics or the spacing of the adjacent harmonics. This can be done using a comb filter given by

$$C(\omega,\omega_0)=\begin{cases}W(k\omega_0); & \omega=k\omega_0, \quad k=1,2,\ldots\frac{\Omega_m}{\omega_0}\\ 0 \quad\quad ; & otherwise\end{cases} \tag{5.5}$$

and correlating it with the speech spectrum. The output of the correlation, $A_c(\omega_0)$, is the summation of the weighted comb peaks

$$A_c(\omega_0)=\frac{\omega_0}{\Omega_m}\sum_{k=1}^{\Omega_m/\omega_0}S(k\omega_0)W(k\omega_0) \qquad \frac{2\pi}{\tau_{\min}}\le\omega_0\le\frac{2\pi}{\tau_{\max}} \tag{5.6}$$

where Ω_m is the maximum frequency considered in the speech spectrum. If ω_0 is equal to the fundamental frequency, the comb response will match the harmonic peaks, and the maximum ouput will be obtained as show in the figure 5.2. To have a better subjective quality, a weighting coefficient can be applied to each individual tooth, normally decreasing weight with an increasing frequency [5]. This method may be considered as a maximum likelihood technique in the presence of additive noise, making it very robust.

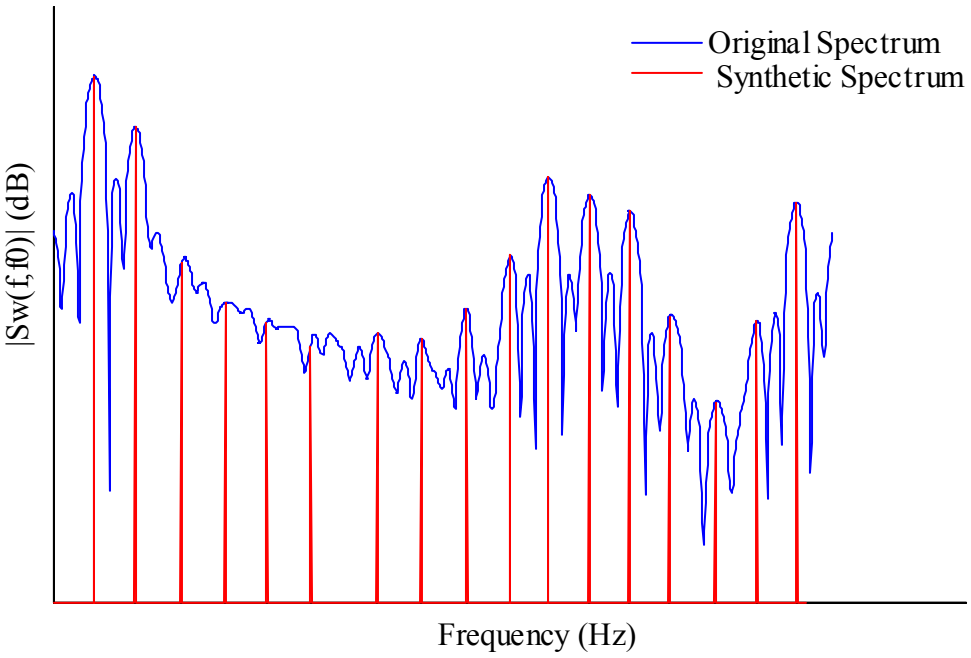


Figure 5.3 Original spectra and Synthetic spectra used in the Harmonic Peak PDA method

5.2.2(b) Spectrum Similarity

The spectrum similarity method determines the pitch by comparing the reconstructed spectrum with original speech spectrum. The error criterion used in this method is mean squared similarity criterion, given by

$$E(\omega_0) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|S_W(\omega) - \hat{S}_W(\omega,\omega_0)\right|^2 d\omega \qquad 5.7$$

where $\omega_0=2\pi/\tau$ and τ is the candidate pitch period, $S_W(\omega)$ is the original spectrum of windowed speech segment, and $S_W(\omega,\omega_0)$ is the reconstructed, pitch-dependent spectrum, which are defined by

$$S_W(\omega) = \sum_{n=-\infty}^{\infty} s(n)\omega(n)e^{-j\omega n} \qquad 5.8$$

$$\hat{S}_W(\omega,\omega_0) = \sum_{m=-M}^{M} A_m(\omega_0) W(\omega - m\omega_0) \qquad 5.9$$

where 2M+1 is the analysis window size,

$$W(\omega) = \sum w(n) e^{-j\omega n} \qquad 5.10$$

$$\text{and } A_m(\omega_0) = \frac{\int_{(m-0.5)\omega_0}^{(m+0.5)\omega_0} S_W(\omega) W^*(\omega - mw_0) d\omega}{\int_{(m-0.5)\omega_0}^{(m+0.5)\omega_0} |W(\omega - mw_0)|^2 d\omega_0} \qquad 5.11$$

This method was used in the MBE vocoder proposed by griffin in 1988 [15]. A typical original and synthetic spectra with correct pitch is shown in figure 5.3

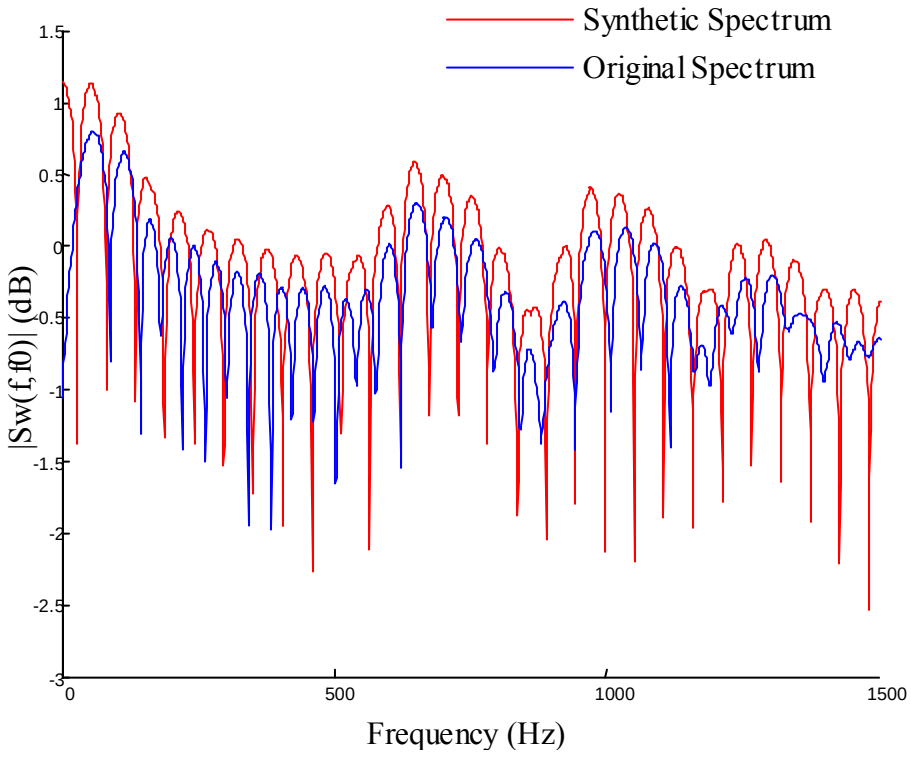


Figure 5.3 Original spectra and Synthetic spectra used in the spectrum similarity PDA method

5.2.2(c) Cepstrum peak detection

The method of separating the vocal tract information and excitation using the cepstrum analysis has already been discussed in the chapter [2]. The original time signal is transformed using a Fast Fourier Transform (FFT) algorithm and the resulting spectrum is converted to a logarithmic scale. This log scale spectrum is then transformed using the same FFT algorithm to obtain the power cepstrum. The power cepstrum reverts to the time domain and exhibits peaks corresponding to the period of the frequency spacings common in the spectrum. The mathematical expressions for estimation using cepstrum analysis are

Cepstral transform of the original signal is

$$C(\tau) = \left| IFFT\left(\log\left|FFT(s(n)\right|^2\right)\right| \qquad 5.12$$

and the fundamental frequency is estimated in the same way as in the autocorrelation method:

$$f_0 = \frac{1}{\tau_{\max}}, \quad \tau > 0 \qquad C(\tau_{\max}) = \max_{\tau} C(\tau) \qquad 5.13$$

Figures 5.4a – 5.4c show the original, log spectrum, and cepstrum respectively. It can be seen that the log spectrum has strong ripples in frequency, due to the pitch. A ripple peaks at about k ms in the cepstrum. This peak can be used as the basis of a powerful pitch detection system.

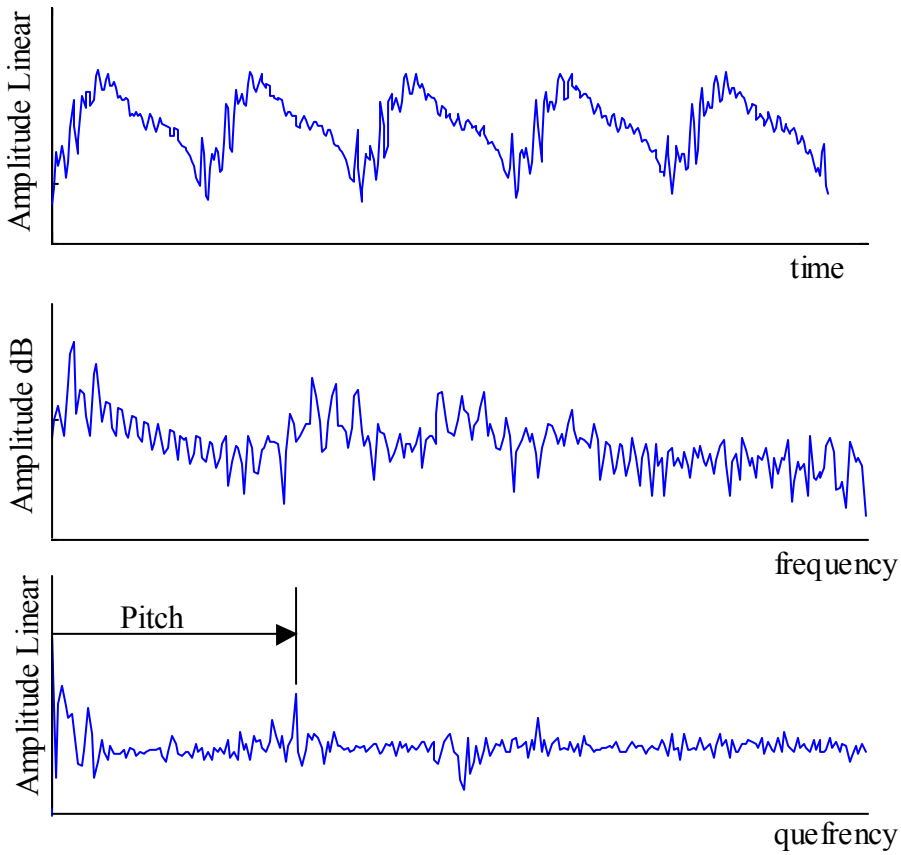


Figure 5.4 (a) Input Speech waveform (b) Log spectrum of the speech waveform and (c) Ceptsrum of the speech waveform

5.3. Event pitch detectors

Event detection pitch detectors estimate the pitch period by locating the instant at which the glottis closes (called an event) and then measuring the time interval between two such events.

5.3.1 Wavelet based PDAs:

Voiced speech is produced through the excitation of the vocal tract with quasi-periodic vibrations of the vocal folds at the glottis. These vibrations produce the glottis opening and

closing within each pitch period, the closing of the glottis being related to the maximum excitation energy. So that at this moment the glottal pulse exhibits a maximum, which after being distorted and delayed by the vocal tract is visible also in the speech waveform. The glottal closure instant (GCI) is a significant moment in a pitch period and was therefore treated as an event or epoch.

Pitch detectors that locate the GCI form the class of event pitch detectors and they were described in several papers**.** These detectors use the original speech signal or the LPC residual as input, the target of the processing being to obtain a waveform with clean maxima, easy to locate, so that the distance between two successive maxima estimates the *instantaneous pitch* period.

The wavelet based method uses the discontinuity of the speech signal at the time glottal closure. The main property of the D_yWT is, if a signal x(t) or its derivatives have discontinuities, then the modulus of the D_yWT of x(t), $|D_yWT(b,2^j)|$, exhibits local maxima around the points of discontinuity [8]. So that the distance between consecutive the local maxima gives the fundamental pitch period of the signals.

The D_yWT of a signal x(t) [16]

$$D_yWT(b,2^j) = \frac{1}{2^j}\int_{-\infty}^{\infty} x(t)g^*\left(\frac{t-b}{2^j}\right)dt \qquad 5.14$$

$$= x(t) \otimes g^*_{2^j}(t)$$

the following section describes the fundamental pitch period estimation using wavelet dyadic transform [8]

The D_yWT event pitch detection algorithm proceeds as follows. The D_yWT of a segment of a speech signal of length L ms is computed at some specified range of scales such as a = 2^3 to 2^5. For each scale, the local maxima that exceed a specified threshold are located with respect to parameter b of $D_yWT(b,a)$. For example, the threshold can be set equal to 80% of the global maximum of the D_yWT of the speech signal segment in [8]. Then, the locations of the local maxima across consecutive scales are compared. If the locations of the local maxima agree

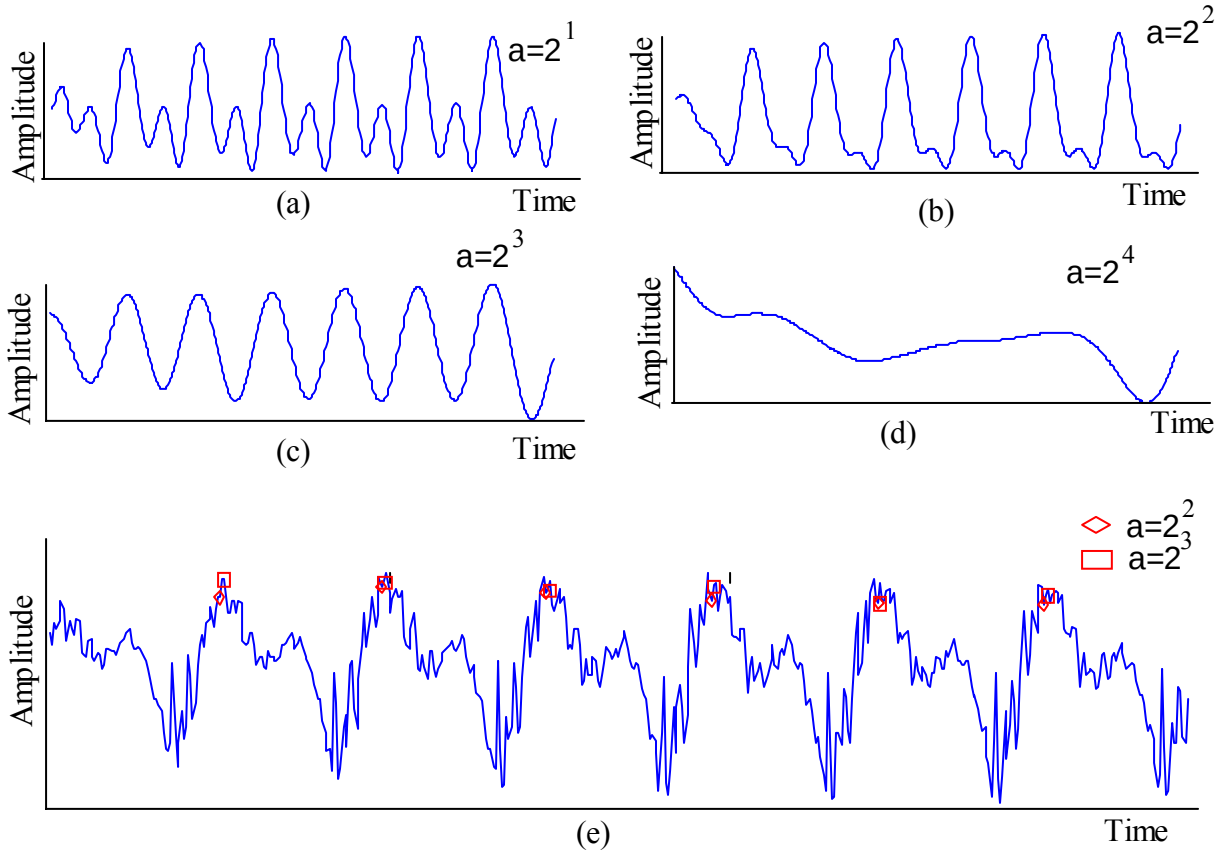


Fig 5.5 D_yWT of the part of the signal /do you/ spoken by the female speaker using SPLINE wavelet (a) computed with scale $a=2^2$ (b) computed with scale $a=2^3$ (c) computed with scale $a=2^4$ (d) computed with scale $a=2^5$ (e) Original signal with stars and square indicates the locations of local maximum which greater than 0.8 time global maximum

across two scales, it is assumed that the locations of these maxima correspond to the time of transients caused by the glottal closure. Finally, the pitch period is estimated by measuring the time interval between two such local maxima.

The method is well illustrated in figure 5.5. Figure 5.5c is the original speech signal waveform used to estimate the pitch period. Figure form 5.5a to 5.5d shows the wavelet transform of the speech signal. After careful examination of these figures, it can be seen that that number of local maxima and their location at scales a= 2^2 and 2^3 are identical. Therefore, the algorithm stops at scale a=2^3. The pitch period estimation is algorithm terminated and estimated by measuring the time interval between any two local maxima.

Chapter 6: A Novel Wavelet-Based Technique for Pitch Detection and Segmentation of Non-Stationary Speech

Pitch detection is an essential task in most speech coding techniques. Pitch is defined as the perceived fundamental period of a signal. Requirements for a successful pitch detection technique include robustness to noise, exactness of pitch estimation, and ability to provide the means for classification of speech segments into different categories, such as voiced/unvoiced (voicing decision).

One of the commonly used techniques for pitch detection is based on the ACR [1] function:

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) \qquad 6.1$$

where $x(m)$ is the original signal. At $k = 0$, ACR is maximum. The next peak's location identifies the signal's fundamental period. The ACR technique can be applied on $x(m)$ or on the residual signal e(n) obtained from LPC [2]. In the Cepstrum estimation technique [3], the log-spectrum of $x(m)$ is obtained and transformed back to time domain. The distance between peaks defines the fundamental frequency.

Most pitch detectors are insensitive to non-stationary pitch period variations over the given speech segment, and are unsuitable for low and high pitched speakers. Wavelets can estimate non-stationary pitch periods, and not simply averages of consecutive periods. In [4], a wavelet is chosen as the derivative of a smoothing function. Then, the local maxima of the wavelet transform identify abrupt changes or transients in speech caused by the glottal closure. The

algorithm attempts to find two consecutive scales at which the pitch period is equal. The distance between two successive local maxima equals the pitch period. MAWT's wavelet stage is more robust than [4] as shown in the results section.

Most traditional methods make voicing decisions using a threshold on features extracted from speech segments, mostly relying on single approaches, such as ACR, Cepstrum, maximum likelihood estimation (MLE) [5], or wavelets [4]. Thus, they are sensitive to the threshold selection and noise. Moreover, decisions are usually absolute; segments are classified as voiced/unvoiced, and sometimes, as transitional speech. One technique that uses multiple features is presented in [6]. MAWT combines multiple feature extraction, ACR, and an additional wavelet-based refinement step to solve the aforementioned problems. Features are used in a cascade. Each feature uses a relaxed threshold to classify segments as either one of the known categories (voiced/unvoiced) or as uncertain. If a feature fails to make a certain decision, consecutive features attempt to do so. Even if the final decision is uncertain, the wavelet stage partitions the signal appropriately as it is described in sections 2.4 and 3.

The letter is organized as follows. Section 2 introduces MAWT. In section 3, MAWT is compared with existing pitch detection techniques. Section 4 closes with some concluding remarks.

6.1. Proposed Technique

This section introduces MAWT, including feature extraction, ACR, the wavelet based stage, and the segmentation approach. The overall technique is presented in Figure 6.1.

6.1.1 The feature extraction and ACR stages

The features are used in a cascade. Each feature uses a relaxed threshold to classify segments as either voiced/unvoiced or as uncertain. If the feature fails to make a certain decision, consecutive features attempt to do so. The following features are used for the voicing decisions:

A. *Short-time Energy (STE)*: A low, relaxed threshold T_{STE} is used to separate the low STE unvoiced segments from the uncertain segments. Define $w(n)$ as a window of length L. Then, the STE is defined as:

$$STE = \sum_{n=-\infty}^{\infty} x^2(n)w(n) \tag{6.2}$$

B. *Zero-Crossing Rate (ZCR)*: Classification of high STE unvoiced segments, or silence with strong background noise may not be correct for a low T_{STE}. Although, the ZCR is larger for high STE unvoiced than voiced segments, noisy voiced segments may also have a high ZCR. The ZCR at time t is defined as:

$$\mathrm{ZCR}(t) = \sum_{m=-\infty}^{\infty} \left|\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\right| w(t-m) \tag{6.3}$$

Here, the difference between two ZCRs estimated with and without the use of a noise reduction filter is used. Generally, this difference is smaller for voiced compared to unvoiced segments. Thus, a large ZCR-difference threshold T_{ZCRD} identifies unvoiced segments in both noisy and clean speech.

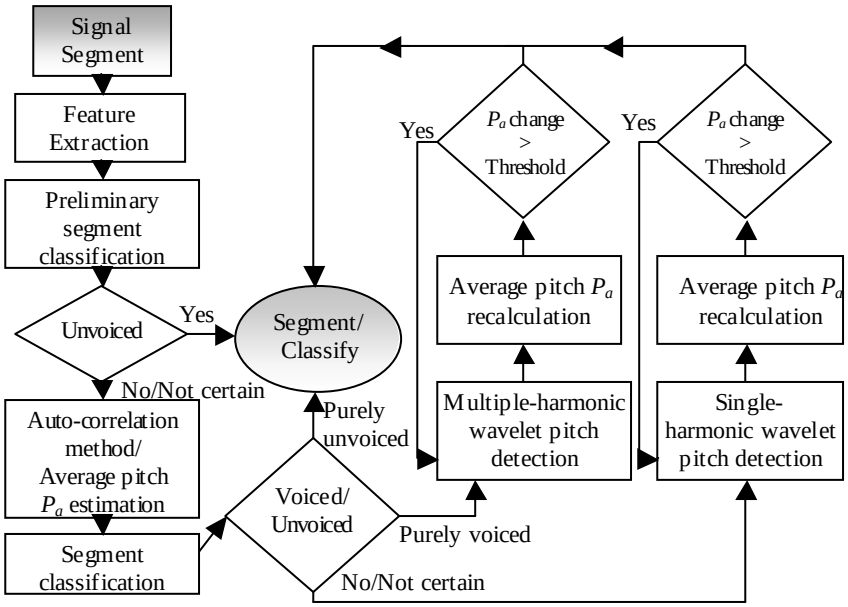
C. *Auto-Correlation (ACR)*: In this work, ACR is applied on the residual signal obtained from LPC. The following ACR-based features are used:

C1. *Percentage of Peak*: It is the ratio between ACR value at pitch period and total segment energy:

$$Per = \frac{\phi(P)}{\sum_{n=0} x^2(n)} \qquad 6.4$$

A low threshold $T_{PP\text{-}L}$ identifies unvoiced while a high threshold $T_{PP\text{-}H}$ voiced segments.

C2. *Pitch Tracking*: If the particular segment is voiced, the pitch period is expected to be within the range of either previous or future segments. A large difference between consecutive periods T_{PT} identifies unvoiced segments.



6.1.2 The wavelet based stage

The next stage is a wavelet based pitch period refinement step. In this work, the exponential wavelet $g(t,a)$ is used due to its optimal time-frequency localization:

$$g(t,a) = e^{-(t/a)^2} \cos(\omega_c t / a) \qquad 6.6$$

where a represents the scale. Experimental results presented in [4] illustrated that the spline wavelet was more suitable for pitch detection than $g(t,a)$. On the other hand, the method presented here is fundamentally different, and frequency-time localization is of more importance.

Here, an average pitch period P_a is already estimated in a speech segment $y_{seg}(t)$ using ACR. The wavelet stage allows estimation of a pitch period P taking into account non-stationary variations in $y_{seg}(t)$. Since P_a is known, the wavelet's frequency response $G(\omega,a)$ can be centered at the estimated fundamental frequency $\omega_o = 2\pi/P_a$. The frequency domain signal segment after filtering with $G(\omega,a)$ is:

$$Y_G(\omega,a) = G(\omega,a)\, Y_{seg}(\omega) \qquad 6.7$$

It is expected that the slowly varying pitch period in $y_{seg}(t)$ causes a widening of the peak at ω_o. Using Equation6.4, frequencies around ω_o are preserved, thus the varying fundamental frequency component contained in $Y_G(\omega,a)$ is retained. Then, P is estimated as the distance between two consecutive local maxima in the filtered time domain segment $y_G(t,a)$. In contrast to [4], estimation is based on finding consecutive local maxima whose distance is comparable to P_a. In order to assure isolation of the frequency components around ω_o without including significant harmonic information, the scale a is chosen as:

$$a = \omega_o/2 \qquad 6.8$$

Moreover, the wavelet's input central frequency ω_c is selected so that the wavelet's output frequency ω_c/a is equal to the fundamental $\omega_c/a = \omega_o$ or equivalently using Equation 6.8:

$$\omega_c = \omega_o^2 /2 \qquad 6.9$$

In order to make the algorithm more robust, multiple harmonics may be used in a similar manner. It is expected that the n^{th} harmonic can detect a period n times the actual pitch period P. Thus, P may be estimated as a weighted average of periods detected from several harmonics:

$$P = \sum_n C(n\omega_o)\, P_n \qquad 6.10$$

where P_n is the pitch period estimated from the wavelet centered at the n^{th} harmonic, and $C(n\omega_o)$ is defined as the signal's energy concentrated around the n^{th} harmonic:

$$C(n\omega) = \int_{n\omega_o-\varepsilon}^{n\omega_o+\varepsilon} Y_{seg}(\omega)d\omega \qquad 6.11$$

The multiple-harmonic approach is used only for certainly identified voiced segments. Otherwise, involvement of multiple harmonics may lead to artifacts.

6.1.3 Pitch detection under noisy conditions

An iterative procedure is used for further pitch refinement in order to consider noisy speech. Since the algorithm is ACR based, it is possible that the estimated P_a may not be accurate under noisy environment. Thus, after initial period P_a is found, the approach presented in 2.2 is applied iteratively: P_a is recalculated as the average of the estimated non-stationary P, and the center ω_c/a and scale a of the wavelet's Gaussian envelope are readjusted accordingly. The procedure stops if the P_a does not change significantly.

6.1.4 Advantages of MAWT in speech segmentation and modeling

In general, pitch estimation is meaningful only for voiced speech. Here, relaxed thresholds are used to avoid misclassification of speech segments, and an "uncertain" class is introduced. Thus, a "pitch period" may be estimated for unvoiced or transitional speech marked as "uncertain". Next, it is explained why this does not impose a negative effect on modeling, which is also illustrated in the results section.

a. Segments identified as purely unvoiced/voiced: Relaxed thresholds have been used, thus this classification is almost certainly correct. In particular, pitch detection is applied to voiced segments, while a pitch is not estimated for the unvoiced segments.

b. Voiced segments identified as uncertain: In this case, ACR has already assigned an average pitch period P_a to the segment. The wavelet stage refines P_a into a non-stationary period P as mentioned in 2.2.

c. Purely unvoiced segments identified as uncertain: Although this may be unlikely to occur, assuming the existence of a pitch period is not critical. An unvoiced segment can be modeled as a sum of sinusoids with known magnitude but random phase. MAWT segments the unvoiced region by preserving frequencies around the hypothetical fundamental frequency ω_h. The time domain local maxima correspond to sinusoids with frequencies around ω_h. Since the phase of the sinusoids changes randomly, the local maxima appear at random locations. Therefore, the segment is not modeled as a periodic signal.

d. Combination of voiced and transitional speech identified as uncertain: In this case, a strong periodic component due to the voiced sub-segment exists; therefore, the voiced part is successfully segmented. The transitional part may be loosely modeled as a gain-varying, oscillating signal with varying oscillation period (see Figure 3). The period of oscillations may not differ significantly from the pitch period in the voiced sub-segment. Hence, the transitional part is segmented into these oscillation periods. This result is appropriate since transitional speech sub-segments appear as resampled versions of each other.

e. Combination of unvoiced and transitional speech identified as uncertain: As mentioned above, transitional speech appears as a gain-varying, oscillating signal. Hence, a wide peak may exist in the spectrum around the frequency of oscillation. This frequency is identified as the assumed fundamental frequency ω_h. Thus, transitional speech is segmented as in *d*, while the unvoiced part is segmented as in *c*.

The advantages of MAWT are that (a) voicing decisions are not crucial, and (b) signals are not regarded as stationary. Modeling using MAWT could be as follows: the excitation signal for voiced and unvoiced segments is a sequence of impulses and white noise, respectively. In contrast to traditional modeling, in the case of voiced signals, an impulse is placed in each sub-

segment’s starting point. Thus, consecutive impulses are placed apart by a variable distance equal to the sub-segment's length. Consecutive sub-segments are grouped together based on similarity, and all sub-segments in the group are resampled to the average length. Then, LPC may be applied on the resampled sub-segments. The additional information required is the individual sub-segment’s gain and original length.

6.2. Results

Results illustrate the superiority of MAWT over three pitch estimation techniques, namely, ACR on the LPC residual signal, Cepstrum, and wavelet-based [4]. A noise reduction filter is used for MAWT, ACR and Cepstrum. The wavelet technique does not require such a filter since it uses low frequency wavelets.

Table 1 compares the four methods in terms of pitch detection error defined as $E_P = \sum|(P - P')| / P$, where P and P' are the true and estimated pitch periods, respectively. It is clearly shown that MAWT provides a significantly smaller E_P than all other techniques for all noise levels. The wavelet technique appears to be more robust to noise than Cepstrum and ACR.

Figure 2 presents comparison results between the wavelet method [4] and MAWT. Both techniques, in contrast to ACR and Cepstrum, can estimate non-stationary pitch. Nevertheless, MAWT is more stable. Figure 6.2(a) presents the fundamental frequency component obtained from MAWT’s wavelet stage, and Figure 6.2 (b) shows the original signal and the estimated pitch period, where “ ” indicates the start and the end. Figures 6.2(c) and 6.2(d) depict two consecutive scales of the wavelet transform for the same signal. Since both scales give the same pitch period, the algorithm converges. Figure 6.2(e) shows the estimated pitch period. Figure 6.2 illustrates that both techniques work properly for a strongly periodic segment with almost

constant gain. Although the periods' starting points are different for the two techniques what is of importance is the pitch period length and not the exact location.

Figure 3 presents similar results for a gain varying segment that partially contains transitional speech. While MAWT is successful (Figure 6.3(b)), the wavelet technique in [4] fails to converge, since there are no two consecutive scales for which the pitch period is the same; Figure 6.3(e) shows that the pitch period differs for the two consecutive scales represented by "+" and "o". Furthermore, Figure 6.3(b) illustrates that the transitional part using MAWT is appropriately segmented: the 800-length segment is split into variable length sub-segments. In particular, it can be clearly observed that sub-segments labeled as A to F appear to be resampled and attenuated versions of sub-segment G. This is even true for sub-segments A, B, C although they differ significantly from G in terms of gain and length. As mentioned in section 2, segmentation using variable lengths may lead to a new modeling approach in which standard coding techniques such as LPC may be used.

	20 dB	5 dB	0 dB	-5 dB
Proposed MAWT	1.89	3.93	3.48	6.49
ACR with LPC	2.44	8.70	6.00	16.84
Cepstrum	4.33	6.76	8.62	11.60
Wavelet Based	4.11	5.15	5.17	8.96

Table 1

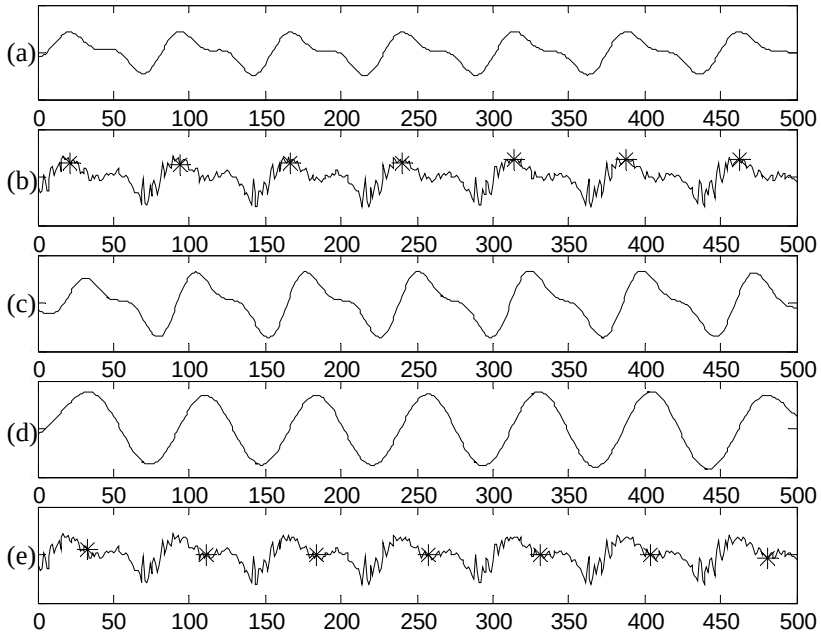


Figure 6.2

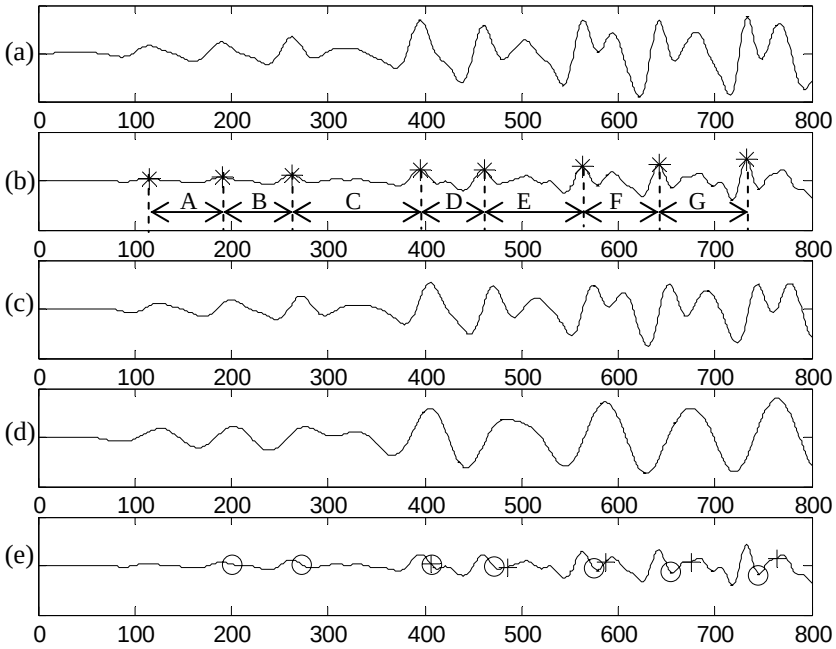


Figure 6.3

Conclusions

The importance of coding in speech communication has been presented. The basic theory behind speech coding, the various methods and the importance of pitch estimation in relation to speech coding have been discussed in detail. In addition, various traditional and recent pitch detection algorithms were presented.

A novel pitch detection and speech segmentation method has also been discussed in detail. It is shown that the proposed method is superior to previous techniques in terms of number of correct detections, percentage of pitch detection error, and stability. Moreover, the segmentation results illustrate that the resulted speech segments can contain non-stationary pitch sub-segments. Thus, the technique is also appropriate for fast changing and transition speech. Furthermore, this technique can lead to a novel approach for speech modeling. Traditional techniques can be used on the variable length speech sub-segments by resampling consecutive sub-segments into an average length.

Reference:

[1] J. R. Deller and Jr., J.G. Proakis, J.H.L. Hansen, "*Discrete-time Processing of speech signals*," Macmillan, New York, 1993.

[2] L. R. Rabiner and R. W. Scahfer, "*Digital Processing of Speech Signals*," Prentice-Hall, Inc., Enaglewodd Cliffs, New Jersey 07632

[3] A. M. Kondoz, "*Digital Speech: Coding for Low Bit Rate Communications Systems*," John Wiley and Sons, New York, 1994.

[4] F. A. Westall, R.D. Johnston and A. V. Lewis, "*speech Technology for Telecommunications*," BT telecommunication series, Chapman and Hall, New york, 1998.

[5] W. Hess, "Pitch *Determination of speech signals*," Springer Verlag, Berlin 1983.

[6] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, Vol. 63, No 4, 1975.

[7] A.M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, Vol. 41, pp. 293-309, Feb 1967.

[8] S. Kadambe and F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Transactions on Information Theory*, Vol. 38, No. 2, March 1992.

[9] J.D. Wise, J.R. Caprio, and T.W. Parks, "Maximum likelihood pitch estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 24, pp 418-423, Oct 1976.

[10] S. Ahmadi, and A.S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 3, pp. 333 –338, May 1999.

[11] O Rioul and M Vetterli, “wavelets and signal processing,” *IEEE Signal Processing magazine*, Vol. 8, No 4, pp 14-38, 1991.

[12] A V Oppenhiem and R W Schafer, “*Digital signal processing,*” Prentice Hall, New Jersey 1975.

[13] T. Tremain, “The government standard linear predictive coding algorithm: LPC-10,” *speech technology*, 1(2), pp 40-49, April 1982.

[14] DVSI, INMARSAT M voice codec. USA version 1.3, Feb 1991.

[15] D.W. Griffin and J. S. Lim, “Multi-band Excitation Vocoder,” *IEEE transactions on Acoustics, Speech and Signal Processing*, pp 418-423, Oct 1976.

[16] S. G. Mallat and S. Shong, “Compact Signal Representation with Multiscale Edges,” Technical Report, RRT-483-RR-219, Counrant Institute of Mathematics and Science, Dec 1989.

[17] Delprat M, Urie A and Evci C, “Speech coding requirements from the perceptive of the future mobile systems,” *Proceedings IEEE workshop on speech coding for Telecommunications*, pp 89-90, Oct 1993.

[18] ITU-T, 5-, 4-, 3-, and 2- bits per sample Embedded Adaptive Differential Pulse Code Modulation (ADPCM), Technical Report G.727, International Telecommunication Union, Geneva, 1990.

[19] ITU-T, Pulse Code Modulation (PCM), Technical Report G.711, International Telecommunication Union, Geneva, 1993.

[20] D. Kemp, R. A. Sueda and T. E. Tremain, “An Evaluation of 4800BPS voice coders,” *Proceedings of ICASSP*, pp 200-203, May 1989.

[21] J. Tribolet and R. Crocherie, “Frequency Domain coding of speech,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol 27, No 5, pp 512, Oct 1979.

[22] C. K. Un and D. T. Magill, “The residual-excited linear prediction vocoder with transmission rate below 9.6 kbits/s,” *IEEE transactions on communications*, Vol 23, No 12, pp 1466, Dec 1975.

[23] J. Chung and R. Schafer, “A 4.8 kbps homomorphic vocoder using analysis-by- synthesis excitation analysis,” *in proceedings ICASSP*, Vol 89, pp 122, 1989.

[24] ---, “excitation modeling in homomorphic vocoder,” *in proceedings ICASSP*, New Mexico, Vol-90, pp.25, Apr 1990.

[25] J. N.Homes, “The JRSU channel vocoder,” *Proceedings Institute of Electrical Engineers*, vol 27, pt. F no 1, pp, 53-60, Feb 1980.

[26] M. Schroeder and B. Atal, “ code excited linear prediction: High quality speech at low bit rates”, *IEEE transactions on Acoustics, Speech, and Signal Processing*, pp 247-254, 1979.

[27] B. Atal, J. Remde, "A New Model of LPC excitation for producing natural sounding speech at low bit rates," *Proc. ICASSP*, pp 614-617, 1982.

[28] P. Kroon and E.F. Deprettere, "A Class of Analysis-by-Synthesis Predictve Coders for High Quality Speech Coding at Rates Between 4.8 and 16 kbit/s," *IEEE Journal on Selected Areas in Communications*, Vol 6, pp 334-363, February 1988.

[29] J. H. Chen, “High quality 16bit/s speech coding with a one-way delay less than 2ms,” *proceedings of ICASSP*, pp 453-456, 1950.

[30] D. O’Shaughnessy, *Speech Communication: Human and Machine*. Addison-WesleyPublishing Company, 1987.

[31] ETSI TC-SMG, GSM 06.90 Version 7.1.0 Release 1998, Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi-Rate (AMR) speech transcoding, 1998.

[32] J. L. Flanagan, “*Speech Analysis, Synthesis and Perception*,” 2nd Ed, Springer-Verlag, New York, 1972

[33] W. Koenig, H. K. Dunn, and L. Y. Lacy, “The Sound Spectrograph,” *J Acoustic Society of America*, Vol 17, pp 19-49, July 1946.

[34] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, “A comparative performance study of several pitch detection algorithms,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, No 24(5), pp 399-418, Oct 1976.

[35] S. Singhal and B. Atal, “Aplitude optimization and Pitch prediction Multipulse coders,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp 317-327, Mar 1989.

[36] P. Kroon, E. Deprettere and R. Sluyter, “Regular-pulse Excitation: A novel approach to effective and efficient multipulse coding of speech,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp 1054-1063, Oct 1986.

Vita

Vijay Kura was born in India and received his B.Tech. in Electrical and Electronics Engineering form Jawaharlal Nehru Technological University at Hyderabad, India. He defended his Master's thesis in speech processing in October 2003. His research interests are mainly speech and Image signal processing, Power systems and Logic Design. He has a student membership in IEEE since 2001.