1-1-2003

Adaptive variable bit-rate speech coder for wireless applications

Junhong Chen
Ryerson University

Adaptive Variable Bit-rate Speech Coder for Wireless Applications

by

Junhong Chen

A Project
Presented to Ryerson University
in partial fulfillment of the
requirement for the degree of

Master of Engineering

in the Department of
Electrical and Computer Engineering

Toronto, Ontario, Canada

Author's Declaration

I hereby declare that I am the sole author of this Research Paper

I authorize Ryerson University to lend this Research Paper to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this Research Paper by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Abstract

In wireless communication, speech is an essential service for which low rate speech coding is a key technique. Therefore, reducing the transmission bandwidth and achieving higher speech quality are primary concerns in developing new speech coding algorithms. The goal of this report is to develop joint channel and source controlled adaptive variable bit-rate algorithm to achieve a high speech quality and maintain an efficient spectrum usage. To realize channel controlled bit rate variability, a new method of smooth variable frame length instead of the adaptive multi-rate speech coder (AMR) is introduced. In addition, the smooth bit rate switch concept is proposed. To realize source controlled bit rate variability, voice activity detection, novel voice/unvoiced segment detection and adaptive Forward-Backward quantizer algorithms are discussed. At last, the joint channel and source controlled bit rate variability algorithm has been evaluated in a CELP (FS1016) speech coder and the measurement result are presented.

Keywords: variable rate speech coding, source controlled coding, channel controlled coding, CELP, linear predictive coding, wireless communications

Borrow List

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign, and give address and date.

Acknowledgments

The completion of this master report would not have been possible without the valuable advice, continual guidance and technical expertise of my supervisor, Prof. Krishnan.

I am grateful to my fellow graduate students in the Signal and Image Processing Group for their stimulating discussions, companionship, and for creating a fruitful and pleasant work atmosphere. I also thank the other members of my committee for their valuable comments and feedback.

My gratitude goes to my wife Heng Wang for her love and understanding support and encouragement throughout my studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As we know, if speech is to travel the information highways of the future, efficient transmission and storage will be an important consideration. With the advent of the digital age, the analog speech signals can be represented digitally. There is an inherent flexibility associated with digital representations of speech. However, there are drawbacks such as a high data rate when no compression is used. Thus, speech coders are necessary to reduce the required transmission bandwidth while maintaining high quality. There is ongoing research in speech coding technology aimed at improving the performance of various aspects of speech coders.

From the primitive speech coders developed early in the twentieth century, the study of speech compression has expanded rapidly to meet current demands. Recent advancements in coding algorithms have found applications in cellular communications, computer systems, automation, military communications and biomedical systems. Although high capacity optical fibers have emerged as an inexpensive solution for wire-line communications, conservation of bandwidth is still an issue in wireless cellular and satellite communications. However, the bandwidth must be minimized while meeting other requirements discussed in the next section.

1.1 Attributes of Speech Coders

Given the extensive research done in the area of speech coding, there are a variety of existing speech coding algorithms. In selecting a speech coding system, the following attributes are typically considered:

• *Complexity*: This includes the memory requirements and computational complexity of the algorithm. In virtually all applications, real-time coding and decoding of speech is required. To reduce costs and minimize power consumption, speech coding algorithms are usually implemented on DSP chips. However, implementations in software and embedded systems are not uncommon. Thus, the performance of the hardware used can ultimately select among potential speech coding algorithms based on their complexity.

• *Delay*: The total one-way delay of a speech coding system is the time between a sound is emitted by the talker and when it is first heard by the listener. This delay comprises of the algorithmic delay, the computational delay, the multiplexing delay and the transmission delay. The algorithmic delay is the total amount of buffering or look-ahead used in the speech coding algorithm. The computational delay is associated with the time required for processing the speech. The delay incurred by the system for channel coding purposes is termed the multiplexing delay. Finally, the transmission delay is a result of the finite speed of electro-magnetic waves in any given medium.

In most modern systems, echo-cancellers are present. Under these circumstances, a one-way delay of 150 ms is perceivable during highly interactive conversations, but up to 500 ms of delay can be tolerated in typical dialogues [1]. When echo-cancellers are not present in the system, even smaller delays result in annoying echoes [2]. Thus, the speech coder must be chosen accordingly, with low-delay coders being employed in environments where echoes may be present.

• *Transmission bit rate:* The bandwidth available in a system determines the upper limit for the bit rate of the speech coder. However, a system designer can select from fixed-rate or variable-rate coders. In mobile telephony systems (particularly CDMA based ones), the bit rate of individual users can be varied; thus, these systems are well suited to variable bit-rate coders. In applications where users are allocated dedicated channels, a fixed-rate coder operating at the highest feasible bit rate is more suitable.

• *Quality:* The quality of a speech coder can be evaluated using extensive testing with human subjects. This is a tedious process and thus objective distortion measures are frequently used to estimate the subjective quality .The following categories are commonly used to compare the quality of speech coders: (1) commentary or broadcast quality describes wide-bandwidth speech with no perceptible degradations; (2) toll or wireline quality speech refers to the type of speech obtained over the public switched telephone network; (2) communications quality speech is completely intelligible but with noticeable distortion; and, (4) synthetic quality speech is characterized by its ‘machine-like’ nature, lacking speaker identifiableness and being slightly unintelligible. In general, there is a trade-off between high quality and low bit rate.

- *Robustness:* In certain applications, robustness to background noise and/or channel errors is essential. Typically, the speech being coded is distorted by various kinds of acoustic noise ▪ in urban environments, this noise can be quite excessive for cellular communications. The speech coder should still maintain its performance under these circumstances. Random or burst errors are frequently encountered in wireless systems with limited bandwidth. Different strategies must be employed in the coding algorithm to withstand such channel impairments without unduly affecting the quality of the reconstructed speech.

- *Signal bandwidth:* Speech signals in the public switched telephone network are bandlimited to 300 Hz ▪ 3400 Hz. Most speech coders use a sampling rate of 8 kHz, providing a maximum signal bandwidth of 4 kHz. However, to achieve higher quality for video conferencing applications, larger signal bandwidths must be used.

Other attributes may be important in some applications. These include the ability to transmit non-speech signals and to support speech recognition.

1.2 Classes of Speech Coders

Speech coding algorithms can be divided into two distinct classes: waveform coders and parametric coders. Waveform coders are not highly influenced by speech production models; as a result, they are simpler to implement. The objective with this class of coders is to yield a reconstructed signal that matches the original signal as accurately as possible. The reconstructed signal converges towards the original signal with increasing bit rate. However, parametric coders rely on speech production models. They extract the model parameters from the speech signal and code them. The quality of these speech coders is limited due to the synthetic reconstructed signal. However, as seen in Fig. 1.1, they provide superior performance for lower bit rates. Many waveform-approximating coders employ speech production models to improve the coding efficiency. These coders overlap into both categories and are thus termed hybrid coders.

1.2.1 Waveform Coders

Since the ultimate goal of waveform coders is to match the original signal sample for sample, this class of coders is more robust to different types of input. Pulse code modulation (PCM) is the simplest type of coder, using a fixed quantizer for each sample of the speech signal. Given the non-uniform distribution of speech sample amplitudes and the logarithmic sensitivity of the human auditory system, a non-uniform quantizer yields better quality than a uniform quantizer with the same bit rate. Thus, the CCIT standardized G.711 in 1972, a 64 kbs logarithmic PCM toll quality speech coder for telephone bandwidth speech.

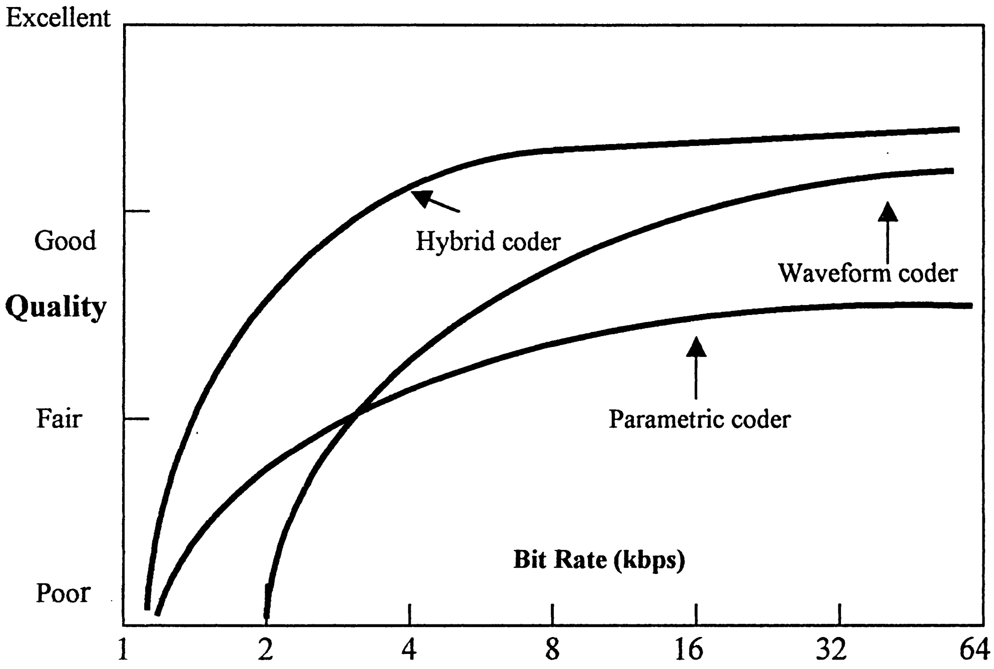


Fig. 1.1 Subjective performance of waveform and parametric coders. Redrawn from [1].

In exchange for higher complexity, toll quality speech can be obtained at much lower bit rates. With adaptive different pulse code modulation (ADPCM), the current speech sample is predicted from previous speech samples; the error in the prediction is then quantized. Both the predictor and the quantizer can be adapted to improve performance. G.727, standardized in 1990, is an example of a toll quality ADPCM system which operates at 32 kbs. Another possibility is to convert the speech signal into another domain by a discrete cosine transform (DCT) or another suitable transform. The transformation compacts the energy into a few coefficients which can be quantized efficiently. In adaptive transform coding (ATC), the quantizer is adapted according to the characteristics of the signal [3].

1.2.2 Parametric Coders

The performance of parametric coders, also known as source coders or vocoders, is highly dependent on accurate speech production models. These coders are typically designed for low bit rate applications (such as military or satellite communications) and are primarily intended to maintain the intelligibility of the speech. Most efficient parametric coders are based on linear predictive coding (LPC), which is the focus of this thesis. With LPC, each frame of speech is modeled as the output of a linear system representing the vocal tract, to an excitation signal. Parameters for this system and its excitation are then coded and transmitted. Pitch and intensity parameters are typically used to code the excitation and various filter representations are used for the linear system. Communications quality speech can currently be achieved at rates below 2 kbps with vocoders based on LPC [4].

1.2.3 Hybrid Coders

The speech quality of waveform coders drops rapidly for bit rates below 16 kbps, whereas there is a negligible improvement in the quality of vocoders at rates over 4 kbps. Hybrid coders are thus used to bridge this gap, providing good quality speech at medium bit rates. However, these coders tend to be more computationally demanding. Virtually all hybrid coders rely on LPC analysis to obtain synthesis model parameters. Waveform coding techniques are then used to code the excitation signal and pitch production models may be incorporated to improve the performance.

Code-excited linear prediction (CELP) coders have received a lot of attention recently and are the basis for most speech coding algorithms currently used in wireless telephony. In CELP coders, standard LPC analysis is used to obtain the excitation signal. Pitch modeling is used to efficiently code the excitation signal. Standardized in 1996, G.729 is a CELP based speech coder which produces toll quality speech at a rate of 8 kbps [5].

1.3 Wireless Channel Properties

Wireless communication applications and services have undergone enormous development recently due to the continuing growth of wireless communication, especially the emergence of 3G wireless network. However, wireless communication poses many challenges. It is known that the mobile wireless channel has limited bandwidth and is usually impaired due to multi-path fading, shadowing, inter-symbol interference and noise disturbances. So, compared to the wired links, the wireless channels are typically much more noisy and have a higher bit error rate [6]. As a result, random and burst errors can have devastating effect on speech quality. Typically, the channel error rate varies with the time varying channel environment.

Large-scale path loss

Large-scale path loss describes the variation in mean received signal strength as a function of distance from the transmitter. The Friis transmission equation gives the received power in a free space environment as follows:

$$P_r = P_t G_t G_r (\frac{\lambda}{4\pi R})^2 \tag{1.1}$$

where *Pr* is the received power, Pt is the transmitted power, *Gt* is the gain of the transmitting antenna, and *Gr* is the gain of the receiving antenna. The remaining term is the inverse of the path loss, and accounts for spherical spreading loss of the transmitted wave due to propagation over the transmit-receive distance *R*. So, the strengths of the waves as the distance between the transmitter and receiver increases.

Shadowing

At a given distance from the transmitter, variations about the mean path loss will occur due to obstruction by objects in the environment. This can be modeled using a lognormal distribution about the mean value of large-scale path loss that is predicted by a distance-dependent model like the one in (1.2).

$$PL = \left(\frac{4\pi d_0}{\lambda}\right)\left(\frac{R}{d_0}\right)^r \tag{1.2}$$

where the first term is the free-space path loss at reference distance d_0 and the exponent Y(sometimes n is used) is determined empirically by a curve fit to measured data.

Multipath effects: fading, intersymbol interference, and Doppler spread

One of the distinctive features of a mobile radio channel is multipath propagation, in which the received signal consists of multiple reflected, diffracted, and scattered components, as well as (possibly) a direct line-of-sight component. Because all these components travel different distances and encounter different reflections, their phases are different. The relative phases of the received signals change as the mobile moves. Depending on the relative phases of the signals, they can reinforce each other or cancel each other. In the latter case a fade results. As the receiver is moved the received signal power undergoes variations, resulting in a fading envelope that can be measured.

Diversity systems that use signals received by two or more antennas can combat this effect.

The difference in path length between multipath components causes them to arrive with different delays. This causes intersymbol interference in digital systems, if the difference is significant in relation to the symbol period. The amplitudes of the multipath components also differ because they undergo different path losses. The received signal can be represented as a superposition of all the received components as follows (1.3)

$$x(t) = \sum_{n=1}^{N} \alpha_n(t) e^{j\theta_n} s[t - \tau_n(t)] \tag{1.3}$$

where *x(t)* is the received signal, $\alpha_n(t)$ is s the time-varying attenuation coefficient of the n path, $e^{j\theta_n}$ is time varying phase shift associated with the nth path, *s(t)* is the original transmitted signal, and $\tau_n(t)$ is the time-varying delay of the *n*th path

1.4 Report Contribution

This report focuses on improving the performance of variable bit rate speech coders for achieving very high capacity while maintaining an acceptable level of speech quality in wireless communication network. Due to the wireless channel property which we discussed before, the new Adaptive Multi rate speech coding (AMR) allows almost wire-line speech quality even for poor channel conditions by dynamically splitting the gross bit rate between source and channel coding according to the channel quality. In 1999, 3GPP released a speech coding standard for the WCDMA-Adaptive Multi-Rate (AMR) vocoder. The standard consists of a multi-mode variable rate coder and a source controlled rate scheme including a voice activity detector, a comfort noise generation system. However, multi-mode variable rate coder consists of many different mode of encoding for each bit rate option. This method increases the complexity of encoding and limited multi-mode can not exactly track channel condition.

In this report, we introduce a novel approach to yield the multi-rate performance without increasing complexity and also to achieve smooth and gradual switch between different bit rate. Our method is based on varying frame length in order to change bit rate while keeping the same bit number per frame. Large frame length means more average, less bit rate; Small frame length means less average, more accurate tracking of voice characteristic and more bit rate. In effect, this method can easily realize variable bit rate by simply varying frame length; also, with our new smooth switch algorithm, it can completely remove any artifact due to sharp variation in bit switch and accurately characterize channel condition.

For source controlled variable rate, besides the voice activity detection (VAD), we introduce different bit allocation for voice and unvoiced segment based on the algorithm in which more bits are allocated for voice segment and less bits for unvoiced segment. To enhance voice/unvoiced segment detection against noise, we propose a novel spectral correlation algorithm in frequency domain with adaptive threshold to decide voiced and unvoiced segment. With this improved scheme, it is more robust against noise even under high noise condition. Secondly, we modify an adaptive forward-backward quantizer algorithm by using mean variance to detect the similarity of the current and previous LPC coefficient instead of calculating LSD to reduce the computation complexity and this further reduce the bit rate. With the above source controlled variable rate algorithms, the bit rate can be significantly reduce with little degradation of speech quality. Finally, we implement our scheme with standard CELP (FS1016) speech coder and investigate the performance.

1.5 Report Organization

The report is organized as follows: The fundamentals of Code excitation linear predictive (CELP) speech coders are reviewed in Chapter 2. Conventional methods of short-term linear prediction, long-term linear prediction and stochastic codebook search by analysis and synthesis approach are presented. Some basic speech quality measures used to evaluate the performance of speech coders are overviewed. Chapter 3 introduces the idea of using a variable frame length to achieve channel controlled variable bit rate with a smooth switch algorithm and it compare with the current adaptive multi-rate speech coder (AMR). The source controlled variable bit rate speech coding based on speech characteristic is presented and analyzed in Chapter 4 including two novel schemes of robust voice/unvoiced segment detection algorithm and modified adaptive forward-backward quantizer. These algorithms can significantly reduce bit rate with little degradation of speech quality. The integrated algorithms comprising of channel and source controlled variable bit coding is then implemented in a CELP speech coder (FS1016) and the simulation results are presented in Chapter 5. Also, the report is concluded with a summary of work and some suggestions for future work.

Chapter 2

Code Excitation Linear Predictive Speech Coding

Most current speech coders are based on LPC analysis due to its simplicity and high performance [2]. This chapter provides an overview of LPC analysis for speech application. Among many LPC speech coders, code excited linear predictive (CELP) coder is one of popular speech coders in wireless communication [1]. Therefore, standard LPC analysis, long term adaptive code book and stochastic codebook search in CELP coders are introduced in this chapter. Also, quality measures used to measure the performance of speech coding algorithms are examined.

2.1 Speech Production Model

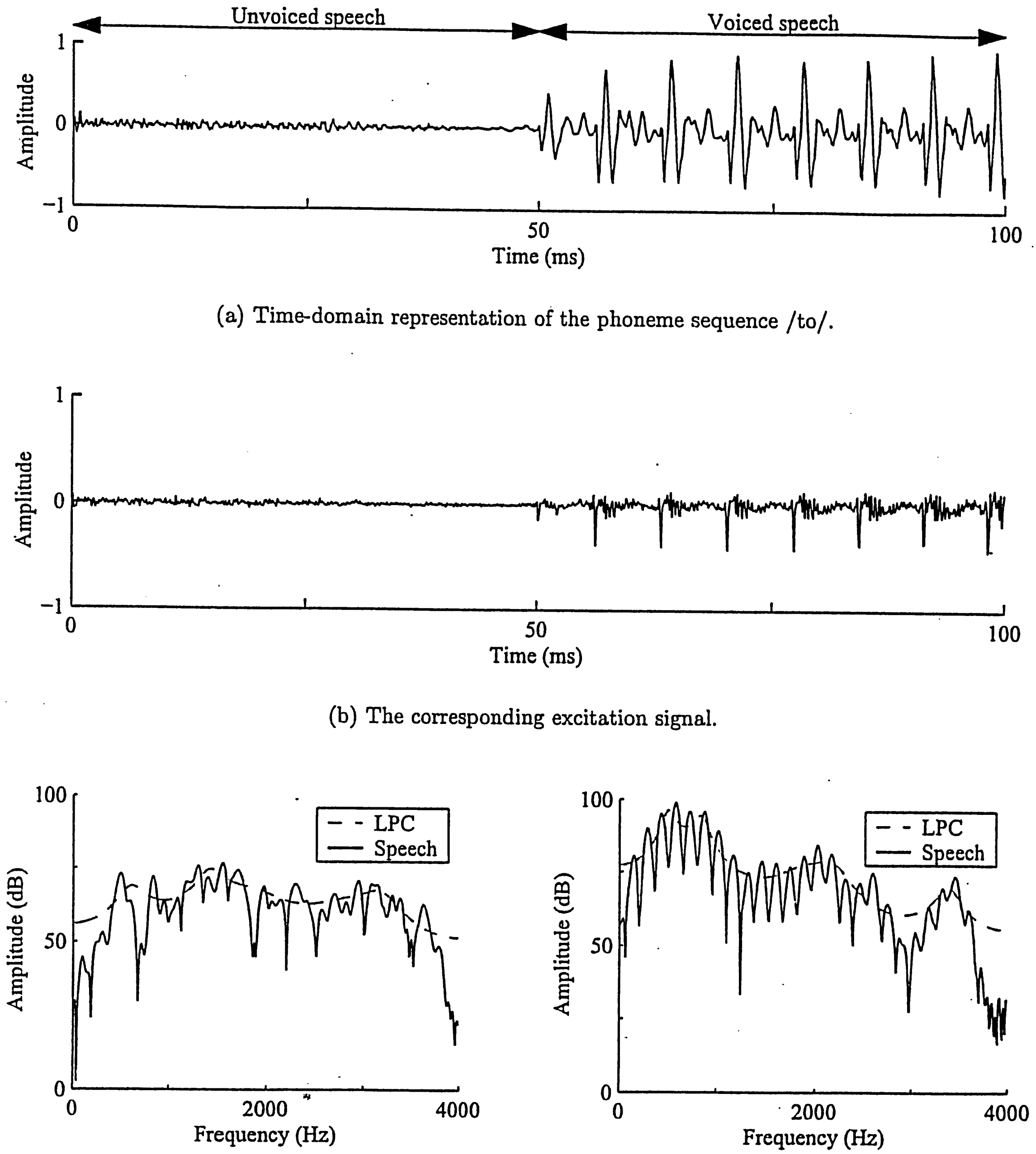
Due to the inherent limitations of the human vocal tract, adjacent samples of the speech signals are highly redundant. These redundancies allow speech coding algorithms to compress the signal by removing the irrelevant information contained in the waveform. Knowledge of the vocal system and the properties of the resulting speech waveform is essential in designing efficient coders. The properties of the human auditory system, although not as important, can also be exploited to improve the perceptual quality of the coded speech. Speech consists of pressure waves created by the flow of air through the vocal tract. These sound pressure waves originate in the lungs as the speaker exhales. The vocal folds in the larynx can open and close quasi-periodically to interrupt this air-flow. This results in voiced speech (e.g., vowels) which is characterized by its periodic and energetic nature. Consonants are an example of unvoiced speech — aperiodic and weaker; these sounds have a noisy nature due to turbulence created by the flow of air through a narrow constriction in the vocal tract. The positioning of the vocal tract articulators acts as a filter, amplifying certain sound frequencies while attenuating others

A time-domain segment of voiced and unvoiced speech is shown in Fig. 2.1(a).

A general linear discrete-time system to model this speech production process, known as the terminal-analog model [4], is shown in Fig. 2.2. In this system, a vocal tract filter *V(z)* and radiation model *R(z)* (to account for the radiation effects of the lips) are excited by the discrete-time excitation signal $u_G[n]$. The lips behave as a first order high-pass filter and thus *R(z)* grows at 6 dB/octave. Local resonance and anti-resonances are present in the vocal tract filter, but *V (z)* has an overall flat spectral trend. The glottal excitation signal $U_G[n]$ is given by the output of a glottal pulse filter *G(z)* to an impulse train for voiced segments; *G(z)* is usually represented by a 2nd order low-pass filter, falling off at 12 dB/octave. For unvoiced speech, a random number generator with a flat spectrum is typically used. The z-transform of the speech signal produced is then given by:

$$s(z) = \theta_0 U_G(z) V(z) R(z) \tag{2.1}$$

where $U(z) = \theta_0 E(z)$ is the gain adjusted excitation signal. Fig. 2.1(b) shows the estimated excitation signals for voiced and unvoiced speech segments using a 10th order all-pole filter for *H(z);* the autocorrelation method was used with a 25 ms Hamming window (see Section 2.3). Note that the excitation signal for the unvoiced speech segment seems like white noise and that for the voiced speech closely resembles an impulse train. The power spectra for voiced and unvoiced speech are shown in Fig. 2.1(c) with the corresponding frequency responses of the vocal tract filter *H(z).* The periodicity of voiced speech gives rise to a spectrum containing harmonics of the fundamental frequency of the vocal fold vibration. A truly periodic sequence, observed over an infinite interval, will have a discrete-line spectrum but voiced sounds are only locally quasi-periodic.



(a) Time-domain representation of the phoneme sequence /to/.

(b) The corresponding excitation signal.

(c) The power spectrum (solid line) and LPC spectral envelope (dashed line) of the unvoiced segment (left) and voiced segment (right).

Fig.2.1 An unvoiced to voiced speech transition, the underlying excitation signal and short-time spectra. [2]

The resonance evident in the spectral envelope of voiced speech, known as formants in speech processing, are a product of the shape of the vocal tract. The -12 dB/octave for *E(z)* gives rise to the general -6 dB/octave spectral trend when the radiation losses from *R(z)* are considered. The spectrum for unvoiced speech ranges from flat spectra to those lacking low frequency components. The variability is due to place of constriction in the vocal tract for different unvoiced sounds ▪ the excitation energy is concentrated in different spectral regions.

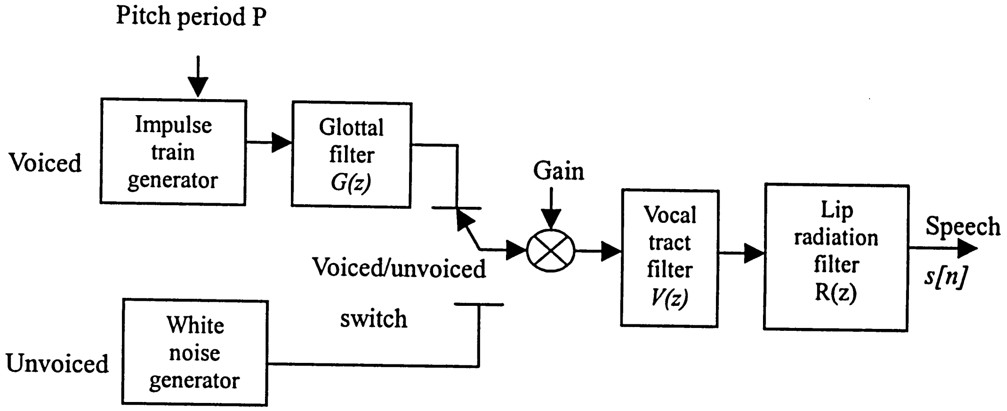


Fig. 2.2 The terminal-analog model for speech production [4]

Due to the continuous evolution of the shape of the vocal tract, speech signals are non-stationary. However, the gradual movement of vocal tract articulators results in speech that is quasi-stationary over short segments of 5 to 20 ms. This slow change in the speech waveform and spectrum is evident in the unvoiced-voiced transition shown in Fig. 2.1. However, a class of sounds called stops or plosives (e.g., /p/, /b/, etc.) result in highly transient waveforms and spectra. An obstruction in the vocal tract allows for the buildup of air pressure; the release of this vocal tract occlusion then creates a brief explosion of noise before a transition to the ensuing phoneme. The resulting transient waveform, such as the one shown in Fig. 2.3, generally poses difficulty to speech coders which operate under the assumption of stationarity over frames of typically 10 to 40 ms. Another class of sounds that typically impedes the performance of speech coders is voiced fricatives. The excitation for these sounds consists of a mixture of voiced and unvoiced elements, and thus the vocal tract model of Fig. 2.2 does not provide an accurate fit to the actual speech production process.

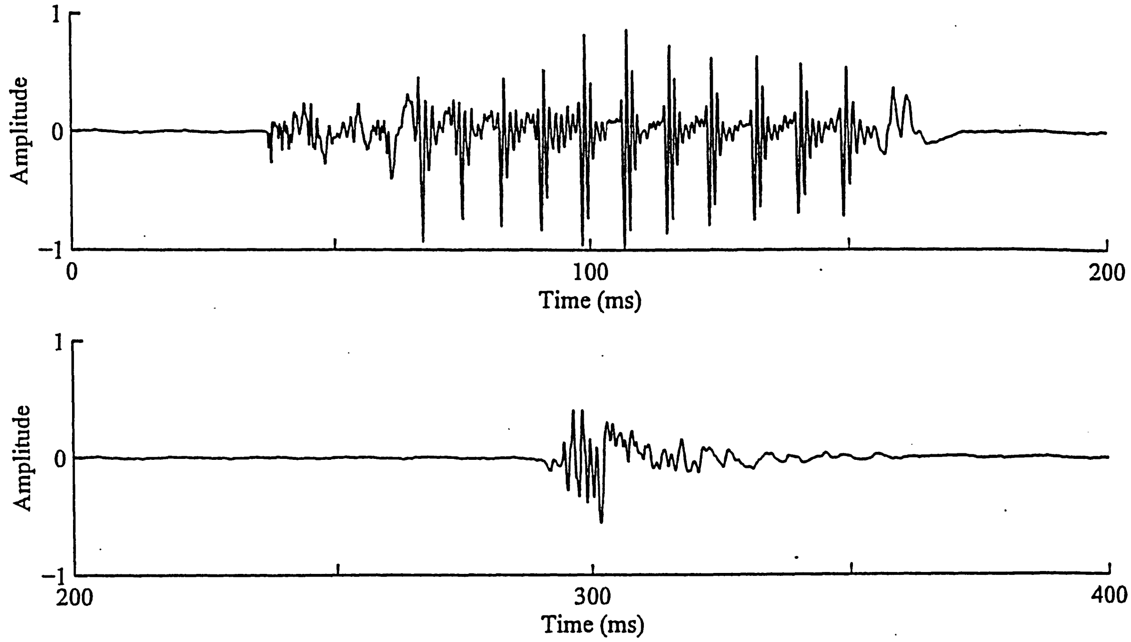


Fig. 2.3 The time-domain waveform of the word 'top' showing the transient nature of the plosives /t/ and /p/.

2.2 Linear prediction speech coding

LPC starts with the assumption that a dynamic speech signal can be viewed as a stationary waveform for short period of time, in other words, the biological speech forming mechanisms remain constant during this short period of time. This mechanism is modeled as a buzzer at the end of a tube. The space between the vocal cords, glottis, produces the buzz, which is characterized by its intensity and frequency (pitch). The vocal tract (the throat and the mouth) forms the tube, which is characterized by its resonance, which are called formants.

The most general predictor form in linear prediction is the autoregressive moving average (ARMA) model where a speech sample *s[n]* is predicted from p past predicted speech samples *s[n - 1],..., s[n - p]* with the addition of an excitation signal *u[n]* according to:

$$s[n]=\sum_{k=1}^{p}(a_k).s[n-k]+G\sum_{l=0}^{q}(b_l).u[n-l] \quad , \qquad b_0=1 \qquad (2.2)$$

where G is a gain factor for the input speech and a_k and b_l are sets of filter coefficients.

Equivalently, in the frequency domain, the transfer function of the linear prediction speech model is

$$H(z) = \frac{B(z)}{A(z)} = G\frac{1+\sum_{k=1}^{p} b_l.z^{-l}}{1-\sum_{k=1}^{p} a_k.z^{-k}} \tag{2.3}$$

H(z) is referred to as a pole-zero model in which the polynomial roots of the denominator and the numerator are, respectively, the poles and zeros of the system. When $a_k = 0$ for $1 \le k \le p$, *H(z)* becomes an all-zero or moving average (MA) model since the output is a weighted average of the q prior inputs. Conversely, when $b_l = 0$ for $1 \le l \le q$, *H(z)* reduces to an all-pole or autoregressive (AR) model in which case the prediction operation is written as:

$$s[n] = \sum_{k=1}^{p} (a_k).s[n-k] \tag{2.4}$$

and its frequency domain transfer function simply as:

$$H(z) = \frac{G}{1-\sum_{k=1}^{p} a_k.z^{-k}} = \frac{1}{A(z)} \tag{2.5}$$

The spectral pattern can be modeled by *1/A(z).*

Prediction error is: $E = \sum_{n} (S[n+r] - \sum_{k=1}^{p} a_k.z^{-k})$ (2.6)

The following figure illustrates the block diagram of prediction

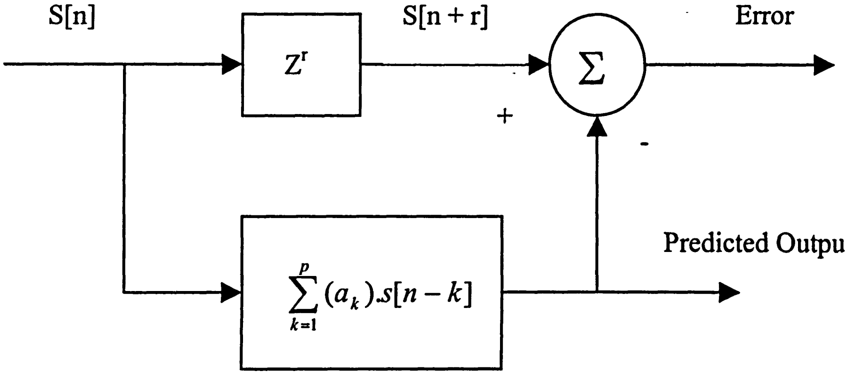


Fig. 2.4 Block diagram of prediction

If r = 0 the predictor attempts to match the present value, If r > 0 it tries to predict a future value of x [n].

LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of formants is called inverse filtering, and the remaining signal is called the residue. The number that describes the formants and the residue can be stored or transmitted somewhere else. LPC synthesizes the speech signal by reversing the process: use the residue to create a source signal, use the formants to create an all-pole filter (which represents the tube), and run the source through the filter, resulting in speech.

2.3 CELP speech coder

The most widely documented scheme for speech coding operating at under 8 kbps is Code Excited Linear Prediction (CELP) technique, which compresses sampled speech by analysis by-synthesis of incoming speech. Opposed to waveform encoding, the goal of CELP algorithm is to code and transmit the minimal amount of information necessary to synthesize speech which is audibly perceived to be accurate. In 1991, the US General Services Administration published the Federal Standard 1016 (FS1016) [7], which specifies the conversion of analog voice to digital data by a method of 4.8 Kbps CELP. In this section, we will introduce the CELP algorithm based on FS1016 which is made up of three major parts: *short-term linear prediction*, *long-term adaptive codebook search* and *stochastic codebook search*.

The input voice stream to FS1016 is segmented into frames of duration 30ms. Each frame is in turn divided into four subframes of length 7.5ms. Accordingly, with the sampling rate of 8 kHz, there are 240 voice samples per frame, and 60 samples per subframe. When speech is being analyzed, short-term linear prediction is first performed over the entire speech frame to extract the 10 linear prediction coefficients. Afterwards, long-term adaptive codebook search and stochastic codebook search are applied in sequence to each of the four subframes. The procedure for speech synthesizing is simply a reverse of the speech analyzing. (Fig. 2.5)

2.3.1 Short-Term Linear Prediction

In the short-term linear prediction, the speech signal is reasonably assumed stationary within a small observation window [8]. An infinite impulse response (IIR) filter of order ten, named *Linear Predictive Coding* (LPC), is used to track the spectrum envelope of the original speech signal (Fig. 2.5), which emulates the filtering effect of the vocal tract. The formula of the LPC filter is

$$F_{LPC}(z) = \frac{1}{A(z)} \tag{2.7}$$

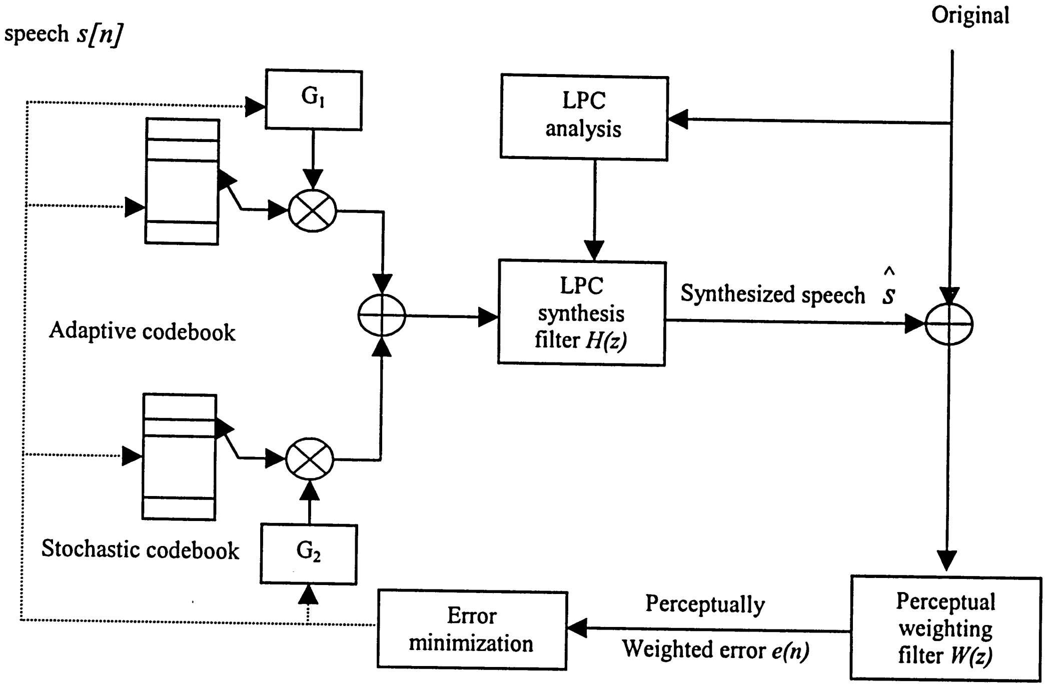


Fig 2.5 CELP analyzer.

where

$$A(z) = 1 - \sum_{j=1}^{10} a_j z^{-j}$$

(2.8)

In time domain, the synthesized output speech signal $s^{(syn)}(n)$ (i.e., synthesizing the true speech from the lips) due to the excitation input $r(n)$ (i.e., signals from the glottal excitation) via the filter can be expressed by

$$s^{(syn)}(n) = r(n) + \sum_{j=1}^{10} a_j s^{(syn)}(n-j) \tag{2.9}$$

Notably, at this stage, the functional blocks for adaptive codebook search and stochastic codebook search in Fig. 2.1(b) are both disabled; hence, $r(n)$ simply emulates an impulse excitation input to the Linear Predictor filter. The objective of the linear prediction is to make the best estimate of the 10 linear coefficients $\alpha_1, \alpha_2, \ldots, \alpha_{10}$ such that the true speech $s(n)$ can be well-approximated by the synthesized speech $s^{(syn)}(n)$ under an impulse glottal excitation input $r(n)$.

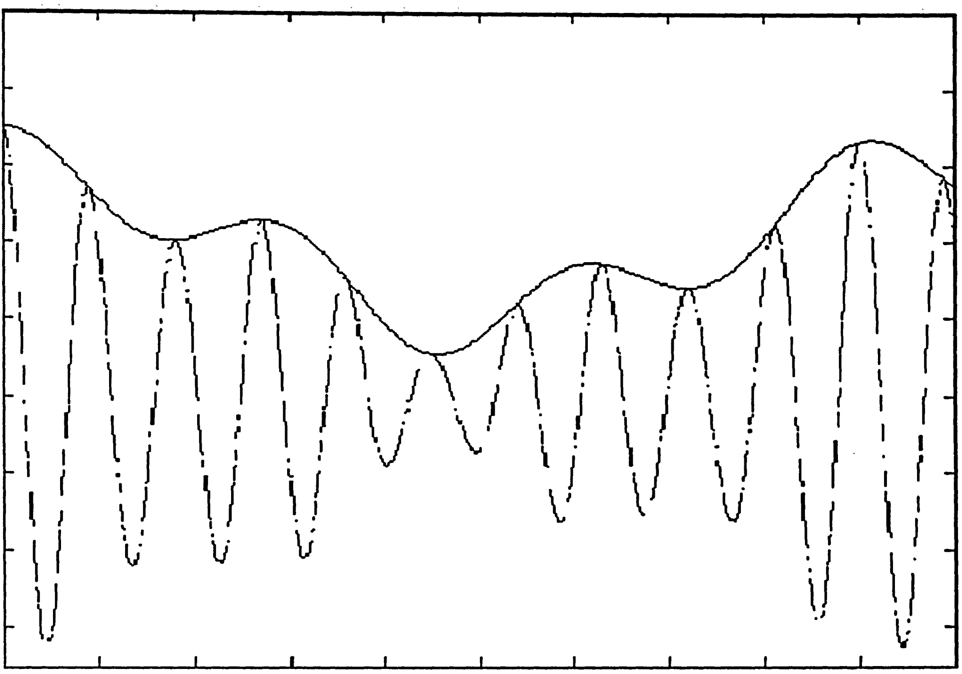


Fig 2.6 Spectral envelop of the original speech signal

To cope with the human-ear perpetual effect on speech, an "adjustment" on the synthetic speech, as well as the true speech, is performed before the optimization of the 10 linear coefficients, which is named the *Perpetual Weighted Filter.* Hence, the residual signal $e(n)$ is equal to $w(n)*[s(n)*s^{(syn)}(n)]$, where $w(n)$ is the impulse response of the perpetual weighted filter, and "*" denotes the convolution operation. As a result of the adjustment of the perpetual weighted filter, the best estimation of $\alpha_1, \alpha_2, \ldots, \alpha_{10}$ is defined based on the minimization of *e(n).*

$$
\begin{aligned}
\sum_n e^2(n) &= \sum_n \left[w(n) * (s(n) - s^{(syn)}(n))\right]^2 \\
&= \sum_n \left[w(n) * (s(n) - \sum_{j=1}^{10} \alpha_j s^{(syn)}(n-j))\right]^2
\end{aligned}
\qquad (2.10)
$$

Since the 10 LPC coefficients are computed on a frame basis by an open-loop estimation, its computational complexity, when being compared to the adaptive and stochastic codebook searches, is quite low. The challenge for this step is the quantization loss. As only limited number of bits is reserved for each coefficient, additional quantization errors are unavoidably introduced to the residual signal. In addition, the quantized LPC coefficients by no means guarantee the stability of the resultant IIR filter.

In order to amend the above problems, FS1016 chooses to quantize the *Line Spectral Frequency* (LSF) coefficients, which are conceptually one-to-one mappings of the LPC coefficients [9]. The LSFs are roots of a system function, and are located on the unit circle in the *z*-domain. Thus, they have the same amplitude, and are only different in their phases. The quantization applied to the LSFs therefore only affects the resultant phases. As a result, the stability of the system (filter) is less vulnerable through quantization.

Although for most of the time, the speech is short-term or frame-wisely stationary in its nature, it is still possible that the coefficients obtained from two consecutive frames are quite different. In order to generate a smooth speech at the decoding phase, FS1016 specifies a weighted interpolation to make a gentle migration in LSF coefficients between two consecutive frames. As illustrated in Fig. 2.7, the computed LPC coefficients for the proceeding frame (frame *i*) and the next frame (frame *i* + 1) are transformed to their equivalent LSF coefficients, f_{1i}, f_{2i}, ..., f_{10i} and f_{1i+1}, f_{2i+1}, ..., f_{10i+1}, for transmission. Then at the decoding phase, the respective LSF coefficients of the subframes of the current frame (Fig 2.5) are derived from:

The LSFs of subframe 1 = $(7/8)f_{ji} + (1/8)f_{ji+1}$ for $j = 1, \ldots .\ 10.$ (2.11)

The LSFs of subframe 2 = $(5/8)f_{ji} + (3/8)f_{ji+1}$ for $j = 1, \ldots, 10.$ (2.12)

The LSFs of subframe 3 = $(3/8)f_{ji} + (5/8)f_{ji+1}$ for $j = 1, \ldots, 10.$ (2.13)

The LSFs of subframe 4 = $(1/8)f_j + (7/8)f_j$ for $j = 1, \ldots, 10$ (2.14)

The speech of each subframe is thereafter synthesized based on the corresponding interpolated LSF coefficients.

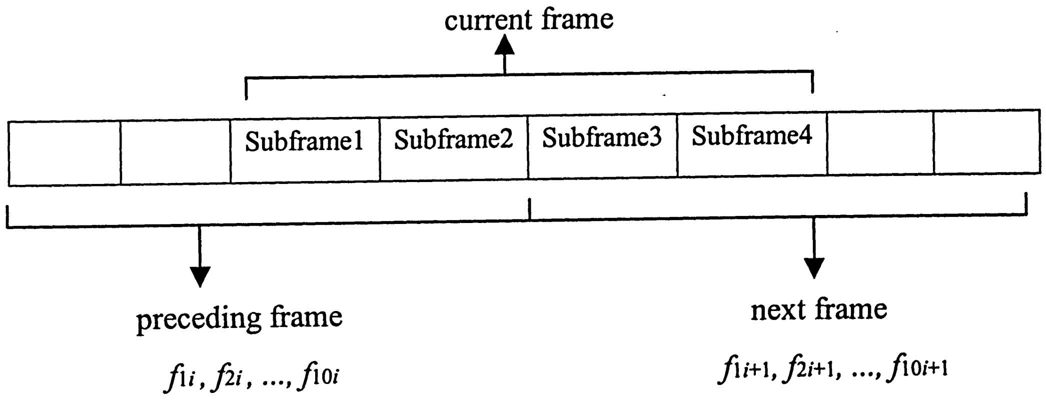


Figure 2.7 Interpolation of LSF coefficients.

Among the 10 LSF coefficients, 4 LSF coefficients are more perceptually sensitive to human ears. Accordingly, FS1016 reserves 4 bits for each of these 4 sensitive LSF coefficients, and puts 3 bits for each of the remaining 6 LSF coefficients. This sums to 34 bits for the 10 LSF coefficients, which consumes a bandwidth of 34 bits/30 ms = 1.133 kbps.

2.3.2 Long-Term Adaptive Codebook Search

After finding the best LSF coefficients, the adaptive codebook search will then be activated for further minimization of the residual signal. In principle, if the vocal tract filter can be accurately modeled by the linear predictor filter, then the residual signal presents exactly the glottal excitation signal. The glottal excitation signal is periodic in nature. Its period is named the *pitch period* or *pitch delay*, and the estimator of the pitch period is called the *Long-Term Predictor* (LTP) or simply the *Pitch Predictor*

In FS1016, the synthetic glottal excitation signal that corresponds to optimal pitch period for each subframe is selected from an *adaptive codebook* through a closed-loop scheme.

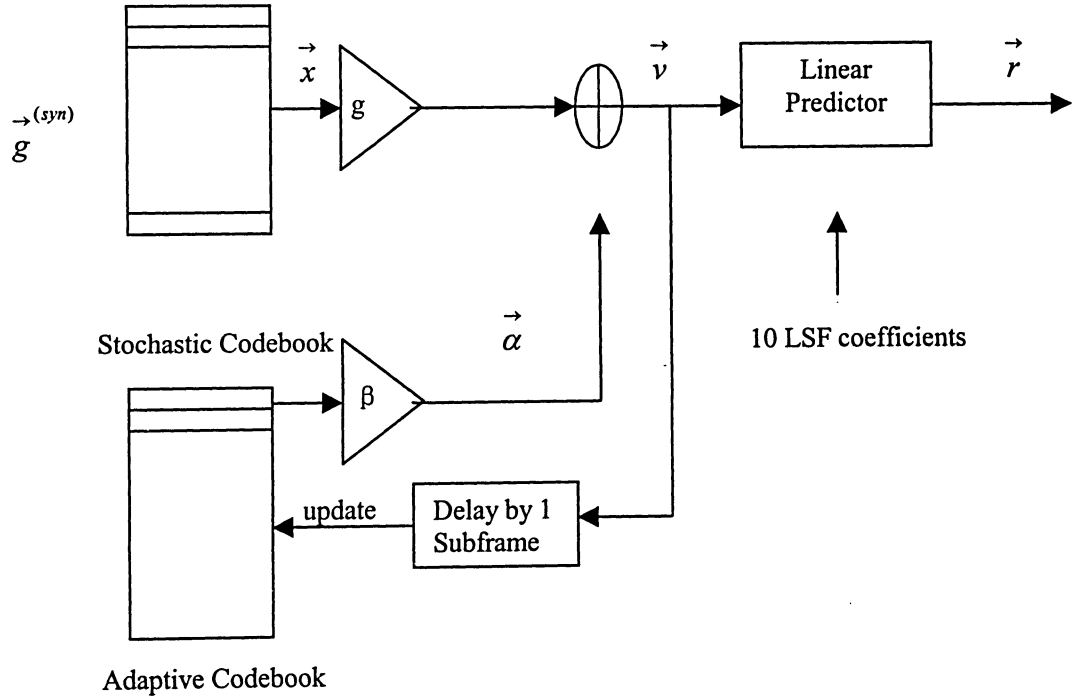


Fig 2.8 The detailed functional block of the CELP synthesizer

The procedure is *adaptive* to the previous $r(n)$, that is, the combined output of the stochastic codebook search and the adaptive code book search due to the previous subframe (Fig. 2.5). The relation between the previous $r(n)$ and the current pitch predictor output $r(n)$ can be in concept characterized by the LTP filter:

$$F_{LTP}(z) = \frac{1}{1-\beta . z^{-T}} \tag{2.15}$$

where T is the pitch period and β is the *pitch gain.* The time-domain expression due to input $r_{\text{initial}}(n)$ is therefore:

$$r(n) = r_{initial}(n) + \beta r(n-T) \tag{2.16}$$

FS1016 then searches the best estimates of pitch period T, among those pre-specified 256 candidates, and its corresponding pitch gain β such that the minimum mean square of

$$\begin{aligned}\sum_n e^2(n) &= \sum_n \left[w(n) * (s(n) - s^{(syn)}(n))\right]^2 \\ &= \sum_n \left[w(n) * (s(n) - \beta \cdot r(n-T) - \sum_{j=1}^{10} a_j s^{(syn)}(n-j) \right]\end{aligned} \tag{2.17}$$

is achieved.

All 256 adaptive codewords for selecting to minimize (2.11) are pre-made according to a 147 dimensional vector that is subframe-wisely updated according to the previous $r(n)$. Specifically, the update procedure is to remove the oldest 60 components of the 147-dimensional vector, followed by shifting in the 60 components of the previous $r(n)$,

Additional 128 non-integer-valued-delay codewords are obtained by interpolating the two nearest integer-valued-delay codewords. The pitch gain ranges from -1 to 2.0, is quantized discordantly with equal number of bits assigned for each subframe.

In FS1016, the approaches to search the optimal pitch delays for odd numbered subframes and even-numbered subframes are different. Based on the maximum match score criterion, the optimal pitch delay (and the corresponding optimal gain) for each odd-numbered subframe is first selected from the 128 integer-valued delays. Denote the resultant optimal integer-valued pitch delay and the corresponding match score by T^* and m^*, respectively. Then, FS1016 tests whether the largest match score corresponding to the sub-multiple pitch delays $(1/2)^*$, $(1/3)T^*$ and $(1/4)T^*$ is within 1 dB of m^*. If so, update $T_.$ and $m_.$ by the respective sub-multiple pitch delay and match scores. In the end, FS1016 examines the match scores of those non-integer-valued-delay codewords whose pitch delays are within I^*-3 and $I^* + 3$, where $I_.$ is the corresponding codeword index of T^*, and update the optimal pitch delay once a larger match score than the previous optimal match score is located. The above procedure is specified in FS1016 as *nonfull search mode*. As anticipated, if *full-search mode* is adopted, all the 256 candidate codewords are truly examined.

As for the even-numbered subframes, efforts are redirected to locate the optimal pitch delay offset relative to the optimal pitch delay of the previous subframe. Specifically, if the optimal codeword for the previous odd numbered subframe is indexed by i, then FS1016 searches only the codewords whose indices range from j = min[max($i_.$31, 1), 193], j+1, . . ., j+63 for the current even-numbered subframe. Again, in the *non-full search mode*, only the integer-valued pitch delays (belonging to the 64 candidates) are tested, which yields the optimal integer-valued pitch delay $T_.$ with maximum match score m^*. Afterward, the match scores corresponding to those non-integer valued delays within I^*- 3 and $I^* + 3$ are examined, where I^* is the corresponding codeword index of T^*, and the codeword with the largest match score, among those examined ones, is outputted. Notably, no sub-multiple pitch delays are examined for even-numbered subframes. Since taking the *non-full search mode* reduces the computational complexity with negligible loss of speech quality, it is taken as default mode except otherwise stated throughout the thesis. In total, *FS1016 distributes 48 bits for the pitch delays and pitch gains of the four subframes in a frame, in which 8 bits and 6 bits are reserved for pitch delays in odd frame and in even frame, respectively. There resultant transmission rate is thus 48 bits/30 msec = 1.6 kbps.*

2.3.3 Stochastic Codebook Search

After the LPC analysis and the pitch prediction, the residual signal become periodic, which is often referred to as the *innovation signal*. The noise-like innovation signal, although lack of speech information, can not be neglected. In its absence, the speech will sound artificial. In FS1016, a stochastic codebook is employed to approximate the innovation signal. Each codeword in the stochastic codebook has its own index. There are totally 512 codewords in the codebook. To speed up the search process, reduced-size codebooks of 256, 128 and 64 codewords are also specified in FS1016. Nevertheless, better speech quality is achieved by using the codebook with more codewords. Analogous to the adaptive codebook search, the stochastic codebook search is performed per subframe by a closed-loop operation. Also adopted is the minimum squared error criterion. To facilitate the codebook search, the 512 codewords in the stochastic codebook are drawn from a one-dimensional array of (+1, 0, -1) value with approximately 77% zeros. The consecutive codewords overlap except at the first and the last two components. Such a codebook design has several advantages:

- Only two bits are required to represent the ternary values, +1, -1 and 0.
- Multiplying with +1 and -1 can be replaced by sign changes, which greatly reduces the computational complexity.
- Adding the product of a term and zero is equivalent to remain unchanged in the original quantity, and the chance of meeting a zero is as high as 77%.
- When convolution operations are performed for two consecutive codewords, the convolution for the second codeword can retain those convolved results obtained from its overlapped part with the previous codeword, and reduce the computational complexity.

In total, FS1016 distributes 56 bits for the stochastic codebook search. The index for the best codeword in the stochastic codebook requires 9 bits for each subframe. The stochastic gain lies between -1330 and 1330 discordantly, and each of the four gains consumes 5 bits. Therefore to the transmission rate is 56 bits/30 msec = 1.866 kbps.

2.4 Distortion and Performance Measures

A useful distortion and performance measure corresponds well with the subjective quality of the speech: low and high subjective quality speech yields small and large distortions, respectively. Distortion and performance measures are used extensively in speech processing for a variety of purposes [10]. In speech coding, they are typically used to compare the performance of different systems or configurations. The numerous distortion and performance measures can all be divided into two main categories: subjective measures and objective measures.

2.4.1 Subjective Distortion Measures

This class of distortion measures is based on the opinion of a listener or a group of listeners as to the quality or intelligibility of the speech. These measures are time-consuming and costly to obtain, requiring a set of discriminating listeners.

In addition, a consistent listening environment is required since the perceived distortion can vary with such factors as the playback volume and type of listening instrument used (e.g., headphones versus telephone handsets) [11]. However, subjective distortion measures provide the most accurate assessment of the performance of speech coders since the degree of perceptual quality and intelligibility is ultimately determined by the human auditory system. Subjective distortion measures are used to measure the quality or intelligibility of speech. Quality tests strive to determine the naturalness of the speech. The mean opinion score (MOS) and diagnostic acceptability measure (DAM) are the most commonly used subjective quality tests. On the other hand, the prime concern of intelligibility tests is the percentage of words, phonemes or other speech units that is correctly heard. The standard intelligibility test is the diagnostic rhyme test (DRT) [12]. For this project, subjective measures based on MOS is primarily used to evaluate the quality.

2.4.2 Objective Distortion Measures

This category of measures can be evaluated automatically from the speech signal, its spectrum or some parameters obtained thereof. Since they do not require listening tests, these measures can give an immediate estimate of the perceptual quality of a speech coding algorithm. In addition, they can serve as a mathematically tractable criterion to minimize during the quantization stages of a speech coder. The two main factors in selecting an objective distortion measure are its performance and complexity. The performance of an objective distortion measure can be established by its correlation with a subjective distortion measure of the same features (quality or intelligibility). An extensive performance analysis of a multitude of objective distortion measures is given in [12]. Objective distortion measures can be broadly classified into three categories: time-domain, frequency-domain and perceptual-domain measures.

Time-domain distortion measures are most useful for waveform coders which attempt to reproduce the original speech waveform. The most frequently encountered measures of this type are the signal-to-noise ratio (SNR) and the segmental signal-to-noise ratio (SNRseg). Most medium to low bit-rate coders are hybrid or parametric coders. Since the auditory system is relatively phase insensitive, these coders tend to focus on the magnitude spectrum. As a result, the time-domain measures cannot adequately gauge the perceptual quality of these systems. Frequency-domain measures are thus used to determine the performance of these types of speech coders since they are less sensitive to time misalignments and phase shifts between the original and coded signals. They are also useful for the quantization of spectral coefficients the codebook vector which is most perceptually similar, as determined by the distortion measure, to the original spectral envelope would be selected.

2.4.2.1 Signal-to-Noise Ratio

The SNR is the ratio of signal energy to noise energy expressed in decibels dB and is given by:

$$SNR = 10\log_{10}\frac{\sum_{n=-\infty}^{\infty} s(n)^2}{\sum_{n=-\infty}^{\infty}(s(n)-\hat{s}(n))^2} \quad dB \tag{2.18}$$

where *s [n]* is the original signal and *[n]* is the 'noisy' signal. The SNR is characterized by its mathematical simplicity. The drawback is that it is a poor estimator of the subjective quality of speech. The SNR of a speech signal is dominated by the high energy sections consisting of voiced speech. However, noise has a greater perceptual effect in the weaker energy segments [13]. A high SNR value can thus be misleading as to the perceptual quality of the speech.

2.4.2.2 *Segmental Signal-to-Noise Ratio*

The SNR$_{seg}$ in dB is the average SNR (also in dB) computed over short frames of the speech signal. The SNR$_{seg}$ over M frames of length N is formulated as:

$$SNR_{seg} = \frac{1}{M}\sum_{i=0}^{M-1} 10\log_{10}\left[\frac{\sum_{n=iN}^{iN+N-1} s^2(n)}{\sum_{n=iN}^{iN+N-1}(s(n)-\hat{s}(n))^2}\right] \quad dB \tag{2.19}$$

where the SNRseg is determined for *s[n]* over the interval *n = 0, . . . ,NM-1*. This distortion measure weights soft and loud segments of speech equally and thus models perception better than the SNR. The length of frames is typically 10-40 ms corresponding to values of *N* between 120 and 200 samples, assuming a sampling rate of 8 kHz. For this project, SNR$_{seg}$ is used for objective measure of speech quality.

Silent portions of the speech can bias the results by yielding a large negative SNR for the corresponding frames. This problem can be alleviated by removing frames corresponding to silence from the calculations. Another method is to establish a lower threshold (typically 0 dB) and replace all frames with an SNR below the threshold.

Similarly, a deceptively high SNR$_{seg}$ can result when frames have a very high SNR, even though perception can barely distinguish among frames with an SNR greater than 35 dB [23]. Therefore, an upper threshold around 35 dB can be used to prevent a bias in the positive direction.

2.5 Chapter Summary

This chapter overviewed the fundamentals of CELP speech coders, then presented Short-term Linear Prediction, Long-term Linear Prediction, Stochastic Codebook search and distortion, performance measures. In this project, the CELP analysis is done primarily based on FS1016 algorithm. The bit allocation for short-term linear prediction, adaptive codebook and stochastic codebook search are presented. The subjective measures (MOS) and objective measure (SNR, segment SNR) are the chief distortion and performance measures used for speech quality.

Chapter 3

Channel Controlled Bit-rate Variability

Maximum capacity while maintaining an acceptable level of voice quality under traffic and radio propagation conditions is a main objective in the design of a cellular network for mobile or personal communication. Due to the wireless communication channel characteristic, channel coding is necessary to remove most transmission errors as long as the system operates within a reasonable C/I (carrier to interference ratio) range. However, the drawback of this solution is a lower speech quality than achievable for good channel conditions, since a large amount of the gross bit rate is consumed on the channel coding.

The variable bit-rate speech coder solves the problem in an effective manner. The ratio between net bit rate and error protecting redundance is adaptively chosen according to the current channel conditions. When the channel condition is bad, the speech coder operates at low bit rate thus allowing powerful forward error control. In turn, for good channel the speech encoder may use its highest net rate implying high speech quality, as a weak error protection capacity is sufficient. Here we call this type of speech coder the channel controlled variable bit-rate speech coder. It can be divided into two main categories: embedded and multi-mode variable rate coders.

3.1 Embedded Multi-rate Speech Coder

Embedded coding is a technique which allows simple bit dropping in a given bit-stream. In the other words this means that a single coding algorithm generates a fixed-rate data stream from which one of several reduced rate signals can be extracted by a simple bit dropping procedure. This also means that bits can be discarded or dropped between the encoder and decoder. The corresponding decoders fill in the missing bits with zeros and then decodes the resulting (modified) full-rate data signal with a fixed decoder algorithm. Thus, each lower rate data signal is embedded in the higher full rate bit stream.

Embedded coding offers a more elegant approach to external rate control. Since the coder itself generates a fixed rate stream, rate switching is simply achieved by suitable bit-dropping to achieve bit rate variability. Embedded coders can have multiple rates in a hierarchical fashion with each sub-rate signal embedded in the next higher rate signal. Clearly, embedded coding is a special case of multi-mode coders.

3.2 Multi-mode AMR

Comparing with embedded multi-rate coder, we can achieve variable bit rate via an adaptation algorithm whereby the network select one of a number of available speech coders, called codec modes, each with a predetermined source/channel bit allocation. This concept is called adaptive multi-rate (AMR) coding and is a form of channel controlled multi-modal coding of speech [22].

The AMR concept is the centerpiece of ETSI's GSM AMR standardization activity, which aims to define a new European cellular communication system designed to support an AMR mechanism in both the half rate and full rate channels. For each channel mode, the *codec mode*, i.e. bit partitioning between speech and channel bit-rates, can be varied rapidly to track the channel error rates or the channel's C/I. This variation is represented to the right in Figure3.1. By decreasing this coding bit-rate, i.e. switching from codec model 3 to 2 or from codec 2 to 1, the robustness is increased under poor conditions. The changes must occur quite immediately (several times a second), with no perceptible speech degradation. This process is equivalent to Link Adaptation [14].

3.2.1 Multi-mode AMR Speech Coding

Multi-mode AMR vocoder is based on the code-excited linear predictive (CELP) coding model and analysis-by-synthesis method is employed to quantize the excitation, where the encoding is usually based on some simplified algorithm rather than global optimization strategy. For AMR vocoder, the commonly used algorithms are as follows: (1) The structured stochastic codebook is the part of the coder where the main bit-rate

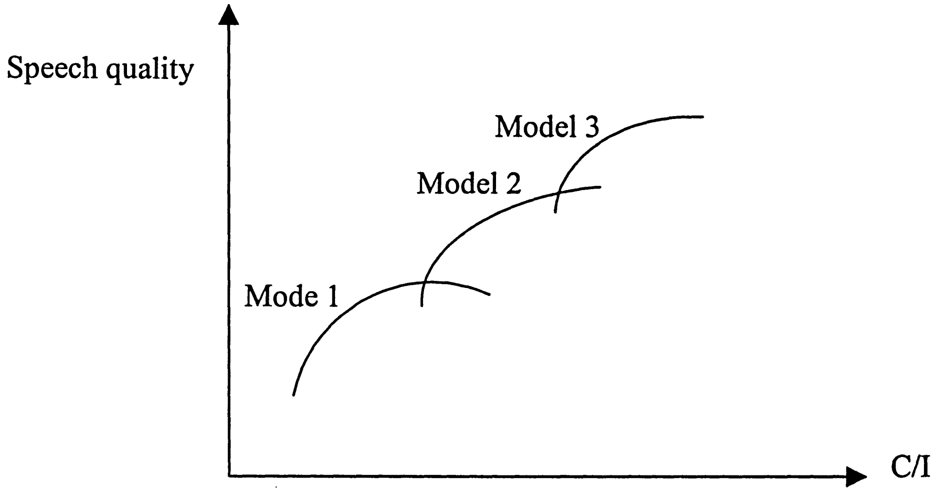


Fig.3.1 Adaptive multi-modes

variation between different code modes; (2) both open-loop and close-loop pitch detection were employed, where open-loop pitch analysis is performed to confine the close-loop pitch search to a small number of lags. Besides pitch of even sub-frame is adjusted based on that of odd sub-frame; (3) sequential search of the pulse position and position is confined in some pre-defined tracks.

AMR coder in GSM is capable of operating at 8 different bit-rate denoted coder modes. Linear prediction (LP) analysis is performed once 20 ms frame. The speech frame is divide into four subframes of 5 ms each. The bit allocation for coder is shown for each mode in Table 3.1 (LSF-- Line Spectral Frequency coefficients; Adapt CB—adaptive code book index; Adapt gain—adaptive gain; Stoch CB—stochastic code book index; Stoch CB gain-- stochastic gain). The Table 3.1 shows that the set of LP coefficients is converted to LSF and vector quantized with 38, 27, 26, 23 bits are used at rates 12.2, 7.95, 10.2, 7.4, 6.7, 5.9, 5.1, 4.75 kbps. The adaptive codebook index is encoded with 9, 8, 4 bits in odd subframes and relatively encoded with 6, 5, 4 bits in even subframes for different rate mode. 35, 31, 17, 14, 11, 9 bits is for the stochastic codebook index on different rate mode respectively. Only 12.2 bps mode use 4 bits for each adaptive gain in each subrame. Stochastic gain in each subframe is encoded with 5, 6, 7, 8 bits for 12.2 and 7.95, 5.95 and 5.1, 10.2, 7.40 and 6.70, 4.75 bps rate mode respectively. So, by different bit allocation for each bit rate mode, a bit rate rang from 12.2 kbps to 4.75 kbps can be achieved.

Mode	Parameter	Subframe				Total
		1	2	3	4	
12.20 kbps	LSF					38
	Adapt CB	9	6	9	6	30
	Adapt gain	4	4	4	4	16
	Stoch CB	35	35	35	35	140
	Stoch CB gain	5	5	5	5	20
10.20 kbps	LSF					26
	Adapt CB	8	5	8	5	26
	Stoch CB	31	31	31	31	124
	Gain	7	7	7	7	28
7.95 kbps	LSF					27
	Adapt CB	8	6	8	6	28
	Adapt gain	4	4	4	4	16
	Stoch CB	17	17	17	17	68
	Stoch CB gain	5	5	5	5	20
7.40 kbps	LSF					26
	Adapt CB	8	5	8	5	26
	Stoch CB	17	17	17	17	68
	Gain	7	7	7	7	28
6.70 kbps	LSF					26
	Adapt CB	8	4	8	4	24
	Stoch CB	14	14	14	14	56
	Gain	7	7	7	7	28
5.90 kbps	LSF					26
	Adapt CB	8	4	8	4	24
	Stoch CB	11	11	11	11	44
	Gain	6	6	6	6	24
5.10 kbps	LSF					23
	Adapt CB	8	4	4	4	20
	Stoch CB	9	9	9	9	36
	Gain	6	6	6	6	24
4.75 kbps	LSF					23
	Adapt CB	8	4	4	4	20
	Stoch CB	9	9	9	9	36
	Gain	8		8		16

Table.3.1 Bits allocation of Multi-mode AMR

3.2.2 Overview of the AMR Coding System

In general, adaptation depends on the current state of the communication channel. Since channel estimation is done at the decoder, the receiver needs to signal to the encoder through the reverse link some information needed for mode selection. The rate control mechanism varies depending on the direction of transmission, due to a constraint that the code mode control mechanism must be located in the base station.

We know that for uplink transmission, the base station monitors the channel condition and decides which mode the mobile station should use. The base station communicates this information in the form of a codec mode command, transmitted in the downlink. Upon reception, the mobile station encoder switch to the indicated mode; for downlink transmission, based on the received bits and possibly other information that may be available, the mobile station computes a downlink channel measurement which is representative of the state of the channel. The mobile station cannot autonomously decide which mode to use. Hence, this measurement is quantized and transmitted back on the uplink to the base station. The base station then decodes which mode it will use for the downlink transmission of the next frame.

3.3 Variable Frame Length AMR

Recent research [15,16] has shown that variable rate techniques can augment fixed rate speech coding systems, producing similar or higher quality speech at lower average bit rate and achieve high network capacity. However, these variable bit rate coding techniques still preserve the fixed frame length structure.

In multi-mode AMR, multi-rate is achieved by using different bit rate mode which has different bit allocation at fixed frame length. Is there any other scheme to realize variable bit rate? The answer is "**YES**". Contrary to the multi-mode AMR, We can also to achieve variable bite rate by changing frame length with the same bit allocation for each frame. For Parametric Coders, the speech is recovered with many parameters by tracking the speech characteristic. The idea behind Variable frame length AMR scheme is that large frame means more average, less accurate; small frame size means less average, more accurate. In terms of bit rate, large frame length use low bit rate, small frame length use high bit rate. Here we introduce a novel variable frame length AMR coder and related other schemes to realize bit rate variability according to the channel condition, then bit rate variations will simply be a function of frame length variation.

3.3.1 Stationary Frames of Speech Signal

To apply the Linear Prediction (LP) method to a signal, it is necessary that the signal be stationary. In fact the speech signal is not stationary in its whole duration, but it consist of almost stationary small intervals. To apply the LP filtering to speech signal, the first step is to decompose the signal to stationary frames. In obtaining the frame duration two factors are considerable.

(1) In data compression scheme using LPC filters, one must transmit the LPC parameters for each frame. Therefore it is desirable to have frames as long as possible, such that the total number of frames be minimized, as a consequence the total number of filter coefficients be minimized. In the other hand, computation of autocorrelation function, which is the basis of filter design, is more accurate for longer durations.

(2) The signal is not stationary in very long and very short frames. In the case of variable frame length, minimum and maximum frame length in which signal is stationary is 10 ms and 40 ms [2].

Considering the above factors there must be a frame length selection range between the smallest frame length (10ms) and the longest frame length (40ms) for stationary speech signal.

Stationary speech condition was determined by three parameters: pitch variation, voicing cut-off frequency variation and the change in energy every time a new set of samples. The pitch deviation of less than two samples, a cut-off frequency variation of less than three times the pitch frequency and less than 40% energy were found to be tolerable changes for frame change.

3.3.2 Variable Bit Rate Function

In this project, the sampling rate is 8000 samples/second and 144 bits per frame for FS1016 CELP is used. By using variable frame length scheme, we can change bit rate according channel condition. Typically, the frame length used in this project is from 10 ms to 30 ms. There are some examples as follows:

Frame length	Bit rate
40 ms	3600 bits/sec
30 ms	4800 bits/sec
20 ms	7200 bits/sec
10ms	14400 bits/sec

Variable bit rate function is $B=[1000/L]*b$ (3.1)

In the above equation, B is current bit rate; b is bit quantity per frame (FS1016 CELP 144bits/frame); L is frame length(ms).

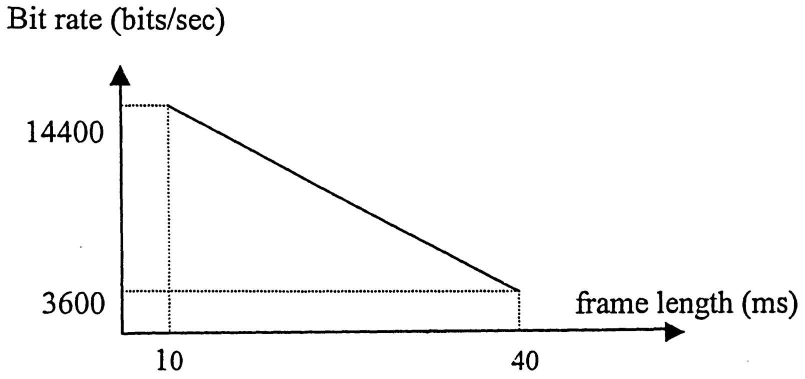


Figure 3.2 Variable bit rate function

3.3.3 Performance Analysis

To compare speech quality of different bit rate by variable length bit rate scheme, we apply FS1016 (CELP) as simulation platform. The result of simulation from 15ms to 30 ms frame length is as follows:

SNR_{seg} (dB)	Bit rate (kbps)	Frame Size (ms)
5.82	4800	30
6.27	5760	25
7.27	7200	20
8.65	9600	15

Table 3.2 Segment SNRs of different frame length

Table 3.3 shows the segmental signal to noise ratios (SNR_{seg}) obtained while testing the coder under different frame length: 30, 25, 20, 15 ms, respectively. Test material consists of 4 sentences (2 mail, 2 female). Simulation results indicate small frame length correspond high bit rate and high speech quality. It proves that different bit rate corresponding different level of speech quality can be simply achieved by changing frame length.

The effects of joint channel and source controlled coding are explained in Chapter 5. Simulation results in Table 5.1, compares bit rate for different frame length when the proposed source controlled variable bit rate algorithms in Chapter 4 are implemented with variable frame length AMR. It shows that more bit rate can be reduced due to less frame length. The bit rate reduce in 15 ms frame length is 30% which is 2 times better than 30 ms frame length when source controlled variable bit rate algorithms is used. It can contribute the joint channel and source controlled variable bi rate speech coder because this feature can compensate the bit rate increase while frame length is reduced to achieve high quality. Details are in Chapter 5.

However, two problems will be encountered when the proposed variable frame length is applied for the speech coder. One problem is that different frame length has different processing time delay. It will cause synchronization problem in the decoder. The other one is artifact due to sudden bit rate switch from low to high bit rate mode. To solve these problems, a smooth switch algorithm for variable frame length AMR will be presented in the next section.

3.3.4 Smooth Switch Different Bit Rate

Compared to the wired links, the wireless channels are typically much more noisy and have a higher bit error rate. Meanwhile, multi-path and shadow fading, time dispersion occurs frequently in wireless channel. As a result, random and burst errors can have devastating effect on speech streaming quality So, the channel condition in wireless communication cannot be predicted accurately. It might change rapidly or change smoothly or will remain stable for a certain time duration. Based on the above wireless channel Property, for AMR speech coder, when we select the bit rate in terms of certain channel condition, the smooth switch must be considered, especially when we do switch from the very high bit rate to very low bit rate or from the very low bit rate to very high bit rate. Otherwise, there will be an artifact and significant speech quality "jump". In our variable frame length AMR, we introduce a simple smooth bit rate switch scheme. The idea behind it is as follows:

There are ten bit rate modes: B_1;B_2; B_3; B_4;B_5; B_6; B_7;B_8; B_9; B_{10}.Assuming we want to switch from B_1 to B_{10}

Hard switch: $\boldsymbol{B_1} \longrightarrow \boldsymbol{B_{10}}$

Smooth switch: $\boldsymbol{B_1} \rightarrow \boldsymbol{B_2} \rightarrow \boldsymbol{B_3}$............$\rightarrow \boldsymbol{B_8} \rightarrow \boldsymbol{B_9} \rightarrow \boldsymbol{B_{10}}$

For our variable frame length AMR, it is simple to apply the above smooth switch scheme (figure 3.4). When we switch B_i to B_k (k>i), the only thing we will do is decrease frame length step by step till we get k_j.For FS1016 CELP one step is 1ms. Based on the simulation test result, our simple smooth switch scheme can completely remove any speech quality "jump", and also there is no any artifact due to bit rate switch. The details are in chapter 5.

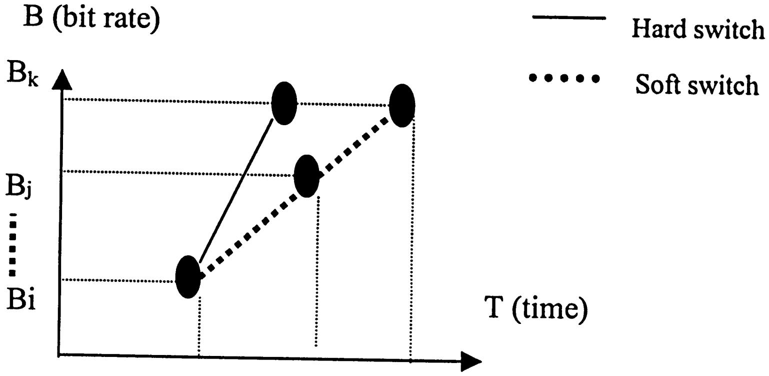


Figure 3.4 Hard switch vs Smooth switch

For synchronization issue, smooth switch scheme can lower the bit rate transform to avoid big processing time delay " jump" since 1 step size of 1 ms is little time duration which human may not notice. Therefore, there are totally 30 bit rate selection modes in the proposed variable frame length AMR.

3.4 Benefits of Variable Frame Length AMR

In fact, variable frame length coding technology has been applied in image and video coding. Through the above analysis, we know we can achieve bit rate variability by using variable frame length. Compared with multi-mode AMR, there are some notable advantages for the proposed variable frame length AMR:

(1) Basically, multi-mode AMR use a family of fixed rate coder to adapt the channel condition. It can be said that the multi-mode AMR consist of several speech coders which has complicated structure and requires large hardware memory to storage codec information. The proposed variable length AMR has advantage regarding its simplicity because the variable bit rate is achieved by changing the frame length. And also there is no need for extra hardware memory for storing codec information

(2) In current GSM AMR standard, Multi-mode AMR only has limited 8 different bit rate modes for half-rate channel and 9 different bit rate mode for full-rate channel. It cannot adapt accurately according to the channel condition. However, the variable length AMR has more bit rate selection range (30 bit rate modes) and also it can track the channel condition better.

(3) Multi-mode AMR cannot apply bit rate switch directly which need other algorithm to deal with smooth switch problem since the bit rate mode is

discrete. In the proposed technique, the rate switch gradually. Such a gradual switch will provide better human perception.

3.5 Chapter Summary

This chapter presented some of the ideas and method for channel controlled variable bit rate speech coding including embedded and adaptive multi-rate (AMR) speech coding. After overview of AMR speech coding scheme which apply multi-mode method in current GSM systems, this Chapter introduced another method to achieve bit rate variability by variable frame length. In addition, a smooth bit rate switch scheme is applied with the variable frame length method. The advantage and benefit of new method to achieve channel controlled bit rate variability is presented in the last section.

Chapter 4

Source Controlled Bit-rate Variability

Maximizing capacity while maintaining an acceptable level of speech quality is a central objective in the design of a mobile network. To achieve this goal, many researchers have developed a number of techniques. Among them, variable bit rate coding has been proved is a high efficient way to increase the network capacity.

Variable bit rate coding clearly allows the long-term average bit-rate R_a for a given level of quality to be substantially less than the peak bit rate R_p that would be required by an equivalent quality fixed rate coder. Suppose that the maximum total data rate from all mobiles in a sectors is R_S based on bandwidth, modulation bandwidth efficiency and interference constraints. If each mobile is sending at a variable bit rate with average rate R_a and the system can somehow pool their data signals into one composite data signal. The number of users would be given by $N=R_p/R_a$.. This assumes an efficiently designed scheme that can fully exploit variable rate coding. As long as we can have equal speech quality with a fixed bit rate coder, the ideal the capacity gain is $G= R_p/R_a$. which can lead to significant increase system capacity when we can reduce average rate R_a.

In the source-controlled coding, the coder in some dynamically allocates bits in respond to the local behavior of speech source. Such coders are generally designed with the intent of maintaining a desired level of quality for each short segment of speech with fewest bits needed.

4.1 Voice Activity Controlled Variable Rate Coding

4.1.1 Overview of VAD

Monitoring the presence or absence of speech is the most natural and simplest way to employ source-controlled variable rate coding. A constant bit rate coder is readily converted into a source-controlled variable rate coder by adding a detector that switches off the coder during periods when no active voice is present. This produces a data signal whose rate switches between the full rate and zero rate in a random manner somewhat like the classical random telegraph signal.

In a classic study of voice activity patterns, Brady observed that one side of two way telephone conversations consists of interminent talk spurts separated by pause or silence. The process of identifying when talk spurs occur is called voice activity detection (VAD). Based on a simple speech detector. Brady found that on an average, a speaker is talking about 44% of the time. Subsequent studies based on more accurate detectors of voice activity have found a lower percent of talking time.

The quality of the VAD algorithm is a very important consideration in the design of systems that enhance capacity by exploiting voice activity. The increase in capacity is determined by the voice activity factor (VAF) which is the fraction (or percent) of the time the detector identifies the presence of active speech. Reliably measuring the VAF of a detector requires averaging the conversation in many calls with many different speakers. If silence is detected as speech, the capacity is reduced; on the other hand, when speech is detected as silence, degradations in the recovered speech quality are introduced.

In GSM, the VAD decisions are usually based on multiple features extracted from the signal including time varying energy, zero-crossing counts, sign bit sequences, and features generated from within the speech coding algorithm. In GSM, the VAD decision is used by the discontinuous transmission (DTX) mechanism [34], which allows the radio transmitter to be switched off most of the time during speech pause.

Hangover time

During speech pause, the acoustic signal is not really "silence" in nature.

Background noise and possible echoes of the far-end speech are always present. The task of VAD algorithm is complicated because certain speech sounds have a very low energy level and are random in character and thereby are readily confused with background noise. Difficult phonetic (phonetic units of speech) for a detector including weak fricative sound such /f/ in *fat* /th/ in *thing*, there are often extremely shorts pauses during active speech, notably with certain phonetic segment called plosives where a detector could prematurely declare the start of silent interval. To avoid this, some hangover time is needed during which the VAD delays its decision of silence and continues to observe the waveform before it declares that a transition has occurred from active speech to silence.

Conservative consideration

For mobile environment, the design of a VAD is complicated by high level of acoustic noise coming to the microphone. The acoustic noise may include such sources as vehicle engine noise, car radios, restaurant background noise, city street sounds, etc. To avoid degrading the speech quality, the VAD algorithm can be designed fairly conservatively so that a lot of the background noise will be classified as active speech rather than silence.

Comfort noise

During the speech pause, synthetic noise similar to the transmit side background noise is generated on the receive side. This synthetic comfort noise is produced by transmitting parameters describing background noise at a regular rate during pause. It can be coded at a very low bit rate or only its power level can be transmitted and random noise regenerated at the receiver. This reproduction of background noise is often called comfort noise.

4.1.2 VAD Algorithm

The block diagram of the VAD algorithm is depicted in Figure 4.1 [17]. The VAD algorithm uses parameters of the speech encoder to compute the Boolean VAD flag (VAD flag). This input frame for VAD is sampled at the 6.4 kHz frequency and thus it contains 256 samples. Samples of the input frame (*s(i)*) are divided into sub-bands and level of the signal (level[n]) in each band is calculated. Input for the tone detection function are the normalized open-loop pitch gains which are calculated by open-loop pitch analysis of the speech encoder. The tone detection function computes a flag (tone flag) which indicates presence of a signaling tone, voiced speech, or other strongly periodic signal. Background noise level (bckr_est[n]) is estimated in each band based on the VAD decision, signal stationarity and the tone-flag. Intermediate VAD decision is calculated by comparing input SNR (level[n]/bckr est[n]) to an adaptive threshold. The threshold is adapted based on noise and long term speech estimates. Finally, the VAD flag is calculated by adding hangover to the intermediate VAD decision. The block diagram of the VAD decision algorithm is shown in Figure 4.1.

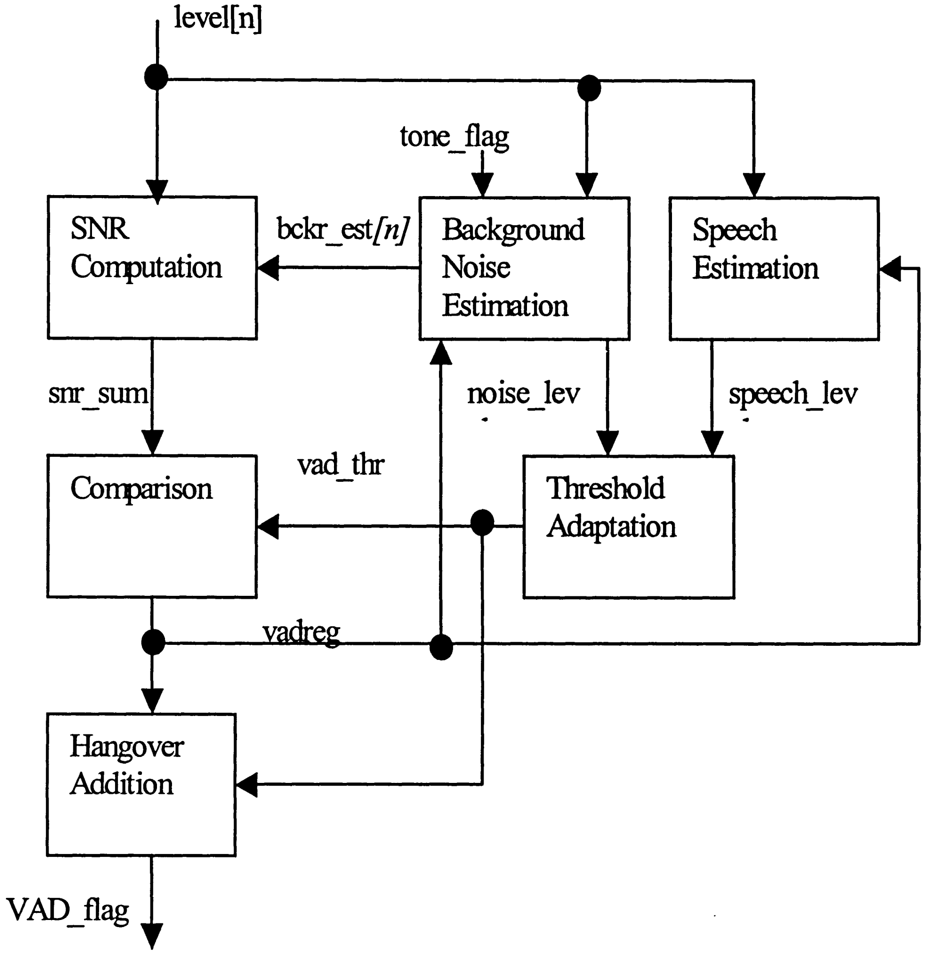


Fig 4.1. Simplified block diagram of the VAD decision algorithm

4.2 Voiced and Unvoiced Segment Detection

Through voice activity patterns offer an important and essential component for source-controlled variable rate coding even during active talk spurs the speech signal has a time varying short-term entropy. In other words, variable rate coding of active speech segment is a natural way to achieve further reductions in average bit-rate for a given reproduction quality. In this section, we discuss voiced and unvoiced segment variable rate algorithm for coding of speech segment that has been declared as active by a VAD algorithm.

4.2.1 Speech Characteristic

In Chapter 2, we have introduced some property of voiced and unvoiced speech segment. Here we will discuss more about it. The two types of speech sounds, voiced and unvoiced, produce different sounds and spectra due to their differences in sound formation. With voiced speech, air pressure from the lungs forces normally closed vocal cords to open and vibrate. The vibrational frequencies (pitch) vary from 50 to 400 Hz (depending on the person's age and sex) and forms resonance in the vocal tract at odd harmonics. These resonance peaks are called formants and can be seen in the voiced speech Figures 4.2 and 4.3 below.

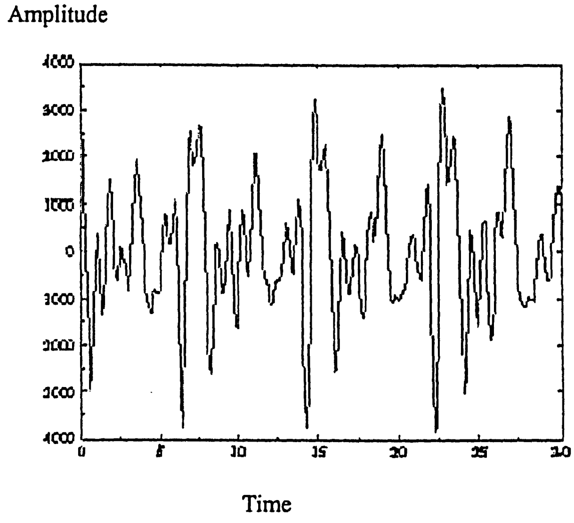


Fig. 4.2 Voiced Speech Sample

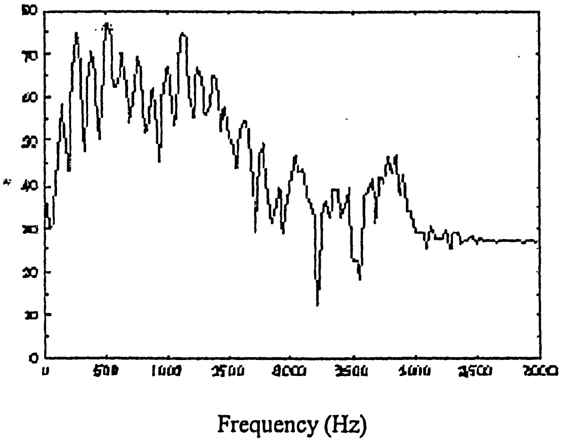


Fig. 4.3 Power spectral density of voiced Speech

Unvoiced sounds, called fricatives (e.g., s, f, sh) are formed by forcing air through an opening (hence the term, derived from the word "friction"). Fricatives do not vibrate the vocal cords and therefore do not produce as much periodicity as seen in the formant structure in voiced speech; unvoiced sounds appear more noise-like (see Figures 4. 4 and 4.5). Time domain samples lose periodicity and the power spectral density does not display the clear resonant peaks that are found in voiced sounds. In addition, the energy of voiced segment is generally higher than the energy of unvoiced segments.

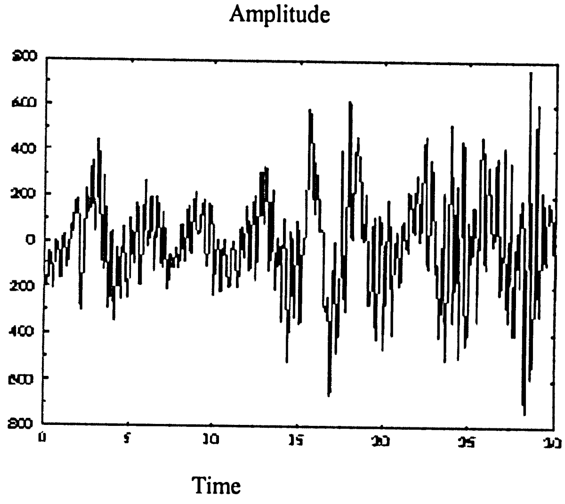


Fig. 4.4 Unvoiced Speech Sample

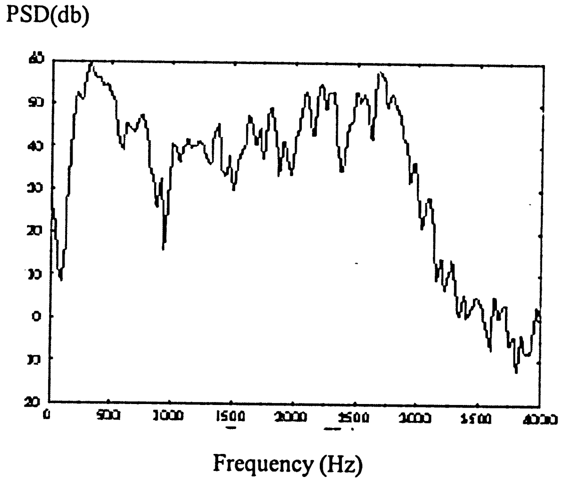


Fig 4.5: Power spectral density of unvoiced speech

The spectrum for speech (combined voiced and unvoiced sounds) has a total bandwidth of approximately 7000 Hz with an average energy at about 3000 Hz. The auditory canal optimizes speech detection by acting as a resonant cavity at this average frequency. Note that the power of speech spectra and the periodic nature of formants drastically diminish above 3500 Hz. Speech encoding algorithms can be less complex than general encoding by concentrating (through filters) on this region. Furthermore, since line quality telecommunications employ filters that pass frequencies up to only 3000-4000 Hz, high frequencies produced by fricatives are removed. A caller will often have to spell or otherwise distinguish these sounds to be understood (e.g., "F as in Frank").

4.2.2 Conventional Voiced/Unvoiced Segment Detection Algorithm

The need for deciding whether a given segment of a speech waveform should be classified as voiced speech or unvoiced speech, arises in many speech analysis systems. A variety of approaches have been described in the literature for making this decision. In this section, firstly, we introduce a number of conventional approach for classifying a given speech segment, which is described in [2]. Also, the joint approach provides an effective method of combining the contributions of a number of approach which individually may not be sufficient to discriminate the voiced speech or unvoiced speech.

The following five measurements have been used in the implementation described as follows:

- Energy of the signal
- Zero–crossing rate of the signal
- Autocorrelation coefficient at unit sample delay
- First predictor coefficient
- Energy of the prediction error

The choice of these particular parameters is based on the experimental evidence that the parameters vary consistently from one class to another, which we will discuss later in this part.

Speech Measurements

A block diagram of the analysis and decision algorithm is shown in the following figure. Prior to analysis, the speech signal is high pass filtered to remove any dc, low frequency hum, or noise components which might be presented. The formula of the high pass filter is given below.

$$H(z) = \frac{1-2z^{-1}+z^{-2}}{1-e^{-\alpha T}\cos(bT)z^{-1}+e^{-2\alpha T}z^{-2}} \tag{4.1}$$

Then the five parameters mentioned in the previous section are computed for each block of samples. Following we state in detail the definitions of them.

1. Zero-crossing count N_z, the number of zero crossing in the block

The zero crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced as a result of excitation of the vocal tract by the periodic flow of air at the glottis and usually shows a low zero crossing count. Unvoiced speech is produced due to excitation of the vocal tract by the noise-like source at a point of constriction in the interior of the vocal t ract and shows a high zero crossing count. The zero crossing count of silence is expected to be lower than for unvoiced speech, but quite comparable to that for voiced speech.

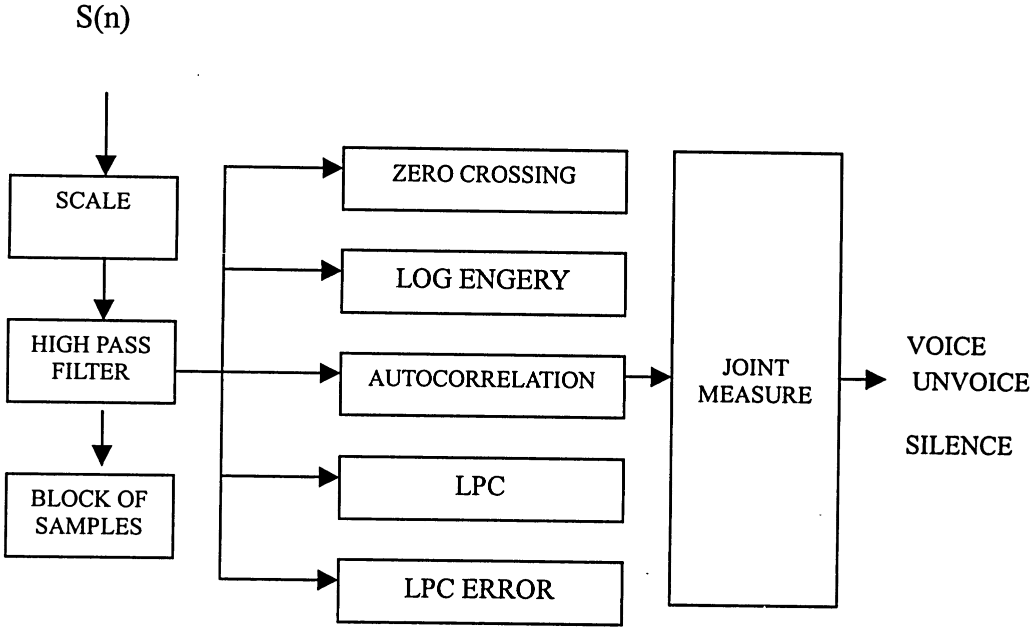


Fig 4.6 Block diagram of voiced/unvoiced detection algorithm

2. Log energy E_s -- defined as

$$E_s = 10\log(e + \frac{1}{N}\sum_{n=1}^{N} S^2(n)) \tag{4.2}$$

Where e is a small positive constant added to prevent the computing of log of zero. Generally speaking, E_s for voiced data is much higher than the energy of silence. The energy of unvoiced data is usually lower than for voiced sounds but higher than for silence.

3. Normalized autocorrelation coefficient at unit sample delay, C_I which is defined as

$$C_1 = \frac{\sum_{n=1}^{N} s(n)s(n-1)}{\sqrt{(\sum_{n=1}^{N} s^2(n))(\sum_{n=0}^{N-1} s^2(n))}} \tag{4.3}$$

This parameter is the correlation between adjacent speech samples. Due to the concentration of low frequency energy of voiced sounds, adjacent samples of voiced speech waveform are highly correlated and thus this parameter is close to 1. On the other hand, the correlation is close to zero for unvoiced speech.

4. First predictor coefficient, a_I of a 12-pole linear predictive coding analysis using the covariance method. It can be shown that this parameter is the negative of the Fourier component of the log spectrum at unit sample delay. Since the spectra of the three classes -- voiced, unvoiced, silence -- differ considerably, so does the first LPC coefficient.

5. Normalized prediction error, E_p, expressed in dB, which is defined as

$$E_p = E_s - 10\log(10^{-6} + \left| \sum_{k=1}^{p} (\alpha_k \phi(0,k) + \phi(0,0)) \right| \tag{4.4}$$

$$\phi(i,k) = \frac{1}{N} \sum_{n=1}^{N} s(n-i)s(n-k) \tag{4.5}$$

Where E_s is defined above and *Ø (i, k)* is the term of the covariance matrix of the speech samples, and a_k's are the predictor coefficients. This parameter is a measure of the non-uniformity of the spectrum.

The five parameters discussed above are correlated with each other. These correlations vary between the parameters and between classes. The decision algorithm discussed in the next section will make use of it to differentiating between the classes.

Decision Algorithm

As mentioned before, the five measurements are used to classify the block of the signal as either silence, unvoiced, or voiced speech. To make this decision, a classical minimum probability–of–error decision is used in which it is assumed that the joint probability density function of the possible values of the measurements for the *i*th class is a multidimensional Gaussian distribution with known mean m_i and covariance matrix W_i. i=1,2,3 corresponds to class 1 (silence), class 2 (unvoiced), and class 3 (voiced), respectively.

For the decision rule, the distribution of the measurement does not need to be necessarily exactly normal. In the case of unimodal distributions, it is sufficient that the distribution be normal in the center of its range, which is often true for physical measurements.

Let X be an L dimensional column vector (in our case L=5) representing the measurements, that is the *k*th component is the *k*th measurement. The L-dimensional Gaussian density function for x with mean vector m_i and covariance matrix W_i is given by

$$g_i(X) = (2\pi)^{-L/2}|W_i|^{-1/2} \exp(-\frac{1}{2}(X - M_i)^H W_i^{-1}(X - M_i)) \qquad (4.6)$$

The decision which minimizes the probability error states that the measurement vector X should be assigned to class i if

$$pig_i(X) \geq pig_j(X) \qquad (4.7)$$

where Pi is the a priori probability that X belongs to the *i*th class. This decision rule, by throwing away some insignificant parts and manipulations, can be further simplified: the quantity distance $\hat{d}_i$ defined as

$$\hat{d}_i = (X - M_i)^H W_i^{-1}(X - M_i) \qquad (4.8)$$

is computed and the index i is chosen such that $\hat{d}_i$ is minimized.

In order to use the above decision algorithm, a training set of data is required to obtain the mean vector and the covariance matrix for each class. This training set is created by manually segmenting natural speech into regions of silence, unvoiced speech and voiced speech. The measurements mentioned above are made on each block of data. Let x_i denote the measurement vector for *n*th block for class (i=1, 2, 3) and Ni denotes the number of blocks manually classified as class i in the training set, then

$$W_i = \frac{1}{N_i}\sum_{n=1}^{N} x_i(n)x_i^r(n) - M_i M_i^H \tag{4.9}$$

$$M_i = \frac{1}{N_i}\sum_{n=1}^{N} x_i(n) \tag{4.10}$$

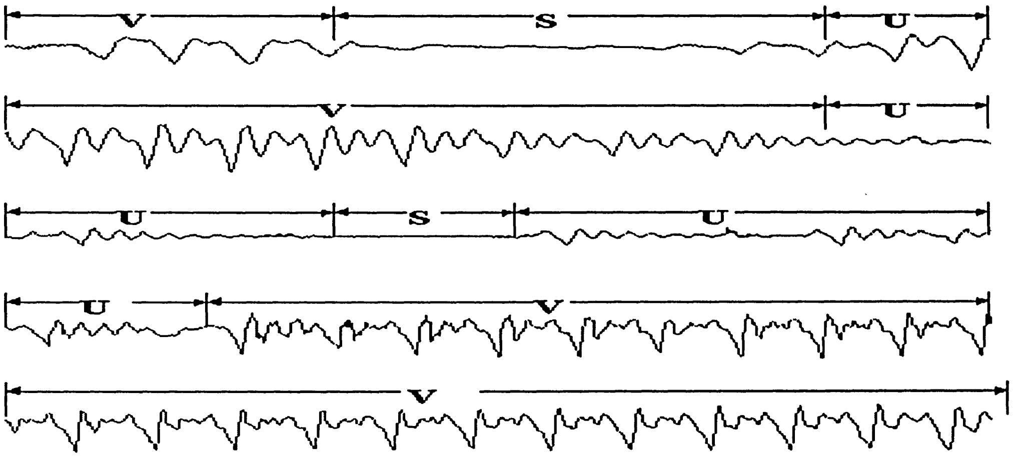


Figure 4.7. Comparison between actual data and V/U/S determination results

V- voiced, U-unvoiced, S-silence

An example of the speech waveform showing the various voiced, unvoiced, and silence regions as determined by the algorithm is shown in Fig. 4.6.

A fairly general framework based on a pattern recognition approach has been described in which a set of measurements are made on the interval being classified, and a minimum non-Euclidean distance measure is used to select the appropriate class. Almost any set of measurements can be used as long as there is some physical basis for assuming that the measurements are capable of reliably distinguishing between these three classes.

The major limitation of the method is the necessity for training the algorithm on the specific set of measurements chosen. Strictly speaking, the training data is particular to one set of recording conditions. Thus, whenever the transmission system varies or the background noise level varies, a new set of training data is required. If the recording conditions differ considerably from one occasion to another, it may be possible to adapt the algorithm by continuously updating the training data based on some measure of the relative distances to each of the classes. It is necessary to develop a new reliable voiced/unvoiced segment detection algorithm which is robust against noise environment.

4.2.3 Spectral Correlation Voiced/Unvoiced Segment Detection Algorithm

In the above section, voiced/unvoiced/silence segment detection algorithm including 5 measurement parameters was performed in the time domain on every frame. Although the detection result was adequate in clean speech frames, its performance reduced significantly under background noisy conditions. In order to solve this problem, we propose a frequency-domain voiced/unvoiced segment determination algorithm based on the spectral correlation and adaptive reference signal, threshold.

Spectral correlation in frequency domain

We consider the case where a single speech source is present is available. We use a window of length N, and assume that each N-sample frame of any signal we consider has been convolved with that window. The frame length N denote the amount of sample within a frame. Time domain input signals are denoted by small letters, e.g., x_n. We denote a frame as $x = (x_0, \ldots, x_n)$. The corresponding frequency domain signals are denoted by capital letters, e.g., X_k, $k = 1, \ldots, N$. The two are related by the Fast Fourier Transform (FFT). Also, a voiced reference signal $a_n (r=1,\ldots,N)$ is required for this algorithm. The corresponding frequency domain signals are A_k, k=$1,\ldots,N$

Then, we calculate the cross-correlation between the input signal and voiced reference signal in frequency domain.

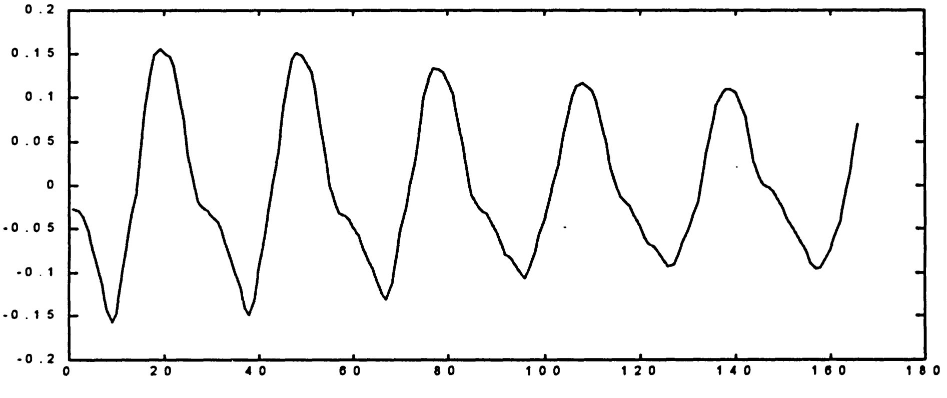


Fig. 4.8 voiced reference signal

$$R_i = \sum_{k=r=1}^{N} X_k * A_r \qquad i=1,\ldots,2N-1 \qquad (4.11)$$

We set H as the maximum of absolute value of R_i

$$H= max(R_i) \tag{4.12}$$

When H is bigger than the predefined threshold T_S, the segment of input signals are defined as voiced segment, otherwise, unvoiced segment.

Adaptive voiced reference signal

Let x_n be the windowed clean speech signal emitted at time n, and let y_n be the windowed noisy speech signal received at the microphone at the same time. Let u_n denote the windowed noise signal. Assuming additive noise, we have

$$y_n = x_n + u_n \tag{4.13}$$

In the frequency domain, (4.11) becomes $Y_k = X_k + U_k$. Denote the frame signals collectively by X, Y, U as in (4.13).

The cross-correlation function:

$$R_i = \sum_{k=1}^{N} (X_k + U_k) A_k \qquad 4.14)$$

According to the above cross-correlation function (4.14), we can see the background noise will affect the R_i, then performance quality will be reduced. To solve this problem, we introduce an adaptive reference signal. In the section 4.2.1, we know usually the energy of voiced segment is bigger than unvoiced segment, so we can use this speech property which the highest energy segment should be voiced segment (we assume the background is less than speech signal, SNR>0). Based on this idea, we will choose the segment with the highest signal energy as the voiced reference signal by comparing the signal energy of each frame.

The energy of frame: $$S = \sum_{n=1}^{N} |Y_n|^2 \qquad n=1,...N \tag{4.15}$$

N denotes the total number of sample each input frame. S denotes the signal energy of frame.

When the system detect that the background noise is less than a threshold T_e, the clean voiced reference signals will be employed again. However, when the background noise is bigger than a threshold T_e, the segment with the highest energy will become the voiced reference signals. So, this concept is called

Adaptive voiced reference signal. It is one of the key parts of the proposed algorithm.

Algorithm structure

The above three parts *(Spectral correlation in frequency domain, Adaptive voiced reference signal)* of spectral correlation voiced/unvoiced/ segment detection algorithm can be integrated for our voiced/unvoiced segment detection. The structure of algorithm is shown in Fig 4.9

According to the input speech signal and background noise, firstly, the background noise level (SNR) is measured to decide that what kind of voiced reference signal will be used based on the threshold T_s. Then, Fast Fourier Transform (FFT) is applied for input speech signal. The next step is to do spectral correlation calculation in frequency domain.

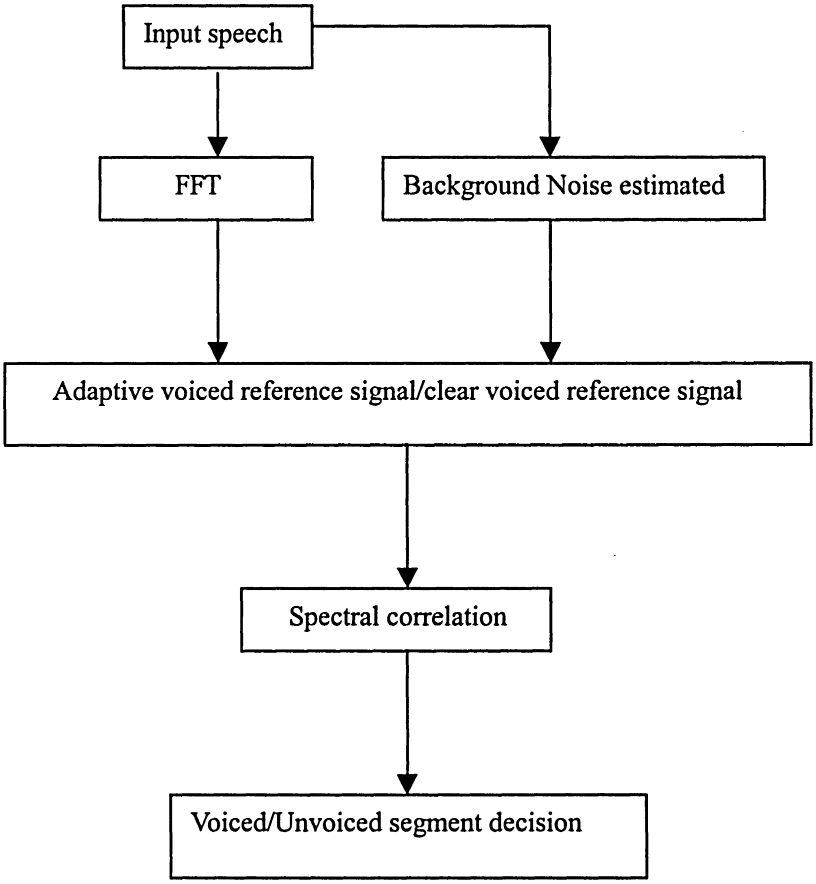

Input speech
FFT
Background Noise estimated
Adaptive voiced reference signal/clear voiced reference signal
Spectral correlation
Voiced/Unvoiced segment decision

Fig. 4.9 Block diagram of spectral correlation voiced/unvoiced detection

Compared with the adaptive threshold T_e, the input speech can be classified into voiced or unvoiced segment.

4.2.4 Robust Performance Measurement

During the test of this spectral correlation followed by voiced/unvoiced segment, it was found that the algorithm is more robust against background noise than the conventional algorithm, even the heavy noise environment. The robustness is useful in wireless communication and a critical component in the implementation of variable bit rate speech coders.

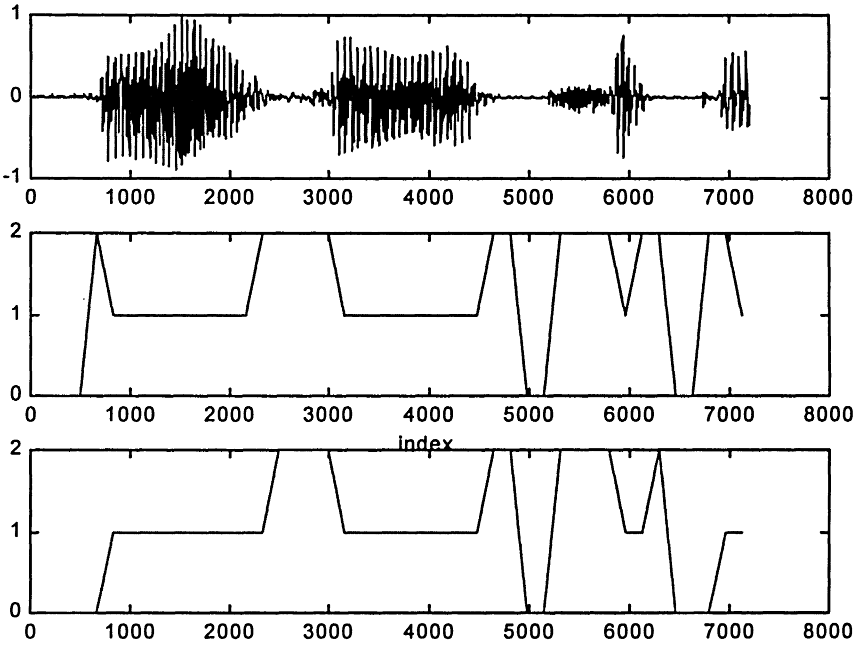


Fig 4.10(a) Performance comparison (clean background)

2---unvoiced 1---voiced 0---silence

In Fig 4.10(a), the top panel is the input clear speech, the second panel is the conventional detection algorithm, the bottom one is the proposed spectral correlation algorithm. From this diagram, it is so obvious that the proposed algorithm is more accurate for voiced or unvoiced segment detection.

The Fig 4.10 (b) shows that the conventional detection algorithm does not work in heavy background noise (10dB). However, our proposed spectral correlation detection in frequency domain algorithm can still detect voiced and unvoiced segment. It proves that our detection algorithm is more robust against noise.

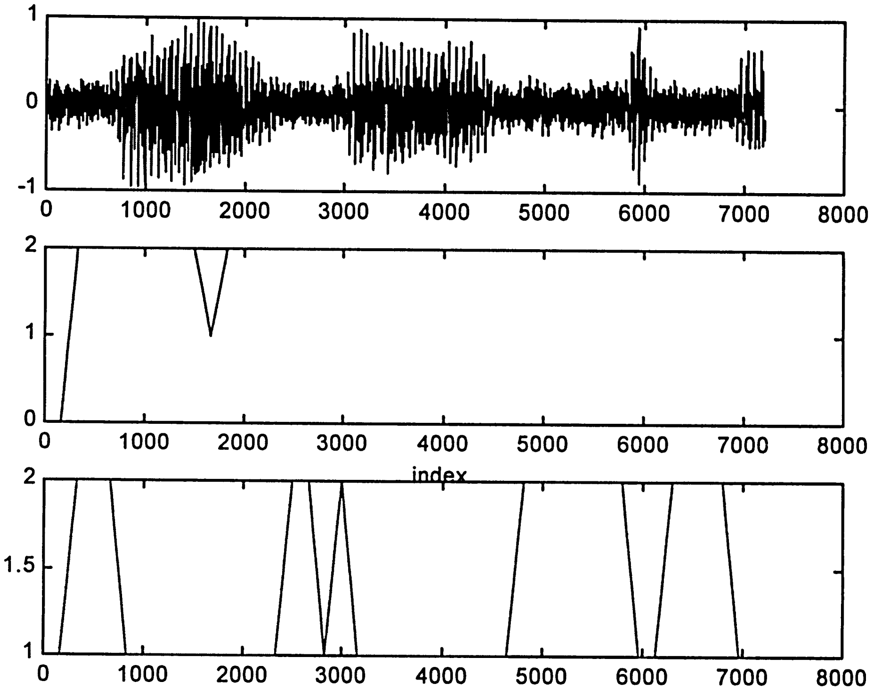


Fig. 4.10 (b) Performance comparison (noise background)

2---unvoiced 1---voiced

4.2.5 Variable Rate Coding of Voiced and Unvoiced Segment

Based on the speech characteristic, each coding frame can be classified into voiced, unvoiced and silence segment. Tests conducted on large speech files show that after silence suppression approximately 62% of the speech frames correspond to voiced speech, around 30% are unvoiced and 5% can be classified as onsets, transitions from unvoiced to voiced speech.

The required bit rate variability is realized by exchanging the adaptive codebooks and their corresponding gain codebooks while leaving the CELP coding scheme for all the other codec parameters invariant. The theory behind this technique is that for unvoiced segment, the adaptive codebook will not affect the quality of the synthesized speech. So, when we encode the unvoiced speech segment, we simply skip adaptive codebooks search and allocate 0 bit for adaptive codebook and corresponding gain. Furthermore, seamless bit rate switching can be realized by simply switching on or off their adaptive excitation codebooks search. The bit allocation for FS1016 CELP is shown in Table 4.1.

	Subframe							
Parameter	1		2		3		4	
	V	UV	V	UV	V	UV	V	UV
LSF	34							
ACB	8	0	6	0	8	0	6	0
G_{ACB}	5	0	5	0	5	0	5	0
SCB	9		9		9		9	
G_{SCB}	5		5		5		5	
other	6							
Total	144							

Table 4.1. Overall Bit allocation(v: voiced; uv: unvoiced)

ACB= Adaptive codebook; G_{ACB}= Gain of adaptive codebook; SCB= Stochastic codebook; G_{SCB}= Gain of stochastic codebook ;

4.3 Adaptive Forward-Backward Quantizer

Linear prediction plays a central role in various low and intermediate bit rate speech coding algorithms [18]. Usually, a new set of linear predictive coding (LPC) coefficients is determined every 20 to 30 ms and, after quantization, transmitted to the decoder as side information. To reduce the degradation of the speech quality caused by a direct quantization of LPC coefficients, Line Spectral Frequency (LSF) parameters are used for an indirect quantization and interpolation of predictor coefficients. Traditionally, scalar quantization of LSF coefficients was used. In FS1016 Federal Standard CELP [19] a total of ten LSF coefficients are scalar quantized to 34 bits-per-frame (bpf). Since the predictor coefficients are updated every 30 ms, the side information required for transmitting LSF parameters needs 1133.3 bps. The overall bit rate of FS1016 coder is 4.8 kbps, so more than 23% of the required bandwidth is spent on transmission of LSF coefficients.

It is well known that the speech signal is often slowly time-varying and non-stationary. The statistics between the current block and some temporally close previous blocks may often be similar leading to close sets of predictor coefficients. By allowing previous blocks to be over-lapped, the chance for statistical matching between the current block, and one of the so constructed temporally close previous block, will surely increase. By adapting the quantizer design to this new strategy, the "global" statistical correlation of speech signals will be more thoroughly exploited and a significant bit rate decrease is expected. To exploit the advantages of both forward and backward linear prediction, we introduce the following adaptive forward-backward coding scheme: A previously decoded and temporally close speech signal is segmented into overlapping blocks. If, and only if, the LPC coefficients calculated from one of those synthetic blocks is sufficiently "close" in some sense to the unquantized LPC coefficients calculated from the current speech block, the backward LPC scheme shall be applied, i.e., the LPC coefficients based on the previously decoded optimal speech block are used to encode the current block and only the time delay shall be transmitted to the decoder.

4.3.1 Adaptive Forward-Backward Quantization Analysis

As usual, the input speech is divided into non-overlapping blocks of L samples. For each block, the LPC coefficients are determined by using, e.g., the Levinson-Durbin algorithm. These LPC coefficients are optimal for the current speech block in the sense that the energy of the prediction residual signal is minimized .In traditional CELP coders, the LPC coefficients based solely on the current block are quantized by using either scalar or vector quantization scheme. In the following, we describe the adaptive forward-backward quantizer [20].

Conventional forward-backward quantization algorithm

The algorithm starts with defining the adaptive forward-backward LPC codebook, which consists of S code vectors each having p entries, where p represents the order of linear predictor. The *i*th code vector is determined by calculating the LPC coefficients, i.e., $\hat{a}_1^{(i)..}$, $\hat{a}_p^{(i)}$, based upon the previously decoded (synthetic) speech block $[y_{n-ik-L},\ y_{n-ik-L+1}, \ldots y_{n-ik-L}]$ that is available at both the encoder and decoder (see Fig. 1), where L is the length of the LPC block and *K* is the time delay unit chosen to be equal to the length of the sub-block, i.e., K = N. Then logarithmical spectral distortion (LSD) measure is used to evaluate the similarity between the previous and current set of LPC coefficients defined above. The LSD introduced by the ith code vector of the adaptive forward-backward LPC codebook is given by

$$LSD^{(i)|dB|} = \frac{10}{\ln 10}\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|V^{(i)}(\varpi)\right|^2 d\varpi} \qquad i=0,....S-1 \qquad (4.15)$$

$v^{(i)}(w)$ being defined by

$$V^{(i)}(w) = \ln\frac{1}{|A(w)|^2} - \ln\frac{1}{\left|\hat{A}^{(i)}(w)\right|^2} \qquad i=0,\cdots,S-1 \qquad (4.16)$$

where $A^{(i)}(z)=1+\sum_{l=1}^{p}\alpha_l^{(i)}z^{-l}$ and $\hat{A}^{(i)}(z)=1+\sum_{l=1}^{p}\hat{\alpha}_l^{(i)}z^{-l}$. As usual, a_1, ..., a_p denote the LPC coefficients based solely on the current speech block.

As seen, the LSD measure is determined for every candidate code vector. Then the one that has the smallest spectral distortion, i.e., $LSD^{(index)}$ with *index* = arg min$_i$ $LSD^{(i)}$, is selected from the adaptive LPC codebook. If $LSD^{(index)} > T$, a predefined threshold, then the current LPC coefficients, i.e., a_1, ..., a_p, are used in speech coding and, after quantization, transmitted to the decoder. If $LSD^{(index)} \leqslant T$, then the corresponding LPC coefficients, i.e., $\hat{a}^{(index)}{}_1$, ..., $\hat{a}^{(index)}{}_p$, are used in speech coding and only the index to the adaptive LPC codebook needs to be transmitted to the decoder. An additional flag bit is required to notify the decoder whether forward or backward linear prediction is applied at the encoder.

The application of this algorithm slightly increases the computational complexity at both the encoder and decoder. At the encoder, the increase in computational complexity is two-fold. First, for every new block, the code vectors in the adaptive LPC codebook need to be updated by the use of the newly decoded (synthetic) speech samples. However, each time, only $L=K=4$ code vectors which involve the most recently determined synthetic speech samples need to be calculated and added to the adaptive LPC codebook to replace the oldest code vectors. Second, the optimal code vector, which has the smallest LSD, needs to be selected from the adaptive LPC codebook. However, the computational complexity raised by the proposed algorithm at the encoder is negligible compared to the closed-loop excitation sequence generation of the CELP algorithm. This is because (1) each time only the four newest LPC code vectors are calculated to replace the four oldest ones and (2) instead of the LSD measure we have used the computationally less expensive COSH measure [], which is an upper bound of the LSD measure. At the decoder, if backward linear prediction is applied, the LPC coefficients are determined based on the previously decoded speech samples [$y_{n\text{-}ik\text{-}L}$, $y_{n\text{-}ik\text{-}L+1}$, ...$y_{n\text{-}ik\text{-}L}$]

The adaptive forward-backward LSF quantization scheme has been integrated into the FS1016 Federal Standard CELP coder. Extensive computer experiments showed that the bandwidth required for transmission of predictor coefficients was reduced by a factor of 2.7 with less then 1 dB drop in the segmental SNR and virtually no degradation in the perceived speech quality [21]. Although this scheme can be effective to reduce bit rate, the LSD measurement is quite complex. It is essential to develop new scheme to achieve adaptive forward-backward LSF quantization.

4.3.2 Improved Adaptive Forward-Backward Quantization Algorithm

Instead of using the LSD to make the decision for the coefficient similarity, here, we proposed a novel mean variance spectral coefficient similarity algorithm. The quantized LSF coefficient of the previous frame is $x_1,x_2,...,x_M$. The vector quantizaed LSF coefficient of the current frame is $y_1,y_2,...,y_M$.

Then use mean variance measure to evaluate the similarity between the previous and current set of LSF coefficients defined above. The mean variance introduced by the ith code vector of the adaptive forward-backward LPC codebook is given by

$$Mean \quad Variance = 10\log_{10}(\sqrt{\sum_{i=1}^{M}(y_i - x_i)^2}) \qquad i = 1,\dots,M \quad M = 10 \quad (18)$$

The threshold T = 18 dB was defined after performing 50 times. If mean variance > T, a predefined threshold, then the current quantized LSF coefficients, i.e., $y_1,y_2,...,y_M$ are used in speech coding and, transmitted to the decoder. If mean variance< T, then the corresponding previous LPC coefficients, i.e., $x_1,x_2,...,x_M$, are used in speech coding and only the flag bit needs to be transmitted to the decoder to notify the decoder whether forward or backward linear prediction is applied at the encoder. The adaptive forward-backward quantization of the LPC coefficients is summarized as follows:

At encoder:

Step 1. Calculate the LPC coefficients on the current speech block.

Step 2. Quantize the LPC coefficient to LSF coefficient. Compare the LSF of previous frame to calculate the mean variance.

Step 3. If Mean Variance<T, a predefined threshold, then the LSF coefficients of previous frame are used for coding the current speech block and the index is encoded and transmitted to the decoder. Set the flag bit to 1 to inform the decoder that backward linear prediction is applied at the encoder. Go to Step 5.

Step 4. If Mean Variance > *T*, then the LPC coefficients based on the current block (determined in Step 1) are used for coding the current speech block and, after scalar or vector quantization, transmitted to the decoder. The flag bit is set to 0 to inform the decoder that forward linear prediction is applied at the decoder.

Step 5 Encode the speech by using the LPC coefficients calculated for the current speech block in either Step 1 or Step 4.

At the decoder:

Step 1. If backward linear prediction is applied at the encoder (the received flag bit is 1), determine the LPC coefficients based on the previously decoded speech samples Go to Step 3.

Step 2. If the flag bit shows that forward linear prediction is applied at the encoder, receive the current LPC coefficients.

Step 3. Decode the speech by using the LPC coefficients determined in either Step 1 or Step 2.

Comparing the original Adaptive Forward-Backward Quantization algorithm, the proposed adaptive forward-backward quantization algorithms has the following advantages : (1) we use the quantized LPC coefficient LSP to make the adaptive forward –backward decision instead of unquantized LPC coefficients. (2) we employ mean variance instead of LSD to determine use the current or the previous frame coefficient.(3) The computation complexity of the proposed technique is significantly less than the original adaptive forward –backward quantization technique

We integrated this adaptive forward-backward quantization scheme into FS1016 CELP coder. The frame lengths of both 30 ms and 20 ms are used, which correspond to L = 240 and L = 160 samples, respectively, when the sampling rate is fs = 8 kHz. As is well known, the performance of the CELP coding technique depends on the block length. As a matter of fact, both the segSNR and the decoded speech quality improve as the coder parameters are updated more frequently. The results shown there are more backward quantization when the block length is reduced from 30 ms to 20 ms. It shows when we reduce the frame length, more bit rate can be saved because of more backward quantization in CELP coder.

4.4 Chapter Summary

In this chapter we have described the source controlled variable bit rate speech coder which is being implemented by detecting voice activity to switch on or off for speech signal transmission, splitting bit allocation for voiced and unvoiced subframes and calculate mean variance of LSP coefficient to adaptively apply forward/backward quantizer is discussed. Firstly the method for voice activity detection (VAD) of input signals. In order to make our coder very scalable, we employ a GSM wideband VAD algorithm. Next we gave the details of our proposed novel voiced and unvoiced segment detection algorithm we employed, as well as the comparison of the conventional corresponding algorithm. We discussed one of the main contributions of this novel algorithm is the robustness feature which is very important for our source controlled variable bit rate speech coder in real application; In third section of this Chapter we introduced the adaptive forward –backward quantizer scheme. In order to apply this algorithm, we modified the algorithm by using mean variance instead of LSD coefficient to detect the similarity of LPC coefficient between the previous and current frame. This approach also reduces the computation complexity significantly. Finally, we integrated the above three algorithm with the FS1016 CELP to realize the source controlled variable rate speech coding.

Chapter 5

Joint Channel and Source controlled Coding

The integration of the channel controlled and source controlled algorithm into a speech coder, and the simulation results are presented in this chapter. The popular speech codec in wireless communication, CELP was chosen as a platform for the simulation experiments. In the section 5.1, the structure of channel and source controlled AMR speech coding algorithm is briefly explained along with the fundamentals of CELP coders. The experimental setup used for evaluation of the performance for the proposed variable bit rate speech coding is described in section 5.2. The objective and subjective tests used to measure the speech coding quality are also presented in the section 5.3. In the section 5.4, results from the variable bit-rate AMR speech coder are presented.

5.1 The structure of Proposed Variable Bit Rate Speech Coder

Comparing with the original CELP coder, the characteristics of the proposed integrated channel and source controlled variable bit rate speech coder are as follows:

1. The length of a frame is adjusted according to the channel condition (C/I), and variable bit rate is achieved by allocating fixed number of bits to the variable length frames.
2. By using a voice activity detection method (VAD), the speech coder separates voice and silence speech frames. This produces a data signal whose rate switches between the full rate and severa.
3. For unvoiced speech subframe, CELP coding method is different from that of a voiced speech subframe. There is no adaptive code book search and the corresponding gains for the unvoiced subframe, and this results in significant bits saving.
4. For the LPC coefficients of CELP, the adaptive forward-backward quantizer is employed to decide if the speech decoder can reuse the previous frame coefficient or not. This technique also allows to achieve saving bit saving in bit allocation.

5.2 Simulation Experiment

For the channel controlled scheme, the variable frame length scheme was implemented in the FS1016 CELP (4800 kbps) speech coder using the same framework presented in Section 3.3. Compared to the FS1016 in which 30ms frame used, our variable frame length AMR speech coder uses 4 kinds of frame length: 15ms, 20ms, 20ms and 30ms according to the 4 channel condition model. In addition, the smooth switch algorithm is applied to eliminate the artifact caused as a result of bit rate switch and different time delay because of the variable frame length. The speech signals used in the simulation are sampled at 8 kbps, quantized at a 16 bits/sample resolution, and is pulse code modulation encoded.

For source controlled scheme, the proposed spectral correlation algorithm is implemented to classified the input speech into voiced and unvoiced segments. When the segment is unvoiced, we only do the stochastic code book search , then the bits for adaptive code book index and adaptive gain can be saved. By LP analysis, LPC coefficients were obtained for a frame. Two quantizer types: forward and backward were used to obtain the LSFs for the frames, and a threshold of 18 dB was used for this method. For less computation complexity, the mean square variance LPC parameters computed by the current and previous frame were implemented. If the mean variance

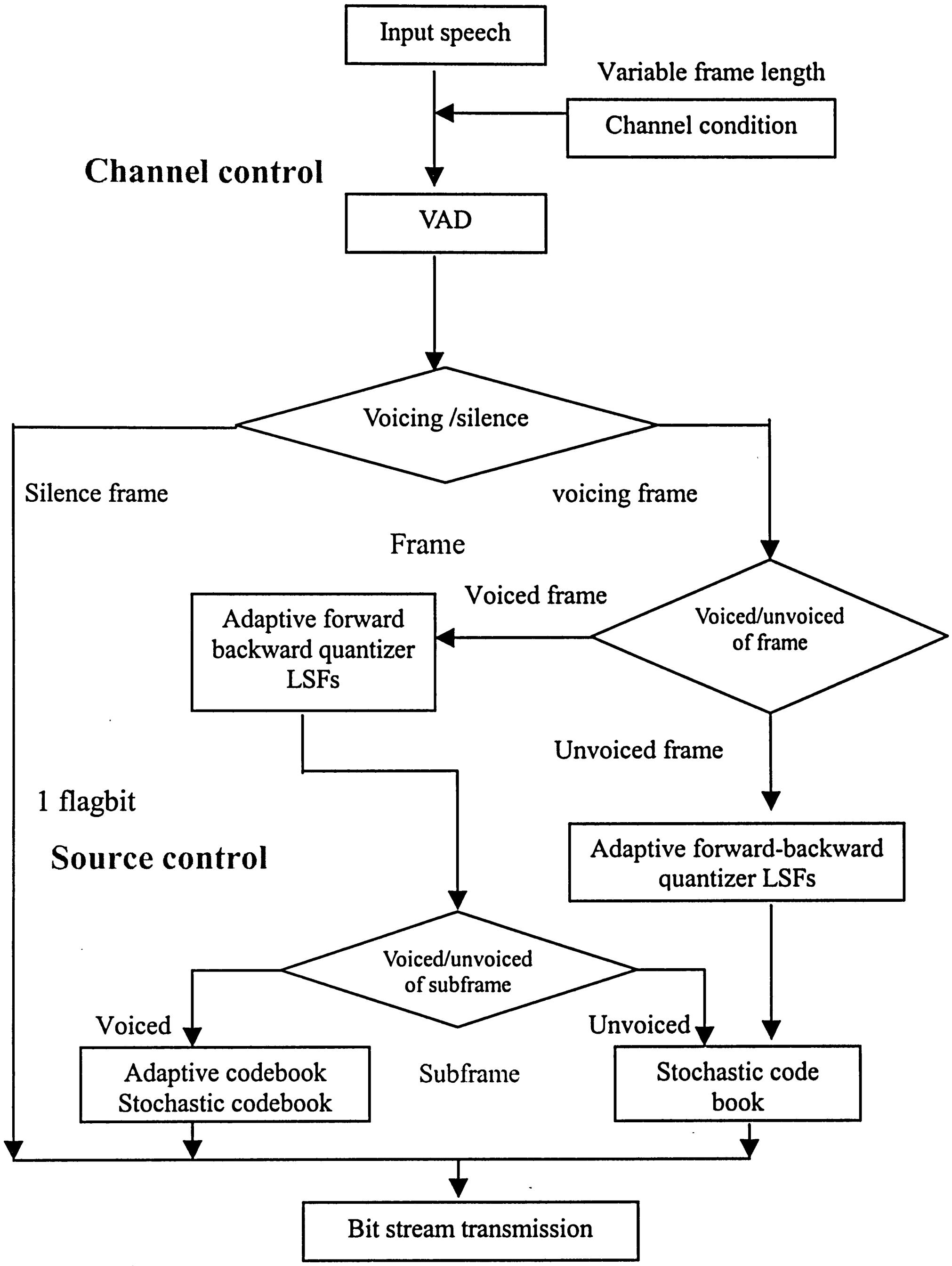


Fig 5.1 The structure of variable bit speech coder

of LSFs is less than the threshold, the backward quantizer will be applied, then the 34 bits for LSFs can be saved instead of 1 flag bit.

Steps for the simulation are as follows:

(1) Source controlled algorithms including variable length AMR and smooth switch algorithm are implemented based on channel condition (4 channel

condition level corresponding 4 kinds of frame length)

(2) Voicing activity detection (VAD) is used to detect if the speech signals are voicing or silence. For silence frame, only several control bits are transmitted in step 6. For voicing frame, the speech coder will go to the next step.

(3) Adaptive forward/backward quantizer is applied for obtaining LSFs coefficients.

(4) Based on the speech character of frame, we use the proposed spectra correlation algorithm which is key part of this project to detect either voiced or unvoiced frames. If the whole frame is detected as unvoiced frame, all of four sunframes will be labeled as unvoiced, then go to step 5.

(5) For voiced frame, the proposed spectra correlation algorithm is employed again to decide either voiced or unvoiced segment for each Subframe. For voiced Subframe, the speech coder will do both adaptive and stochastic code book search. However, for the unvoiced subframe, speech coder only do the stochastic code book search. Then 8 bit for odd subframe or 6 bits for even subframe and 5 bits for adaptive code book gain can be saved.

(6) Decode the speech by using the above parameters.

The simulation is programmed in Matlab. The demo will be shown in my presentation. So far, the primary goal of our project has been accomplished, that the coder described in figure 5.1 has been implemented to realize the bit rate variability. The average bit rate can be reduced significantly. The subjective and objective measures of the proposed joint channel and source controlled speech coder will be presented in the next section.

5.3 Results

In this section, we show the simulation results. For this simulation, 4 speech samples (2 male voice and 2 female voice) are tested by the proposed joint channel and source controlled variable bit rate speech coder (VBR). We compared the performance of our variable bit rate speech coding with that of FS1016 CELP coding. For source controlled algorithm, the minimum number of bits coding one frame is 1 (for LSF); 56 (for unvoiced frame), 6 additional bits In the FS1016 CELP 4800kbps coding, the number of bits per frame is 144. So, we can say that the minimum total bit rate of our variable speech coding is smaller than that of FS1016 CELP by 72 bits which is half of the fixed bit rate FS1016 CELP coding.

As the performance measure, the subjective measure segment SNR is used in this simulation. In the table 5.1 (a), average bit rate and segment SNR are shown. Frome the table 5.1 (a), we can see that (1) the proposed vocoder can significantly reduce average bit rate while maintaining the speech quality with little or not degradation since the maximum segment SNR variance is 0.17 dB. (2) The bits for adaptive forward-backward quantizer could be further reduced by the decreasing the frame length while we employ the channel control variable AMR. The main reason is that for the small frame length, the coefficients of the current and previous frame are almost identical.

		Bit Rate	SNR_{sgn}	Bit rate Reduce(%)
15ms	CELP	9600	8.65	30
	VBR	6670	8.5	
20ms	CELP	7200	7.27	21
	VBR	5663	7.11	
25ms	CELP	5760	6.27	16
	VBR	4829	6.13	
30ms	CELP	4800	5.82	14
	VBR	4143	5.65	

(a) Objective Measure

	Bit rate	MOS
CELP	4800	3.7
VBR	4143	3.5

(b) Subjective Measure (5 listeners)

Table 5.1 Simulation result

The subjective (MOS) measure in Table 5.1 (b) is also used to evaluate the simulation result. The MOS test was completed with 5 listeners. The listening quality and listening effort score is based upon a five point category judgement scale as follows:

Score	Listening Quality	Listening Effort
5	Excellent	Complete relaxation possible; no effort required
4	Good	Attention necessary; no appreciable effort required
3	Fair	Moderate effort required
2	Poor	Considerable effort required
1	Bad	No meaning understood with any feasible effort

Table 5.2 MOS category

We compared the performance of the proposed source controlled variable bit rate speech coding with frame length 30 ms to the fixed bit rate FS1016 CELP. The results show that our variable speech coding can significantly reduce average bit rate while maintaining the speech quality with little degradation since MOS scale for the proposed speech coder is only 0.2 lower than the FS1016 fixed rate coder.

The other advantages of the proposed coder include: (1) Bit-rate variability as a result of variable frame lengths, and this results in less memory requirements as compared to memory requirements of the multi-mode AMR used in GSM systems. why our coder has significant contribution. (2) We simply apply variable frame length scheme to achieve bit rate variability instead of multi-mode AMR scheme which is applying in GSM system. The smooth switch scheme eliminates the artifact due to the sudden bit rate switch between high and low bit rate model. (3) The proposed frequency domain spectra correlation method to detect voiced/unvoiced segment is more robust than the conventional approach based on zero crossing, signal energy. (4) The adaptive forward-backward quantizer algorithm has been modified to reduce the computational complexity.

5.4 Chapter Summary

In this Chapter we discussed the degree of success we obtained in integration with channel and source controlled variable bit rate speech coding to realize bit rate scalability. We show the whole structure of the our speech coding scheme and the simulation result We were satisfied by the performance of our bit rate scalability speech coding, in terms of the bit rate can be significantly reduced with little degradation of speech quality.

Chapter 6

Summary

This project introduced joint channel and source controlled variable bit rate speech coding methods to reduce the average bit rate in CELP-based speech coders. By applying the variable frame length and smooth switch scheme, the channel controlled bit rate variability could be obtained according to the channel condition. By employing the bit splitting allocation for voice and unvoiced segment and adaptive forward –backward quantizer algorithm, the source controlled bit rate variability according to the speech characteristic can be achieved.

6.1 Summary of The Work

After presenting the basic properties and types of speech coders, Chapter 1 outlines the objectives of the work. Chapter 2 provided an introduction to CELP, based on speech, and also gives an overview of different aspects of speech coders. The three aspects CELP: Short-term Linear Prediction, Long-term Linear Prediction, Stochastic Codebook search Distortion and performance measures to evaluate speech coder performance were described Chapter 2.

Chapter 3 builds a framework for the channel controlled variable bit rate algorithm. The overview of the multi-mode AMR was first investigated. Section 3.1 discussed the algorithm for multi-mode AMR by using different bit rate speech coding mode to change bit rate. In Section 3.2, new methods to achieve bit rate variability were examined. These included variable frame length, variable bit rate function and smooth switching technique. Chapter 4 describes how the source controlled variable bit rate coder was implemented. As a first step voicing activity detection (VAD) algorithm is implemented. Even though the conventional scheme accurately detect the voiced and unvoiced segment, but it did not work in a heavy background noise. For this reason, we proposed a spectral correlation voiced/unvoiced segment detection algorithm by calculating the correlation between the input signal and adaptive reference voiced signal in frequency domain then comparing it with a preset threshold. Simulation results suggested the robustness of the technique under heavy background noise conditions. Also the original algorithm of adaptive forward backward quantization was modified by using the mean variance instead of LSD to detect the similarity of LPC coefficient between the current and the previous frame

Chapter 5 presented the simulation results of joint channel and source controlled technique. The quality of the coded speech was assessed both by using subjective and objective measures. The most important thing is that the proposed speech coding algorithm can significantly reduce bit rate with little or no quality degradation, and the algorithm is simple and less complex than the current algorithms.

6.2 Future Work

The real-time utility of the algorithm was not evaluated in this project, and would be worthwhile step to implement the algorithm on a DSP platform. The computation time delay should be evaluated in the future due to the proposed algorithms. The proposed method for both voiced/unvoiced segment detection algorithm and adaptive forward-backward algorithm has to be improved for worse SNR conditions of below 10 dB. Also, the simulation experiments need to be further tested under noise background. Subjective tests based on distributed listening protocols (based on world wide web) would allow us to accurately assess the quality of the coder with a wider population.

References

[1] S. Dimolitsas and J. G. Phipps, Jr., "Experimental quantification of voice transmission quality of mobile-satellite personal communications systems," IEEE J. Select. Areas Commun., vol. 13, pp. 458 ▪ 464, Feb. 1995.

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

[3] N. S. Jayant and P. Noll, Digital Coding of Waveforms: Principles and Applications to Speech and Video. Englewood Clis, New Jersey: Prentice-Hall, 1984.

[4] D. O' Shaughnessy, Speech Communications: Human and Machine. New York: IEEE Press, second ed., 2000.

[5] U-T, coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), Mar.1996. ITU-T Recommendation G.279.

[6] Theodore S. Rappaport, "Wireless communication principles and practice", second edition.

[7] General Services Administration Office of Information Resources Management, "Telecommunications: Analog to Digital Conversion of Radio Voice by 4,800 bit/second Code Excited Linear Prediction (CELP)," February 1991.

[8] T. P. Barnwell, K. Nayebi, and C. H. Richardson, Speech Coding: A Computer Laboratory Textbook, John Wiley & Sons, 1996.

[9] P. Kabal and R. P. Ramachandran, "The Compuattion of Line Spectral 18. Frequencies Using Chebyshev Polynomials," IEEE Transaction on ASSP, pp. 1419-1426, vol. ASSP-34, No.6, December 1986.

[10] S. Dimolitsas, "Objective speech distortion measures and their relevance to speech quality assessments," IEEE Proc. Communications, Speech and Vision, vol. 136, pp. 317-324, Oct. 1989.

[11] S. Dimolitsas, F. L. Corcoran, and C. Ravishankar, "Dependence of opinion scores on listening sets used in degradation category rating assessments," IEEE Trans. coustics, Speech, Signal Pocessing, vol. ASSP-3, pp. 421-424, Sept. 1995.

[12] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, Objective Measures of Speech Quality. Englewood Clis, New Jersey: Prentice Hall, 1988.

[13] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis, Discrete-Time rocessing of Speech Signals. New York: IEEE Press, 2000.

[14] Olivier corbun, Magnus almgren and Krister savanbro, "Capacity and speech quality aspects using adaptive multi-rate (AMR), 1998, IEEE

[15] E. Paksoy, K. Srinivasan, A. Gersho, "Variable Rate Speech Coding with Phonetic Segmentation" Proc. of IEEE ICASSP, pp. 155-158, 1993 [16] S. Villette, M. Stefanovic, I. Atkinson, A Kondoz, "High Quality Split Band LPC Vocoder and its fixed point real-time implementation" Proc.of

EUROSPEECH, pp. 1243-1246, 1997

[17] 3GPP TS 26.194 V5.0.0 (2001-03)

[18] A. Gersho, "Advances in speech and audio compression," Proceedings of IEEE, vol. 2, pp.900—918, 1994.

[19] J.P. Campbell; V.C.Welch, and T.E. Tremain, "An expandable error-protected 4800 BPS CELP coder (U.S. Federal Standard 4800 BPS voice coder)," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1989, pp. 735-38

[20] Zijun Yang, Jozsef Vass, Yunxin Zhaoy, and Xinhua Zhuang. "A novel variable rate LPC quantizer for high performance speech coder", 1995. Internet

[21] A.H. Gray and J.D. Markel, "Distance measures for speech processing," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.24, pp. 380-391, Oct. 1976.

[22] Ekudden, E.; Hagen, R.; Johansson, I.; Svedberg, J "The adaptive multi-rate speech coder " Speech Coding Proceedings, 1999 IEEE Workshop on, 1999 Page(s): 117 –119

[23] Chen Xinfu; Zhang Zhengyang; Yi Kechu; AMR vocoder and its multi-channel implementation based on a single DSP chip Info-tech and Info-net, 2001. Proceedings. ICII 2001 - Beijing. 2001 International Conferences on, Vol.2, 2001 Page(s): 650 –655.

[24] Hindelang, T.; Hagenauer, J.; Schmautz, M.; Xu, W. "Channel coding techniques for adaptive multi rate speech transmission" Communications, 2000. ICC 2000. 2000 IEEE International Conference on, Vol. 2 , 2000 Page(s): 744 –748.

[25] Beritelli, F.; Casale, S.; Ruggeri, G., "Performance comparison between VBR speech coders for adaptive VoIP applications" Communications, 2002. ICC 2002. IEEE International Conference on, Vol.4 , 2002 Page(s): 2578 -2582.

[26] McCree, A.; Unno, T.; Anandakumar, A.; Bernard, A.; Paksoy, E.; "An embedded adaptive multi-rate wideband speech coder" Acoustics, Speech, and Signal Processing, 2001. Proceedings. 2001 IEEE International Conference on, Vol.2, 2001 Page(s): 761 –764.

[27] Bessette, B.; Salami, R.; Lefebvre, R.; Jelinek, M.; Rotola-Pukkila, J.; Vainio, J.; Mikkola, H.; Jarvinen, K.; "The adaptive multirate wideband speech codec (AMR-WB)" Speech and Audio Processing, IEEE Transactions on, Vol.10 No.8, Nov 2002 Page(s): 620 –636.

[28] Bessette, B.; Salami, R.; Lefebvre, R.; Jelinek, M.; Rotola-Pukkila, J.; Vainio, J.; Mikkola, H.; Jarvinen, K.; "Real-time implementation and evaluation of variable rate CELP coders" Speech and Audio Processing, IEEE Transactions on, Vol.10 No.8, Nov 2002 Page(s): 620 –636.

[29] Haskell, B.G.; Reibman, A.R.; " Variable bit-rate video coding for ATM and broadcast applications" Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on, Vol.1, 27-30 Apr 1993 Page(s): 114 –116.

[30] Exford, J.; Towsley, D.; "Smoothing variable-bit-rate video in an

internetwork"Networking, IEEE ACM Transactions on , Vol.7, Issue: 2 , Apr 1999 Page(s): 202 –215

[31] Oshikiri, M.; Akamine, M.; "A 2.4 kbps variable bit rate ADP-CELP speech coder" Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on , Vol.1, 12-15 May 1998 Page(s): 517 –520.

[32] Uvliden, A.; Bruhn, S.; Hagen, R.; " Adaptive multi-rate. A speech service adapted to cellular radio network quality" Signals, Systems & Computers, 1998. Conference Record of the Thirty-Second Asilomar Conference on, Vol.1, 1-4 Nov 1998 Page(s): 343 –347.

[33] Ito, H.; Serizawa, M.; Ozawa, K.; Nomura, T.; "An adaptive multi-rate speech codec based on MP-CELP coding algorithm for ETSI AMR standard" Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on, Vol.1, 12-15 May 1998 Page(s): 137 –140.

[34] Vahatalo, A.; Johansson, I.; "Voice activity detection for GSM adaptive multi-rate codec" Speech Coding Proceedings, 1999 IEEE Workshop on, Page(s): 55 –57.

[35] Paksoy, E.; Carlos de Martin, J.; McCree, A.; Gerlach, C.G.; Anandakumar, A.; Wai-Ming Lai; Viswanathan, V.; "An adaptive multi-rate speech coder for digital cellular telephony" Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on, Vol.1, 15-19 Mar 1999 Page(s): 193 –196.

[36] Ekudden, E.; Hagen, R.; Johansson, I.; Svedberg, J.; "The adaptive multi-rate speech coder" Speech Coding Proceedings, 1999 IEEE Workshop on , 1999 Page(s): 117 –119.

[37] Hindelang, T.; Hagenauer, J.; Schmautz, M.; Xu, W.; "Channel coding techniques for adaptive multi rate speech transmission" Communications, 2000. ICC 2000. 2000 IEEE International Conference on, Vol.2, 2000 Page(s): 744-748.