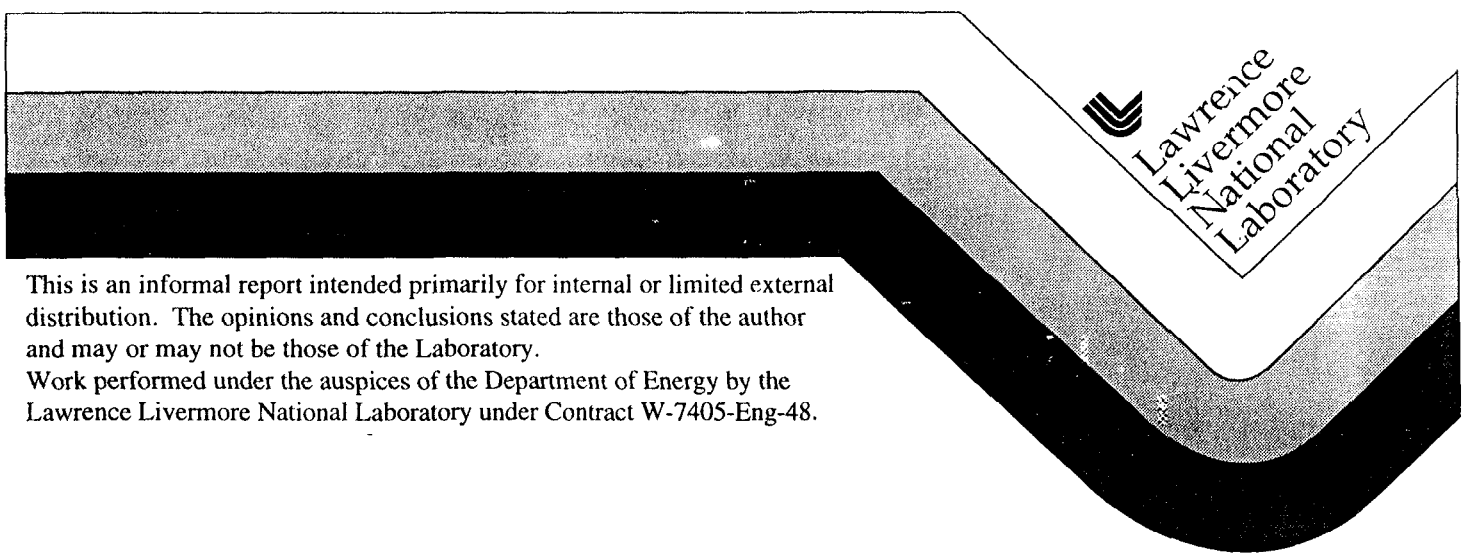
UCRL-ID-130167

Speech Coding

Channasandra Ravishankar, Hughes Network Systems
Spiros Dimolitsas, Lawrence Livermore National Laboratory

May 8, 1998



FILENAME.APP = 6709MS00.DOC

SPEECH CODING

Channasandra Ravishankar
Hughes Network Systems
11717 Exploration Lane
Germantown, Maryland 20876

and

Spiros Dimolitsas
Lawrence Livermore National Laboratory
P. O. Box 808, L-151
Livermore, California 94551

Speech is the predominant means of communication between human beings and since the invention of the telephone by Alexander Graham Bell in 1876, speech services have remained to be the core service in almost all telecommunication systems. Original analog methods of telephony had the disadvantage of speech signal getting corrupted by noise, cross-talk and distortion Long haul transmissions which use repeaters to compensate for the loss in signal strength on transmission links also increase the associated noise and distortion. On the other hand digital transmission is relatively immune to noise, cross-talk and distortion primarily because of the capability to faithfully regenerate digital signal at each repeater purely based on a binary decision. Hence end-to-end performance of the digital link essentially becomes independent of the length and operating frequency bands of the link Hence from a ***transmission point of view*** digital transmission has been the preferred approach due to its higher immunity to noise.

The need to carry digital speech became extremely important from a ***service provision point of view*** as well. Modern requirements have introduced the need for robust, flexible and secure services that can carry a multitude of signal types (such as voice, data and video) without a fundamental change in infra-structure. Such a requirement could not have been easily met without the advent of digital transmission systems, thereby requiring speech to be coded digitally.

The term "Speech Coding" is often referred to techniques that represent or "code" speech signals either directly as a waveform or as a set of parameters by analyzing the speech signal. In either case, the codes are transmitted to the distant end where speech is reconstructed or synthesized using the received set of codes. A more generic term that is applicable to these techniques that is often interchangeably used with "speech coding" is the term "voice coding". This term is more generic in the sense that the coding techniques are equally applicable to any voice signal whether or not it carries any intelligible information, as the term "speech" implies. Other terms that are commonly used are "speech compression" and "voice compression" since the fundamental idea behind speech coding is to reduce (compress) the transmission rate (or equivalently the bandwidth) and/or reduce storage requirements In this document the terms "speech" and "voice" shall be used interchangeably.

1.1 Digital Speech : An Introduction

Here speech, which is a continuous time, continuous amplitude signal is typically sampled at a given rate and each sample is represented (or quantized in speech processing terminology) using a certain number of bits. Sampling is the process of converting a continuous time signal such as speech to a discrete time signal. The duration between samples, or, inversely the sampling rate that is used for speech is governed by Nyquist sampling theorem which depends upon the speech spectral characteristics as described below. It is observed that speech energy falls off rapidly after 4 kHz and that, the intelligibility and practically all of the naturalness and talker features are present in the speech spectrum below 3.5 kHz. Thus according to the Nyquist sampling theorem, if speech is filtered by a sharp cutoff analog filter prior to sampling so that the maximum frequency component is 4 kHz, then a sampling rate of 8 kHz (or 8000 samples per second) can be used without losing any information. Practically all speech coders used for telephony applications use 8 kHz sampling rate. For specialized applications such as audio/videoconferencing, wideband speech coding techniques are used where sampling rates are as high as 20 kHz. Unless otherwise specified, it is assumed in this article that for digital speech coding techniques, a sampling rate of 8 kHz shall be used.

Quantization is the process whereby the discrete time continuous amplitude samples are converted to discrete-time discrete-amplitude samples; and the discrete amplitude signal is represented using a certain number of bits. Conversion from continuous amplitude sample to a discrete amplitude sample is performed simply by dividing the dynamic range of the signal into discrete levels and approximating the input sample to the level closest to it. It is obvious then, that more the number of bits that are used, the better will be the representation or lower will be the error due to quantization in the reconstructed signal. The process of sampling and quantization is illustrated in Figure 1. Figure 1 also illustrates the effects of increased quantization distortion due to decreased number of bits. It will be shown in Section 2.1 that the signal to quantization noise ratio increases (or decreases) by 6 dB for every increase (or decrease) of one bit. If f_s is the sampling rate (in number of samples per second) and B is the number of bits per sample, then the

channel capacity required to transmit digitally represented speech is $C = f_s B$. It can be shown using Shannon's channel capacity theorem that larger the value of C, larger will be the bandwidth required to transmit on a channel with a given signal-to-noise ratio As described above, for a given speech bandwidth (4 kHz for telephony applications), the sampling frequency f_s is fixed at 8000 samples per second and the only other variable that controls the channel bandwidth required is B, the number of bits per sample.

For a given channel bandwidth it is highly desirable to reduce C (or B) as low as possible in order to accommodate bits from multiple users to be multiplexed on the same channel bandwidth. However, as noted above, B cannot be arbitrarily reduced since the signal to quantization noise ratio (and hence quality of reconstructed signal) degrades significantly with every reduction in one bit. ***It is then obvious that the objective of digital speech coding is to satisfy two conflicting requirements : On the one hand it is required to maintain good speech quality and on the other hand it is required to reduce the bit rate as low as possible.*** To achieve this objective, most low bit rate speech coding techniques rely on a parametric approach, whereby a block (usually called a "frame" in speech coding terminology) of speech samples are represented or "coded" by a minimal set of parameters that will permit reconstructing speech with a desired quality. While the set of parameters that speech coders use vary from one technique to another, practically all modern day speech coding techniques have relied heavily upon the vast amount of research that was conducted for the past several decades in areas of speech production, perception, analysis, and, synthesis [2] for the choice of a given set of parameters. The set of parameters are then quantized and transmitted digitally to the remote decoder where speech is reconstructed

In this article, speech coders that use the digitization technique described above where each sample is represented using a certain number of bits will be referred to as ***waveform coders*** since the objective there is to achieve a reconstructed signal whose waveform is as close to the original as possible. On the other hand, speech coders that parameterize speech signals purely based on extracting parameters of an assumed model without necessarily having the objective of reproducing a signal whose waveform looks like the

original speech waveform shall be referred to as ***Parametric Speech Coders*** or even more simply referred to as "Vocoders".

It is noted that speech coders have been categorized into various other categories in the literature depending upon the criterion used for classification. For example speech coders have been classified as high bit rate, medium bit rate and low bit rate using bit rate as the criterion; wireline or toll quality, cellular quality, communications quality, intelligible quality and synthetic quality using speech quality as criterion; time domain, frequency domain, quefrency domain and time-frequency domain using the domain in which speech processing is performed.

1.2 The "Vocoder" : A Historical Perspective

As described above, the concept of speech coding has been an area of study for several decades and the first "vocoder" apparatus was reported as early as 1939 by Homer Dudley in his paper titled "Remaking Speech" as described in [1]. The vocoder apparatus consisted of electrical circuits which first analyzed speech signals that extracted certain parameters and then synthesized or "remade" using those set parameters. These parameters that were extracted to synthesize speech were based on the understanding gained from a significant amount of work in the areas of speech analysis and synthesis prior to Dudley's automatic (and almost instantaneous) vocoder apparatus. These efforts led to the understanding that in order to produce intelligible speech, the analyzer had to extract pitch (determined by fundamental frequency) information of the talker, spectrum information in terms of the relative power in different frequency bands and an intensity determined by the total sound power.

At the synthesizer two stream of sounds were generated based on pitch extraction by the analyzer. Properly controlled variations of these two streams generated intelligible synthetic speech. The first type of sound stream was generated when the analyzer determined a non-zero pitch in the talker's speech signal and the synthesized speech signal was hence characterized by the fundamental frequency of the talker, the spectrum shape and the intensity. The second type of sound stream was generated when the analyzer observed a zero pitch in talker's speech and the synthesized speech signal was characterized by random

frequency components (independent of the talker or speech material), the spectrum shape and intensity. Figure 2 illustrates the plan of the circuit [1] used by Homer Dudley for his vocoder apparatus used to remake speech. It is noted that long before Dudley's real-time vocoder apparatus, serious efforts were put into development of manually operated speech synthesizer (or speaking machines) as early as 1791 by Von Kempelen, and, later on an, electrically operated vowel synthesizer in 1922 by Stewart. However, as noted above, these efforts led to "making" of speech unlike Dudley's vocoder apparatus which performed electrical analysis of human speech and in real-time performed electrical "remaking" (or synthesis) of speech

A significant amount of research was simultaneously being conducted to understand the human speech production mechanism that would serve to benefit workers in multitude of disciplines in a variety of ways. For phoneticians and linguists these studies would provide them a tool to describe in simple ways the acoustical features associated various phonemes in different languages. For physiologists, laryngologists, and, physicists these studies would help them detect, diagnose and isolate problems related to organs involved in human speech production mechanism. For communication engineers these studies would help them determine the essential features that need to be preserved in speech events in order to reconstruct intelligible speech. In this way, only the essential features of speech events need to be transmitted rather than the speech signal itself, thereby achieving significant bandwidth compression.

1.3 Speech Production Mechanism

The vocal system consisting of the vocal tract, the nasal tract and the lungs is responsible for producing the various sounds in human beings [3]. A schematic is illustrated in Figure 3. The vocal tract begins at the opening between the vocal cords, or glottis, and ends at lips The nasal tract begins at the velum and ends at the nostrils. The lungs is the source of energy for the production of speech. Speech is simply the acoustic wave that is radiated from the vocal system when air is expelled from the lungs. Physiological features such as lengths and cross sections of the vocal tract and nasal tracts and the tensions in vocal cords distinguishes different talkers for the same sound. The relative positions and cross sections within the vocal

tract and nasal tract, as well as the positions of lips and velum distinguishes different sounds generated by the same talker.

Speech sounds are broadly classified into three different classes depending on their mode of excitation, namely voiced sounds, unvoiced sounds and plosive sounds. Voiced sounds are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses of air which excite the vocal tract. Unvoiced sounds or fricatives are generated by forming a constriction at some point in the vocal tract (usually towards the mouth end), and forcing air through the constriction at a high enough velocity to produce turbulence. This creates a broad spectrum noise source to excite the vocal tract. Plosive sounds results from making a complete closure at or near the lip region, building up pressure behind closure, and abruptly releasing it

The vocal tract and nasal tract in Figure 3 have been modeled as tubes of non-uniform cross sectional areas for the purposes of analysis [3]. As the air expelled from the lungs propagates along these two tubes, the frequency spectrum of the resulting sound is determined by the resonant frequencies of the tubes. These resonant frequencies are popularly known as formant frequencies in speech processing. Different sounds are formed by varying the shape of the vocal tract, thereby yielding different formant frequencies for different sounds. Thus spectral properties of speech signal vary with time as the vocal tract shape varies

The time-varying spectral characteristics of speech signal was first graphically displayed using a spectrograph [4], whereby speech energy at different frequencies as a function of time could be observed. The two dimensional pattern (horizontal time axis and vertical frequency axis) produces dark patches in regions where signal energy is high and light patches when signal energy is low. Voiced regions are typically characterized by dark striated appearance due to periodicity in the time waveform, whereas unvoiced regions are characterized by lighter and uniformly filled patches. Unlike unvoiced regions plosives appear darker on the spectrogram with a sharp transition from lighter bands, and unlike voiced regions plosives typically do not exhibit the same amount of periodicity.

The article is organized as follows : Waveform coding techniques are described in Section 2 and parametric coding techniques using linear prediction coding (LPC) are discussed in Section 3. Section 4 deals with modeling the excitation of the human speech production mechanism. This will be used as the input to LPC synthesis filter at the remote speech decoder. Section 5 deals with some important speech coding techniques that are not processed in time-domain; rather speech is transformed into a different domain and then analyzed to extract parameters of a given speech model. Section 6 describes some important international and regional speech coding standards that are in use today based on techniques described in Sections 2, 3, 4 and 5. Section 7 provides a qualitative overview of the methods involved in assessment of speech codec performance. The effects of having multiple speech coding technologies as a result of having multiple links in an end-to-end connection is discussed in Section 8. Future trends in speech coding is discussed in Section 9 followed by conclusions in Section 10.

2. Waveform Coders

As discussed in the introduction, waveform coders strive to achieve the objective of encoding speech in a manner that will permit reconstructing speech that is close to original sampled speech waveform on a sample-by-sample basis. The first and most popular waveform coding technique that is being predominantly used in most national digital telephone networks is the Pulse Code Modulation (PCM) technique. Another waveform coding technique that is being increasingly used on international satellite links and in some large national networks is Adaptive Differential PCM or ADPCM. Other waveform coding techniques (that are successors of PCM and predecessors of ADPCM) such as Adaptive Delta Modulation (ADM) and Continuously Variable Slope Delta (CVSD) Modulation are being used in some private networks such as military communication. These techniques are simply special cases of ADPCM. In the sequel, some fundamentals of PCM, ADM and ADPCM techniques are described.

2.1 Pulse Code Modulation

Here speech samples are quantized in the speech encoder using B bits per sample and these B bits are transmitted to the PCM decoder. The decoder uses the received bits to reconstruct speech as shown in Figure 4. The choice of B depends on the available channel capacity. The choice of B determines the quantization step size and hence signal-to-quantization noise ratio as described below.

Let $s(n)$ represent a speech sample at time instant n and $\tilde{s}(n)$ be its quantized value. Let Δ be the step size of the quantizer. The value of Δ satisfies the equation

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

where $S_{\max}$ is the maximum amplitude of the speech signal into the quantizer.

Then

$$\tilde{s}(n) = s(n) + e(n)$$

where $e(n) \in (-\frac{\Delta}{2}, \frac{\Delta}{2}]$

The signal-to-quantization noise ratio SNR_q in decibels (dB) for such a configuration is defined by

$$SNR_q = 10\log_{10} \frac{E\{s^2(n)\}}{E\{e^2(n)\}} \tag{2-2}$$

where $E\{\ \}$ denotes the mathematical expectation or mean or the first moment of the random variable under consideration. In practice, the $E\{\ \}$ is replaced by an unbiased estimate of the mean based upon short segments of speech. As discussed in the introduction, speech signal can be considered as a nonstationary random process whose characteristics changes slowly in time depending on the type of sound that is being produced and the talker that produces it. Therefore, in reality $E\{s^2(n)\}$ and hence SNR_q is a time-varying quantity.

Assuming that e(n) is uniformly distributed between $(-\frac{\Delta}{2}, \frac{\Delta}{2}]$ we have

$$E\{e^2(n)\} = \frac{\Delta^2}{12}$$

Substituting the value of $E\{e^2(n)\}$ in equation (2-2) above and substituting for Δ from equation (2-1) above it can be shown that SNR_q can be written in the form

$$SNR_q = 10\log_{10} E\{s^2(n)\} + 6B + f(S_{\max})$$

From the above analysis, it is clear that (i) the signal-to-quantization ratio decreases by 6 dB when the number of bits B per sample is reduced by one bit and (ii) for an assumed $S_{\max}$ and a chosen number of bits B per sample, the short-term SNR_q is a monotonic function of $E\{s^2(n)\}$ implying that SNR_q is poor for durations where speech signals have smaller amplitudes as compared to durations where speech signal has larger amplitudes. Item (ii) above is highly undesirable since studies on speech signal characteristics have repeatedly shown that speech signal amplitudes are less than $S_{\max}/4$ for most of the time Hence a quantization scheme is desirable whose step size is smaller for lower amplitude speech signals and larger for larger amplitude signals. Such a quantizer is called non-uniform quantizer. In practice, non-uniform quantization is achieved by first transforming the signal amplitudes in a manner that will result in approximately a uniform distribution and then perform uniform quantization. At the decoder, an inverse transformation is applied to retain its original distribution. In this manner, the SNR_q can be significantly increased, or in alternate terms, it is possible to use lesser number of bits with non-uniform quantization than with uniform quantization for obtaining the same SNR_q. Two types of signal transformation that are being widely used worldwide are called the μ-law and A-law The μ-law which is used in North American telephone networks is defined as follows:

$$s_c(n) = S_{\max} \frac{\log\left(1 + \mu \frac{|s(n)|}{S_{\max}}\right)}{\log(1+\mu)} sign[s(n)]$$

The A-law method of compressing speech signal which is employed in European telephone networks is defined as

$$s_c(n) = \frac{As(n)}{1+\log A} \qquad for\ \ 0 \le |s(n)| \le \frac{S_{max}}{A}$$

$$= S_{max} \frac{1+\log\left(\frac{A|s(n)|}{S_{max}}\right)}{1+\log A} sign[s(n)] \qquad for\ \frac{S_{max}}{A} < |s(n)| \le S_{max}$$

In both cases $s_c(n)$ represents the compressed version of original speech s(n). A mapping table corresponding to these laws is provided in International Telecommunications Union - Telecom Sector (ITU-T, formerly CCITT) Recommendation G.711 for PCM signals

Due to the nature of the logarithmic curves in both A-law and μ-law transformations which tend to COMpress larger amplitude signals, and the inverse transformations (exponential in nature) which tend to exPAND large amplitude signals, the two functionalities together have been popularly called COMPANDING and speech signals that have undergone companding are called companded PCM. It has been found that the signal quality of a 13 bit uncompanded PCM is equivalent to that of a 8 bit companded PCM [5]. Today, μ-law and A-law companding are used in almost all telephone networks and each speech sample is represented using 8 bits so that the transmission bit rate of PCM signal is 64 kbps.

2.2 Adaptive Differential PCM (ADPCM)

PCM technique is extremely robust across a variety of speech signals, since it does not make any inherent assumptions about the time-varying spectral characteristics of speech signal. However, the bit rate of 64 kbps can be prohibitively high, especially in bandwidth-scarce links such as satellite links. Hence there arose a need for a lower bit rate speech coding technique. This led to the development of speech coders that encoded differences between adjacent samples or difference between the current sample and a predicted

value of the current sample (based on previous samples), broadly referred to as Differential PCM or DPCM The DPCM speech coder is based upon (i) the observation that adjacent speech samples are highly correlated and (ii) correlated speech sample can be predicted whose associated prediction error has a significantly small variance as derived below :

let s(n) be the speech sample at time instant n. For simplicity, let the prediction formulation be of the form

$$s(n) = \alpha\ s(n-1) + d(n)$$

Then it can be shown that mean-square value of prediction error d(n), or $E\{d^2(n)\}$ is minimum when

$$\alpha = \alpha^* = E\{s(n)s(n-1)\}/E\{s^2(n)\}$$

i.e, when α is equal to the correlation coefficient (assuming s() to be zero mean and identically distributed) between adjacent samples s(n) and s(n-1). The corresponding mean square error between predicted value and actual value is then given by

$$E_{\min} = E\{s^2(n)\}(1-\alpha^{*2})$$

It is therefore easy to see that when correlation α^* is large, mean square error between actual speech sample and predicted sample becomes small, and hence fewer bits are needed to represent d(n) The prediction formulation above is often referred to as a linear first order predictor, since the current sample is being predicted based upon one previous sample and relationship between s(n-1) and s(n) is linear. In practice, for speech signals, the prediction formulation is usually is of a higher order of the form

$$s(n) = \sum_{i=1}^{N} \alpha_i s(n-i) + d(n) \qquad (2\text{-}3)$$

It is important to note that, in general, it can be shown that the mean-square value of the prediction error $E\{(s(n) - f(s(n-1), s(n-2), .., s(n-N))^2\}$ is minimum if

$$f() = f^*() = E\{s(n) / s(n-1), s(n-2), .., s(n-N)\}$$

In practice however, $f^*()$ is assumed to be linear in s(n-1),s(n-2),..,s(n-N) as shown in equation (2-3) primarily because of the simplicity and analytical tractability that a linear formulation provides as compared to that of a nonlinear formulation as shown above. A secondary but important reason for formulation of linear predictors in most systems stems from the fact that $f^*()$ in equation above is indeed linear in s(n-1),s(n-2) ,. , s(n-N) if joint probability distribution of s(n), s(n-1), .. , s(n-N) obeys a Normal distribution.

Thus, if α^* and d(n) are transmitted and s(0) is known, then it is possible for the decoder to exactly reproduce s(n) for any n > 0. However, in practice d(n) has to be quantized to $d_q(n)$ before transmission

Hence decoder output $\tilde{s}(n)$ will not be equal to s(n). In such an event if the formulation at the decoder is of the form

$$\tilde{s}(n) = \alpha \tilde{s}(n-1) + d_q(n)$$

Such a formulation however leads to a situation where the difference between reconstructed speech sample $\tilde{s}(n)$ at the output of the ADPCM decoder and input speech sample $s(n)$ at the input of ADPCM encoder is an accumulation of quantization errors $d(m) - d_q(m)$, $0 \le m \le n$. For the single order predictor formulation it can be shown that

$$\begin{aligned} s(n) - \tilde{s}(n) &= \alpha(s(n-1) - \tilde{s}(n-1)) + d(n) - d_q(n) \\ &= \alpha^n(s(0) - \tilde{s}(0)) + \sum_{i=0}^{n-1} \alpha^i (d(n-i) - d_q(n-i)) \end{aligned}$$

To avoid accumulative effect of quantization errors, the encoder replicates the operation of the decoder and estimates α based on $\tilde{s}(n-1)$ or in the N-th order predictor case of equation above, estimate $\underline{\alpha} = [\alpha_1\ \alpha_2\ ....\alpha_N]^t$ based on $\tilde{s}(n-i), 1 \le i \le N$ as shown in Figure 5. Such a configuration ensures

that the error between original sample and reconstructed sample is simply the quantization error associated with $d(n)$. For example for the first order predictor,

At the encoder

$$s(n) = \sum_{i=1}^{N} \alpha_i \tilde{s}(n-i) + d(n)$$

At the decoder

$$\tilde{s}(n) = \sum_{i=1}^{N} \alpha_i \tilde{s}(n-i) + d_q(n)$$

Therefore which $s(n) - \tilde{s}(n) = d(n) - d_q(n)$ is the quantization error associated with $d(n)$ alone and has no contributions from $d(n-1), d(n-2)$ etc. It is noted that the prediction formulation in equations above essentially is an all-pole formulation since the current sample is predicted based on previous output samples and not on previous inputs. However, many ADPCM speech coders employ a pole-zero prediction of the form

$$\tilde{s}(n) = \sum_{i=1}^{N_1} \alpha_i \tilde{s}(n-i) + \sum_{j=0}^{N_2} \beta_j d_q(n-j) \qquad (2\text{-}4)$$

Such a pole-zero formulation permits higher prediction gain (in other words lower variance or dynamic range for prediction residual) for nasal sounds. Such a pole-zero adaptive predictor is employed in ADPCM based ITU-T speech coding standards G.726 and G.727 operating at bit rates of 40, 32, 24 and 16 kbps. These will be discussed in Section 5.

2.3 Adaptive Delta Modulation

The ADM technique is simply a special case of ADPCM, in the sense that α of first order prediction formulation is usually constrained to equal to 1 and $d_q(n)$ is constrained to be equal to $\pm\Delta$, where Δ is known at encoder and decoder. This implies that only one bit per sample needs to be transmitted to the remote decoder depending on the sign of Δ Although such a scheme is sub-optimal as compared to

ADPCM, the significant reduction in bit rate (equal to sampling rate) has rendered itself useful in some military applications with sacrifice in voice quality. A major source of impairment comes from the inability of the model to track very large (larger than Delta) and very small (less than Delta) variations thereby leading to noticeable distortions in reconstructed speech samples. To improve voice quality at the output of the decoder, the scheme is made adaptive whereby the value of Δ is made adaptive to track the variations in the input speech signal. One such scheme is based on adjusting delta in the following recursive manner

$$\Delta(n) = \Delta(n-1)\, K^{d_q(n)d_q(n-1)}$$

This is illustrated by the step-size adaptation box in Figure 6. It is noted here that such a scheme does not require any additional information at the decoder as compared to the non-adaptive approach, since the adaptation on Delta is based upon parameters known to the remote decoder

One drawback of Adaptive Delta Modulation is that transmission errors can cause degradation of speech quality that can last for a long time, especially when α is constrained to 1. To recover from transmission errors, it is necessary to introduce a leakage factor in both prediction and Δ adaptation. One such method is the Continuously Varying Slope Delta method, popularly called CVSD.

3. Parametric Speech Coders

Parametric coders typically analyze speech signal (in blocks or frames) and extract parameters that are deemed necessary to synthesize speech with a given quality objective under the constraint of a given bit rate. Among all the parametric coders that have been investigated and reported in the literature, the most widely used technique for speech analysis is the linear predictive coding technique. Here the prediction model is similar to that used in ADPCM described in Section 2.2, but the order of the model and its objective are different from that used in ADPCM coders. In ADPCM coders, the prediction filter is used to reduce the variance of the difference between the actual and predicted signal, thereby reducing the number of bits necessary to represent and transmit the prediction residual. Parametric coders that will be described below use the predictive model to estimate the poles of the vocal tract transfer function and hence obtain the spectral envelope of the speech signal. A L-th order linear predictive model is of the form

$$\hat{s}(n) = \sum_{i=1}^{L} \alpha_i s(n-i) \tag{3-1}$$

where $\hat{s}(n)$ is the predicted value of current speech sample s(n) based on previous speech samples. The prediction error, e(n), is defined as

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{L} \alpha_i s(n-i)$$

or in terms of the transfer function

$$E(z) = S(z)A(z)$$

or

$$S(z) = \frac{E(z)}{A(z)} \tag{3-2}$$

where

$$A(z) = 1 - \sum_{k=1}^{L} \alpha_k z^{-k} \tag{3-3}$$

The coefficients α_i in equation (3-3) is popularly known as Linear Predictive Coding (LPC) coefficients. The basic problem of linear prediction is determination, representation and quantization of LPC coefficients α_i. Equation (3-2) can be interpreted as the vocal tract being modeled as an all-pole system $1/A(z)$ whose input is e(n) and output is speech sample s(n). e(n), then would represent the excitation source to the vocal tract system as shown in Figure 3.

One of the earliest objectives of linear prediction coding was to model the speech spectral envelope with such a prediction (all-pole) filter whose input e(n) would be a quasi-periodic sequence (whose period is equal to the pitch period) for voiced speech and random noise for unvoiced speech. This is illustrated in Figure 7. It is extremely important to note that the acoustic theory of speech production points to the fact

that in order to obtain intelligible speech, it is sufficient to identify five dominant resonant frequencies (poles) of the vocal tract function. Hence a large number of vocoders that use linear predictive coding to analyze speech signal use a 10-th order linear prediction (L = 10 in equation (3-1)) that will result in an all-pole filter with five complex poles corresponding to the five dominant resonant frequencies of the vocal tract. Furthermore, it is interesting to note that the model described in Figure 7 is not different from the model which Homer Dudley's "vocoder" apparatus was based upon, except for the fact that digital computers were not available then to determine the coefficients of the linear predictive model; instead Homer Dudley used an electrical circuit to extract the spectral envelope of speech signal.

The LPC coefficients are determined by minimizing the mean squared prediction error over a short segment of speech waveform. Because of the time-varying nature of the speech signal, the LPC coefficients have to be determined based on short segments called "frames" of speech signal. In most speech coders the LPC coefficients are computed approximately every 10 - 20 ms. The frequency with which LPC coefficients are determined (or updated) is based upon the observation of spectrogram of speech which indicates that the spectral envelope changes slowly with time, and that for a duration of about 10 - 20 ms, it can be assumed that the spectral envelope remains reasonably constant. Furthermore, the optimal choice of LPC coefficients is computationally expensive and hence computed only as often as necessary. In order to retain the smoothness with which the spectrum changes are registered on the spectrogram, most speech coders perform linear interpolation of LPC parameters in between LPC coefficient updates

The mean squared prediction error E_n is given by

$$E_n = E\{(s(n) - \hat{s}(n))^2\}$$

where E{} denotes the expected value. However, because of the time varying nature of speech, the mathematical expectation is replaced by a short term average $\hat{E}_n$ defined as

$$\hat{E}_n = \frac{1}{m_{high} - m_{low}} \sum_{m=m_{low}}^{m_{high}} \left(s_n(m) - \sum_{i=1}^{I} \alpha_i s_n(m-i) \right)^2 \tag{3-4}$$

where $s_n(m) = s(n+m),\ m_{low} \le m \le m_{high}$ is a segment of speech surrounding the speech sample s(n) of interest, and, speech samples representing between $s_{m_{low}}$ to $s_{m_{high}}$ denotes a frame of speech.

Minimizing $\hat{E}_n$ with respect to α_i by setting $\partial \hat{E}_n / \partial \alpha_i = 0,\ i = 1,..,L$ leads to a set of L simultaneous equations given by

$$\sum_{m=m_{low}}^{m_{high}} s_n(m-i)s_n(m) = \sum_{k=1}^{L} \alpha_k \sum_{m=m_{low}}^{m_{high}} s_n(m-i)s_n(m-k) \quad 1 \le i \le L$$

or more compactly represented as

$$\varphi_n(i,0) = \sum_{k=1}^{L} \alpha_k \varphi_n(i,k) \quad 1 \le i \le L \tag{3-5}$$

where

$$\varphi_n(a,b) = \sum_{m=m_{low}}^{m_{high}} s_n(m-a)s_n(m-b) \tag{3-6}$$

Solution to the set of simultaneous equations will yield the set of optimal LPC coefficients. There are two fundamental approaches, namely autocorrelation method and covariance method, that have been used to arrive at a solution. The basic difference in the two approaches is the choice of the limits m_{low} and m_{high} that are used in solving the set of simultaneous equations. These will be explained below.

3.1 Autocorrelation Method of LPC Analysis

In this method, it is assumed that the waveform segment $s_n(m)$ is zero outside the interval $0 \le m \le N-1$. Such an assumption is equivalent to a windowing operation on the original speech samples, where the window w(n) is represented as

$$w(n) = 1 \quad 0 \le n \le N-1$$
$$= 0 \quad elsewhere$$

Such a window is called a rectangular window. Such a window has the effect of spreading the spectrum of the speech segment since it has high sidelobes. Some of the more common windows that have been used for speech coding include

Hamming Window :

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & otherwise \end{cases}$$

Hanning Window :

$$w(n) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & otherwise \end{cases}$$

Bartlett Window :

$$w(n) = \begin{cases} \frac{2n}{N-1} & 0 \le n \le (N-1)/2 \\ 2 - \frac{2n}{N-1} & (N-1)/2 \le n \le N-1 \\ 0 & otherwise \end{cases}$$

Blackman Window :

$$w(n) = \begin{cases} 0.42 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) + 0.08\cos\left(\frac{4\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & otherwise \end{cases}$$

The choice of the window shape is critical in terms of handling sidelobe effects in the windowed speech in the spectral domain The Blackman window has the least frequency leakage in terms of sidelobe contribution, but at the same time has the lowest frequency resolution. The performance of Hamming and Hanning window are in-between the two extremes of rectangular window and Blackman window and hence the most popular in speech coding.

Since $s_n(m)$ is zero for $0 \le m \le N-1$ when any of the above windows is used, it is easy to verify that the prediction residual $e(m)$ is nonzero over the interval $0 \le m \le N-1+L$, thereby implying that lower and upper limits for $\hat{E}_n$ in equation (3-4) have to be $m_{low} = 0$ and $m_{high} = N-1+L$ It is also noted that the prediction error $e(m)$ for $0 \le m \le L-1$ is likely to be large since equation implies that non-zero samples are being predicted from zero samples in this region Similarly, $e(m)$ is likely to be large in the region $N \le m \le N+L-1$ since zero samples are being predicted from non-zero samples Substituting the values of m_{low} and m_{high} in equation (3-6) we get

$$\varphi_n(a,b) = \sum_{m=0}^{N-1+L} s_n(m-a)s_n(m-b)$$

and since $s_n(m) = 0$ for $m < 0$ and $m > N-1$, we can rewrite $\varphi_n(a,b)$ as

$$\varphi_n(a,b) = \sum_{m=0}^{N-1-(a-b)} s_n(m)s_n(m+a-b)$$

From the above equation, $\varphi_n(a,b)$ can be viewed as a short-term autocorrelation function of $s_n(m)$ evaluated at a lag of $(a-b)$, i.e,

$$\varphi_n(a,b) = R_n(a-b)$$

where

$$R_n(\tau) = \sum_{m=0}^{N-1-\tau} s_n(m)s_n(m+\tau)$$

It can also be shown that under the assumption of $s_n(m) = 0$ for $m < 0$ and $m > N-1$, that $R_n(\tau)$ is an even function and hence $R_n(\tau) = R_n(-\tau)$. This property is exploited in the autocorrelation method of LPC analysis to reduce the computational complexity of the algorithm, and this is evidenced when the set of simultaneous equations in equation (3-5) are represented in a matrix form

$$\begin{bmatrix} \varphi_n(1,0) \\ \varphi_n(2,0) \\ \\ \cdots\cdots \\ \varphi_n(L,0) \end{bmatrix} = \begin{bmatrix} \varphi_n(1,1)\ \varphi_n(1,2)\ \varphi_n(1,3) \ldots & \varphi_n(1,L) \\ \varphi_n(2,1)\ \varphi_n(2,2)\ \varphi_n(2,3) \ldots & \varphi_n(2,L) \\ \\ \cdots\cdots & \\ \varphi_n(L,1)\ \varphi_n(L,2)\ \varphi_n(L,3) \ldots & \varphi_n(L,L) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \\ \cdots \\ \alpha_L \end{bmatrix}$$

Substituting for φ_n in terms of R_n and using the symmetric property of R_n we get

$$\begin{bmatrix} R_n(1) \\ R_n(2) \\ \\ \cdots\cdots \\ R_n(L) \end{bmatrix} = \begin{bmatrix} R_n(0) & R_n(1) & R_n(2) \ldots & R_n(L-1) \\ R_n(1) & R_n(0) & R_n(1) \ldots & R_n(L-2) \\ \\ \cdots\cdots & & & \\ R_n(L-1) & R_n(L-2) & R_n(L-3) \ldots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \\ \cdots \\ \alpha_L \end{bmatrix}$$

From this it is easy to see that the LxL matrix has a Toeplitz structure and that, the number of correlation computations necessary to solve the above equation is simply $L+1$. Several efficient recursive procedures have been devised to solve equation above, the most efficient being the Durbin's recursive procedure which is outlined below. The recursive procedure entails computing the following quantities recursively L times and superscripts indicate the recursion number.

$$E_n^{(0)} = R_n(0)$$

$$k_i = \frac{\left(R_n(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_n(i-j) \right)}{E^{(i-1)}} \qquad 1 \le i \le L$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i-1$$

$$E_n^{(i)} = \left(1 - k_i^2\right) E_n^{(i-1)} \tag{3-7}$$

The above equations are solved for $1 \le i \le L$ after which the solution to matrix equation above is given by

$$\alpha_i = \alpha_i^{(L)}$$

It is noted that $\alpha_j^{(i)}, 1 \le j \le i$ above actually are the LPC coefficients of an i-th order LPC model. Such a property has been exploited by some speech coders to extract lower order LPC coefficients from a higher order model as well as to make voiced/unvoiced decisions on speech segments.

3.2 Covariance Method of LPC Analysis

In this approach, rather than windowing the speech segment, the duration over which the prediction error is computed is windowed. The limits on E_n for the covariance method is $0 \le m \le N-1$ and hence $m_{low} = 0$ and $m_{high} = N-1$ which leads to

$$\hat{E}_n = \frac{1}{N} \sum_{m=0}^{N-1} \left(s_n(m) - \sum_{i=1}^{L} \alpha_i s_n(m-i) \right)^2$$

$$\varphi_n(a,b) = \sum_{m=0}^{N-1} s_n(m-a) s_n(m-b)$$

It can be shown in this case that $\varphi_n(a,b) = \varphi_n(b,a)$, however $\varphi_n(a,b) \ne f(a-b)$ and hence does not truly represent the autocorrelation function. Instead, $\varphi_n(a,b)$ represents cross correlation between similar, but not identical sequences. Hence, in matrix form, the covariance method leads to

$$\begin{bmatrix} \varphi_n(1,0) \\ \varphi_n(2,0) \\ \\ \cdots\cdots \\ \varphi_n(L,0) \end{bmatrix} = \begin{bmatrix} \varphi_n(1,1) & \varphi_n(1,2) & \varphi_n(1,3) \ldots & \varphi_n(1,L) \\ \varphi_n(1,2) & \varphi_n(2,2) & \varphi_n(2,3) \ldots & \varphi_n(2,L) \\ & & & \\ \cdots\cdots & & & \\ \varphi_n(1,L) & \varphi_n(2,L) & \varphi_n(3,L) \ldots & \varphi_n(L,L) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \\ \cdots \\ \alpha_L \end{bmatrix}$$

written compactly as

$$\Omega = \Phi\alpha$$

where the LxL matrix Φ is symmetric, but not Toeplitz. The above matrix equation is solved using Cholesky decomposition where Φ can be decomposed into a product of upper triangular (U), diagonal (D) and lower triangular (U') which results in the following

$$\Omega = UDU'\alpha$$

Letting

$$DU'\alpha = \alpha',$$

it is easy to see that since U is an upper-triangular matrix, the solution for each element of α' vector can be recursively obtained. After obtaining α' vector, α vector is obtained in a similar manner by noting from above equation that $U'\alpha = D^{-1}\alpha'$, and exploiting the fact that U' is a lower-triangular matrix.

It was seen from autocorrelation method and covariance method of LPC analysis, that determination of LPC coefficients vector α involved first precomputing R_n or φ_n and then solving the matrix equation. However, there exists a third technique called Lattice method of LPC analysis that eliminates the precomputation of correlation values and directly obtains LPC coefficients from speech samples. This is described in [9].

3.3 Quantization and alternate representations of LPC parameters

In the above discussion, techniques to estimate the LPC coefficients were described. Since the ultimate goal is to reduce the bit rate, the LPC coefficients have to be quantized using as few bits as possible. The LPC coefficients may be scalar quantized or vector quantized. In scalar quantization, each LPC coefficient is quantized independent of each other in a manner similar to that performed in PCM, except, here different coefficients may be represented using different number of bits depending on their importance. In vector quantization (VQ), the entire vector of coefficients (or sub-vectors) are quantized jointly and the

quantization is based on finding an element (or vector) of a codebook that is close (with respect to a defined distortion measure) to the vector of LPC coefficients to be quantized. The index of the codebook is transmitted to the voice decoder. In such systems, both encoder and decoder are expected to have a copy of the same codebook. The codebook itself is generated based on a large training set. VQ of LPC parameters has received considerable attention (and perhaps the only alternative) for low bit rate speech coders operation at 4 kbps or below While original VQ techniques involved having a single large codebook and an exhaustive search procedure, modern day VQ typically have multiple smaller codebooks which are searched in multiple stages, thereby reducing search complexity and storage. A classical example can be found in [10].

Let $1/\hat{A}(z)$ represent the synthesis filter with quantized LPC coefficients, i.e ,

$$\frac{1}{\hat{A}(z)} = \frac{1}{1 - \sum_{i=1}^{L} \hat{\alpha}_i z^{-i}}$$

where $\hat{\alpha}_i$ represents the j-th LPC coefficient after quantization (scalar or vector). It is extremely important to note that in order for the synthesis filter $1/\hat{A}(z)$ to be stable, the roots of $\hat{A}(z)$ have to lie inside the unit circle. Even if the LPC coefficients were such that $1/A(z)$ was stable, the quantized set of LPC coefficients (which will be the one that will be used by the voice decoder) may lead to an unstable synthesis filter. Furthermore, checking the stability (or finding the roots of a 10-th order polynomial) is no trivial task. Hence alternative methods of representing LPC coefficients that guarantees stability of the LPC synthesis filter after quantization have been uncovered, the popular ones being reflection coefficients, log area ratios, arcsines of partial correlation coefficients, and, line spectral frequencies.

Observing equation (3-7), it is easy to see that the values of $|k_i| < 1.0$, and it is also observed that the LPC coefficients were derived from k_i. Hence k_i (called the reflection coefficients or Partial Correlation

coefficients, PARCOR) form a suitable alternative for LPC coefficients, since k_i can be checked for stability by simply verifying whether $|k_i| < 1.0$. It can however be shown that the LPC spectral distortion introduced by quantizing k_i depends on $|k_i|$ and that, values of $|k_i|$ closer to unity are more sensitive than those reflection coefficients that have smaller magnitudes. In order to normalize this, a nonlinear transformation is performed on k_i such as

$$l_i = \log\left(\frac{1 - k_i}{1 + k_i}\right)$$

where l_i are called Log-Area Ratio (LAR) coefficients. Another nonlinear transformation that is quite often used is the arcsine transformation $s_i = \sin^{-1}(k_i)$ and s_i are called arcsine coefficients. Both LAR and arcsine functions tends to emphasize larger values of $|k_i|$ and de-emphasize smaller values of $|k_i|$ thereby performing a transformation similar to the expansion function in PCM systems. Such a transformation allows more accurate quantization of larger values of $|k_i|$ compared to smaller values of $|k_i|$.

Perhaps, the most popular alternative to LPC coefficients is the Line Spectral Frequency (LSF) representation. Here, odd and even polynomials are formed using A(z) as follows :

$$P(z) = A(z) + z^{-(L+1)}A(z^{-1})$$

$$Q(z) = A(z) - z^{-(L+1)}A(z^{-1})$$

It can be shown that the roots of P(z) and Q(z) lie on the unit circle, and that the roots of P(z) and Q(z) alternate on the unit circle in the complex z-plane. Furthermore $z = -1$ and $z = +1$ are roots of P(z) and Q(z) respectively. Hence P(z) and Q(z) can be written as

$$P(z) = (1 + z^{-1})P_1(z)$$

$$Q(z) = (1 - z^{-1})Q_1(z)$$

where

$$P_1(z) = \sum_{i=0}^{L} p_i z^{-i}$$

$$Q_1(z) = \sum_{i=0}^{L} q_i z^{-i}$$

where, it can be shown that

$p_0 = 1,\ q_0 = 1$

$p_j = \alpha_j + \alpha_{L-j+1} - p_{j-1}; \qquad 1 \le j \le L$

$q_j = \alpha_j - \alpha_{L-j+1} + q_{j-1}; \qquad 1 \le j \le L$

The roots of $P_1(z)$ and $Q_1(z)$ form the line spectral frequencies for the set of LPC coefficients and they are on the unit circle. Such an ordering is essential to ensure stability of the synthesis filter. It can be shown that the coefficients of $P_1(z)$ and $Q_1(z)$ are such that $p_j = p_{L-j}$ and $q_j = q_{L-j}$. Hence $P_1(z)$ and $Q_1(z)$ take the form

$$\begin{aligned} P_1(z) &= 1 + p_1 z^{-1} + ..+ p_{L/2-1} z^{L/2-1} + p_{L/2} z^{L/2} + p_{L/2-1} z^{L/2+1} + ..+ p_1 z^{L-1} + z^L \\ &= z^{-L/2}\left[(z^{L/2} + z^{-L/2}) + p_1(z^{L/2-1} + z^{-(L/2-1)}) + ... + p_j(z^{L/2-j} + z^{-(L/2-j)}) + ... + p_{L/2}\right] \end{aligned}$$

$$\begin{aligned} Q_1(z) &= 1 + q_1 z^{-1} + ..+ q_{L/2-1} z^{L/2-1} + q_{L/2} z^{L/2} + q_{L/2-1} z^{L/2+1} + ..+ q_1 z^{L-1} + z^L \\ &= z^{-L/2}\left[(z^{L/2} + z^{-L/2}) + q_1(z^{L/2-1} + z^{-(L/2-1)}) + ... + q_j(z^{L/2-j} + z^{-(L/2-j)}) + ... + q_{L/2}\right] \end{aligned}$$

Since the roots of $P_1(z)$ and $Q_1(z)$ lie on the unit circle, $P_1(z)$ and $Q_1(z)$ need to be evaluated only at $z = e^{jw}$. On observing that $e^{jx} + e^{-jx} = 2\cos(x)$, it is clear that it is necessary to find the roots of the following two equations

$$\cos(wL/2) + p_1 \cos(w(\tfrac{L}{2} - 1)) + ... + p_j \cos(w(\tfrac{L}{2} - j)) + .. + p_{L/2} = 0$$

$$\cos(wL/2) + q_1 \cos(w(\tfrac{L}{2} - 1)) + ... + q_j \cos(w(\tfrac{L}{2} - j)) + .. + q_{L/2} = 0$$

The above two equations are solved for L/2 values of w in the range of 0 to pi radians. Let $w_{p1}, w_{p2}, .., w_{pL/2}$ be the roots of $P_1(z)$ in the range $0 \le w_{pi} \le \pi$ and $w_{q1}, w_{q2}, .., w_{qL/2}$ be the roots of $Q_1(z)$ in the range $0 \le w_{qi} \le \pi$. The w_{pi} and w_{qi}, which are expressed in radians are converted to a vector of line spectral frequencies as $\left[2\pi w_{p1}, 2\pi w_{q1}, 2\pi w_{p2}, 2\pi w_{q2}, .., 2\pi w_{pL/2}, 2\pi w_{qL/2}\right]$.

It is noted here from equations above that it is only necessary to find roots of two L/2-th order polynomials rather than one L-th order polynomial. Several methods of finding roots of the above equations have been reported in literature, such as finding zero crossing point, phase reversal tracking and use of Chebychev polynomials [11]. In all methods, the complexity of finding roots of the above two equations is significantly less than finding the roots of a L-th order polynomial. Furthermore, after quantization of the line spectral frequencies, it is only necessary to check for the ordered property to verify whether the synthesis filter is stable or not.

In addition to the ordered property of LSFs, another attractive feature of LSFs is its localized spectral sensitivity property, i.e., a small quantization error in a particular LSF element will result in deviation from unquantized LPC spectrum only in the frequencies surrounding the quantized LSF. Such a property has been very effectively exploited in split-vector quantization of line spectral frequencies [12] to achieve transparent quantization. Here the L element LSF vector is split into sub-vectors and each sub-vector is independently vector quantized. Such a scheme permits different size codebooks to be used for different sub-vectors depending on importance placed to the sub-band of frequency encompassed by the sub-vector. It is to be noted that the LSF vector that is formed after quantization has to obey the ordered property in order to retain stability of the LPC synthesis filter. Such a constraint forces a partial search on the individual codebooks thereby resulting in increased quantization error. Attempts to circumvent this problem include training the codebook vectors in a constrained fashion or populating the codebooks after training in a manner such that complete search is possible[13].

4. Excitation Modeling

The LPC analysis described serves to model the vocal tract parameters (namely, A(z) of equation (3-2)) of the human speech production mechanism. In order to reproduce speech at the remote decoder, in addition to vocal tract parameters, it is necessary to model, digitize and transmit the excitation parameters (namely, E(z) of equation (3-2)) of the human speech production mechanism using as few bits as possible. From a speech encoder point of view, E(z) can be treated as the output of a system A(z) whose input is speech S(z), in other words E(z) is the LPC residual. From a decoder point of view, E(z) can be interpreted as the input to LPC synthesis filter 1/A(z) whose output is speech signal S(z). Hence the usage of the terms "LPC residual" and "excitation sequence" are used interchangeably.

Perhaps, the simplest excitation model that can be used to synthesize intelligible speech is the two state excitation model as was demonstrated by Homer Dudley in his vocoder apparatus in 1939. The model is illustrated in Figure 7. Here a periodic excitation (typically a set of equally spaced impulses) whose period is equal to the inverse of the fundamental frequency of the segment of speech under consideration is used for voiced speech, and, for unvoiced speech, a noise-like excitation is used which is independent of talker or speech material under consideration.

The fundamental frequency or pitch is typically estimated by determining the peak of the autocorrelation of the LPC residual signal for a given range of autocorrelation lags. Typically the range of autocorrelation lags over which the pitch period is determined is between 20 and 120 corresponding to pitch frequencies between 400 Hz and 66 Hz for 8 kHz sampled speech signals. In order to reduce the complexity associated with the computation of autocorrelation function for about a 100 lag values (between 20 and 120), the LPC residual signal is low-pass filtered to less than 100 Hz and down sampled or decimated by 4. The Simplified Inverse Filtering Technique (SIFT) of pitch estimation is based on this approach [14]. Other approaches have been reported in literature such as picking the peak of the cepstrum of speech signal

within a given range; computing the average of the magnitudes of the differences (known as AMD Function or AMDF) between speech samples with different offsets (belonging to a given range) and picking the offset with least AMD as the pitch value [14]. A more comprehensive view of pitch estimation techniques are provided in [15]. The decision as to whether a speech frame is voiced or unvoiced is usually classified using features such as energy measurements, zero crossing rate, peak value of the autocorrelation function, and, magnitude of the first reflection coefficient and performing and arriving at a weighted distortion measure and decision thresholds for obtaining a low misclassification error [21,22].

The two state excitation model illustrated in Figure 7 can produce intelligible quality speech at very low bit rates of 2400 bps or less, but has serious limitations in terms of producing high quality speech that will preserve the naturalness of speech and characteristics of the talker. The reasons are multi-fold : (1) The LPC residual for voiced sounds are not impulses and hence when the LPC synthesis filter is driven by a series of impulses, the resulting speech gets noticeably degraded. (2) The LPC residual for unvoiced sounds is not truly bandlimited white noise; it has a spectral tilt to it depending on the inadequacies of the LPC modeling. Hence when the LPC synthesis filter is driven by bandlimited white noise , the resulting speech gets noticeably degraded. (3) The LPC technique is often unable to model poles that are close to each other (4) The all-pole model is not accurate for nasal sounds which are characterized by zeroes as well (5) The LPC coefficients need to be quantized before transmitting them on a digital channel and hence suffers quantization errors, resulting in shifting of the resonant frequencies. (6) The model heavily relies upon voice/unvoiced classification of segments of speech and hence a function of the accuracy of such a classification - for some plosives and voiced fricatives, an accurate classification is difficult since they do not bear the characteristics of completely voiced or unvoiced sounds.

It is clear that in order to produce naturally sounding speech, the LPC residual signal has to be adequately represented and transmitted to the voice decoder; which would then be used as the excitation sequence to the LPC synthesis filter at the remote voice decoder. (It is noted that the if the prediction residual obtained

using quantized LPC parameters at the encoder was used at the remote decoder, then exact reconstruction of original speech is possible). Bulk of the research in speech coding over the last decade was devoted to arriving at good excitation models rather than improve the deficiencies of the LPC analysis. Part of the motivation behind such an approach was that, it would lead to a single unified technique that would provide a solution to all of the inadequacies mentioned above.

The inadequacies of the two state excitation model (in conjunction with LPC analysis) were initially addressed by representing the LPC residual (and hence the excitation sequence to the voice decoder) as a combination of periodic and a-periodic signals. Such schemes not only led to elimination of excitation sequences that were purely based on pitch estimation and voiced/unvoiced decisions, but also consequently led to reduction of some of the buzzy artifacts in voiced segments of reconstructed speech signal due to excessive periodicity [16,17,18,19,20]. Among these two schemes, namely Residual Excited Linear Predictive (RELP) coder and Mixed Excited Linear Predictive (MELP) coders gained prominence in this effort which are described briefly below.

4.1 Open Loop Excitation Modeling

The RELP and MELP coders mentioned above belong to a class of excitation modeling technique called open-loop modeling since the objective here is to best represent the LPC residual signal No feedback as to whether the model parameters will yield speech that will yield good quality speech in a perceptual sense is provided.

4.1.1 Residual Excited Linear Predictive (RELP) Vocoder

Here the LPC residual signal is low pass filtered and the decimated residual signal is then encoded using Adaptive Delta Modulation (ADM) techniques described in Section 2.3 and transmitted to the remote voice decoder [16]. A block diagram of the RELP encoder and decoder is shown in Figure 8. The low pass filter has a cut-off frequency that is at least high enough to accommodate the second harmonic of the highest possible fundamental frequency. Hence for low bit rate applications, RELP coder typically uses cutoff frequencies in the vicinity of 800 to 1000 Hz, based on the assumption that the highest possible human fundamental frequency is about 400 Hz. At the receiver, the low-pass filtered residual signal is recovered

using an adaptive delta demodulator and then processed using a nonlinear device to generate higher harmonics of the residual signal. The spectral flattener shown in Figure 8 actually contains a nonlinear device which generates higher harmonics of the residual signal with decreasing strengths; hence a double differencer is used to enhance the higher harmonics (effectively provide a flat spectrum at higher harmonics); this is followed by a high pass filter so as to only use the spectrally flattened excitation spectrum for higher harmonics. This is illustrated in Figure FIG_RELP2. The lower harmonics will be the same as that decoded using adaptive delta demodulator.

Finally a controlled amount of white noise is mixed with the spectrally flattened excitation signal before using it to drive the LPC synthesis filter so as to reduce any buzziness in voiced segments of reconstructed speech.

It is extremely important to note that there is no explicit pitch extraction or transmission associated with RELP coding, thereby making the scheme pitch independent and immune to pitch determination and tracking errors.

4.1.2 Mixed Excitation Linear Predictive (MELP) Coder

One of the shortcomings of RELP coder is the bit rate necessary to perform an ADM coding of the low pass filtered decimated residual signal. Typical bit rates in the range of 6 to 7 kbps have been reported in literature as being necessary to encode the residual signal in order to produce natural sounding speech. For speech coders that are required to operate at bit rates of 4.8 kbps and below, RELP coders render themselves unsuitable. The Mixed Excitation Linear Predictive coders [17,18] provides a low bit rate alternative to RELP, and at the same time removes the inadequacies of the two-state excitation modeling, both in terms of removing buzziness in reconstructed speech as well as providing robustness to voicing decision errors. Unlike RELP coders, where pitch estimation was absent, MELP coders perform pitch

estimation and incorporates frequency dependent voicing measures to model the LPC residual signal. Here the residual signal is modeled as a sum of aperiodic and periodic components. The periodic component is generated by an impulse train passed through a low pass filter, and the aperiodic component is generated using white noise and a high pass filter. The gain of the two components are chosen such that the overall excitation spectrum is flat. A more sophisticated mixed excitation scheme (see Figure 9) includes a method of introducing controlled amount of jitter in the impulse train to reflect the amount of periodicity in the LPC residual signal. Furthermore, an adaptive spectral enhancer is used to emphasize the energy in formant regions of speech. A MELP coder operating at 2.4 kbps was recently standardized by the U.S. Department of Defense for use in the military.

4.2 Closed Loop Excitation Modeling - Analysis-by-Synthesis Coding

One of the disadvantages of the open-loop excitation modeling schemes described in Section 4.1 is that, the excitation parameters are extracted such that they best represent the LPC residual signal, and not necessarily something that will result in reconstructed speech that is close to original speech signal. Furthermore, the perception based residual modeling in RELP and MELP are optimal in LPC residual domain. A desirable alternative is to replicate the voice decoder operation in the encoder and extract encoder parameters that will minimize a perceptually weighted distortion measure between original speech and reconstructed speech. Such a technique, whereby the analysis parameters are optimized based on synthesizing speech in the encoder is referred to as analysis-by-synthesis coding. This is a powerful technique since it not only optimizes parameters in speech domain, but also permits perceptual weighting to be performed in speech domain, which is highly desirable.

The perceptual weighting of the difference between the original speech signal and reconstructed speech signal is a key feature of most analysis-by-synthesis coders. Here the difference signal is typically passed through a time-varying pole-zero filter that will shape the spectral distortions (due to imprecise modeling and quantization) to be strong in formant regions and weak in valley regions, thereby striving to achieve a

constant signal-to-noise ratio across the frequency spectrum of interest Such perceptual weighting essentially exploits noise masking properties of the human auditory system, thereby making quantization noise in spectral valleys inaudible.

In general the perceptual weighting filter that is used in most speech coders are of the form

$$W(z) = \frac{A\left(z/\gamma_1\right)}{A\left(z/\gamma_2\right)} \tag{4-1}$$

where $A(z)$ is the unquantized LPC spectrum. While the actual effect of $\gamma_1, \gamma_2 \in [0,1]$ is to broaden the poles and zeroes (peaks and valleys of the spectrum) in a controlled manner, the real purpose of using such a weighted filter is to perform optimization that will match the synthesized speech more closely at spectral valleys compared to spectral peaks, thereby achieving the desired constancy in signal to noise ratio.

It has to be noted that perhaps the most optimal choice of a weighted error criterion would have been to perform-by-synthesis that optimizes the difference between original speech signal and synthesized speech signal after passing through the human speech perception mechanism. This is illustrated in Figure 10. In the figure $f(X)$ is in general, a nonlinear function of the vector X; $f(X)$ represents the transfer function of the human speech perception mechanism, including the cochlea of the human ear [33]. $d(X_1, X_2)$ is a decision function which, ideally should represent the decision making process in the human brain. However, in order to reduce the complexity such a search procedure and for the sake of analytical tractability, a simplified weighted distortion measure such as the weighted linear minimum mean square error is used as optimization criterion. Here $f(X)$ is replaced by WX where W is a NxN matrix whose entries represent the impulse response the filter $W(z)$ of equation (4-1) which coarsely represents the human speech perception mechanism. The distortion measure $d(.,.)$ is typically chosen to be a L_2 norm, which again is a highly simplified version of the complex decision making process of the human brain

In practice, analysis-by-synthesis coders are implemented as a sequential optimization process (rather than a joint optimization process) whereby closed loop optimization is only performed to extract parameters of the excitation model. This is illustrated in Figure 11 where the LPC analysis is performed in an open loop fashion and then the excitation parameters are extracted using closed loop analysis-by-synthesis approach.

The earliest pioneering and practical work in analysis-by-synthesis coding that used perceptual weighting criterion was reported by Atal and Remde [34] wherein a multipulse excitation scheme was proposed to represent the LPC residual signal. Here the LPC residual signal is represented by a sparse sequence of pulses separated by zeroes. This is described below.

4.2.1 Multi-Pulse LPC Method

The fundamental principle behind this technique is that there are only a fraction of prediction residual samples that are perceptually important that yield a high degree of naturalness in reconstructed speech and hence it is not necessary to transmit all prediction residual samples. It is indeed true that if all prediction residual samples were transmitted remote voice decoder with a very high signal-to-quantization noise ratio, then the reconstructed speech would be very close to the original speech. The objective here however is to reduce the bit rate and still achieve natural-sounding speech.

Here a sub-frame duration of N prediction residual samples $\{r_0, r_1, .., r_{N-1}\}$ that is obtained using quantized LPC coefficients is represented by a set of P (P << N) pulses with positions $p_1, p_2, .., p_P$, $p_i \in [0, N-1]$ and non-zero amplitudes $g_1, g_2, .., g_p$. It is noted that there are $\binom{N}{P}$ possible combinations of P pulse positions from among N positions The pulse positions p_i and their amplitudes g_i for a given combination are determined in a closed-loop analysis by synthesis method using the weighted mean-square error described above. Specifically, the cost function

$$J(s_w(n), \tilde{s}_k(n)) = \sum_{j=0}^{N-1} (s_w(n) - \tilde{s}_k(n))^2$$

is minimized with respect to pulse positions and amplitudes, where $s_w(n)$ is the perceptually weighted input speech signal and $\tilde{s}_k(n)$ is the perceptually weighted synthesized speech signal for k-th, $k \in \left[0, \binom{N}{p} - 1\right]$ combination of pulse positions and amplitudes. First, the perceptually weighted speech signal $s_w(n)$ is obtained by passing original speech signal s(n) through the weighting filter $A(z/\gamma_1)/A(z/\gamma_2)$ as follows

$$s_w(n) = s(n) + \sum_{i=1}^{L} \alpha_i \gamma_1^i s(n-i) - \sum_{i=1}^{L} \alpha_i \gamma_2^i s_w(n-i)$$

and subtracting the zero-input response of the weighting filter.

$\tilde{s}_k(n)$ is obtained by passing the k-th combination of pulse positions and amplitudes through the LPC synthesis filter $1/\hat{A}(z)$ and then performing perceptual weighting on the synthesized speech using the weighting filter $A(z/\gamma_1)/A(z/\gamma_2)$ in a manner similar to that of original speech. In practice, however, these two steps are combined and hence the chosen set of amplitudes are passed through a linear FIR filter whose impulse response represents combination of $\frac{A(z/\gamma_1)}{A(z/\gamma_2)} \frac{1}{\hat{A}(z)}$. Typically the truncated impulse response $h(n),\ n = 0,1,..,N-1$ of the combined synthesis and weighting filter is obtained by filtering a signal consisting of the coefficients of the filter $A(z/\gamma_1)$ extended by zeroes through two filters $1/\hat{A}(z)$ and $1/A(z/\gamma_2)$.

$$\tilde{s}_k(n) = \sum_{j=1}^{p} g_{jk} h(n - p_{jk})$$

where p_{jk} and g_{jk} are the pulse positions and amplitudes corresponding to the k-th combination. Minimizing $J(s_w(n), \tilde{s}_k(n))$ individually w.r.t g_{ik}, i=1,..,P

$$\frac{\partial}{\partial g_{ik}} \sum_{n=0}^{N-1} (s_w(n) - \tilde{s}_k(n))^2 = 0$$

yields a set of P simultaneous equations

$$\sum_{n=0}^{N-1} s_w(n) h(n - p_{ik}) = \sum_{j=1}^{P} g_{jk} \sum_{n=0}^{N-1} h(n - p_{jk}) h(n - p_{ik}) \qquad 1 \le i \le P$$

which in matrix form can be written as

$$\underline{Y}_k = A_k \underline{G}_k$$

where

$$\underline{Y}_k = \begin{bmatrix} \sum_{n=0}^{N-1} s_w(n) h(n - p_{1k}) \\ \sum_{n=0}^{N-1} s_w(n) h(n - p_{2k}) \\ \vdots \\ \sum_{n=0}^{N-1} s_w(n) h(n - p_{Pk}) \end{bmatrix}$$

and

$$A_k = \begin{bmatrix} \sum_{n=0}^{N-1} h(n-p_{1k})h(n-p_{1k}) & \sum_{n=0}^{N-1} h(n-p_{1k})h(n-p_{2k}) & \cdots & \sum_{n=0}^{N-1} h(n-p_{1k})h(n-p_{Pk}) \\ \sum_{n=0}^{N-1} h(n-p_{2k})h(n-p_{1k}) & \sum_{n=0}^{N-1} h(n-p_{2k})h(n-p_{2k}) & \cdots & \sum_{n=0}^{N-1} h(n-p_{2k})h(n-p_{Pk}) \\ & \vdots & & \\ \sum_{n=0}^{N-1} h(n-p_{Pk})h(n-p_{1k}) & \sum_{n=0}^{N-1} h(n-p_{1k})h(n-p_{2k}) & \cdots & \sum_{n=0}^{N-1} h(n-p_{Pk})h(n-p_{Pk}) \end{bmatrix}$$

and

$$\underline{G}_k = \begin{bmatrix} g_{1k} \\ g_{2k} \\ \vdots \\ g_{Pk} \end{bmatrix}$$

Hence the optimal set of amplitudes for the k-th combination is given by

$$\underline{G}_k^* = A_k^{-1} \underline{Y}_k$$

and the resulting mean square error is given by

$$E_k = \sum_{n=0}^{N-1} s_w^2(n) - \underline{Y}_k^t A_k^{-1} \underline{Y}_k \qquad (4\text{-}2)$$

Since A_k is a symmetric positive definite matrix, it is easy to see that the second term $Y_k^t A_k^{-1} Y_k$ is positive.

Hence minimizing $J(s_w(n), \tilde{s_k}(n))$ is equivalent to maximizing the second term $Y_k^t A_k^{-1} Y_k$ in the E_k expression in equation (4-2) Hence in practice, the second term is evaluated for all possible $\binom{N}{P}$ combinations of P pulses and the combination $k = k^*$ that yields the maximum value of $Y_k^t A_k^{-1} Y_k$ is chosen as the optimal combination which yields the least mean square error in a perceptually weighted sense.

The MPLPC technique has the distinct advantage in that it does not depend on the pitch estimation and/or voiced/unvoiced decision. Hence it provides a unified framework for representing excitation to decoder for all types of speech segments. The two drawbacks however are the computational complexity necessary to determine the optimal pulse positions and their amplitudes, and the bit rate necessary to transmit them. As an example, if it is determined that five non-zero samples (P = 5) have to be identified in a duration of 5 ms (N = 40 at 8 kHz sampling rate), then the possible number of combinations for which E_k has to be computed is $\binom{40}{5} = 658008$

Hence less complex sub-optimal schemes have been proposed in the literature.

The earliest proposal was to perform sequential search, i.e., determine one pulse location and amplitude at a time. Here the optimal first pulse position p_1^* (and its amplitude g_1^*) is determined from among N possible choices by computing the second term of equation () with P = 1, namely

$$W_k = \frac{\left(\sum_{n=0}^{N-1} s_w(n) h(n-k)\right)^2}{\sum_{n=0}^{N-1} h^2(n-k)}$$

for k = 0, 1, .., N-1 and determining the value of k ($= p_1^*$) for which W_k is maximum. The corresponding amplitude g1* is determined using

$$g_1^* = \frac{\sum_{n=0}^{N-1} s_w(n)h(n-p_1^*)}{\sum_{n=0}^{N-1} h^2(n-p_1^*)}$$

Subsequent pulse positions p_m^* and g_m^*, $2 \le m \le P$ are obtained one by one by minimizing the cost function

$$J(s_w(n), \tilde{s}_m(n)) = \sum_{j=0}^{N-1} (s_w(n) - \tilde{s}_m(n))^2$$

where

$$\tilde{s}_m(n) = \sum_{j=1}^{m-1} g_j^* h(n-p_j^*) + g_m h(n-p_m)$$

with respect to g_m and p_m or equivalently computing

$$W_{km} = \frac{\left(\sum_{n=0}^{N-1} s_{wm}(n)h(n-k)\right)^2}{\sum_{n=0}^{N-1} h^2(n-k)}; \quad 0 \le k \le N-1, k \notin (p_1^*, p_2^*, .., p_{m-1}^*)$$

where

$$s_{wm}(n) = s_w(n) - \sum_{j=1}^{m-1} g_j^* h(n-p_j^*)$$

and finding the value of k (= p_m^*) for which W_{km} is maximized. The corresponding amplitude g_m^* is obtained using

$$g_m^* = \frac{\sum_{n=0}^{N-1} s_{wm}(n)h(n-p_m^*)}{\sum_{n=0}^{N-1} h^2(n-p_m^*)}$$

It is noted that the total number of pulse positions searched in this sequential procedure is NP + P(P-1)/2. For the above example of N= 40 and P=5, this turns out to be 210 as compared to 658008 for the optimal full-blown search. Hence a significant saving in complexity is achieved in sequential search. This savings is in addition to the

significant savings achieved by not needing to invert matrices using Cholesky decomposition for every possible combination of pulse position like in optimal search

4.2.2 Regular Pulse Excitation Coding

Another popular analysis-by-synthesis technique that has been reported in literature that reduces the computational complexity and reduces the bit rate (which is used as the basis for the design of the Global System for Mobile (GSM) Communications Full Rate speech coder) is the Regular Pulse Excitation (RPE) coding technique Here the spacing between the non-zero pulses are held constant [35] This implies that the only position that needs to be transmitted to the decoder is the position of the first pulse relative to the start of a speech sub-frame, thereby achieving a significant reduction in bit rate or equivalently bandwidth. It is reasonable to expect then that in MPLP coding the pulse positions that get chosen to be transmitted in voiced speech segments include pitch pulses, in the absence of which a large mean square error would result. In RPE coding, the constraint imposed by equal spacing regardless of pitch period causes a severely sub-optimal grid of pulses to be selected. Hence in both types MPLP and RPE coding, there is a strong incentive (in terms of optimal selection of pulses) to first perform long-term prediction that removes periodicity in the LPC residual signal (and hence eliminates the strong pitch pulses) and then perform MPLP or RPE coding. This would then require fewer number of pulses and fewer bits to transmit each pulse amplitude, thereby resulting in bit rate reduction.

4.2.3 Modeling Periodic Component of Excitation - Long Term Prediction (LTP)

As described above, long term prediction is performed essentially to remove the periodic component from the residual signal and then model the long term prediction residual using techniques such as MPLP and RPE techniques described above. The problem of long term prediction is typically formulated as follows :

Let e(n) denote the LPC residual at time instant n, and let e(n) be predicted from e(n-D-M), e(n-D-M+1),.., e(n-D-1), e(n-D), e(n-D+1), .. , e(n-D+M) and let the long term prediction residual be denoted by w(n).

$$e(n) = \sum_{i=-M}^{M} \beta_i e(n - D - i) + w(n) \tag{4-3}$$

It is noted that if a signal is truly periodic with period equal to ΔT where T is the sampling interval (125 microseconds for 8 kHz sampling rate) and Δ a positive integer, then $w(n) = 0$ when $\beta_0 = 1.0$, $D = \Delta$ and $M = 0$. Since speech is a slowly varying nonstationary process, it is required to estimate D and β_i on short segments (typically every 5 ms, corresponding to 40 samples) of residual signal. While it is desirable to optimize for value of M, for reasons of complexity, M is chosen to be less than or equal to 1. Estimation of D and β_i can be performed in an open-loop or closed loop fashion. In open loop method, the objective is to estimate D and β_i that minimizes mean square value of long term prediction error, w(n) In closed loop method, the objective is to estimate D and β_i which when used in voice decoder will minimize a perceptually weighted distortion between reconstructed speech and original speech at the input of voice encoder. As will be described later, some speech coders (open loop and closed loop) try to estimate D with a fractional resolution (as opposed to integer resolution) by either interpolating the residual signal itself or interpolating the autocorrelation function of the residual signal.

4.2.3.1 Open Loop LTP

For the predictor formulation in equation (4-3) with M = 1, the equivalent of Yule-Walker equations can be written as

$$\begin{bmatrix} \sum_n e(n)e(n-D-1) \\ \sum_n e(n)e(n-D) \\ \sum_n e(n)e(n-D+1) \end{bmatrix} = \begin{bmatrix} \sum_n e^2(n-D-1) & \sum_n e(n-D)e(n-D-1) & \sum_n e(n-D+1)e(n-D-1) \\ \sum_n e(n-D)e(n-D-1) & \sum_n e^2(n-D) & \sum_n e(n-D)e(n-D+1) \\ \sum_n e(n-D+1)e(n-D-1) & \sum_n e(n-D)e(n-D+1) & \sum_n e^2(n-D+1) \end{bmatrix} \begin{bmatrix} \beta_{-1} \\ \beta_0 \\ \beta_1 \end{bmatrix}$$

based on which optimum value of $\underline{\beta}^* = \left[\beta_{-1}\ \beta_0\ \beta_1\right]^t$ is obtained as

$$\underline{\beta}^* = \Phi^{-1} \sum_n \underline{e}_{-D}(n) e(n) \tag{4-4}$$

for a given value of D where

$$\underline{e}_{-D}(n) = \begin{bmatrix} e(n-D-1) \\ e(n-D) \\ e(n-D+1) \end{bmatrix}$$

and

$$\Phi = \sum \underline{e}_{-D}(n)\, \underline{e}_{-D}{}^t(n)$$

The resulting mean square error is given by

$$E_D = \sum_n e^2(n) - \left(\sum_n \underline{e}_D(n) e(n) \right)^t \Phi^{-1} \left(\sum_n \underline{e}_D(n) e(n) \right)$$

Since Φ (and hence Φ^{-1}) is a positive definite matrix, minimizing mean square error E_D is equivalent to maximizing $\left(\sum_n \underline{e}_D(n) e(n) \right)^t \Phi^{-1} \left(\sum_n \underline{e}_D(n) e(n) \right)$. Hence in practice, $\left(\sum_n \underline{e}_D(n) e(n) \right)^t \Phi^{-1} \left(\sum_n \underline{e}_D(n) e(n) \right)$ is computed for all possible values of D and the value of D that yields the maximum value is chosen as the optimal delay value. For this optimal value of delay D, the LTP coefficient vector $\underline{\beta}^* = \left[\beta_{-1}\ \beta_0\ \beta_1\right]^t$ is computed as per equation (4-4).

4.2.3.2 Closed Loop LTP

Long-term prediction, using closed loop search is based an analysis-by-synthesis approach, whereby the long term prediction parameters (D, $\underline{\beta}$) are optimized based on reconstructing speech for permissible values of D (and corresponding $\underline{\beta}$) and comparing with original speech using a perceptual weighted filter

$$\frac{A(z/\gamma_1)}{A(z/\gamma_2)}.$$

Essentially, the cost function

$J(D, \underline{\beta}) = \| s_w(n) - z(n) - s_{D\,\underline{\beta}^{(n)}} \|^2$ is minimized with respect to D and $\underline{\beta}$, where $s_w(n)$ is the perceptually weighted input speech, typically derived as

$$s_w(n) = s(n) + \sum_{i=1}^{L} \alpha_i \gamma_1^i s(n-i) - \sum_{i=1}^{L} \alpha_i \gamma_2 s_w(n-i)$$

s_W is input speech signal and α_i (i=1, 2, .., L) are the unquantized linear predictive coding (LPC) coefficients. z(n) is the zero input response of the perceptually weighted synthesis filter

$$\frac{A(z/\gamma_1)}{A(z/\gamma_2)} \bullet \hat{A}(z)$$

which is subtracted from $s_w(n)$ in the equation above to remove the contribution of the previous frame in the optimization process. $\hat{A}(z)$ is the quantized LPC filter that is actually used at the remote decoder to synthesize speech.

$\tilde{s}_{D,\beta}(n)$ is the convolution of the truncated impulse response of the perceptually weighted synthesis filter and the past residual at signal delay D computed as

$$\tilde{s}_{D,\underline{\beta}}(n) = \sum_{i=-1}^{1} \beta_i \sum_{j=0}^{N-1} \tilde{e}(n-D-i-j)\, h(j) \qquad n = 0,1.2,.., N-1$$

h(0), h(1), , h(N-1) is the truncated impulse response of the weighted synthesis filter $\frac{A(z/\gamma_1)}{A(z/\gamma_2)} \bullet \hat{A}(z)$ which is computed in a manner similar to that described for multipulse coding in Section 4.2.1.

From an analysis similar to that for open loop, long-term prediction, it can be shown that minimization of J(D,$\underline{\beta}$) above is equivalent to the maximization of

$$\left(\sum_{n=0}^{N-1} \underline{e}_w^{(D)}(n)\left(s_w(n)-z(n)\right)\right)^t \Phi_D^{-1} \left(\sum_{n=0}^{N-1} \underline{e}_w^{(D)}(n)\left(s_w(n)-z(n)\right)\right) \qquad (4\text{-}5)$$

for all possible values of D (typically between D = 20 to D = 128), where

$$\underline{e}_w^{(D)}(n) = \begin{bmatrix} \sum_{j=0}^{N-1} \tilde{e}(n-D-1-j)\,h(j) \\ \sum_{j=0}^{N-1} \tilde{e}(n-D-1-j)\,h(j) \\ \sum_{j=0}^{N-1} \tilde{e}(n-D-1-j)\,h(j) \end{bmatrix}$$

and

$$\Phi_D = \sum_{n=0}^{N-1} \underline{e}_w^{(D)}(n)\left(\underline{e}_w^{(D)}(n)\right)^t .$$

Once the value of D=D* that maximizes equation (4-5) is obtained, the optimal vector

$\underline{\beta}^{*} = \left[\beta_{-1}^{*}, \beta_{0}^{*}, \beta_{1}^{*}\right]$ is obtained as

$$\underline{\beta}^{*} = \Phi_{D^{*}}^{-1} \sum_{n=0}^{N-1} e_{W}^{(D^{*})}(n)\left(s_{W}(n) - z(n)\right)$$

In most speech coders, in order to reduce the complexity of determining the vector of long-term coefficients, $\underline{\beta}$, corresponding to delays of D, D-1 and D+1, a scalar coefficient corresponding to delay D alone is computed In this case, for an open loop, long-term prediction, the solution for β_0 is obtained by first maximizing

$$\frac{\left(\sum_{n=0}^{N-1} \tilde{e}\,(n-D)\,e(n)\right)^{2}}{\sum_{n=0}^{N-1} \tilde{e}^{2}\,(n-D)} \qquad (4\text{-}6)$$

for permissible values of D and then computing optimal $\beta_0 = \beta^{*}_{0}$ as

$$\beta_{0}^{*} = \sum_{n=0}^{N-1} \tilde{e}\,(n-D^{*})\,e(n) \Big/ \sum \tilde{e}^{\,2}(n-D^{*})$$

where D* is the value of D that maximizes equation (4-6).

For closed loop, long-term prediction, the solution for β_0 is obtained by first maximizing

$$\frac{\left(\sum_{n=0}^{N-1}\left(\sum_{j=0}^{N-1}\tilde{e}(n-D-j)\quad h(j)\ (s_w(n)-z(n))\right)\right)^2}{\sum_{n=0}^{N-1}\left(\sum_{j=0}^{N-1}\tilde{e}(n-D-j)\quad h(j)\right)^2} \tag{4-7}$$

and computing optimal $\beta_0 = \beta_0$* as

$$\beta_0^* = \frac{\left(\sum_{n=0}^{N-1}\left(\sum_{j=0}^{N-1}\tilde{e}(n-D^*-j)\,h(j)\left(S_w(n)-z(n)\right)\right)\right)}{\sum_{n=0}^{N-1}\left(\sum_{j=0}^{N-1}\tilde{e}(n-D^*-j)\ h(j)\right)^2}$$

where D* is the value of D that maximizes equation (4-7).

One advantage of having a vector of long-term coefficients (corresponding to delay values of D, D-1 and D+1) compared to a single coefficient, β_0 is that it covers the case where the delay D is not an integer value but fractional. It is important to note that the role of D is to represent the fundamental frequency, f_0, of the talker for the speech segment under consideration.

However, as noted above, D is computed with resolution of the sampling period. In reality, the LPC prediction residual is periodic with duration ($1/f_0$) that is not an integer multiple of sampling period ($1/f_s$) where f_s is sampling frequency, typically 8 kHz). Hence, many speech coders attempt to estimate the correct fundamental frequency by computing D with a fractional resolution rather than an integer resolution. Computation of fractional values of D is typically performed in one of the following two ways In the more rigorous method, the LPC residual signal (actually the previous excitation sequence) is

upsampled using interpolation filters and then maximizing equations above for all fractional values within the permissible delay range (fractional resolution is typically 1/4 or 1/8). In a less rigorous method, the terms in the above equation itself are interpolated around an integer value of D, and the fractional value of D, for which the expression above is maximized, is obtained. A typical interpolation filter is implemented using a FIR filter based on a truncated windowed *sinc* function such as a Hamming Windowed *sinc* function.

4.2.4 Modeling Aperiodic Component of Excitation

As described above, the periodic component modeling using long-term prediction permits efficient modeling of long-term prediction residual using MP-LPC and RPE techniques. In fact, multipulse modeling of an aperiodic component of the excitation sequence is employed in the Inmarsat Full Rate Aeronautical Speech-Coding standard, operating at 9.6 kbps. The full rate GSM Speech-Coding standard, operating at 13 kbps, uses the RPE technique after long-term prediction to model the periodic component of the LPC residual signal.

However, by far, the most popular analysis-by-synthesis technique that has gained widespread importance and been employed by many speech coders, including many regional and international speech-coding standards, is the code excited linear prediction (CELP) technique. Here, short segments of the long-term prediction residual signal is approximated by an entry in the codebook of stored vectors. Essentially, a VQ of the long-term prediction residual is performed, but the choice of a codebook entry is based upon a synthesizing speech signal using each entry of the codebook and comparing it with original speech in a perceptually weighted domain. It is noted that the even the estimation of periodic component of LPC residual in an analysis-by-synthesis coder can be treated as selection from an a codebook whose entries for any subframe of speech consists of previous excitation signals delayed by different amounts (integer and/or fractional). Obviously the entries of such a codebook changes from one subframe to another and is hence popularly called as an *adaptive codebook*

The power of the CELP technique is that it encompasses most of the excitation modeling techniques, including multipulse and RPE modeling as special cases The CELP technique has also been sometimes referred to as vector excitation coding (VXC) and stochastically excited linear prediction (SELP), depending on the nature of the codebooks.

A typical CELP encoder block diagram is shown in the figure below. The aperiodic component of the excitation model is selected from a codebook of stored vectors, whose K-th vector is denoted by $\underline{C}_k = [C_{ok}, C_{ik}, .. C_{n-1,k}]^t$, where N is the dimension of vectors in the codebook (typically N=40 equivalent to 5 ms). Each N-dimensional vector in the stored codebook represents the shape of an N sample signal. Depending on the signal strength and the LTP prediction gain, the energy in the N sample LTP prediction residual varies. Therefore, the appropriate gain value is computed after the shape is computed Such a procedure is popularly referred to as a gain-shape representation of signals.

The objective here is to find an entry (vector) in the codebooks (and the associated gain) that when used in conjunction with the periodic component of an excitation sequence is input to the LPC synthesis filter, producing speech that is close to original speech in a perceptually weighted sense. The problem is formulated as the minimization of $J(\underline{C}_k, g_c)$ with respect to $\underline{C}_k$ and g_c

$$J(\underline{C}_k, g_c) = \left\| \underline{S}_w Z - H(g_c \underline{C}_k + \hat{\beta}_0^* \ \underline{\tilde{c}}^{(D^*)}) \right\|^2$$

where

$$\underline{S}_w = \left[s_w(0)\, s_w(1) \cdots s_w(N-1) \right]^t$$

$$\underline{Z}w = \left[z(0)\, z(1) \quad z(N-1) \right]^t$$

H is an NxN lower triangular matrix whose j-th row contains the truncated impulse response of the weighted synthesis filter, i.e ,

$$H = \begin{bmatrix} h(0) & 0 & 0..... & 0 \\ h(1) & h(0) & 0..... & 0 \\ h(2) & h(1) & h(0).... & 0 \\ h(N-1) & h(N-2) & ..... & h(0) \end{bmatrix}$$

β^*_0 is the quantized value of β^*_0 as computed using the above equation corresponding to a delay D* and

$\tilde{e}^{(D^*)} = [\tilde{e}(-D^*), \tilde{e}(1-D^*), ..., \tilde{e}(N-1-D^*)]$ is an aperiodic component of excitation vector based on past excitation. When D* is fractional (noninteger), $\tilde{e}(j-D^*)$ is obtained using an interpolation filter on past excitation. $s_w(n)$, z(n) and h(n) are the same as those obtained for the long-term prediction described above. Minimization of $J(\underline{C}_k, g_c)$ can be shown to be the same as the maximization of

$$\frac{\left(\underline{C}_k^t H^t (\underline{S}_w - \underline{Z} - H\hat{\beta}_0^* \underline{\tilde{e}}^{(D^*)})\right)^2}{\underline{C}_k^t H^t H \underline{C}_k}$$

Therefore, the above expression is computed for each vector $\underline{C}_k$ (k=0,1,..., 2^{B-1}) in the codebook and the entry $\underline{C}_k = \underline{C}_k^*$, which maximizes the above expression, is chosen as the shape vector that best represents the aperiodic component of the excitation sequence. The corresponding gain g_c^* is computed as

$$g_c^* = \frac{\underline{C}_{k^*}^* H^t \left(\underline{S}_w - \underline{Z} - H\hat{\beta}_0^* \underline{\tilde{e}}^{(D^*)}\right)}{\underline{C}_{k^*} H^t H \underline{C}_{k^*}}$$

Therefore, in the CELP coder, the optimal excitation to the LPC synthesis filter $1/\hat{A}(z)$ is given by

$$\tilde{e}(n) = \hat{\beta}_0^* \tilde{e}(n - D^*) + g_c^* C_{nk^*} \quad n = 0, 1, 2, .., N\text{-}1$$

The original proposal of CELP which is attributed to Atal and Schroeder [36] suggested the use of unstructured codebooks whose entries were Gaussian random numbers. This resulted in codebook search complexities which were prohibitively high but the promise that the technique held intrigued many researchers. As a result, numerous articles were published on CELP whose codebook search complexities were reduced usually by having structured codebooks Overlapped codebooks have been proposed whereby a given entry of the codebook is formed by a cyclic shift of the previous entry. Sparse excitation codebooks have also been proposed where many entries of the codebooks are zero with some constraints on the positions and magnitudes and signs of non-zero entries. A further simplification was achieved when the amplitude of nonzero pulses were constrained to have a magnitude of 1. Another family of codebook which has gained significant attention because of low complexity, no storage requirement and which spans a significant portion of signal space is the algebraic codebook which is used in the ITU-T 8 kbps toll quality speech coding standard. Another approach of generating excitation codebooks uses centroids of vectors obtained from a large corpus of speech material, very similar to the generation of vector quantizer (VQ) codebook; this led to sophisticated codebook generation principles similar to that used in VQ such as use of multi-stage VQ (or equivalently multiple codebooks) which inherently has a reduced search complexity and reduced storage as opposed to a full-VQ with a single codebook. One such coder is the 8 kbps Vector Sum Excited Linear Predictive (VSELP) coder which was selected for the Full Rate North American Digital Cellular TDMA standard. Here, the two stochastic codebooks are used to model the aperiodic component of the LPC residual signal. The excitation sequence in VSELP is formed by adding vectors from the two stochastic excitation codebooks.

5. Transform (non-time) domain speech coding

The waveform coders and parametric coders based on LPC analysis described above perform processing of speech signals in the time domain. However, there are many speech coders that have gained widespread usage that perform processing in a transformed domain such as frequency domain, quefrency (log-magnitude) domain and other unitarily transformed domains Here speech is first transformed into (or

represented in) the desired domain using the appropriate transform (Discrete Fourier Transform (DFT) or Fast Fourier Transform (FFT) for frequency domain, log-magnitude of frequency domain spectrum for quefrency domain, and, Discrete Cosine Transform (DCT) or Walsh-Hadamard Transform (WHT) or Karhunen-Loeve Transform (KLT) for unitarily transformed domain) and then analyzed accordingly. A primary motivation behind adopting transform domain coding is to exploit the human perception mechanism which is better understood in transform domain rather than time-domain. Sub-band coding, Multi-Band Excited (MBE) coding and Sinusoidal Transform Coding (STC) are popular examples of frequency domain speech coding technique. The Adaptive Transform Coder (ATC) with Discrete Cosine Transform is a popularly referred unitary transform speech coding technique. The Homomorphic vocoder is a good example of the quefrency domain coding

5.1 Sub-band coding

Here speech signal is first transformed into frequency domain and the spectrum is divided into frequency bands, which in general have unequal width [26]. Division of speech spectrum into bands is achieved using bandpass filter-banks such as the lossless Quadrature Mirror Filter [27] banks. Depending on the width of the bandpass filter, the output of the filter is down-sampled or decimated after transforming each band into baseband. Depending on the energy in the pass-band region of the filter-bank, the down-sampled speech samples are encoded like PCM/ADPCM (as described in Sections 2.1 and 2.2) with different number of bits per sample. For example, the down sampled speech samples belonging to the lower frequency bands are usually allocated more bits per sample than higher frequency bands since the lower frequency bands usually carry more energy than higher frequency bands. Furthermore, from a human perception mechanism, proper representation of the lower frequency bands are more critical than that of higher frequency bands. It is noted that the SBC can still be considered as a sample-by-sample processing technique because of the way in which encoding is performed. A simplified block diagram of a typical sub-band coder is illustrated in Figure 13. An excellent treatise on sub-band coding techniques is provided in [25]. ITU-T has standardized a wideband speech coder operating at 64 kbps and below which uses sub-band coding as described in ITU-T G.722.

5.2 Multi-Band Excitation (MBE) Coding

Here frames of speech (typically 20 ms duration) are represented in the frequency domain as a set of parameters that describe the fundamental frequency, magnitudes and phases of harmonics of fundamental frequency as well as a decision about whether a given harmonic is voiced or unvoiced. The voice/unvoiced decision on a per-harmonic (or per-band encompassing several harmonics) is unique to the MBE coder as compared to other traditional vocoders where the entire frame of speech is declared as voiced or unvoiced [49]. It is noted that the MELP coder described in Section 4 1.2 can be treated as a special case of MBE coding where frequency band below a given cut-off frequency is treated as voiced and frequency band above cut-off frequency is treated as unvoiced. At the decoder of MBE coder, speech is synthesized using the parameters received from the encoder; specifically the voice decoder generates speech samples whose spectrum will comprise of periodic and noisy contributions as indicated by the voiced/unvoiced decisions on a per-harmonic basis. A typical MBE voice encoder is illustrated Figure 14.

It is noted that the fact that human speech has sounds that has periodic and aperiodic components at the same instant of time was recognized as early as 1939 [1]. The MBE concept effectively utilizes this feature to produce good quality speech at very low bit rates. As mentioned in Section 5.1 above for sub-band coding, the MBE coder also has the distinct advantage of being capable of perceptually enhancing speech since the human speech perception mechanism is much well understood in frequency domain than in time domain. The MBE coders have also typically exhibited good performance in presence of background noise. The MBE coders are being used in many mobile satellite communication systems, including the Inmarsat-M and Inmärsat Mini-M systems [29,30]

5.3 Sinusoidal Transform Coding (STC)

The basic idea behind Sinusoidal Transform Coding (STC) [31] is that speech is reconstructed as a sum of sinusoids whose amplitudes, frequencies and phases are interpolated between sets of encoded parameters. These parameters are regularly updated by applying short-term Fourier analysis to representative speech

segments at the encoder. The resulting spectra will normally exhibit magnitude peaks located, in principle, at harmonics of the pitch frequency for voiced speech and randomly for unvoiced speech. The required parameters are derived from the spectra at the frequencies identified by the peaks. Similar to MBE coder, STC coder is a parametric frequency domain coder. Unlike MBE coder, the STC declares an entire frame of speech as voiced or unvoiced. In the general form of STC, the set of frequencies are not necessarily harmonically related and hence has the capability to produce naturally sounding speech at moderate bit rates. A simplified block diagram of STC is shown in Figure 15.

5.4 Adaptive Transform Coding

Here a block of speech signals is transformed using DCT, WHT or KLT, and the resulting block is quantized and transmitted [32]. Each transformed element is quantized with different number of bits depending on its perceptual importance. The KLT yields the maximally decorrelated transformed sequence and hence the most optimal. However, the derivation of the basis vectors for KLT is computationally expensive and data dependent. Hence less computationally expensive, although sub-optimal, that are data independent such as DCT is popularly used in practical speech coding implementations. A significant advantage of this approach is the uncorrelatedness of the transformed sequence which allows quantization effects of each transformed element uncorrelated with each other. Furthermore, bit assignment to the transformed vector can be made adaptive based on its perceptual importance. A simplified block diagram if ATC coder is shown in Figure 16.

5.5 Homomorphic Coding

The homomorphic vocoder is a transform domain vocoder where speech is processed in quefrency domain or equivalently, a cepstral domain. A signal y(n) is said to be a cepstral domain representation of x(n) if x(n) has undergone the following transformation

$$y(n) = \Im^{-1}\left\{\log\left(\left|\Im\{x(n)\}\right|\right)\right\}$$

where $\Im$ represents the Fourier transform and $\Im^{-1}$ represents the inverse Fourier Transform. The principle behind homomorphic vocoding is to separate the vocal tract spectrum from excitation spectrum. It is noted that under the assumptions of linearity of the vocal tract system, the speech output of human speech production system can be written as

$$S(\omega) = E(\omega)V(\omega)$$

where $S(\omega)$ is the speech spectrum, $E(\omega)$ is the excitation spectrum and $V(\omega)$ is the vocal tract spectrum. The cepstral domain transformation above permits separation of spectrum because of the logarithmic transformation involved. At the decoder, an inverse transformation is applied to bring it back to time domain

6. Speech Coding Standards

The commercialization of digital speech coding has accelerated in the last decade with the adoption of new speech coding standards and the introduction of major new technologies into commercial networks. This has been motivated by capacity limitations on major international and trans-continental transmission facilities, explosive growth of wireless communications, higher demand for integrated services such as voice, video and data, and increased interest in communication privacy. In the previous sections, various speech coding techniques were discussed during which it was mentioned that several of these techniques were part of regional and international speech coding standards. Some of these standards will be discussed briefly in this section.

6.1 Speech Coder Attributes : Quality and Bit Rate

Before embarking into a review of speech coding technology standards and their evolution, it is important to define attributes to be employed in determining the state of technology at a given time. As discussed in Section 1 in introduction, two conflicting attributes are most important in this regard· the transmission rate of the technique or class of techniques in question; and the end-to-end transmission quality.

Often, the term telecommunications-quality, toll-quality, or more recently wireline-quality, is given to speech coding technology which introduces no perceptible distortion. This does not mean that degradation may not be quantifiable, but simply that distortion is not perceptible. The type of technology used in the network today, such as 64 kb/s Pulse Code Modulation (PCM) and 32 kb/s Adaptive Differential Pulse Code Modulation (ADPCM) as described in Sections 2.1 and 2.2, respectively, are good examples of wireline-quality coding.

In a similar fashion, the term cellular-quality is used in this paper to indicate transmission performance that is less than wireline, and specifically quality that is associated with a perceptible degradation to users. Cellular quality is not annoying in most instances, and typically, all speaker features such as identity and intonation are preserved. Simply, in this paper, cellular-quality is defined as being equivalent to that of full-rate digital standards, such as North American Full Rate Digital Cellular speech coding standard - the 8 kb/s Vector Sum-Excited Linear Prediction (VSELP) and Full Rate GSM speech coding standard - 13 kb/s Regular Pulse Excited with Long-Term Prediction (RPE-LTP). These technologies were discussed in Section 4.

Communications quality denotes performance that is associated with perceptible degradation that can be annoying in some instances. With communications quality, speaker features are occasionally lost, but intelligibility is preserved in most instances. Unlike wireline and cellular quality, some perceptible loss in naturalness is also experienced. Here, communications-quality is defined as being perceptually equivalent to that of Federal Standard 1016, the 4.8 kb/s Code Excited Linear Predictive (CELP) Coder mentioned in Section 4.2.4. Typically, communications quality has been considered as the lower-bound of commercial acceptability.

Intelligible-quality provides yet another demarcation into discernible service performance, whereby a further reduction in quality is manifested by a typical loss of speaker identity and a measurable, but not unacceptable, loss in intelligibility. Naturalness is also typically lost with intelligible-quality coders. It is useful to think of intelligible-quality as being equivalent to that of Federal Standard 1015, the 2.4 kb/s Linear Predictive Coder-10e (LPC-10e) mentioned in Section 4.

Finally, for the sake of completion, it is useful to define one more performance range, this being synthetic-quality. In synthetic quality coders, reproduction of input speaker naturalness is not possible and these coding techniques are typically both speaker- and vocabulary-dependent Coders exhibiting synthetic-quality operate (today) by encoding speech with a few hundred bit/s.

A pictorial representation of the first four of the five quality descriptors against the Mean Opinion Scores (MOS) and Equivalent-Q scales is given in Figure 17. The actual MOS scores or equivalent Q-values for the different quality coders could be different from those shown in Figure 17 for any given test [39], depending on factors such as input speech spectral shaping, input speech level, and type of listening instrument used [40]. However, the important thing to be observed from Figure 17 is the ***relative*** performance of the different voice coding technologies (ordinal presentation) and their relative performance difference, both of which are less dependent on the factors mentioned above.

MOS represents averaged opinions of circuit quality by mapping expressed rating of excellent, good, fair, poor and bad, to 5, 4, 3, 2 and 1, respectively. The "Q value" is the ratio of the speech level to the multiplicative noise level (expressed in dB) that is derived when random noise with an amplitude that proportional to the instantaneous speech amplitude is added to the speech signal as defined in ITU-T Recommendation P.810 For a given speech coder, the equivalent Q-values are obtained by means of subjective tests as described in Section 7. It is worth noting that the technologies shown in Figure 17 belong to a diverse generation of voice coders.

6.2 Evolution of Speech Coding Technology and Standards

The first introduction of digital voice encoding technology into commercial service occurred in the early 1970s with the adoption of 64 kb/s PCM as a standard for the transport of voice and voiceband services over the public switched telephone network (PSTN). Since then, the digitization of international and transcontinental transmission facilities and the associated rapid growth in voice and voiceband data traffic highlighted the need of further efficiency improvements when transmitting of voice signals. This need was fulfilled by the simultaneous evolution of technology which made possible in the early-1980s the delivery of wireline-quality digital voice at one-half the PCM rates using 32 kb/s ADPCM.

Since the early 1980s pressure to improve the transmission efficiency of voice signals has continued to rise, despite the rapid expansion of wireline network capacity As a consequence, in the early 1990s, 16 kb/s Low-Delay Code Excited Linear Prediction (LD-CELP) was developed and adopted by the International Telecommunication Union (ITU) as a new wireline-quality voice compression standard. This was followed by the selection of 8 kb/s Conjugate Structured Algebraic Code Excited Linear Prediction (CS-ACELP) coder as a new world standard by the ITU in 1995. Wireline-quality ITU standards have stimulated and preceded, rather than followed, the evolution of voice coding technology Thus, by noting the year of adoption of different technologies as ITU standards, it is possible to quantitatively observe the evolution of wireline-quality voice coding technology over time This is depicted in Figure 18(a).

From this figure it can be seen that in the early stages of voice coding (early 1970s to early 1980s), technology improvement resulted in the ability to reduce voice coding rates by approximately 3.2 kb/year. In the 1990s this rate slowed down to 1 8 kb/year, although the ability to half the transmission rate actually accelerated. This is more clearly seen in Figure 18(b), where the relationship shown in Figure 18(a) is plotted against a logarithmic scale.

Somewhat of a similar behavior can also be observed when considering the use of communications-quality coding for providing commercial service (Figure 19). In this case, in the mid-1980s, when communications-quality coders were first introduced into commercial service (principally for mobile-satellite applications by Inmarsat), it was possible to improve efficiency at approximately 3 2 kb/year, although this rate has recently slowed down to more like 0.8 kb/year.

From the above it can be seen that over the past decade years it was possible to reduce voice transmission rates while maintaining quality, a fact that is expected to continue in the near future. Nonetheless, even though these relationships appear to monotonically relate transmission rates with time, in reality when examined microscopically, they tend to reveal a series of step functions whereby the ability of technology to deliver lower rates remains constant until some breakthrough causes bit-rate to suddenly drop. Consequently, at any one time, it is not readily obvious whether technology has reached the flat part of step, or is about to make a major breakthrough and permit a steep reduction in bit-rate to occur.

Simultaneously, for wireless applications such as cellular, mobile satellite, aeronautical, maritime, and, military voice communications, where bandwidth is scarce and often expensive, lower bit rate speech coders (as low as 2 4 kb/s) were explored. Such efforts have led to speech coding standards, which among others, include the North American Digital Cellular Standard which uses a 8 kb/s VSELP speech coder, the full-rate European Digital Cellular Standard which employs a 13 kb/s Regular Pulse Excitation with Long Term Prediction (RPE-LTP) speech coder, the Japanese Digital Cellular Standard which employs a modified 6.4 kb/s VSELP speech coder, the International Maritime Satellite (Inmarsat) Aeronautical Standard which employs a 9.6 kb/s Multi-Pulse excited Linear Predictive Coder (MPLPC), the Inmarsat-M Standard which employs a 4.15 kb/s Improved Multi-Band Excited (IMBE) speech coder, the Inmarsat-Mini-M Standard which employs a 3.6 kb/s Advanced Multi-Band Excited (AMBE) speech coder, the US Department of Defense (DoD) Federal Standard FS1016 which employs a 4.8 kb/s CELP speech coder, DoD Federal Standard FS1015 which employs a 2.4 kb/s Linear Predictive Coding (LPC-10) based speech coder, and the newly standardized 2.4 kbps MELP coder as a replacement to FS1016. .

In the sequel, some basic information about the speech coding technologies involved in some ITU, GSM, Inmarsat and DoD standards are discussed. It should be noted that in describing the above mentioned speech coding technologies, emphasis is placed only on the key features associated with each of the technologies and no attempt is made to give all the details entailed in the development of these technologies - details are available in the appropriate references.

6.3 Wireline or Toll Quality Speech Coding Standards

6.3.1 64 kb/s ITU-T Pulse Code Modulated Speech Coder (Recommendation G.711)

The PCM system as described in ITU-T Recommendation G.711 [6] consists of a pre-filter, a sampler, and an analog-to-digital converter at the encoder, and, a digital-to-analog converter and a low pass filter at the decoder. The continuous time speech is typically low pass filtered with a cut-off frequency slightly less than 4 kHz and then sampled at a rate of 8000 samples/sec. Each sample is then quantized using 8 bits and transmitted to the decoder. The decoder then converts the digital stream to the

corresponding amplitude, and the discrete time signal is then passed through a low-pass filter to obtain a reconstructed continuous-time speech signal. As described in Section 2.1 ITU-T Recommendation G.711 provides two encoding laws, namely A-law and μ-law to enhance the dynamic range of the signal without sacrificing the signal-to-quantization noise ratios. Both encoding laws explore the fact that the instantaneous amplitude of the speech signal is less that 25% of its maximum amplitude for more than 50% of the time and hence finer quantizations could be performed on small amplitude samples and coarser quantization on larger amplitude samples. The mapping tables of these encoding algorithms are provided in Tables 1 and 2 of ITU-T Recommendation G.711.

6.3.2 32 kb/s ITU-T Adaptive Differential Pulse Code Modulated Speech Coder (Recommendation G.726)

During the period of 1982-1990, ITU-T (then called as CCITT) adopted several adaptive differential pulse code modulation (ADPCM) algorithms. First the 32 kbps ADPCM algorithm described in G.721 was approved. Later G.723 was standardized which basically was an adaptation of the 32 kbps algorithm in G.721 to 40 kbps to handle voice band data and 24 kbps to handle network congestion. In 1990 CCITT combined G.721 and G.723 and added another ADPCM rate at 16 kbps to handle overload situations resulting in a new recommendation ITU-T G.726 [7] which defines an ADPCM voice coding algorithm operating at 40, 32, 24, and 16 kb/s.

The basic components of the G.726 ADPCM codec are an adaptive sample-by-sample predictor, and adaptive quantizer, and an adaptive inverse quantizer. The difference signal obtained by subtracting the predicted and inverse quantized signals from the original signal is then adaptively quantized and forms the ADPCM output bit-stream. The G.726 ADPCM encoder is similar in principle to that illustrated in Figure 5. As described in Section 2.2, in order to prevent the effect of accumulation of quantization errors, a replica of the remote voice decoder is included in the encoder structure. The adaptive predictor is a pole-zero filter as described in equation (2-4) of Section 2.2, with $N_1 = 2$ and $N_2 = 6$. Such a pole-zero predictor is called as Auto Regressive Moving Average (ARMA) predictor denoted by ARMA(2,6)

denoting the number of coefficients in auto regressive and moving average portions of the predictor It is noted that $\beta_0 = 1$ in equation (2-4) and in G.726. The ARMA coefficients are updated on a sample-by-sample basis (at both encoder and decoder) thereby making it adaptive. Since ADPCM employs backward prediction and sample-by-sample processing the algorithmic delay is equal to 0.125 ms. ADPCM is used in stand-alone codecs, T1 and E1 multi-channel transcoders, and in digital circuit multiplication equipment (DCME) systems such as ITU-T Recommendation G.763 [41].

In addition, an embedded version (ITU-T Recommendation G 727 [8]) of 32 kb/s ADPCM encoding with voice quality indistinguishable from that of 32 kb/s G.726 is used in Packet Circuit Multiplication Equipment (PCME). Embedding permits certain "enhancements" bits to be dropped in the network during congestion without informing the encoder or without any exchange of control information.

The ADPCM techniques defined in ITU-T Recommendations G.726 and G.727 provide the ability to vary the transmission rate among four different bit-rates. The higher (40 kb/s) rate is employed when high-speed voiceband data are being transmitted; while the two lower (24 kb/s and 16 kb/s) rates are employed dynamically as part of an overload traffic control strategy. Consequently, 24 kb/s and 16 kb/s ADPCM are not steady-state encoding rates under normal operating conditions.

6.3.3 16 kb/s Toll Quality ITU-T Standard (Recommendation G.728) : Low Delay CELP (LD-CELP) Coding

CELP coders have been demonstrated to produce very high quality speech at 16 Kb/s. However, like many other parametric speech coders, they contribute a delay typically well above 10ms. In many practical situations such as public switched telephone networks (PSTN) and more complicated networks where tandem encoding are necessary, such long delays contribute to a significant impairment to the performance of the network and in many cases is simply unacceptable. Indiscriminate deployment of long delay parametric speech coders in PSTN trunks could cause substantial revision of echo control procedures in both networks and terminal equipments. The LD-CELP [24] coder was introduced in an effort to meet the

performance requirements specified by CCITT which was to achieve toll-quality speech at 16 Kb/s with a total delay of no greater than 5 msec.

A block diagram of an LD-CELP coder is shown in Figure 20. The essence of CELP which was described in Section 4.2.4 is retained in LD-CELP. The main difference is that, CELP uses forward adaptation for computing the coefficients of the short term prediction filter whereas the LD-CELP coder uses backward adaptive short term predictor. In a backward adaptive configuration, the coefficients of the short term filter are not derived from original speech, but instead from the past reconstructed speech. Since both encoder and decoder have access to the past reconstructed speech, information about the short term filter coefficients is no longer necessary to be transmitted to the decoder. Thus in contrast to CELP where the prediction coefficients, the gain and the excitation sequence have to be transmitted, LD-CELP requires transmission of the excitation sequence only (see Figure 20). The predictor coefficients are obtained by performing LPC analysis on previously quantized speech and the gain is obtained by using the gain information embedded in previously quantized excitation.

For the 16 Kb/s LD-CELP coder, the excitation vector in the excitation codebook has a dimension (or block size) of 5 samples. The long-term predictor or pitch predictor present in the conventional CELP coder is eliminated and a 50-th order LPC analysis is used. The LPC predictor coefficients are updated once every 4 speech vectors (2.5 msec) by performing LPC analysis on previously synthesized speech. The excitation gain is updated once every vector by using a 10-th order adaptive linear predictor in the logarithmic domain. The coefficients of this log-gain predictor are updated once every four vectors by performing LPC analysis on the logarithmic gains of previously quantized and scaled excitation vectors. The 10-th order perceptual weighting filter is also updated once every four vectors by using a 10-th order LPC analysis of input speech. To reduce the complexity (in terms of codebook search time) and algorithmic delay, the 10 bits that are available to represent blocks of 5 samples (at 16 Kb/s) are used to encode a product code of 3 bit gain codebook and a 7 bit shape codebook.

The three gain bits consist of a sign bit and two magnitude bits. This sign bit has the effect of doubling the shape codebook size while retaining the same search complexity. In LD-CELP, the shape codebook is closed-loop optimized by a codebook design algorithm based on the perceptually weighted criterion used

by the LD-CELP encoder. This in contrast to conventional CELP coders which use Gaussian random numbers to populate the codebook. The shape codebook design algorithm is similar to the LBG algorithm for vector quantizer design [42]. After the shape codebook is designed, Pseudo Gray Coding [43] is used to assign codebook indices. With Gray-coded codebook indices, a single bit error will result in a decoded codevector close to the transmitted one. Such a technique significantly improves the performance of the coder under noisy channel conditions.

Finally an adaptive postfilter is used at the decoder to increase the perceptual quality of the synthesized output. The postfilter essentially consists of a short-term an long-term postfilter , the short-term postfilter parameters derived as a result of the LPC analysis performed in decoder for synthesis and the long-term post filter parameter is obtained by performing pitch extraction based on previously reconstructed speech.

The 16 Kb/s LD-CELP was adopted as an ITU-T standard for toll-quality speech coding at 16 Kb/s under Recommendation G.728 in 1992.

6.3.4 8 kb/s Toll Quality ITU-T Standard (Recommendation G.729): Conjugate Structured Algebraic CELP (CS-ACELP)

The 8 kb/s CS-ACELP [44] coder was standardized by the ITU as a new world standard for toll-quality speech coding in 1995. The CS-ACELP, as its name indicates, also belongs to the CELP family of coders. Here the coder operates on speech frames of 10 ms and looks ahead 5 ms for LPC analysis. Hence the algorithmic delay of the coder is 15 ms. Every speech frame is divided into two equal subframes of 5 ms each. Linear Prediction is performed using Levinson Durbin algorithm that uses a bandwidth-expanded autocorrelation coefficients. LPC to LSF conversion is performed using Chebychev polynomials. The 10-th order LSF vector is then quantized using a predictive two- stage vector quantizer (VQ) with 18 bits.

In comparison to the traditional CELP approach, the excitation sequence to the decoder is determined using two codebooks, a fixed codebook and an adaptive codebook (see Figure 21). The fixed codebook has an algebraic structure that helps determine 4 non-zero pulses and their positions per sub-frame of speech using 17 bits As illustrated in table below, every 5 ms (or every 40 samples) three pulses are chosen from three

mutually exclusive sets each of which contain 8 possible positions, thereby requiring 3 bits each to convey the chosen pulse position to the remote decoder. Fourth pulse is allowed to occur in any of the remaining 16 pulse positions, thereby requiring 4 bits to convey its pulse position to the remote decoder. Associated with each pulse position is a sign information that also has to be conveyed to the remote decoder.

Pulse ID	Positions
1	0,5,10,15,20,25,30,35
2	1,6,11,16,21,26,31,36
3	2,7,12,17,22,27,32,37
4	3,4,8,9,13,14,18,19,23,24,28,29,33,34,38,39

The gains of adaptive and fixed codebooks are vector quantized using 7 bits per sub-frame using a conjugate structured codebook. While the algebraic structured codebook significantly reduces the complexity of the algorithm, the conjugate structured codebook increases the robustness of the coder against channel errors.

The adaptive codebook index (or equivalently the optimal delay) for the first sub-frame T_1 is transmitted using 8 bits. The 8 bits represent fractional delay with 1/3 sample resolution in the range $\left[19\frac{1}{3}, 84\frac{2}{3}\right]$ and integer delay in the range [85,143]. For the second subframe, the adaptive codebook index always represents fractional delay with 1/3 sample resolution in the range $\left[\text{int}(T_1) - 5\frac{2}{3}, \text{int}(T_1) + 4\frac{2}{3}\right]$ which is transmitted using 5 bits. As described in Section 4.2.3.2, fractional delays are obtained by interpolating the autocorrelation function of the residual using a Hamming windowed *sinc* function. With an additional parity bit for adaptive codebook indices, a total of 80 bits is transmitted every 10 ms yielding a bit rate of 8 kbps.

It is noted that since 1995 ITU has been active in the process of standardizing a 4 kbps Toll Quality speech coding standard with the objective of standardizing the algorithm in the year 2000. Once again, the

requirements and objectives for such a toll quality was proposed early enough [45] to provide a clear target speech coding researchers.

6.4 Cellular Quality Speech Coding Standards

6.4.1 North American Full-Rate Digital Cellular TDMA Standard (IS 54) : Vector Sum Excited Linear Predictive (VSELP) Coder

The VSELP [46] coder operating at 8 kb/s has been adopted as a standard for North American Time Division Multiple Access (TDMA) Digital Cellular communications The VSELP coder, like the CELP coder, falls into the class of analysis-by-synthesis coders. The VSELP coder was designed to accomplish the highest possible speech quality with robustness to channel errors while maintaining a reasonable computational complexity at 8 Kb/s. The VSELP speech coder achieves these goals through efficient utilization structured excitation codebooks. The structured codebooks contributes to maintaining reasonable computational complexity while increasing robustness to channel errors.

In comparison with the traditional CELP coder structure, the excitation sequence to the decoder in VSELP is derived from three codebooks, namely one adaptive codebook that is associated with the fundamental frequency of the speech signal and two stochastic codebooks. As the name implies, the excitation sequence to the decoder is derived as a weighted sum of the three vectors in the three codebooks. The codewords in the stochastic (fixed) codebooks are formed such that a single bit error in a VSELP codeword on the channel does not affect the output of the vector sum. The frame size for VSELP coder is 20 ms and sub-frame size is 5 ms.

A 10-th order LPC analysis is performed and as described in Section 3.1, the LPC coefficients are represented as reflection coefficients. The 10 reflection coefficients are scalar quantized using 38 bits, the bit allocation being such that the first reflection coefficient represented using 6 bits whereas the last reflection coefficient represented using only 2 bits. Excitation parameters are updated and transmitted every sub-frame of 5 ms. The adaptive codebook is searched for 128 possible lags using the closed loop

search as described in Section 4.2.3.2. Hence the adaptive codebook index is transmitted every 5 ms using 7 bits.

The two stochastic codebooks contain 128 entries each (hence need to transmit 14 bits every 5 ms) and each entry is 40 samples wide. The codebook entries are formed by linearly combining seven basis vectors such that when the codebooks are Gray-coded, a bit error in the transmitted codebook index will only lead to a selection of codebook entry in decoder that was formed which differed in only one basis vector. Thus robustness to channel errors is achieved.

The three codebook gains are (jointly) transmitted every 5 ms using 8 bits and an overall energy of the speech frame using 5 bits per 20 ms is also transmitted. With an additional spare bit, the IS-54 VSELP coder transmits 160 bits every 20 ms thereby achieving a bit rate of 8 kbps

Modified versions of VSELP have also been used in Full Rate Japanese Digital Cellular standard and GSM Half-Rate Cellular standard.

6.4.2 GSM Full Rate Standard : Regular Pulse Excitation Coding with Long Term Prediction (RPE-LTP)

The RPE-LTP coder operating at 13 kb/s [47] was adopted as the Full Rate standard for GSM Time Division Multiple Access (TDMA) Digital Cellular communications. The RPE-LTP coder, like the CELP coder, falls into the class of analysis-by-synthesis coders. RPE-LTP coder processes speech in frames of 20 ms duration and subframes of 5 ms duration.

An eighth order LPC analysis is performed every 20 ms (but interpolated every 5 ms) and the LPC coefficients are represented in LAR domain as described in Section 3.3. The eight LAR coefficients are scalar quantized using a total of 36 bits and similar to VSELP, the bit allocation is different for different LAR coefficients. The first LAR coefficient is quantized using 6 bits where as the last coefficient is quantized using only 3 bits.

Every subframe of 5 ms (40 samples), the long term prediction lag and LTP gain (D^*, β_0^*) in Section 4.2.3.2 are quantized using 7 and 2 bits respectively. Thirteen equally spaced pulses are chosen from 4 possible candidates in a closed loop manner as described in Section 4.2.4 and the chosen candidate is

indicated using 2 bits. As described in Section 4.2.2, because of uniform and known spacing between pulses, the RPE coder has the advantage that the individual pulse positions need not be transmitted unlike the traditional multipulse coders. The normalized pulse amplitudes are quantized using Adaptive PCM technique using 3 bits each. The normalizing factor, which is the maximum of all amplitudes is transmitted using 6 bits. Therefore, a total of 7+2+2+(13*3)+6 = 56 bits are used to represent the excitation sequence every 5 ms.

Hence a total of 260 bits (36 bits for LAR coefficients and 224 (56*4) bits for excitation) are transmitted every 20 ms, thereby achieving a bit rate of 13 kbps A detailed description of the RPE-LTP algorithm can be found in ETSI Recommendation GSM TS 06.10.

6.5 Communications Quality Speech Coding Standards

6.5.1 INMARSAT Full-Rate Aeronautical Standard: The MultiPulse Excited Linear Predictive Coder (MPLPC)

The INMARSAT Aeronautical System employs the Multipulse Excited Linear Predictive (MPLPC) Coder operating at 9.6 kb/s [39].

The Inmarsat Full Rate Aeronautical Standard processes speech in frames of 20 ms duration and models excitation in subframes of 4 ms duration. A 10-th order LPC analysis is performed every 20 ms and the LPC coefficients are represented as reflection coefficients for quantization as described in Section 3.3. The ten reflection coefficients are scalar quantized using 40 bits. Unlike VSELP or GSM, long-term prediction is performed here only every 20 ms rather than every subframe. The LTP lag and gain (D^*, β_0^*) in Section 4.2.3.2 are quantized using 6 and 2 bits, respectively. Multipulse excitation analysis described in Section 4.2.1 is performed on the residual signal after long-term prediction. In order to reduce complexity, the sequential search approach described in Section 4.2.1 is used. Here for every 4 ms (32 samples) duration of the residual signal, three pulses that are subjectively more important are chosen (using analysis-by-synthesis). The position of the three pulses (p_1^*, p_2^*, p_3^* of Section 4.2.1) are quantized using 5 bits each The amplitude of first two pulses (g_1^*, g_2^* of Section 4.2.1) are quantized using 4 bits each and amplitude

of third pulse (g_3^* of Section 4.2.1) is quantized using 3 bits. Overall, the long term prediction residual is quantized using 26 bits every 4 ms.

Hence a total of 192 bits (40 bits for reflection coefficients, 6 bits for LTP lag, 2 bits for LTP gain, 130 (26*5) bits for excitation and 14 bits of error control bits are transmitted every 20 ms, thereby achieving a bit rate of 9.6 kbps.

6.5.2 The United States Department of Defense 4.8 kb/s CELP FS1016 Coder

The FS1016 coder [48] which is primarily used in military applications such as United States Department of Defense operates at 4.8 kb/s and is based on the CELP structure. FS1016 coder uses a 30 ms frame size with four 7.5 ms sub-frames. CELP analysis consists of three basic functions. a) short-term linear prediction; b) long-term adaptive codebook search; and c) innovation stochastic codebook search. CELP synthesis consists of the corresponding three synthesis functions performed in reverse order with the addition of a post-filter to enhance reconstructed speech.

A tenth-order LPC analysis is used to model the speech signal's short-term spectrum, or formant structure. The corresponding LSF parameters are scalar quantized using 34 bits per frame. Every subframe of 7.5 ms, long term signal periodicity or pitch is modeled by an adaptive codebook. The optimal adaptive codebook index D^* for the first and third subframes are represented using 8 bits each, whereas the second and fourth subframe are represented using 6 bits each. The adaptive codebook index represents the optimal pitch in fractional resolution as described in Section 4.2.3.2. The adaptive codebook gain β_0^* is represents using 5 bits per subframe. The residual from the short term LPC parameters and pitch VQ is vector quantized using a fixed stochastic codebook of size 512, thereby requiring 9 bits per subframe to transmit the optimal stochastic codebook index. The optimal scaled excitation vectors from the adaptive and stochastic codebooks are selected by minimizing a time-varying perceptually weighted distortion measure that improves subjective quality by exploiting masking properties of the human ear. The optimal stochastic codebook gain (g_c^* of Section 4.2.4) is quantized using 5 bits per subframe. Hence a total of 104 (28 + 20

+36 + 20) are used every 30 ms to represent the excitation sequence to the LPC synthesis filter at the remote decoder.

This together with 34 bits for LSF quantizer, 1 bit for frame synchronization, 4 bits for error control and 1 bit for future expansion leads to 144 bits every 30 ms leading to a bit rate of 4.8 kbps.

6.5.3 INMARSAT -M and INMARSAT Mini-M Speech Coding Standards : Multi-Band Excitation (MBE) Coding

INMARSAT standardized the Improved Multi-Band Excited (IMBE) Coder operating at 4.15 kb/s for Inmarsat-M service (that uses a briefcase size terminal) and later, the Advanced Multi-Band Excited (AMBE) Coder operating at 3.6 kb/s for Inmarsat-Mini-M [29,30] (that uses a notebook size terminal) service both of which are based on the basic Multi-Band Excitation (MBE) [49] speech model. The principles of the two coders are essentially the same as described in Section 5.2, however they differ in the way the parameters are extracted and quantized. The encoder extracts pitch information every 20 ms and performs voice/unvoiced decision on groups of harmonics. The magnitudes of the harmonics of the pitch frequency are either scalar or vector quantized depending on the number of harmonics and location of the harmonic. While IMBE coder provides cellular quality speech, AMBE vocoder actually achieves close to Cellular quality speech for certain types of filtered speech [29].

6.5.4 The United States Department of Defense 2.4 kbps MELP coder

With increased evidence of rapid advancements in speech coding technology, the US Department of Defense sought for a 2.4 kbps coder during the period of 1994-1996 whose performance would be subjectively equivalent that of the 4.8 kbps FS1016 coder. This resulted in the very recent selection of a 2.4 kbps Mixed Excitation Linear Predictive (MELP) coder [50] among other competing technologies. The structure of the MELP coder is similar to that described in Section 4 1 2 and Figure 9.

Here a frame size of 22.5 ms is used and the LPC coefficients are represented in LSF domain. A 25 bit multi-stage VQ is used to quantize the LSFs. The MSVQ uses joint optimization for both codebook design

and search, using M-best algorithm The 25 bit codebook consists of four stages of 7, 6, 6, and 6 bits, respectively. The gain is transmitted twice per frame of 22.5 ms, the gain for the first subframe is coded with 3 bits covering a small dynamic range based on neighboring subframe values. Gain for second subframe is coded using 5 bits using the full dynamic range of speech. Pitch and overall voicing is quantized using 7 bits per frame. This is true for both voiced and unvoiced speech.

For voiced speech, a Fourier analysis is performed on the LPC residual signal and the magnitudes of the first 10 harmonics are quantized using 8 bits. A bandpass voicing measure to reflect the frequency band over which the speech signal is estimated to be periodic is conveyed to the decoder using 4 bits. Finally a bit indicating the degree of periodicity is also transmitted to the remote decoder which then controls the amount of jitter in the synthesized speech signal.

For unvoiced speech, Fourier magnitudes, bandpass voicing measure and periodicity flag are not transmitted. Instead, the 13 bits are used to perform error control using Hamming codes.

Overall 53 bits are used to quantize the LPC and excitation parameters and with one additional bit for synchronization, 54 bits every 22.5 ms yields a rate of 2.4 kbps.

6.6 Intelligible Quality Speech Coding Standard

6.6.1 The United States Department of Defense 2.4 kb/s LPC-10e FS1015 Coder.

The Department of Defense (DoD) LPC-10 [51] FS1015 vocoder uses a 22.5 ms frame length for analysis and performs a modified covariance analysis to obtain the LPC parameters. It uses the two-state excitation model described in Section 4 and Figure 7. Pitch and voicing decisions are made using the average magnitude difference function algorithm mentioned in Section 4 and a voicing detector. Pitch and voicing decisions are smoothed using dynamic programming techniques which employs two frames of delay.

For voiced speech, the ten LPC coefficients are scalar quantized using a total of 41 bits. Pitch and voicing decisions are quantized using 7 bits and gain information is quantized using 5 bits.

For unvoiced speech, only 4 LPC coefficients are transmitted using 20 bits (5 bits per coefficient). Pitch and gain are quantized using 7 and 5 bits, respectively. Similar to the US DoD MELP coder described above, the unused 21 bits are used for error control during unvoiced speech

With the addition of a synchronization bit and a total of 54 bits per 22.5 ms, the LPC-10 FS1015 coder operates at 2400 bits/s.

7. Speech Coder Performance Assessment

In the previous section it was mentioned that two attributes, namely, ***speech quality*** and ***bit rate*** are predominantly used to characterize speech coder performance. Furthermore it was mentioned that the speech quality produced by a speech coder could be broadly categorized to different categories, namely, wireline or toll quality, cellular quality, communications quality, intelligible quality and synthetic quality. However, the nonlinear nature of low rate parametric coders have rendered analytical or objective methods questionable for classifying speech as wireline, cellular, communications, intelligible and synthetic quality under the variety of source and channel conditions over which the speech coder is to be assessed. As such, subjective tests as described in ITU-T P.800 series Recommendations have remained the only reliable alternative to conduct speech coder performance assessment. While ITU has also recently standardized an objective measurement tool (ITU-T Recommendation P.861) it has still not gained widespread usage to the extent of replacing a subjective test by such an objective tool; the primary reason being that the accuracy objective tool has been recognized to be technology dependent and hence renders itself useless to assess speech coder performance in a variety of conditions.

While the subjective assessment ought to be conducted with the objective of capturing all types of impairments anticipated in the system, the primary intent is to capture communication impairments, if any because of speech coding. In general, communications impairment factors can be divided into three types, depending on the affected direction of the communication link [54]. The first type includes impairments

that cause an increase in listening difficulty when the communications link is unidirectional and no assistance is given to the listener by the talker. The second type comprises impairments that cause difficulty while talking only. Finally, the third type includes impairments that cause difficulty while conversing, or factors associated with the alternation of the talking and listening roles of the participants.

Digital speech coding systems typically give rise to impairments of the first type, in view of the modeling distortions and quantization noise introduced by the encoding and decoding processes [55]. Consequently, listening tests are often used to evaluate the transmission performance of such systems. This will be discussed in further detail in Section 7.1 below. Telephone handsets (particularly with respect to the effect of sidetone), and loading coils with unbalanced 4-to-2 wire terminations without echo control can give rise to the second type of impairments, since the presence of echo and sidetone may increase the talking difficulty experienced by a participant in a telephone conversation [56]. Echo suppressors [57], on the other hand, are an example of devices that introduce impairments of the third type which cause difficulty in conversing, since these devices operate by disallowing fully bi-directional communication to occur simultaneously. Similarly, circuits with long propagation delay are associated with network configurations that introduce impairments of the third type because they alter the perceived dynamics of conversational communication.

7.1 Listener Opinion Tests

For listener opinion tests, the methods recommended by the ITU-T in ***Recommendation P.830*** is frequently employed. Typically naïve (untrained) listeners or subjects are invited to assess the quality of speech material (typically in the form of sentence pairs) passed through the speech coder under consideration. In generating suitable speech material, a set of phonetically balanced sentences uttered by a variety of talkers, both male and female, is normally required [58]. It is common to employ one set of recordings obtained using a microphone appropriate to the various systems under evaluation, and then use the same recordings for several experiments in which the same type of microphone would normally be employed.

According to P.800, listening-only methods can be classified into three groups: Absolute Category Rating (ACR), Degradation Category Rating (DCR), Comparison Category Rating (CCR). The first is an absolute rating method, while the other two are relative rating methods. For subjective tests that require better discrimination accuracy Paired Comparison Rating (PCR) are sometimes used. Among these, ACR and DCR tests are the most commonly employed testing methods.

The ACR test is characterized by a single-stimulus presentation to the subject where a sentence pair is played to a subject through headphones or telephone handsets and is requested to express his opinion on the quality of the speech material on an absolute five-point scale {Excellent, Good, Fair, Poor, Bad} Typically the entire set of phonetically balanced material from large number of talkers are passed through all speech coders under consideration. The performance of each speech coder is typically evaluated by first mapping the five-point scale to {5, 4, 3, 2, 1} and averaging the scores provided by various subjects across all talkers and sentence pairs to yield a Mean Opinion Score (MOS) for the speech coder under consideration.

The DCR test is characterized by a dual-stimulus presentation to the subject . Here the same sentence pair is presented twice (the first being unprocessed or reference sample and second one being processed or test sample) to the subject before he is requested to express his opinion on the relative degradation of processed speech compared to unprocessed speech on a relative five-point scale {Degradation is inaudible, Degradation is audible but not annoying, Degradation is slightly annoying, Degradation is annoying, Degradation is very annoying}. The unprocessed speech sample is essentially the input to the speech coder under consideration and the processed speech is the output of the speech coder under consideration. Similar to MOS of an ACR test, a Degradation MOS (DMOS) score is computed for DCR test.

CCR tests are similar to DCR tests in that it is a dual-stimulus test. However, the CCR method uses a bipolar seven-point scale where the subjects are requested to quantify their preference towards **test OR reference stimuli** While the reference speech sample in DCR tests is always the unprocessed source

signal, in CCR tests the reference may be either processed or unprocessed speech sample. In CCR the dual stimuli can be presented in any random order.

Finally, paired-comparison tests are a simpler case of the CCR method, where a binary scale is used when the subject is asked which sample is preferred (reference or test).

The selection of a suitable experimental approach, particularly the choice between absolute and relative or rank order designs, is very important and is influenced by the type of systems being evaluated, the overall test objective, and the number of total conditions to be assessed [55]. Generally, in the absence of background noise, if the speech coders and test conditions to be assessed result in speech that are degraded in an entirely different manner with respect to each other, then ACR tests are preferable, since they are absolute or single-stimulus by design (i.e., each sample is listened to and rated without a direct comparison with other reference samples). The DCR test is generally a more sensitive test compared to ACR test since minor degradations introduced by the speech coder can be penalized heavily by the subject which in the absence of a reference signal (which is true in real-world) would have gone unnoticed. Hence rating speech on an absolute scale is preferred.

However, for the evaluation of coded speech quality in the presence of acoustic (vehicular) noise, a DCR test is usually chosen. This selection is in accordance with currently revised CCITT procedures, as the distortion of high levels of background noise is believed to be more effectively measured with a dual-stimulus assessments. The primary reason here is that, if ACR were used under high levels of background noise conditions, then even the reference unprocessed noisy sample would not be rated as Excellent or Good and hence the dynamic range provided by the five-point scale is not utilized. Furthermore, most low bit rate speech coders are optimized to work well for speech types of signals and hence the presence of background noises which are not produced by human speech production mechanism can result in distorted output that is annoying to the human ear as compared to situations where low level background noise is

present Hence a reference speech sample that reflects he actual background noise at the input of the speech coder is desirable as provided by the DCR test.

CCR tests are very useful in comparing two systems that are close to each other. Another important application of CCR is the case where a speech coder performs noise cancellation due to which the processed speech sample may actually be perceived to sound better than an unprocessed noisy speech sample.

8. Concatenated Speech Coding

Future long distance, and especially international telephone calls will involve an increasing number of multi-link circuits of cellular, mobile satellite, private and public switched telephone network (PSTN) type of connections. Calls will thus be established over multi-link circuits employing different types of speech coding technologies operation at different bit rates. Since the very early 1980s one of the most unappreciated implications of integrating a variety of voice coding technologies into the network has been the associated reduction in end-to-end quality. It is worth noting that the Integrated Services Digital Network (ISDN) infrastructure was not designed to provide a switching capability at sub-64 kb/s transmission rates, thus in essence, leaving itself outpaced by today's modern coding technology. This means that interconnection of different voice coding technologies is typically only possible after conversion to 64 kb/s PCM. *The implication of this is, that unlike data transmission where throughput is limited by the capacity of the least transparent link, voice quality is reduced disproportionately and below the quality of the least performing link.*

The characteristics of interconnected voice links were recently a subject of investigation [59,60,61] resulting in considerable attention been given to the interconnection characteristics of new voice coding technology. These concerns have, of course been heightened by the wider mix of voice coding technology arising from a proliferation of proprietary, regional, and international standards; by the increasing wireless network access, promising to reach 50% by the end of the century; by the increasing use of wireless local loops in developing countries as a means to accelerate network deployment; and by the accelerating telecommunications network deregulation and privatization resulting in a larger number of network links with disparate technologies being encountered in end-to-end connections.

The proliferation of voice standards and their impact on transmission planning, can be better visualized by considering a few interconnectivity scenarios, and making some assumptions regarding the different types of voice technology that might be used in the network. Several foreseen network scenarios are presented in Figure 22 which involve an international link as part of a multi-link connection. Network configuration A represents a general case of a call initiated from a wireline user in a foreign country that is destined to a North American or European Digital Cellular user. The international link uses a DCME that employs 32 kbps ADPCM speech coding as described in Section 6.3.2. DCME equipments have been slowly migrating to the use of 16 kbps LD-CELP which is shown in configuration B. Configuration C is the case of a call from North American cellular user to a European cellular user. The results of subjective tests [61] using such interconnection indicated that voice quality in the link of Configuration C was degraded by more than 1.0 of a MOS compared to the MOS of weakest portion of the link, namely 13 kbps RPE-LTP. Since the results shown were obtained with an accuracy of 0.1 of an MOS at a 95% level of confidence, and since an MOS drop of more than 0.3 is typically associated with a "new class" of service quality, it can be seen that the impact of PSTN interconnections is to produce an end-to-end connection whose quality differs markedly from that where no such interconnections exist.

An even more interesting scenario is the case of a mobile satellite user calling in country A calling a cellular user local to the mobile satellite user; the Gateway for the mobile satellite system is situated in a country B; the international PSTN link between country A and B is similar to that in configuration C. For this local call voice undergoes a concatenation of three voice coding technologies.

The implication of these observations is that some reduction of the number of different voice coding technologies is likely to occur or, as a minimum, future introduction of voice coding into the network is likely to place some constraints into the use of technologies whose quality and transmission rates are dissimilar. In the past decade, interconnectivity of voice encoding technologies has been increasingly receiving a great deal of attention. Consequently, the development of low-rate coders which remain robust to such interconnected configurations will challenge researchers in the future, just like it is challenging the evolution of technology presently

9. Future Trends in Speech Coding

As evident from previous sections, over the last decade, a significant amount of research effort has been devoted towards better modeling of the human speech production system, better representations of parameters of such model, efficient quantizations of these representations and most importantly better representation of the LPC residual signal. It is very evident that CELP analysis-by-synthesis coders have enjoyed tremendous success in achieving better than communications quality speech at bit rates as low as 4.8 kbps. A good number of candidates for the 4 kbps ITU toll quality standardization efforts are also CELP based. More recently, however, non-CELP based coders have been playing an important role in achieving communications quality (or better) speech below 4 kbps. The 3.6 kbps AMBE coder (described in Section 6.5.3) and 2.4 kbps MELP coder (described in Section 6.5.4) are two classic examples of the trend. In addition, one emerging technology that is receiving considerable attention among many speech coding researchers that has shown significant promise in achieving high quality speech at low bit rates is the Prototype Waveform Interpolation (PWI) technique [62]. PWI uses a powerful model that transforms and decomposes a segment of speech signal into slowly and rapidly evolving waveforms and encodes them separately.

In recent years many researchers are focusing their work towards understanding and utilizing human speech perception models (including the decision making process, namely the human brain), and integrating these models towards development of better quality speech coders. While most of these efforts were initially directed towards obtaining high quality audio coders, these techniques show significant promise to be applicable to speech coders as well. It is the combination of source and auditory coding that perhaps holds the greatest promise in permitting high quality, very low bit rate speech coding to be realized (such as toll quality at 2 kb/s or below).

In summary advancement in four areas (without significantly increasing complexity) will hold the key to the success of speech coding technology in the future (i) A perceptually weighted filter in the analysis-by-synthesis loop of the encoder that better reflects the human speech perception mechanism as described in Section 4.2(ii) Quantization and coding of only those parameters that are important to the human ear based on masking properties of the human ear (iii) Post-filtering of reconstructed speech taking into account the loudness properties of human ear and (iv) providing robustness to the performance of speech coder in the presence of strong background noise for mobile applications

10. CONCLUSIONS

To a large extent voice technology has evolved significantly over the last decade. The ability to transmit voice at 8 kb/s with toll-quality was unthinkable only six years ago when 16 kb/s voice technology delivered this quality, which by itself was considered a breakthrough at the time. The introduction of 4.8 kb/s voice coding as a commercial service and the potential of introducing 2.4 kb/s voice coding for commercial satellite-based mobile services in the future are also remarkable.

However, these achievements did not come without engineering-costs; rather they have been achieved at the expense of higher-complexity, often running into 20–40 Millions of Instructions Per Second (MIPS) with today's fixed-point Digital Signal Processors (DSP).

11. REFERENCES

1. H Dudley, "Remaking Speech", Journal of Acoustical Society of America, vol. 11, No. 2, pp. 169-177, October 1939
2. J. Flanagan, Speech Analysis, Synthesis and Perception, New York, Springler Verlag, 1972
3. G Fant, Acoustic Theory of Speech Production, Gravenhage, The Netherlands, Mounton & Co., 1960
4 W Koemig, H. K. Dunn, and L. Y. Lacy, "The Sound Spectograph", Journal of Acoustical Society of America, vol 17, pp 19-49, July 1946
5 P E Papamichalis, Practical Approaches to Speech Coding, Englewood Cliffs, Prentice Hall, 1987
6. ITU-T Recommendation G.711, "Pulse Code Modulation (PCM) for Voice Frequencies" Red Book, Malaga-Torremolinos, 1984
7. ITU-T Recommendation G.726. "40-, 32-, 24-, and 16-kb/s Adaptive Differential Pulse Code Modulation.", Blue Book, Geneva 1990
8. ITU-T Recommendation G.727. "5-, 4-, 3-, and 2-bits per sample Embedded Adaptive Differential Pulse Code Modulation.", Blue Book, Geneva 1991
9 L R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Englewood Cliffs, Prentice Hall, 1975
10. W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud and V. Cuperman, "Efficient Search and Design Procedures for Robust Multi-Stage VQ of LPC parameters for 4 kbps Speech Coding", IEEE Transactions on Speech and Audio Processing, vol 1, No. 4, pp 373-385, October 1993
11. P. Kabal and R. Ramachandran, "The Computation of Line Spectral Frequencies Using Chebychev Polynomials", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-34, no. 6, pp. 1419-1426, December 1986
12. K. Paliwal and B. Atal, "Efficient vector Quantization of LPC parameters at 24 bits/frame," proceedings of ICASSP, pp. 661-663, 1991

13. C S Ravishankar, B. R. U. Bhaskar, and S. Dimolitsas, "A 1200 bps Voice Coder Based Upon Split VQ of Line Spectral Frequencies", Proceedings of 1993 IEEE Speech Coding Workshop, St. Adele, pp. 37-38, October 1993

14. R. Schafer and J. Markel, Speech Analysis, IEEE Press, New York, 1979

15. W. Hess, Pitch Determination of Speech Signal, Springer Verlag, New York, 1983

16. C. K. Un and D. T. Magill, "The Residual-Excited Linear Prediction Vocoder with Transmission Below 9.6 Kb/s", *IEEE Transactions on Communications,* vol. COM-23, December 1995

17. J. Makhoul et. al., "A Mixed Source Model for Speech Compression and Synthesis," Journal of Acoustic Society of America, vol. 64, pp. 1577-1581, December 1978

18. A. McCree and T. Barnwell III, "A New Mixed Excitation LPC Vocoder, Proceedings of ICASSP, pp. 593-596, 1991

19. V. Viswanathan, M. Berouti, A Higgins and R. Russell, "A Harmonic Deviations Linear Predictive Vocoder for Improved Narrowband Speeech Transmission", Proceedings of ICASSP, 1982

20. C. S. Ravishankar, B. R. U. Bhaskar, and S. Dimolitsas, "A 1200 bps Voice Coder Based Upon Alternate Transmission of LPC and Residual Information", Proceedings of 1995 IEEE Speech Coding Workshop, Annapolis, pp. 111-112, September 1995

21. B. S Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-24, No. 3, pp. 201-212, June 1976

22. J. P. Campbell, and, T. E. Tremain, "Voiced/Unvoice Classification of Speech with Applications to U. S. Government LPC-10E Algorithm" ICASSP, 1986, Tokyo, pp. 472-476

23. M. Yong and A. Gersho, "Vector Excitation Coding with Dynamic Bit Allocation", Proceedings of Globecom, pp. 290-294, 1988

24. J. H. Chen and R. Cox, "The Creation and Evolution of 16 kbps LD-CELP : From Concept to Standard", Speech Communiation, 1993, pp. 103-111, North Holland, Elsiever Science Publishers

25. R. V. Cox et. al., "New Directions in Sub-Band Coding", IEEE Transactions in Selected Areas in Communications, vol.6, No. 2, pp. 391-409, February 1988

26 R Crochiere and L Rabiner, Multirate Digital Signal Processing, Englewood Cliffs, Prentice Hall, 1983

27. P P Vaidyanathan, "Quadrature Mirror Filter Banks for M-Band Extensions and Perfect Reconstruction Techniques", ASSP Magazine, vol. 4, No. 3, July 1987

28. ITU-T Recommendation G.722, 7 kHz Audio Coding within 64 kbps, Blue Book, Melbourne, 1988

29 S Dimolitsas, F L. Corcoran, C. Ravishankar, R. S. Skaland, and A. Wong. "Evaluation of Voice Codec Performance for the Inmarsat mini-M System" Proceedings, 10th *International Digital Satellite Conference,* Brighton, England, May 1995

30 S. Dimolitsas, F.L. Corcoran, and C. Ravishankar, "Voice Transmission Quality of Mobile Satellite Communications Systems," *International Journal of Satellite Communications*, Vol **12**, No. 4, pp 361-368, July-August 1994.

31. R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," IEEE Transactions on Acoustic Speech and Signal Processing, vol. ASSP-34, pp. 744. Aug 1986

32 R Zelinski and P. Noll, "Adaptive Transform Coding Speech Signals, IEEE Transactions on Acoustic Speech and Signal Processing, vol. ASSP-25, 1977

33 J. B. Allen and S T Neely, "Micromechanical Models of the Cochlea", Physics Today, pp. 40-47, July 1992

34 B. Atal and J. Remde, "A New Model for LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates", Proceedings of ICASSP, pp. 614-617, April 1982

35. P Kroon, E. F. Deprettere and R. J. Sluyter, "Regular-Pulse Excitation - A Novel Approach to Effective and Efficient Multipulse Coding of Speech", *IEEE Transactions on ASSP,* vol. ASSP-34, pp. 1054 - 1063, October 1986

36 M. R. Schroeder and B. S. Atal, "Code Excited Linear Prediction (CELP) : High quality Speech at Very Low Bit Rates", *Proceedings of International Conference on Acoustics, Speech and Signal Processing,* pp 937-940, March 1985

37. S. Miki, K. Mano, H. Ohmuro, and, T. Moriya, "Pitch Synchronous Innovation CELP (PSI-CELP)," *Proceedings of European Conference on Speech and Communication Technology*, pp. 261-264, September 1993

38. R. C. Rose and T. P. Barnwell, "The Self Excited Vocoder - an Alternate Approach to Toll-Quality Speech at 4800 bits/s", *Proceedings of International Conference on Acoustics, Speech and Signal Processing,* Tokyo, Japan, pp. 453-456, 1986

39. C. S. Ravishankar and S. Dimolitsas, "Voice Coding Technology for Digital Aeronautical Communications", Air Traffic Control Quarterly, Vol. 4(3), 197-221, 1997

40. S. Dimolitsas, F. L. Corcoran and C. Ravishankar. "Correlation Between Headphone and Telephone-Handset Listener Opinion Scores for Single Stimulus Voice Coder Assessments." *IEEE Letters on Signal Processing,* 1995.

41. ITU-T Recommendation G.763 "Digital Circuit Multiplication Equipment Using 32 kb/s ADPCM and Digital Speech Interpolation", Geneva, 1991

42. Y. Linde, A. Buzo, and A. Gray, "An algorithm for Vector Quantizer Design", IEEE Transations on Communications, vol. COMM 28, pp. 84-95, Jan 1980

43. J. R. B. De Marca and N. S. Jayant, "An Algorithm for Assigning Binary Indices to the Codevectors of a Multi-Dimensional Quantizer", *Proceedings of International Conference on Communications,* Seattle, Washington, pp.1128-1132, June 1987

44. R. Salami et. al., "Description of the Proposed ITU-T 8 kb/s Speech Coding Standard," Proceedings of 1995 IEEE Speech Coding Workshop, Annapolis, Maryland, pp. 3-5, September 1995

45. Spiros Dimolitsas, C. S. Ravishankar, and Gerhard Schröder, "Current Objectives for 4 kbit/s Wireline-Quality Speech Coding Standardization". *IEEE Letters on Signal Processing,* Vol. 1, No. 11, November 1994

46. I. Gerson and M. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8 Kb/s", *Proceedings of International Conference on Acoustics, Speech and Signal Processing,* Albuquerque, New Mexico, pp. 461-464, 1990

47 K. Hellwig, P. Vary, D Massaloux, J. P. Petit, C. Galand and M. Rosso, "Speech Codec for the European Mobile Radio System", *Proceedings of GLOBECOM,* Dallas, Texas, pp. 1065-1069, 1989

48. J P Campbell, Jr., T. E. Tremain, and V. C. Welch, "The DoD 4.8 kb/s Standard (The Proposed Federal Standard FS1016)", *in Advances in Speech Coding, B S Atal, V. Cuperman, and A. Gersho, Editors*, Norwell, MA : Kluwer, 1991, pp. 121-133

49. D. W. Griffin and J. S. Lim, "A New Model Based Speech Analysis/Synthesis System", *Proceedings of International Conference on Acoustics, Speech and Signal Processing,* pp. 513-516, 1985

50 A McCree, K. Truong, E B. George, T. P. Barnwell, and, V. Viswanathan, "A 2.4 kbps MELP Coder Candidate for the New U. S. Federal Standard", Proceedings of ICASSP, pp. 200-203, 1996

51. T. E. Tremain, "The Government Standard Linear Predictive Coding Algorithm : LPC-10", *Speech Technology*, pp. 40-49, April 1982

52. ITU-T Recommendation P.800, Methods of Subjective Determination of Transmission Quality, 1996

53. ITU-T Recommendation P.861, Objective Quality Measurement of Telephone Band (300 - 3400 Hz) Speecs Codecs, 1996

54. D. L. Richards, Telecomminications by Speech, New York : John Wiley, 1973

55. S. Dimolitsas, "Subjective Assessment Methods for the Measurement of Digital Speech Coder Quality," in Speech and Audio Coding for Wireless Applications, edited by B. S. Atal, V. Cuperman and A. Gersho, Kluwer Academic Publishers, 1992

56. ITU-T Recommendation G.131, Stability and Echo, Red Book, vol. III.1, pp. 183-194, Malaga Torremonilos, 1984

57. ITU-T Recommendation G.164, Echo Suppressors, Red Book, vol. III.1, pp. 225-258, Malaga Torremonilos, 1984

58. IEEE Recommended Practice for Speech Quality Measurements, IEEE Transactions on Audio and ElectroAcoustics, vol. AU-17, No. 3, pp. 225-246, September 1969

59. S. Dimolitsas, F. L. Corcoran, and M. Baraniecki, "Transmission Quality of North-American Cellular, Personal Communications, and Public Switched Telephone Networks," *IEEE Transactions on Vehicular Technology*, Vol. **32**, No. 2, pp. 245–251, May, 1994

60. S. Dimolitsas, F. L. Corcoran, and C. Ravishankar, "Voice Quality of Interconnected PCS, Japanese Cellular, and Public Switched Telephone Networks." Proceedings, *IEEE Intl Conference on Acoustics, Speech and Signal Proc*, International Conference on Acoustics, Speech and Signal Processing'95, May 1995, Detroit, MI

61. S. Dimolitsas, F. L. Corcoran, and C. Ravishankar, "Voice Quality of Interconnected North American Cellular, European Cellular, and Public Switched Telephone Networks". Proceedings, *IEEE Vehicular Technology Conference,* VTC'95, Chicago, IL, July 1995

62. W. B. Kleijn, "Encoding Speech using Prototype Waveforms", *IEEE Transactions on Speech and Audio Processing*, pp. 386-399, October 1993

FILENAME.APP = 6709FL00.DOC

Figure 1a. Illustration of a continuous time continuous amplitude speech signal of duration of about 60 ms

Figure 1b. Discrete time (sampled) continuous amplitude version of the speech waveform in Figure 1a; Sampling Rate = 8000 samples/sec; only first 64 samples shown for clarity

Figure 1c. Illustration of 3 bit quantization scheme

Figure 1d. Discrete time, Discrete amplitude (sampled and quantized) version of samples in Figure 1b using 8 bits (256 levels)

Figure 1e Reconstructed speech signal after 8 bit quantization

Figure 1f. Discrete time, Discrete amplitude (sampled and quantized) version of samples in Figure 1b using 3 bits (8 levels)

Figure 1g. Reconstructed speech signal after 3 bit quantization

Figure 2 Homer Dudley's Vocoder Apparatus built in 1939 using analog electronic circuits

Figure 3 Schematic of the human speech production system

Figure 4 Block diagram of a PCM codec

Figure 5 Block diagram of a typical ADPCM codec

Figure 6 A simplified block diagram of an ADM coder; note the 1 bit quantizer

Figure 7 Two-stage excitation model which can produce intelligible quality speech

Figure 8 Residual Excited Linear Prediction (RELP) coder

Figure 9 Mixed Excitation Linear Prediction (MELP) coder

Figure 10 The ideal analysis-by-synthesis coding method

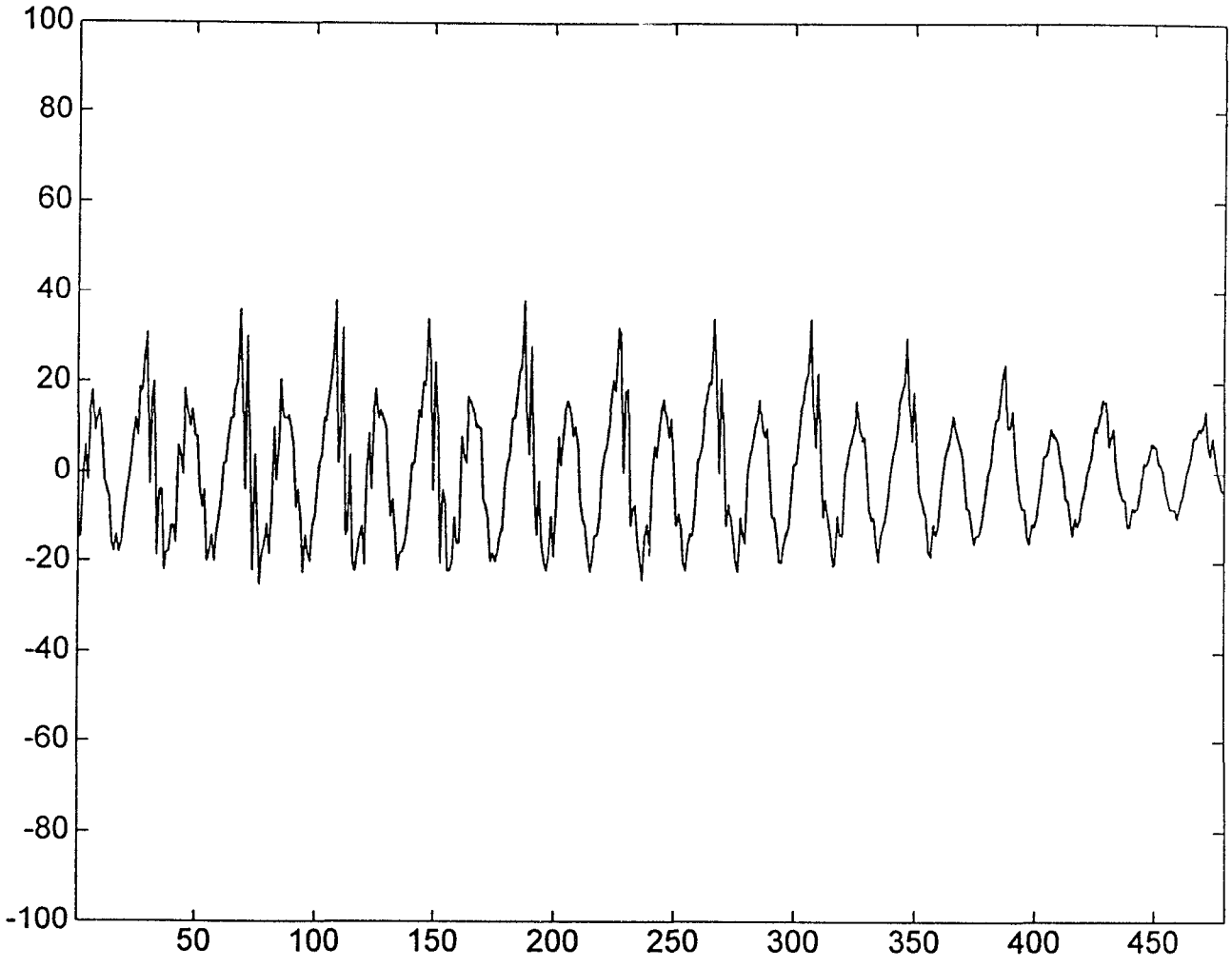
Figure 11 A practical analysis-by-synthesis coding method

Figure 12 Code Excited Linear Predictive (CELP) coder

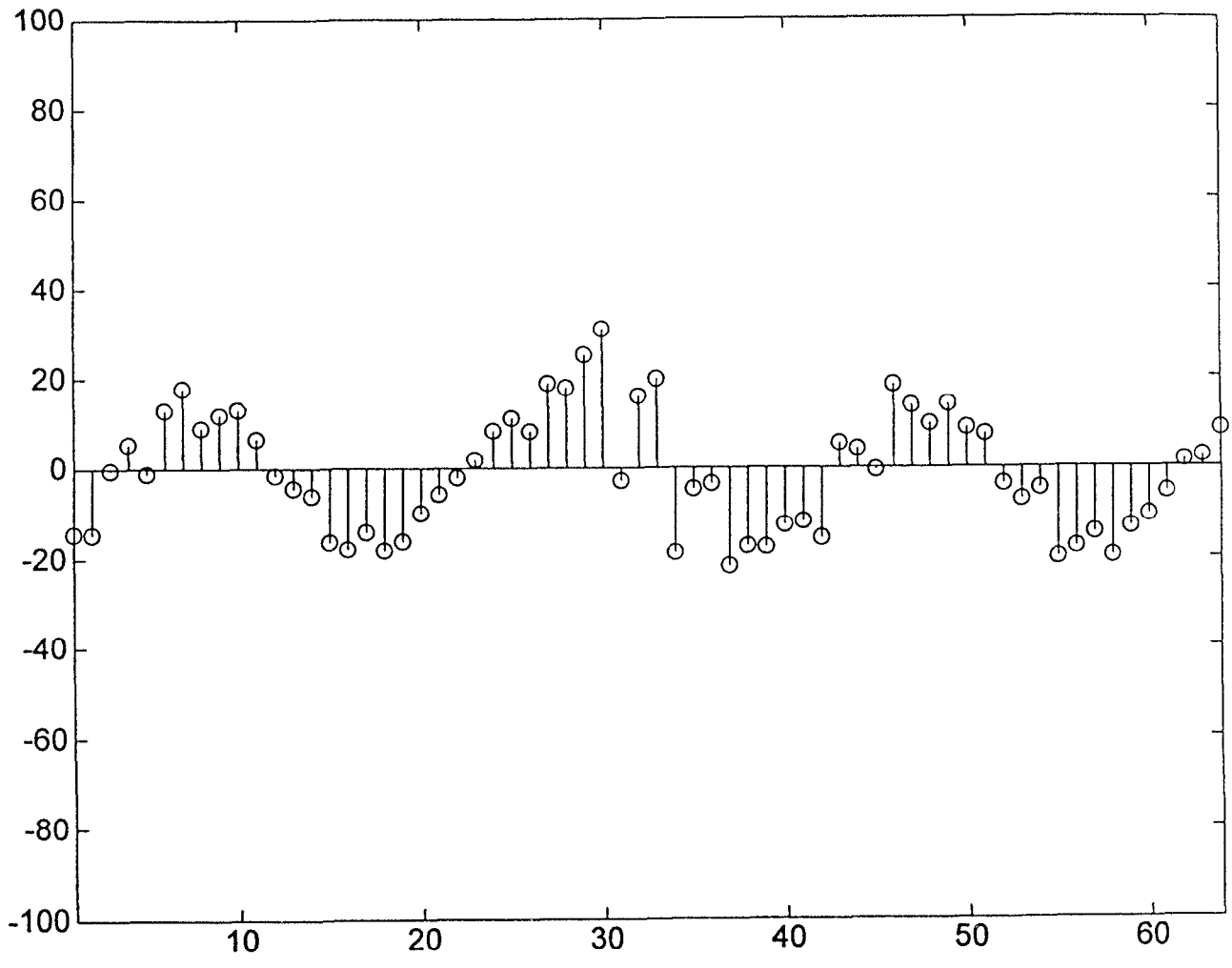
Figure 13 Block Diagram of a sub-band coder

Figure 14 Block Diagram of a Multi-Band Excitation Coder

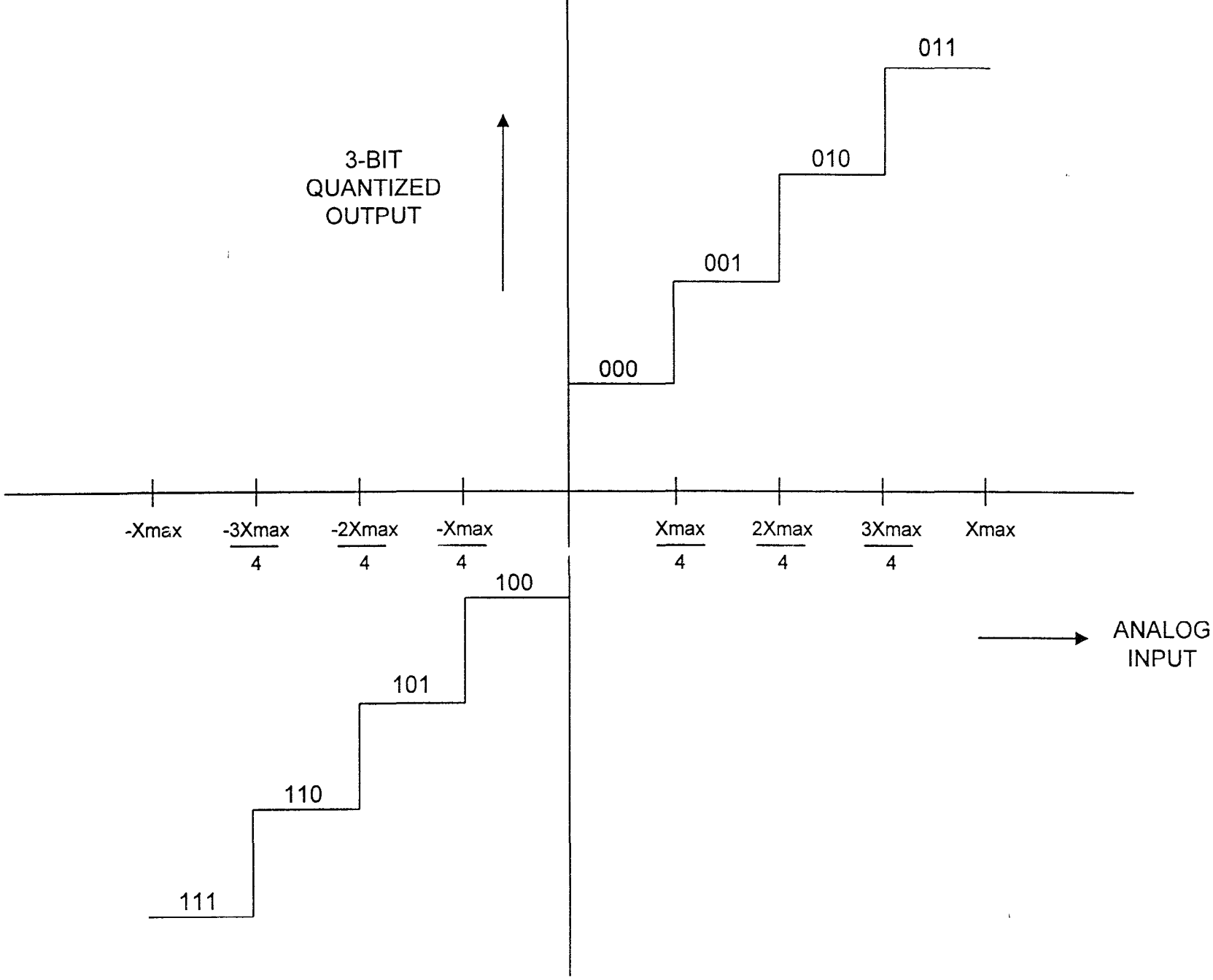
Figure 15 Block Diagram of Sinusoidal Transform Coder



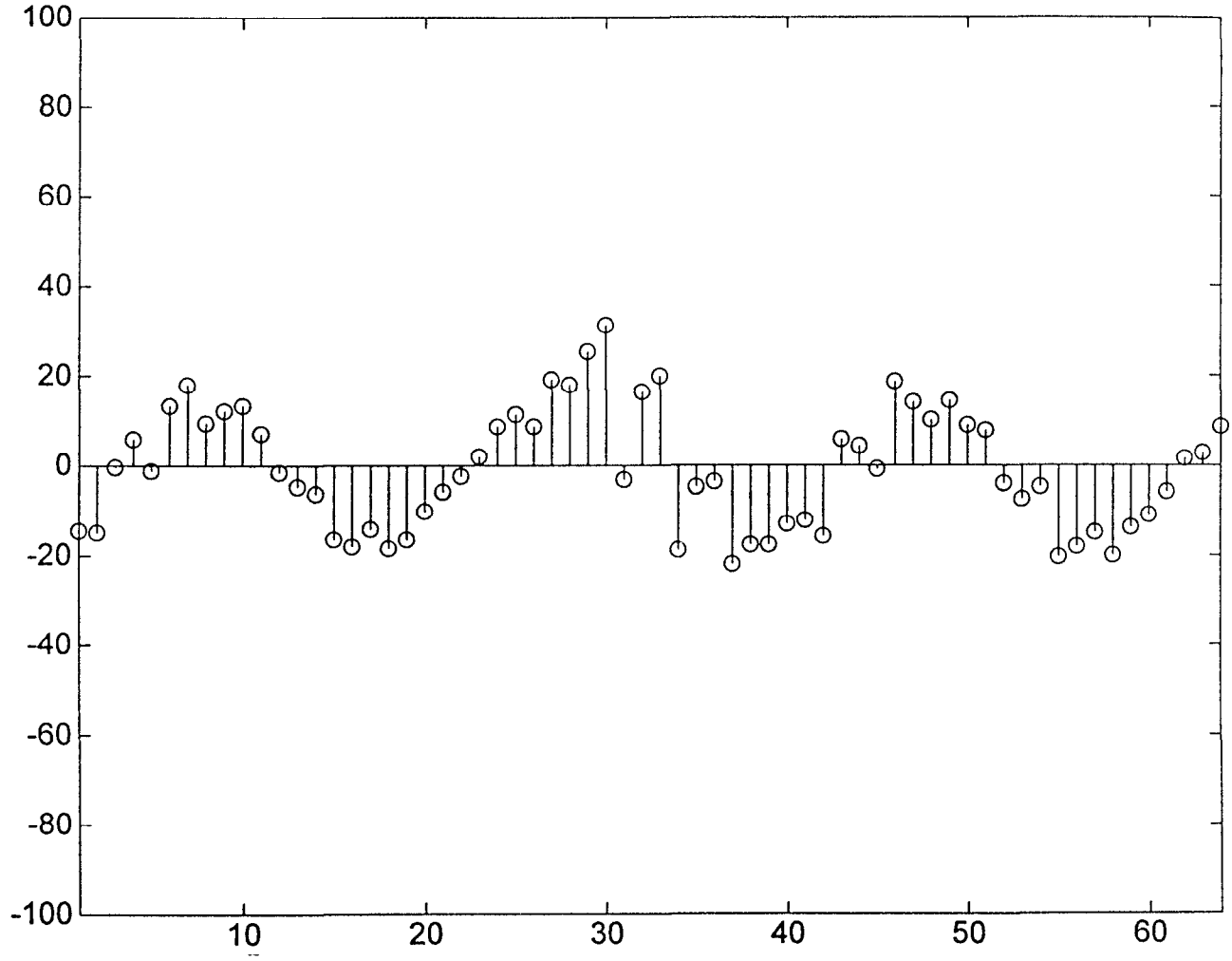
1(a)
6709fg01.VSD



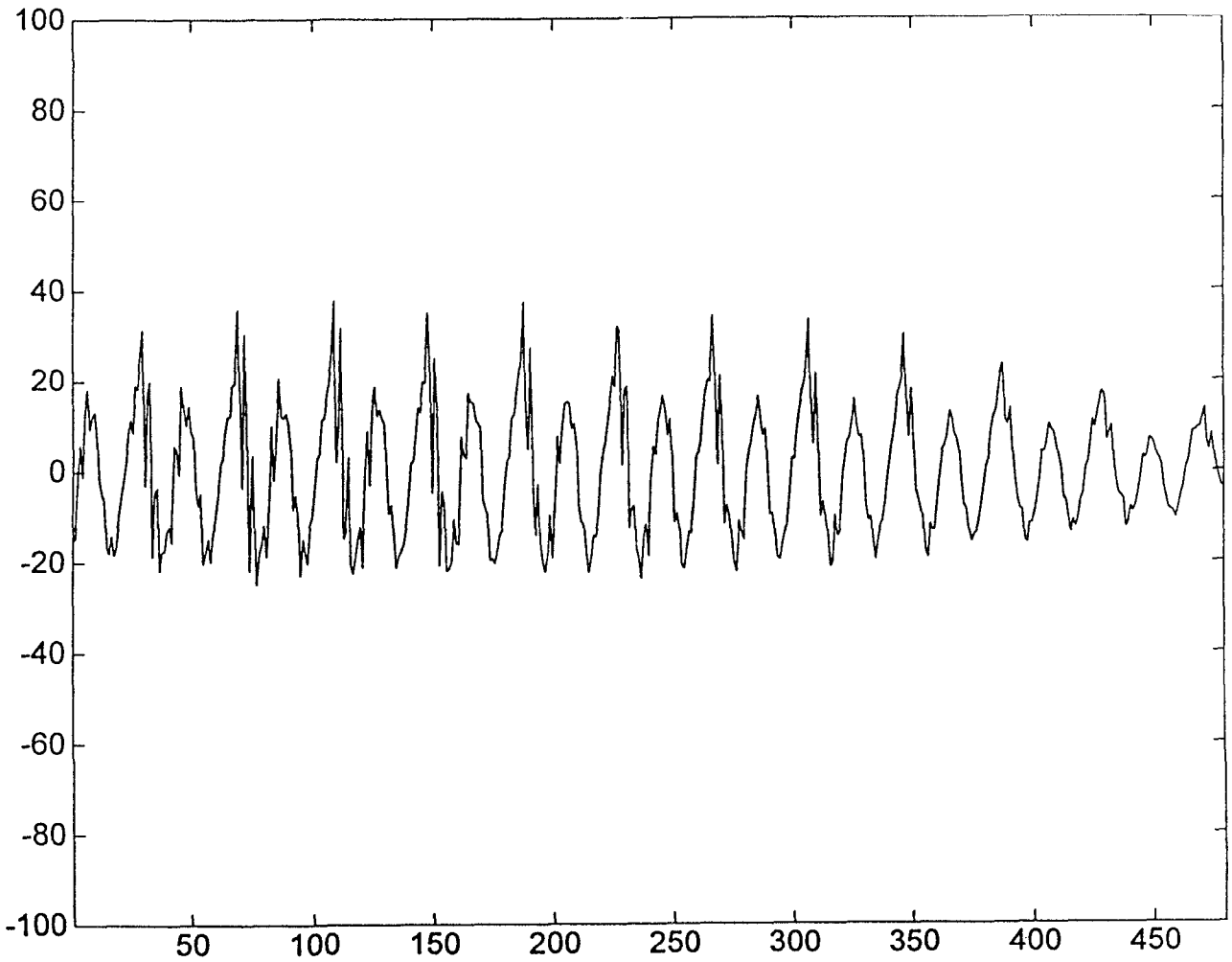
1(b)
6709fg01.VSD



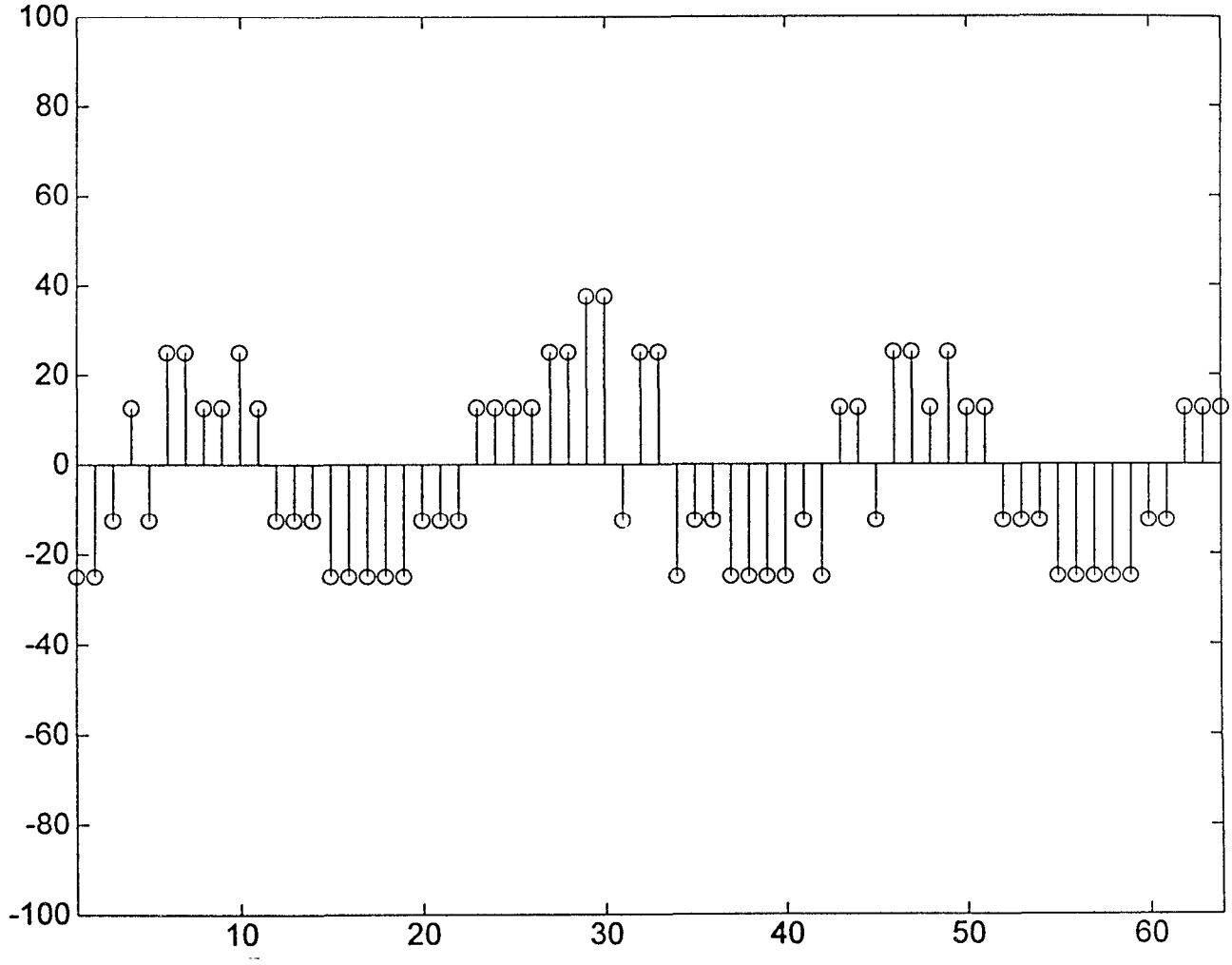
1(c)
6709FG01.VSD



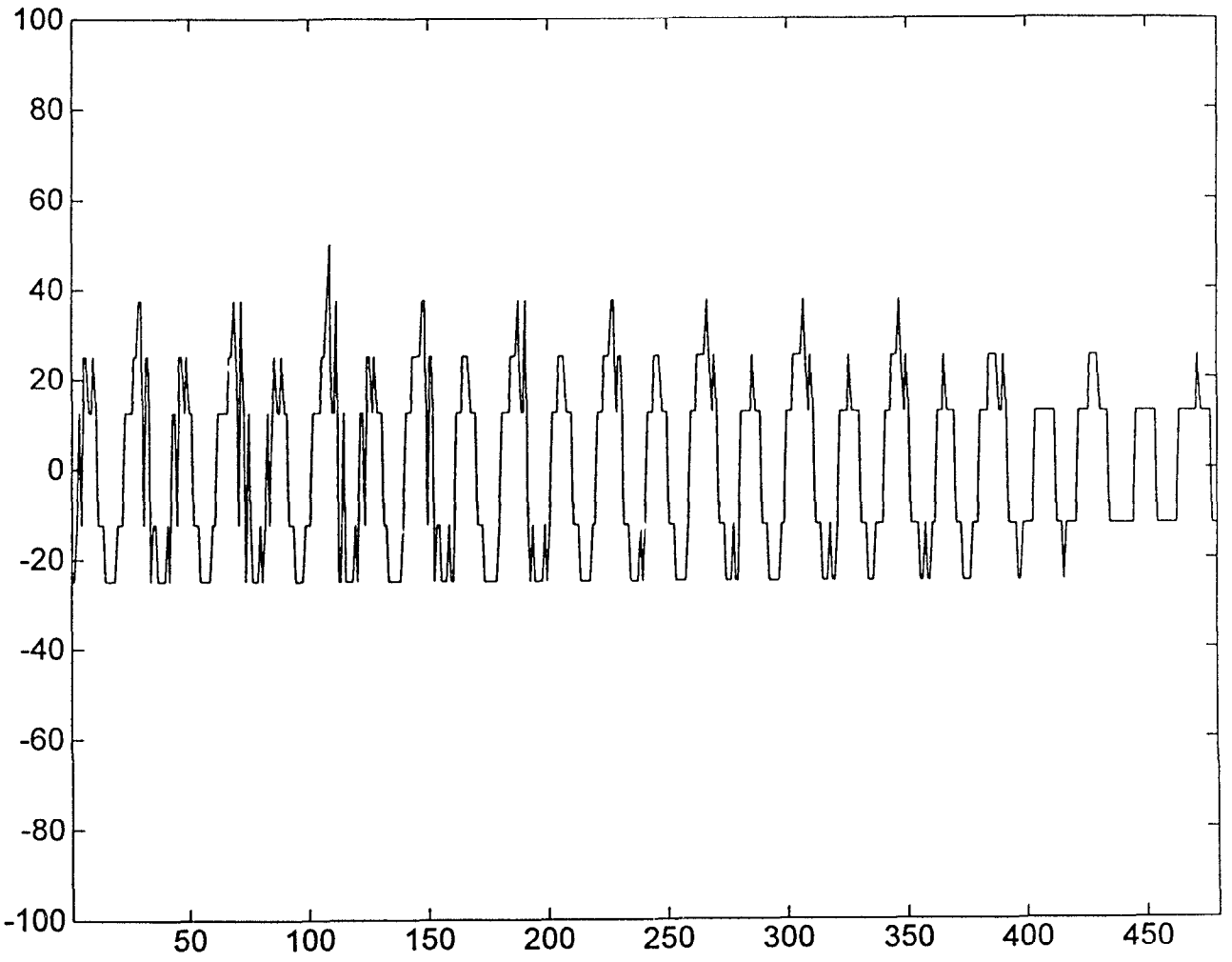
1(d)
6709fg01.VSD



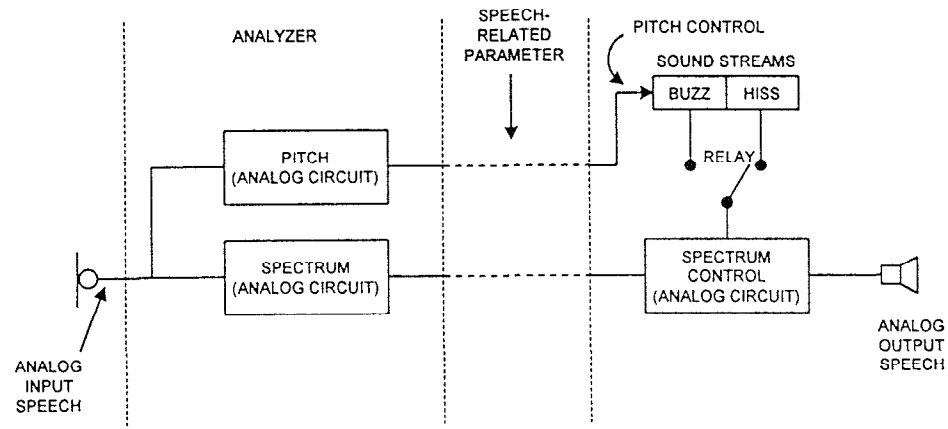
1(e)
6709fg01.VSD



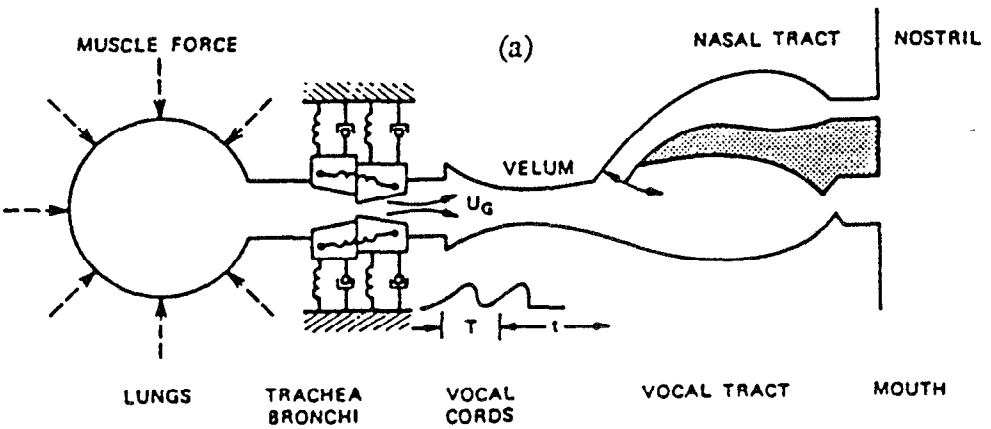
1(f)
6709fg01.VSD



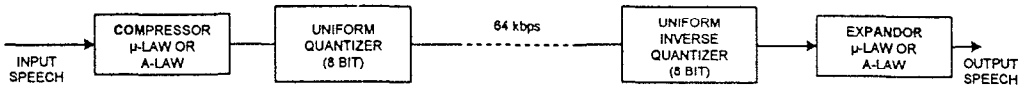
1(g)
6709fg01.VSD



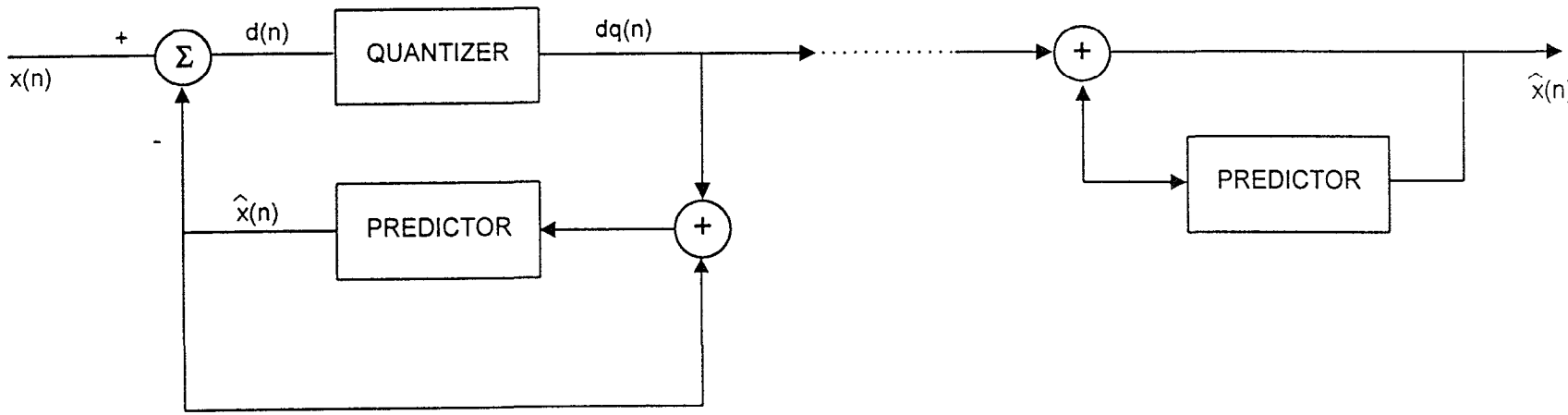
6709FG02.VSD



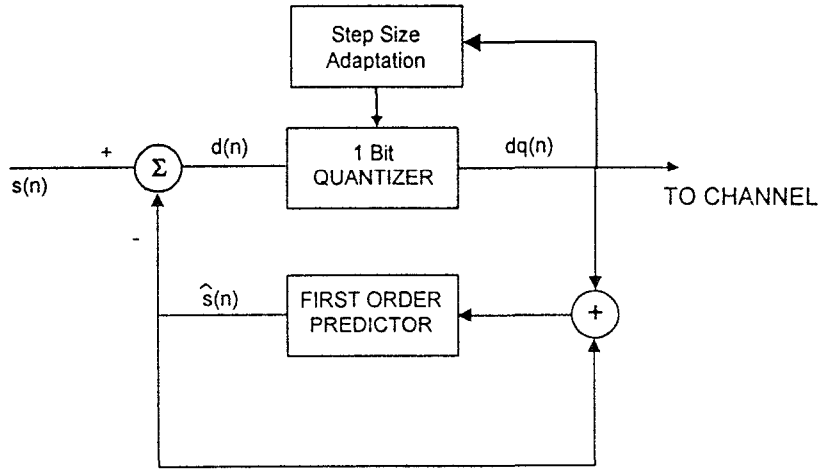
6709fg03.VSD



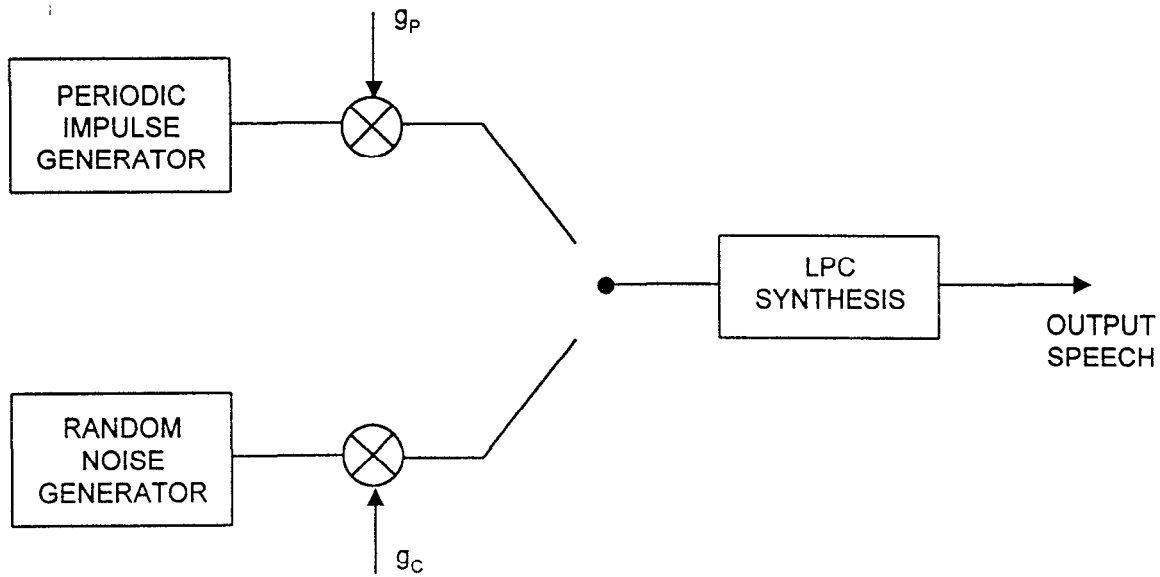
6709fg04.VSD



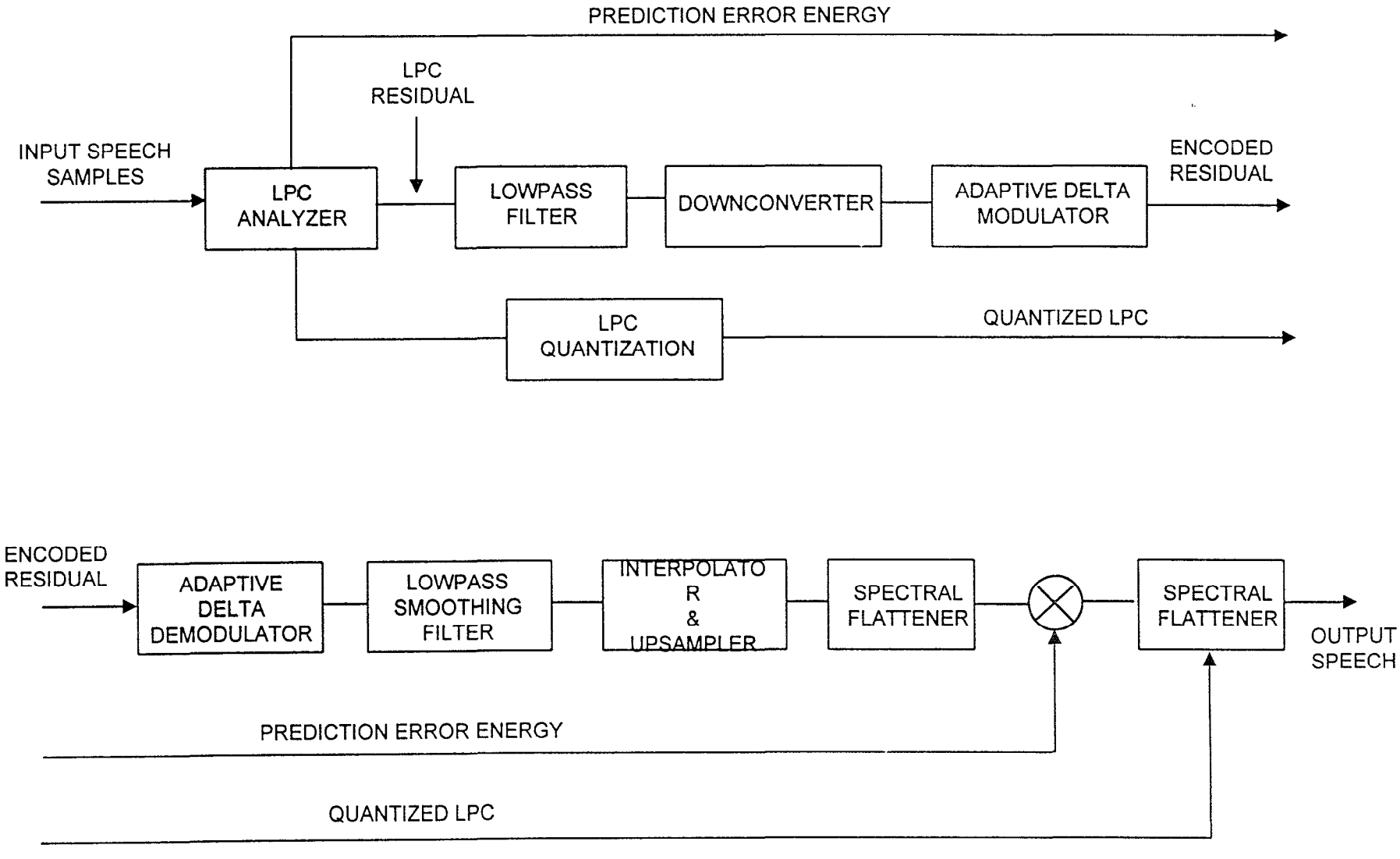
6709fg05.VSD



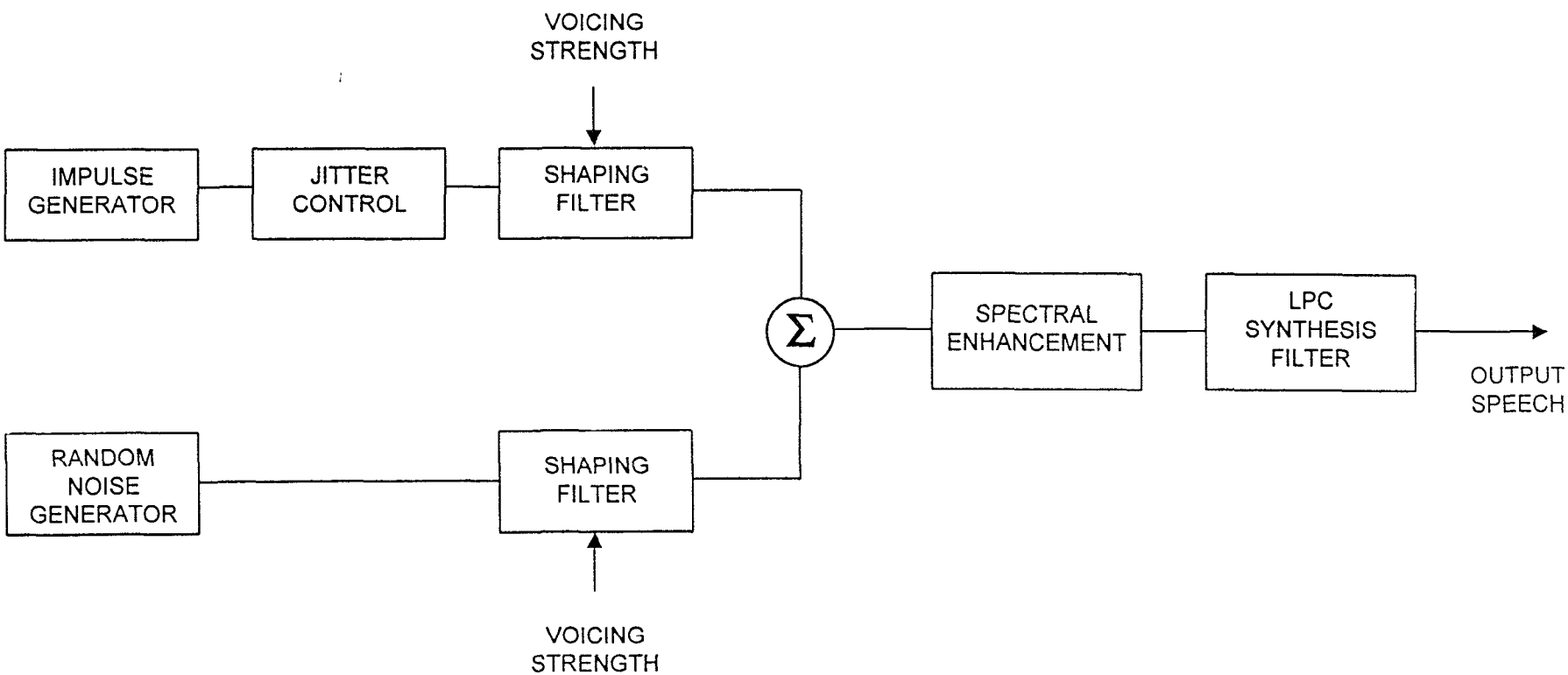
6709fg06.VSD



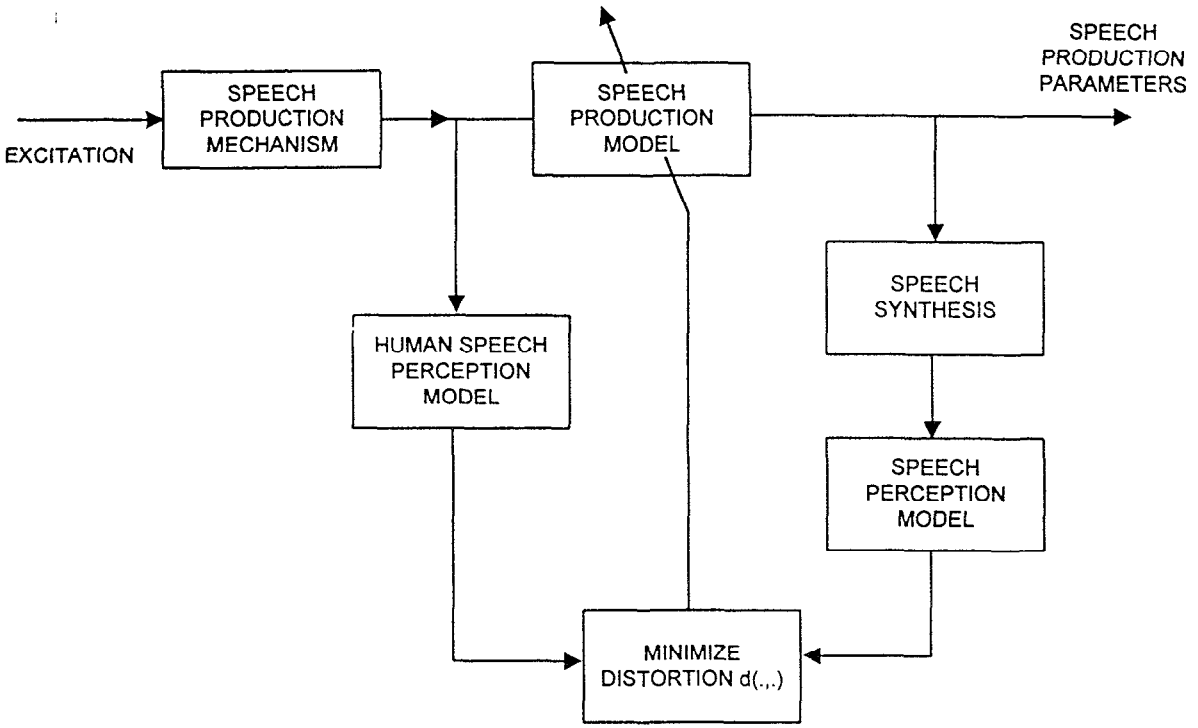
6709fg07.VSD



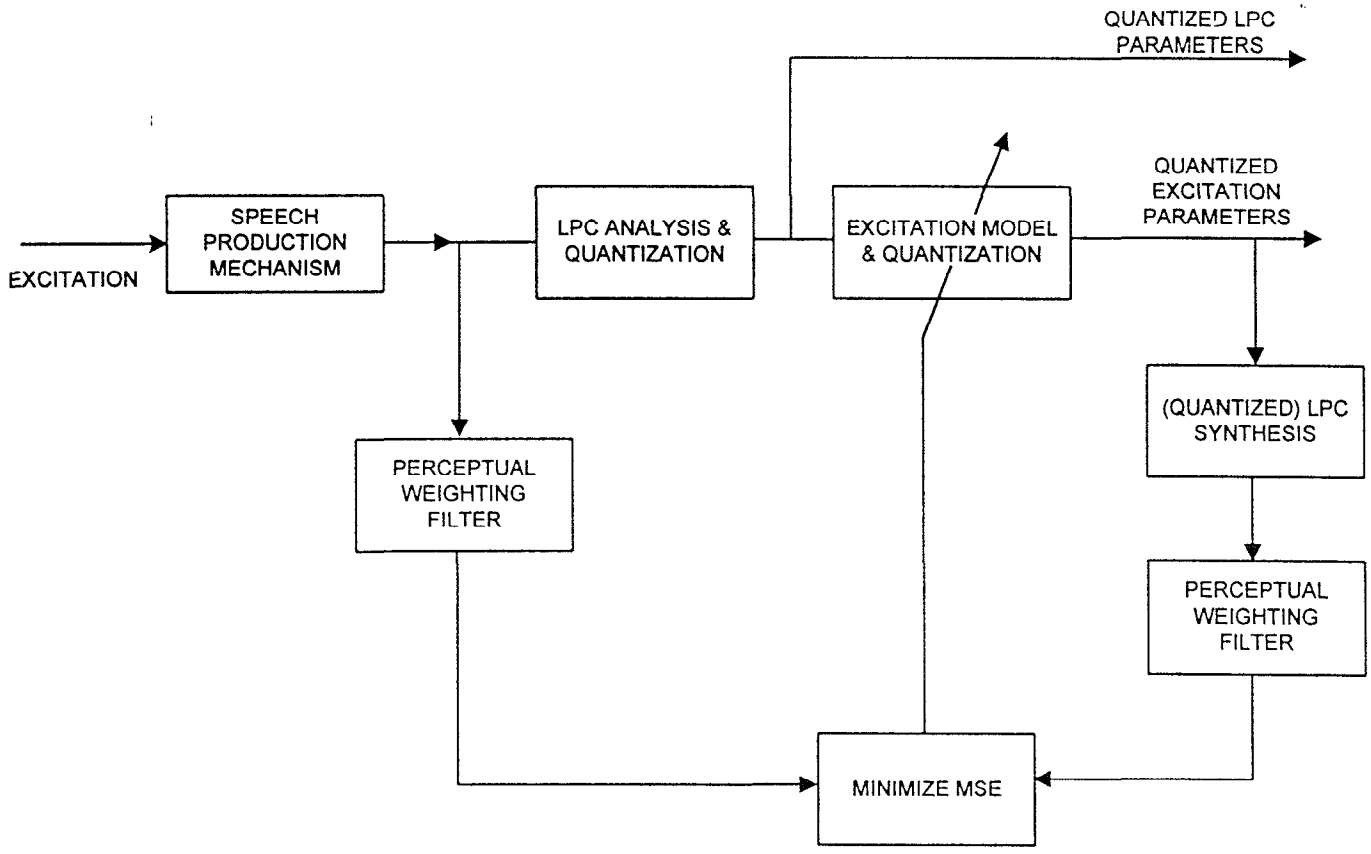
6709fg08.VSD



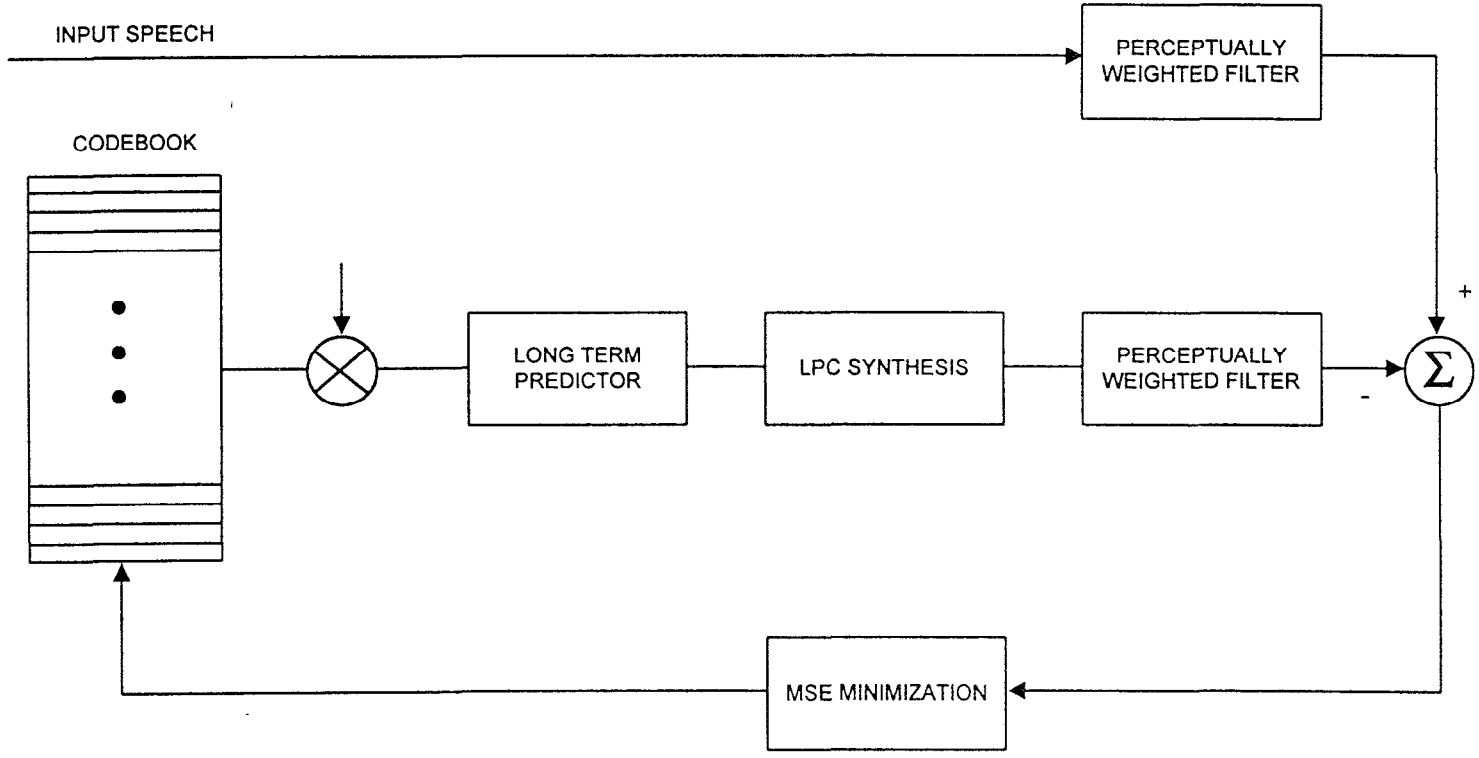
6709fg09.VSD



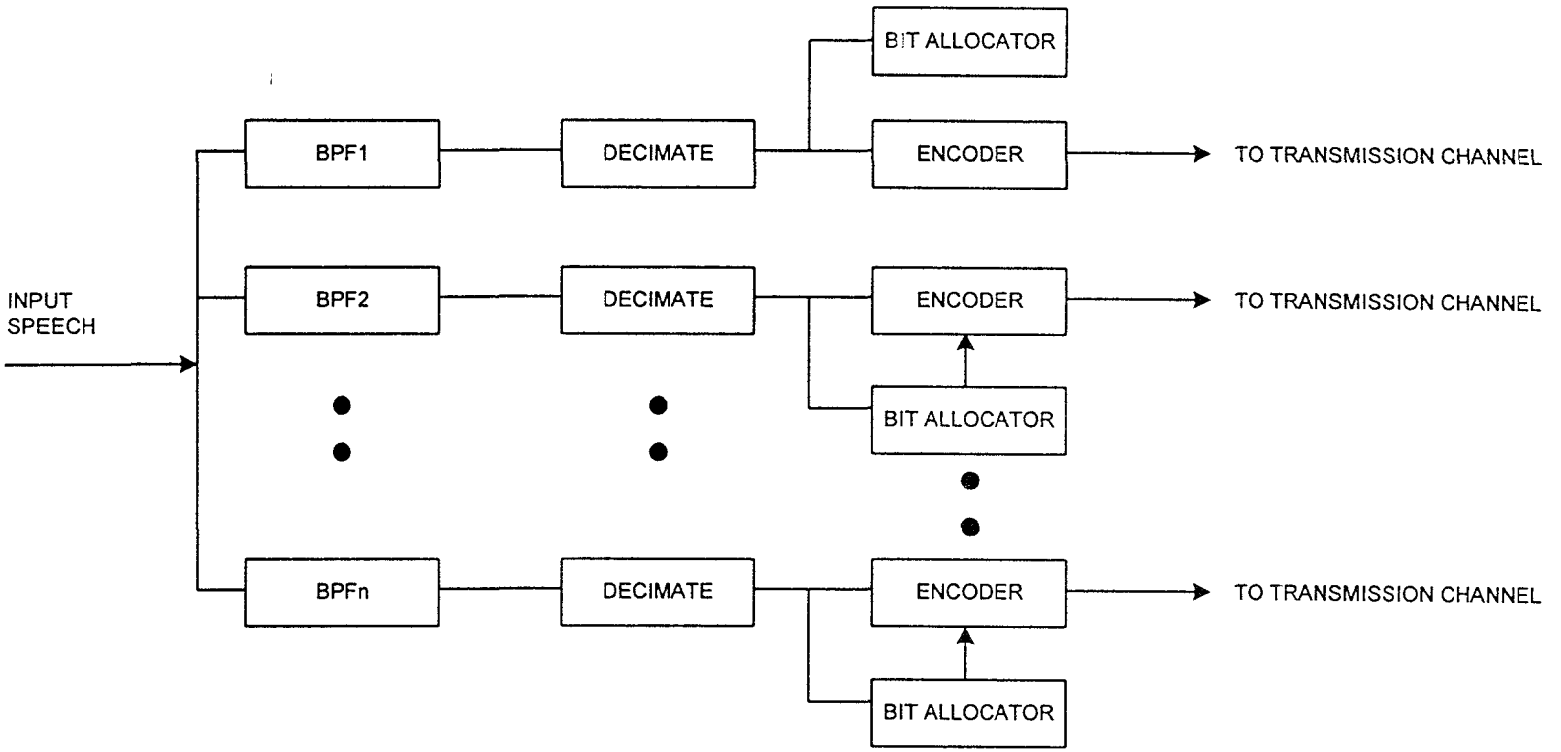
6709fg10.VSD



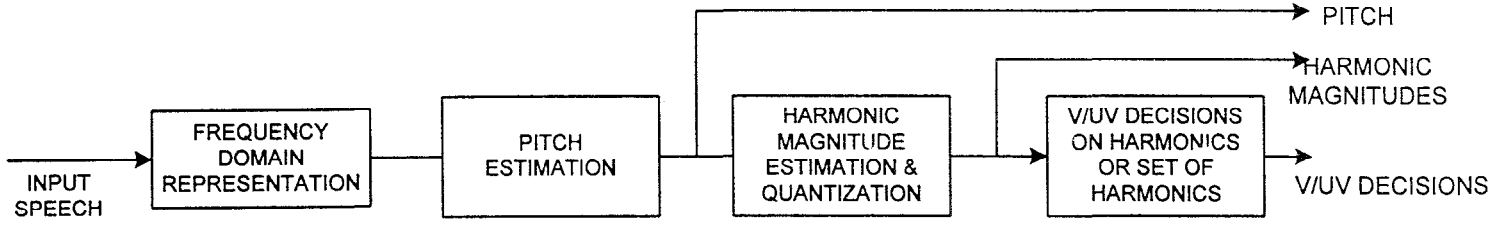
6709fg11.VSD



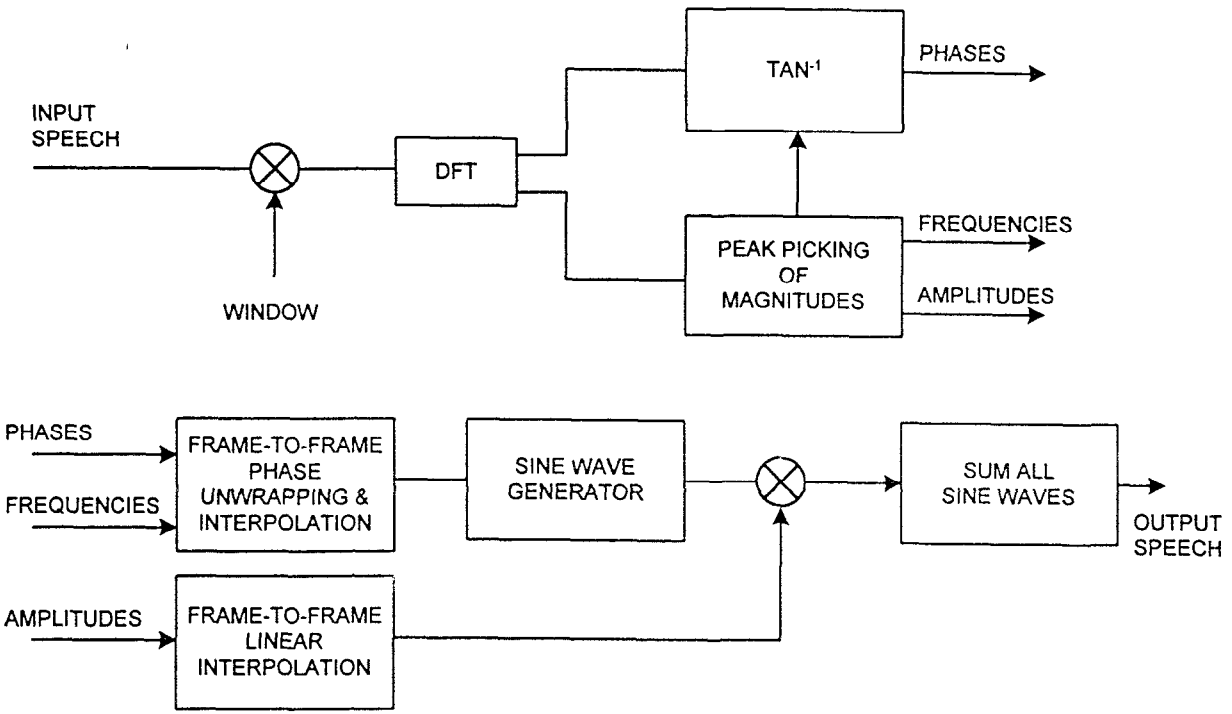
6709fg12.VSD



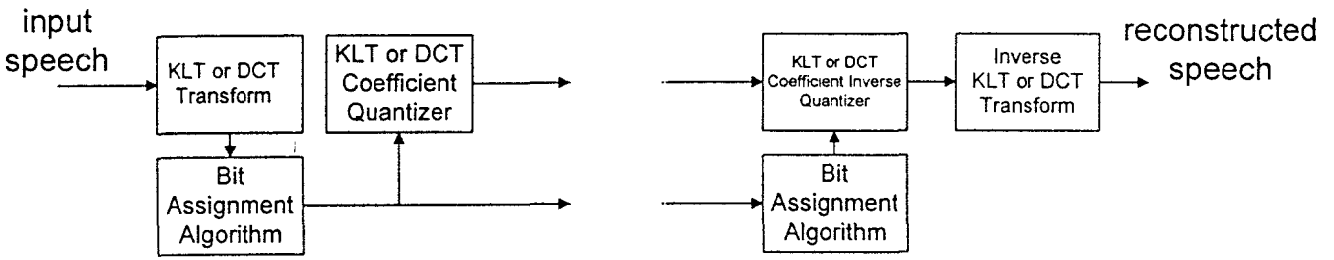
6709fg13.VSD



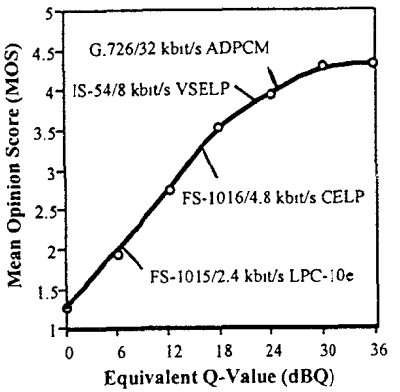
6709fg14.VSD



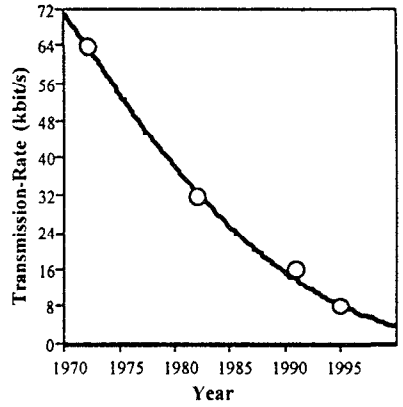
6709fg15.VSD



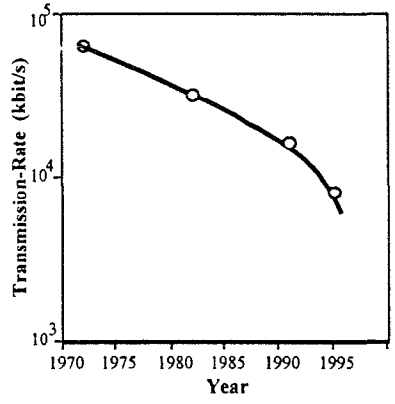
6709fg16.VSD



6709fg17.VSD

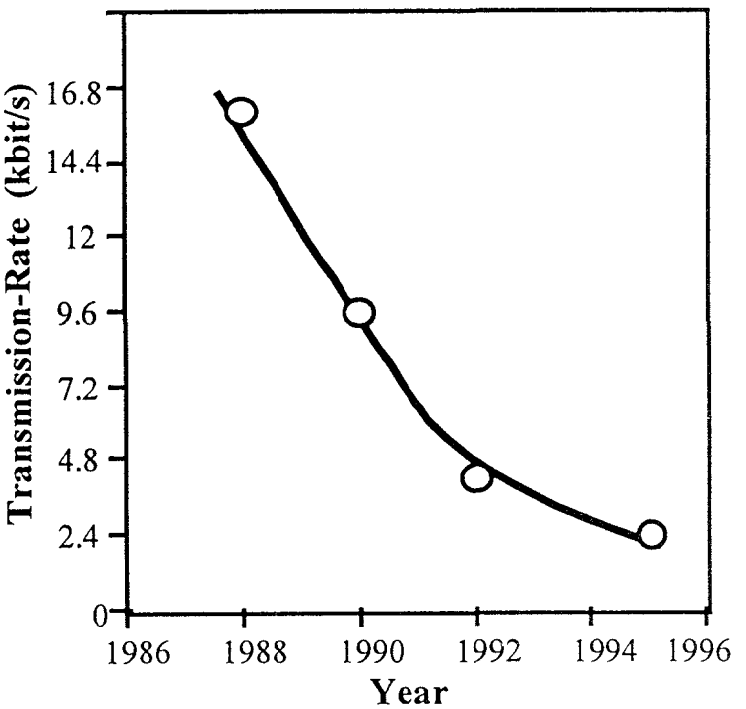


18(a)

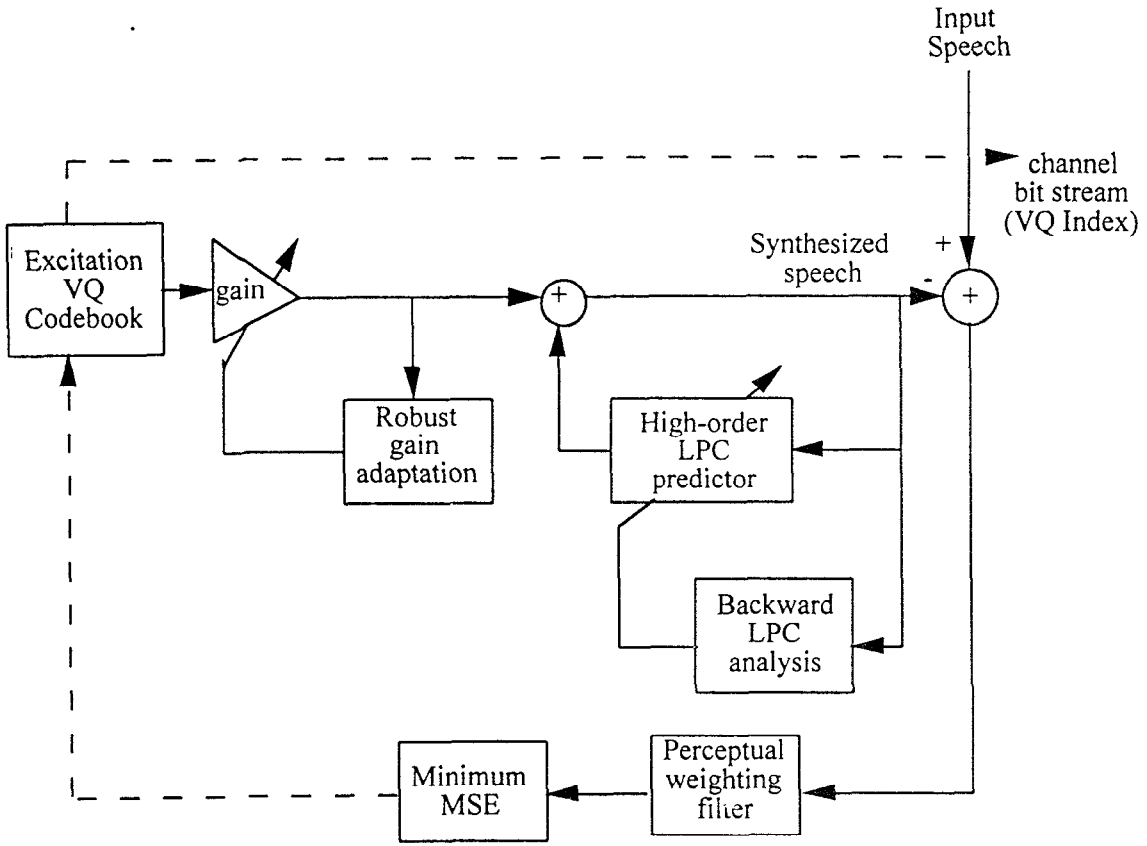


18(b)

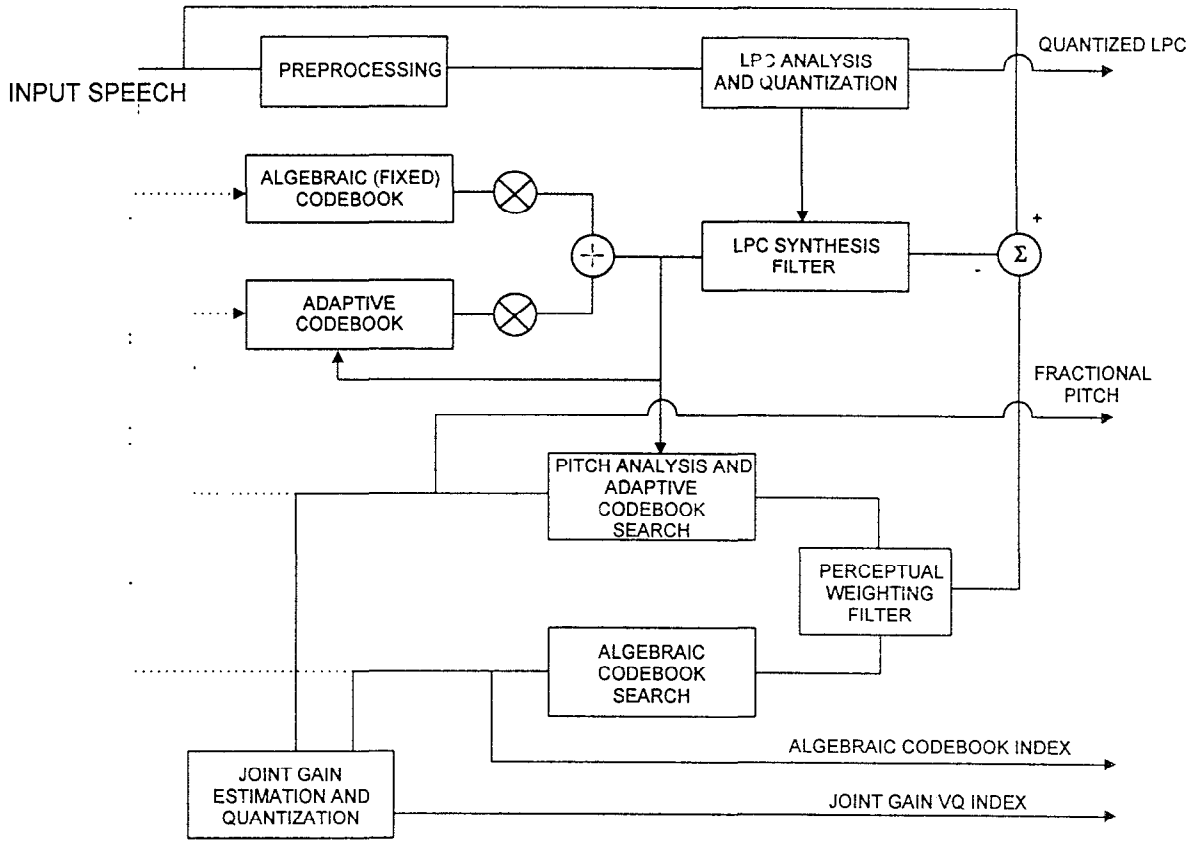
6709fg18.VSD



6709fg19.VSD

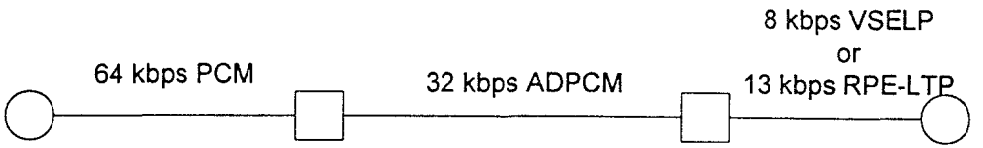


6709fg20.VSD

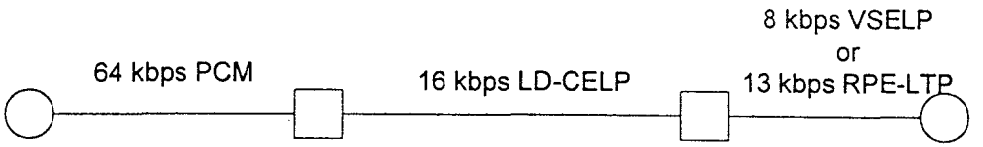


6709fg21.VSD

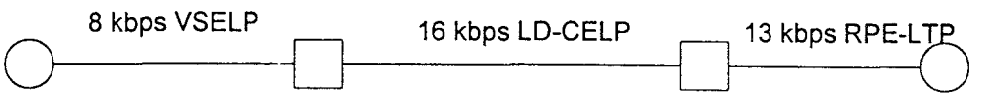
Configuration A



Configuration B



Configuration C



6709fg22.VSD

Technical Information Department • **Lawrence Livermore National Laboratory**
University of California • Livermore, California 94551