Abstract

The past decade has witnessed substantial progress towards the application of low-rate speech coders to civilian and military communications as well as computer-related voice applications. Central to this progress has been the development of new speech coders capable of producing high-quality speech at low data rates. Most of these coders incorporate mechanisms to: represent the spectral properties of speech, provide for speech waveform matching, and "optimize" the coder's performance for the human ear. A number of these coders have already been adopted in national and international cellular telephony standards.

The objective of this paper is to provide a tutorial overview of speech coding methodologies with emphasis on those algorithms that are part of the recent low-rate standards for cellular communications. Although the emphasis is on the new low-rate coders, we attempt to provide a comprehensive survey by covering some of the traditional methodologies as well. We feel that this approach will not only point out key references but will also provide valuable background to the beginner. The paper starts with a historical perspective and continues with a brief discussion on the speech properties and performance measures. We then proceed with descriptions of waveform coders, sinusoidal transform coders, linear predictive vocoders, and analysis-by-synthesis linear predictive coders. Finally, we present concluding remarks followed by a discussion of opportunities for future research.

1. Introduction

Although with the emergence of optical fibers bandwidth in wired communications has become inexpensive, there is a growing need for bandwidth conservation and enhanced privacy in wireless cellular and satellite communications. In particular, cellular communications have been enjoying a tremendous worldwide growth and there is a great deal of R&D activity geared towards establishing global portable communications through wireless personal communication networks (PCNs). On the other hand, there is a trend toward integrating voice-related applications (e.g., voicemail) on desktop and portable personal computers - often in the context of multimedia communications. Most of these applications require that the speech signal is in digital format so that it can be processed, stored, or transmitted under software control. Although digital speech brings flexibility and opportunities for encryption, it is also associated (when uncompressed) with a high data rate and hence high requirements of transmission bandwidth and storage. *Speech Coding* or *Speech Compression* is the field concerned with obtaining compact digital representations of voice signals for the purpose of efficient transmission or storage. Speech coding involves sampling and amplitude quantization. While the sampling is almost invariably done at a rate equal to or greater than twice the bandwidth of analog speech, there has been a great deal of variability among the proposed methods in the representation of the sampled waveform. The objective in speech coding is to represent speech with a minimum number of bits while maintaining its perceptual quality. The quantization or binary representation can be direct or parametric. Direct quantization implies binary representation of the speech samples themselves while parametric quantization involves binary representation of speech model and/or spectral parameters.

With very few exceptions, the coding methods discussed in this paper are those intended for digital speech communications. In this application, speech is generally bandlimited to 4 kHz (or 3.2 kHz) and sampled at 8kHz. The simplest non-parametric coding technique is Pulse Code Modulation (PCM) which is simply a quantizer of sampled amplitudes. Speech coded at 64 kilobits per second (kbits/s) using logarithmic PCM is considered as "non-compressed" and is often used as a reference for comparisons. In this paper, we shall use the term *medium-rate* for coding in the range of 8-16 kbits/s, *low-rate* for systems working below 8 kbits/s and down to 2.4 kbits/s, and *very-low-rate* for coders operating below 2.4 kbits/s.

Speech coding at medium-rates and below is achieved using an *analysis-synthesis* process. In the analysis stage, speech is represented by a compact set of parameters which are encoded efficiently. In the synthesis stage, these parameters are decoded and used in conjunction with a reconstruction mechanism to form speech. Analysis can be *open-loop* or *closed-loop*. In closed-loop analysis, the parameters are extracted and encoded by minimizing explicitly a measure (usually the mean square) of the difference between the original and the reconstructed speech. Therefore closed-loop analysis incorporates synthesis and hence this process is also called *analysis-by-synthesis*. Parametric representations can be speech or non-speech specific. Non-speech specific coders or waveform coders are concerned with the faithful reconstruction of the time-domain waveform and generally operate at medium-rates. Speech specific coders or voice coders (*vocoders*) rely on speech models and are focussed upon producing perceptually intelligible speech without necessarily matching the waveform. Vocoders are capable of operating at very-low rates but also tend to produce speech of synthetic quality. Although this is the generally accepted classification in speech coding, there are coders that combine features from both categories. For example, there are speech-specific waveform coders such as the Adaptive Transform Coder [303] and also hybrid coders which rely on analysis-by-synthesis linear prediction. Hybrid coders combine the coding efficiency of vocoders with the high-quality potential of waveform coders by modeling the spectral properties of speech (much like vocoders) and exploiting the perceptual properties of the ear, while at the same time providing for waveform matching (much like waveform coders). Modern hybrid coders can achieve communications quality speech at 8 kbits/s and below at the expense of increased complexity. At this time there are at least four such coders that have been adopted in telephony standards.

A. Scope and Organization

In this paper, we provide a survey of the different methodologies for speech coding with emphasis on those methods and algorithms that are part of recent communications standards. The paper is intended both as a survey and a tutorial and has been motivated by advances in speech coding which have enabled the standardization of low-rate coding algorithms for civilian cellular communications. The standardizations are results of more than fifty years of speech coding research. Until recently, low-rate algorithms were of interest only to researchers in the field. Speech coding is now of interest to many engineers who are confronted with the difficult task of learning the essentials of voice compression in order to solve implementation problems, such as fitting an algorithm to an existing fixed-point signal processor or developing low-power single-chip solutions for portable cellular telephones. Modern speech coding algorithms are associated with numerical methods that are computationally intensive and often sensitive to machine precision. In addition, these algorithms employ mathematical, statistical, and heuristic methodologies. While the mathematical and statistical techniques are associated with the theory of signal processing, communications, and information theory, many of the heuristic methods were established through years of experimental work. Therefore the beginner not only has to get a grasp of the theory but also needs to review the algorithms that preceded the standards. In this paper we attempt to sort through the literature and highlight the key theoretical and heuristic techniques employed in classical and modern speech coding algorithms. For each method we give the key references and, when possible, we refer first to the article that the novice will find more accessible.

The general notation adopted in this paper is as follows. The discrete-time speech signal is denoted as $s(n)$, where n is an integer indexing the sample number. Discrete-time speech is related to analog speech, $s_a(t)$, by $s(n) = s_a(nT) = s_a(t)\,|_{t=nT}$, where T is the sampling period. Unless otherwise stated lower-case symbols denote time-domain signals and upper-case symbols denote transform-domain signals. Bold characters are used for matrices and vectors. The rest of the notation is introduced in subsequent sections as necessary.

The organization of the paper is as follows. The first section gives a brief description of the properties of speech signals and continues with a historical perspective and a review of performance measures. In Section 2, we discuss waveform coding methods. In particular, we start with a general description of scalar [55][82][152] and vector quantization [81][98][115][192] methods and we continue with a discussion of waveform coders [48][52]. Section 3 presents sinusoidal analysis-synthesis methods [205] for voice compression and Section 4 presents vocoder methods [11][162][163]. Finally, in Section 5 we discuss analysis-by-synthesis linear predictive coders [96][100][123][272] and in Section 6 we present concluding remarks. Low-rate coders, and particularly those adopted in the recent standards, are discussed in more detail. The scope of the paper is wide and although our literature review is thorough is by no means exhaustive. Papers with similar scope [12][23][82][83][96][104][109][150][154][155][157][191][270][279], special journal and magazine editions on voice coding [18][19][131][132][134][135][136][138][139], and books on speech processing [62][86][90][91][99][113][152][199][232][234][236][251][275] can provide additional information. There are also six excellent collections of papers edited by Jayant [156], Davidson and Gray [61], Schafer and Markel [269], Abut [1], and Atal, Cuperman and Gersho [9][10]. For the reader, who wants to keep up with the developments in this field, articles appear frequently in IEEE transactions and symposia associated with the areas of signal processing and communications (see references section) and also in specialized conferences, workshops, and journals, e.g., [133][137][140][291].

B. Speech Properties

Before we begin our presentation of the speech coding methods, it would be useful if we briefly discuss some of the important speech properties. First, speech signals are non-stationary and at best they can be considered as quasi-stationary over short segments, typically 5-20 ms. The statistical and spectral properties of speech are thus defined over short segments. Speech can generally be classified as voiced (e.g., /a/, /i/, etc), unvoiced (e.g., /sh/), or mixed. Time and frequency domain plots for sample voiced and unvoiced segments are shown in Fig. 1. Voiced speech is quasi-periodic in the time-domain and harmonically structured in the frequency-domain while unvoiced speech is random-like and broadband. In addition, the energy of voiced segments is generally higher than the energy of unvoiced segments.

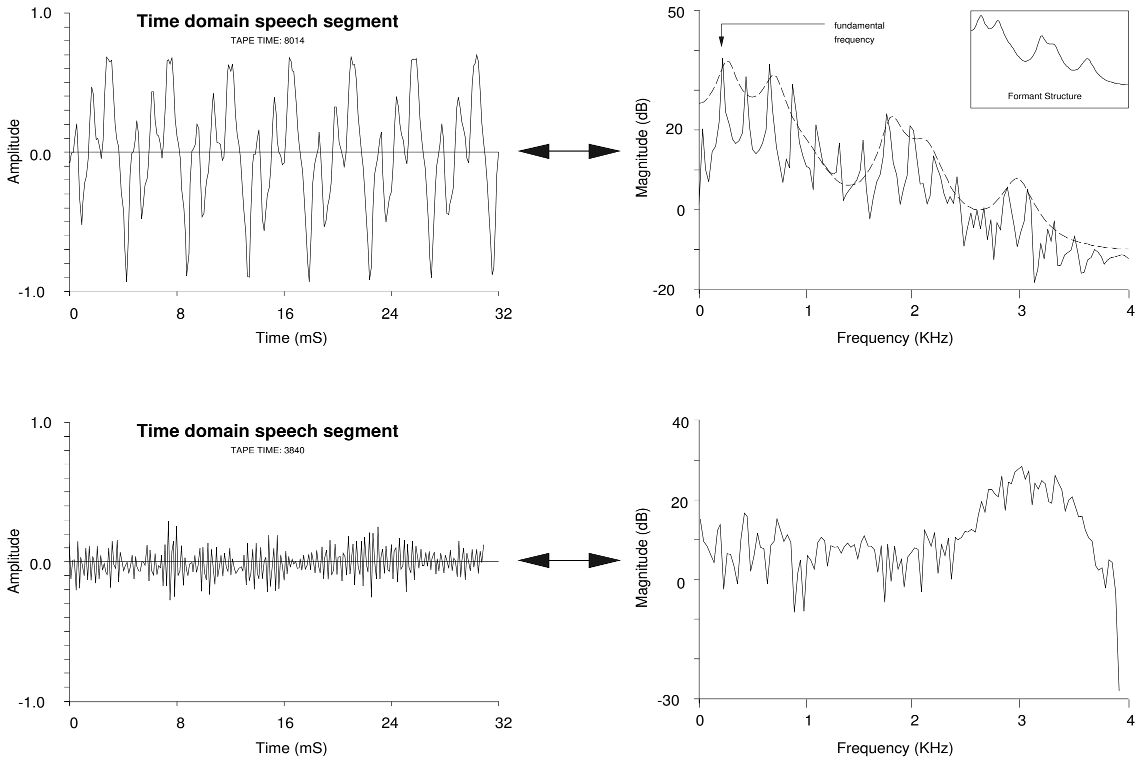


Fig. 1. Voiced and Unvoiced Segments and their short-time spectra.

The short-time spectrum[1] of voiced speech is characterized by its *fine* and *formant* structure. The fine harmonic structure is a consequence of the quasi-periodicity of speech and may be attributed to the vibrating vocal chords. The formant structure (spectral envelope) is due to the interaction of the source and the vocal tract. The vocal tract consists of the pharynx and the mouth cavity. The shape of the spectral envelope that "fits" the short-time spectrum of voiced speech, Fig. 1, is associated with the transfer characteristics of the vocal tract and the spectral tilt (6 dB/octave) due to the glottal pulse [261]. The spectral envelope is characterized by a set of peaks which are called formants. The formants are the resonant modes of the vocal tract. For the average vocal tract there are three to five formants below 5 kHz. The amplitudes and locations of the first three formants, usually occuring below 3 kHz, are quite important both in speech synthesis and perception. Higher formants are also important for wideband and unvoiced speech representations. The properties of speech are related to the physical speech production system as follows. Voiced speech is produced by exciting the vocal tract with quasi-periodic glottal air pulses generated by

1. Unless otherwise stated the term spectrum implies power spectrum the vibrating vocal chords. The frequency of the periodic pulses is referred to as the fundamental frequency or pitch. Unvoiced speech is produced by forcing air through a constriction in the vocal tract. Nasal sounds (e.g., /n/) are due to the acoustical coupling of the nasal tract to the vocal tract, and plosive sounds (e.g., /p/) are produced by abruptly releasing air pressure which was built up behind a closure in the tract.

More information on the acoustic theory of speech production is given by Fant [75] while information on the physical modeling of the speech production process is given in the classic book by Flanagan [86].

C. Historical Perspective

Speech coding research started over fifty years ago with the pioneering work of Homer Dudley [66][67] of the Bell Telephone Laboratories. The motivation for speech coding research at that time was to develop systems for transmission of speech over low-bandwidth telegraph cables. Dudley practically demonstrated the redundancy in the speech signal and provided the first analysis-synthesis method for speech coding. The basic idea behind Dudley's voice coder or vocoder (Fig. 2) was to analyze speech in terms of its pitch and spectrum and synthesize it by exciting a bank of ten analog band-pass filters (representing the vocal tract) with periodic (buzz) or random (hiss) excitation (for voiced and unvoiced sounds respectively). The channel vocoder received a great deal of attention during World War II because of its potential for efficient transmission of encrypted speech. Formant [223] and pattern matching [68] vocoders along with improved analog implementations of channel vocoders [221][292] were reported through the fifties and sixties. In the formant vocoder the resonant characteristics of the filter-bank track the movements of the formants. In the pattern matching vocoder the best match between the short-time spectrum of speech and a set of stored frequency response patterns is determined and speech is produced by exciting the channel filter associated with the selected pattern. The pattern matching vocoder was essentially the first analysis-synthesis system to implicitly employ vector quantization.

Although early vocoder implementations were based on analog speech representations, digital representations were rapidly gaining interest due to their promise for encryption and high fidelity transmission and storage. In particular, there had been a great deal of activity in Pulse Code Modulation (PCM) in the forties (see [156] and the references therein). PCM [228] is a straightforward method for discrete-time, discrete-amplitude approximation of analog waveforms

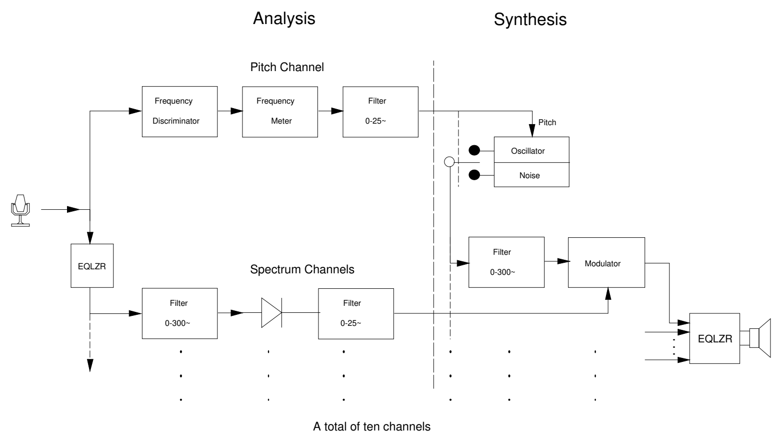


Fig. 2. Dudley's Channel Vocoder [67].

and does not have any mechanism for redundancy removal. Quantization methods that exploit the signal correlation, such as Differential PCM (DPCM), Delta Modulation (DM) [153], and Adaptive DPCM were proposed later and speech coding with PCM at 64 kbits/s and with ADPCM at 32 kbits/s eventually became CCITT[2] standards [32].

With the flexibility offered by digital computers, there was a natural tendency to experiment with more sophisticated digital representations of speech [266]. Initial efforts concentrated on the digital implementation of the vocoder [112]. A great deal of activity, however, concentrated on the linear speech source-system production model developed by Fant [75] in the late fifties. This model later evolved into the familiar speech production system shown in Fig. 3. This model consists of a linear slowly time-varying system (for the vocal tract and the glottal model) excited by periodic impulse train excitation (for voiced speech) and random excitation (for unvoiced speech).

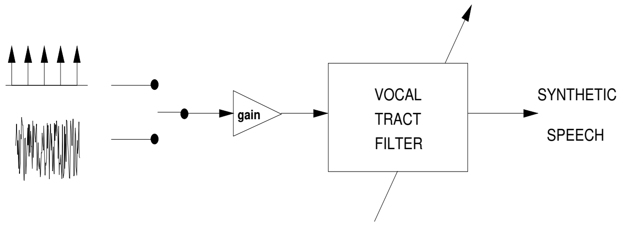


Fig. 3. The engineering model for speech synthesis.

2. International Consultative Committee for Telephone and Telegraph currently called International Telecommunications Union-Telecommunication Standardization Sector (ITU-TSS)

The source-system model became associated with Autoregressive (AR) time-series methods where the vocal tract filter is all-pole and its parameters are obtained by Linear Prediction analysis [189]; a process where the present speech sample is predicted by the linear combination of previous samples. Itakura and Saito [143][264] and Atal and Schroeder [14] were the first to apply Linear Prediction (LP) techniques to speech. Atal and Hanauer [11] later reported an analysis-synthesis system based on LP. Theoretical and practical aspects of linear predictive coding (LPC) were examined by Markel and Gray [199] and the problem of spectral analysis of speech using linear prediction was addressed by Makhoul and Wolf [190].

LP is not the only method for source-system analysis. Homomorphic analysis, a method that can be used for separating signals that have been combined by convolution, has also been used for speech analysis. Oppenheim and Schafer were strong proponents of this method [229][230]. One of the inherent advantages of homomorphic speech analysis is the availability of pitch information from the cepstrum [41][227].

The emergence of VLSI technologies along with advances in the theory of digital signal processing during the sixties and seventies provided even more incentives for getting new and improved solutions to the speech coding problem. Analysis-synthesis of speech using the Short-Time Fourier Transform (STFT) was proposed by Flanagan and Golden in a paper entitled "Phase Vocoder" [87]. In addition, Schafer and Rabiner designed and simulated an analysis-synthesis system based on the STFT [267][268], and Portnoff [240][242][243] provided a theoretical basis for the time-frequency analysis of speech using the STFT. In the mid- to late seventies there was also a continued activity in linear prediction [304][310], transform coding [268][303], and sub-band coding [52]. An excellent review of this work is given by Flanagan et al [82], and a unified analysis of transform and sub-band coders is given by Tribolet and Crochiere [303]. During the seventies, there were also parallel efforts for the application of linear prediction in military secure communications (see the NRL reports by Kang et al [162]-[166]). A federal standard (FS-1015) which is based on the LPC-10 algorithm, was developed in the early eighties (see the paper by Tremain [301]).

Research efforts in the eighties and nineties have been focused upon developing robust low-rate speech coders capable of producing high-quality speech for communications applications. Much of this work was driven by the need for narrowband and secure transmission in cellular and military communications. Competing methodologies promoted in the eighties included: sinusoidal analysis-synthesis of speech proposed by McAulay and Quatieri [205][206], multiband excitation vocoders proposed by Griffin and Lim [117], multi-pulse and vector excitation schemes for LPC proposed by Atal et al [13][272], and vector quantization (VQ) promoted by Gersho and Gray [98][99][115], and others [1][47][192]. Vector quantization [1] proved to be very useful in encoding LPC parameters. In particular, Atal and Schroeder [17][272] proposed a linear prediction algorithm with stochastic vector excitation which they called "Code Excited Linear Prediction" (CELP). The stochastic excitation in CELP is determined using a perceptually weighted closed-loop (analysis-by-synthesis) optimization. CELP coders are also called hybrid coders because they combine the features of traditional vocoders with the waveform matching features of waveform coders. Although the first paper [17] on CELP addressed the feasibility of vector excitation coding, follow-up work [37][100][170][171][176][177][276][315] essentially demonstrated that CELP coders were capable of producing medium-rate and even low-rate speech adequate for communications applications. Real-time implementation of hybrid coders became feasible with the development of highly structured codebooks.

Progress in speech coding, particularly in the late eighties, enabled recent adoptions of low-rate algorithms for mobile telephony. An 8 kbits/s hybrid coder has already been selected for the North American digital cellular standard [100], and a similar algorithm has been selected for the 6.7 kbits/s Japanese digital cellular standard [102][103][217][314]. In Europe, a standard that uses a 13 kbits/s regular pulse excitation algorithm [307] has been completed and partially deployed by the "Group Speciale Mobile" (GSM). Parallel standardization efforts for secure military communications [169] have resulted in the adoption of a 4.8 kbits/s hybrid algorithm for the Federal Standard 1016 [9]. In addition, a 6.4 kbits/s improved multiband excitation coder [121] has been adopted for the International Maritime Satellite (INMARSAT-M) system [322] and the Australian Satellite (AUSSAT) system. Finally we note that there are plans to increase the capacity of cellular networks by introducing half-rate algorithms in the GSM, the Japanese, and the North American standards.

D. Performance

A speech coding algorithm is evaluated based on the bit rate, the quality of reconstructed ("coded") speech, the complexity of the algorithm, the delay introduced, and the robustness of the algorithm to channel errors and acoustic interference. In general high-quality speech coding at low-rates is achieved using high complexity algorithms. For example, real-time implementation of a low-rate hybrid algorithm must be typically done on a digital signal processor capable of executing 12 or more million instructions per second (MIPS). The one-way delay (coding plus decoding delay only) introduced by such algorithms is usually between 50 to 60 ms. Robust speech coding systems incorporate error correction algorithms to protect the perceptually important information against channel errors. Moreover, in some applications coders must perform reasonably well with speech corrupted by background noise, non-speech signals (such as DTMF tones, voiceband data, modem signals, etc), and a variety of languages and accents.

In digital communications speech quality is classified into four general categories, namely: broadcast, network or toll, communications, and synthetic. Broadcast wideband speech refers to high quality "commentary" speech that can generally be achieved at rates above 64 kbits/s. Toll or network quality refers to quality comparable to the classical analog speech (200-3200 Hz) and can be achieved at rates above 16 kbits/s. Communications quality implies somewhat degraded speech quality which is nevertheless natural, highly intelligible, and adequate for telecommunications. Synthetic speech is usually intelligible but can be unnatural and associated with a loss of speaker recognizability. Communications speech can be achieved at rates above 4.8 kbits/s and the current goal in speech coding is to achieve communications quality at 4.0 kbits/s. Currently, speech coders operating well below 4.0 kbits/s tend to produce speech of synthetic quality.

Gauging the speech quality is an important but also very difficult task. The signal-to-noise ratio (SNR) is one of the most common objective measures for evaluating the performance of a compression algorithm. This is given by:

$$SNR = 10 \ \log_{10} \left\{ \frac{\sum_{n=0}^{M} s^2(n)}{\sum_{n=0}^{M} (s(n) - \hat{s}(n))^2} \right\} \tag{1}$$

where $s(n)$ is the original speech data while $\hat{s}(n)$ is the coded speech data. The SNR is a long term measure for the accuracy of speech reconstruction and as such it tends to "hide" temporal reconstruction noise particularly for low level signals. Temporal variations of the performance can be better detected and evaluated using a short-time signal-to-noise ratio, i.e., by computing the SNR for each N-point segment of speech. A performance measure that exposes weak signal performance, is the segmental SNR (SEGSNR) which is given by

$$SEGSNR = \frac{10}{L} \sum_{i=0}^{L-1} \ \log_{10} \left\{ \frac{\sum_{n=0}^{N-1} s^2(iN+n)}{\sum_{n=0}^{N-1} (s(iN+n) - \hat{s}(iN+n))^2} \right\} \tag{2}$$

Since the averaging operation in (2) occurs after the logarithm, the SEGSNR penalizes more the coders whose performance is variant. Other objective measures often mentioned in the literature include the articulation index, the log spectral distance, and the Euclidean distance (see [244] for definitions). Objective measures are often sensitive to both gain variations and delays. More importantly, they typically do not account for the perceptual properties of the ear [222]. Therefore, subjective evaluations [59][224] using phonetically balanced speech records [93][142] are required since the design of most low-rate algorithms is based on perceptual criteria. Subjective test procedures such as the Diagnostic Rhyme Test (DRT), the Diagnostic Acceptability Measure (DAM) [312], and the Mean Opinion Score (MOS) are based on listener ratings. The DRT is an intelligibility measure where the subject's task is to recognize one of two possible words in a set of rhyming pairs (e.g., meat - heat). The rhyme test was originally introduced by Fairbanks [74] and the modified rhyme test was developed later by House [129]. Details on the evaluation of speech using the DRT can be found in an article by Voiers [313]. The DAM scores are based on results of test methods evaluating the quality of a communication system based on the acceptability of speech as perceived by a trained normative listener. The MOS is a measure which is widely used to quantify coded speech quality. The MOS usually involves 12 to 24 listeners [180] (formal CCITT and TIA tests typically involve 32-64 listeners) who are instructed to rate phonetically balanced records according to a 5-level quality scale, Table. 1. Excellent speech quality implies that coded speech is indistinguishable from the original and without perceptible noise. On the other hand, bad (unacceptable) quality implies the presence of extremely annoying noise and artifacts in the coded speech.

Table. 1. The MOS scale.

MOS Scale	Speech Quality
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

In MOS tests listeners are "calibrated" in the sense that they are familiarized with the listening conditions and the range of speech quality they will encounter. Ratings are obtained by averaging numerical scores over several hundreds of speech records. The MOS range relates to speech quality as follows: a MOS of 4-4.5 implies network quality, scores between 3.5 and 4 imply communications quality, and a MOS between 2.5 and 3.5 implies synthetic quality. We note here that MOS ratings may differ significantly from test to test and hence they are not absolute measures for the comparison of different coders.

Formal subjective evaluations, such as the ones described above, can be lengthy and very costly. Recent efforts in speech quality assessment are focussed upon developing automatic test evaluation procedures [180] and objective measures that are capable of predicting the subjective quality of speech [316]. More information on objective and subjective measures can be found in [152][244][312].

2. Waveform Coders

Waveform coders focus upon representing the speech waveform as such without necessarily exploiting the underlying speech model. Waveform coders are generally more robust than vocoders, in the sense that they work well with a wider class of signals, however, they also generally operate at higher data rates relative to vocoders. Our discussion on waveform coders includes descriptions of some of the fundamental scalar and vector quantization schemes. In addition, we present the fundamental concepts associated with sub-band coders and transform coders. Selected well known algorithms are also presented.

2.1 Scalar and Vector Quantization

Signal coding can be non-parametric or parametric depending upon whether the actual signal, or its parametric representation, is quantized. Parametric representations are generally based on signal transformations (often unitary), or on signal models (often linear source-system). Quantization can be memoryless or with memory depending upon whether the encoding rules depend on past inputs or outputs. In this section, we describe time-domain quantization methods. We classify quantization methods into two general classes, namely scalar and vector quantization. More emphasis is placed on vector quantization which has been studied intensively the last fifteen years and exhibited enormous potential for high quality speech coding at low rates.

2.1.1 Scalar Quantization Methods

Scalar quantization [149][153] methods include Pulse Code Modulation (PCM), Differential PCM (DPCM), and Delta Modulation (DM). Uniform PCM is a memoryless process that quantizes amplitudes by rounding off each sample to one of a set of discrete values. The difference between adjacent quantization levels, i.e., the step-size, is constant in non-adaptive uniform PCM. Although uniform PCM is the simplest method for digital encoding, it is also the most expensive in terms of data rates since it has no mechanism for exploiting signal redundancy (correlation). Its performance, in terms of the signal-to-noise ratio (SNR), can be predicted by the expression [152],

$$SNR_1 = 6B + K_1 \qquad \text{(dB)} \tag{3}$$

where B is the number of bits per sample and K_1 is a step-size dependent parameter. Nonuniform PCM, as the name implies, uses a nonuniform step-size. Nonuniform quantizers generally employ a fine quantizing step-size for frequently occurring amplitudes and a coarse step-size for rarely occurring amplitudes. The step-sizes may also be optimally designed by exploiting the shape of the signal's probability density function (PDF) [274]. Another class of nonuniform PCM relies on log-quantizers, so called μ-law and A-law. These are quite common in speech applications and are described well in [152] and [274]. A 7-bit log-quantizer for speech achieves the performance of a 12-bit uniform quantizer. Dynamic range variations in uniform or nonuniform PCM can be handled by using an adaptive step-size. A PCM system with an adaptive step-size is called Adaptive PCM (APCM). The step-size in a *feedforward* system is transmitted as side information while in a *feedback* system the step-size is estimated from past coded speech samples.

A more efficient scalar quantizer is the Differential PCM (DPCM) which utilizes the redundancy in the speech waveform by exploiting the correlation between adjacent samples. In its simplest form a DPCM transmitter encodes only the difference between successive samples and the receiver recovers the signal by integration. Practical DPCM schemes incorporate a time-invariant short-term prediction process, $A(z)$, where

$$A(z) = \sum_{i=1}^{p} a_i z^{-i} \tag{4}$$

and z is the complex variable of the z-transform. This DPCM scheme is also called predictive differential coding (Fig. 4) and reduces the quantization error variance by reducing the variance of the quantizer input. The DPCM system works as follows. The sample $\tilde{s}'(n)$ is the estimate of the current speech sample, $s(n)$, and is obtained from past sample values. The prediction error, $e(n)$, is then quantized $(e_q(n))$ and transmitted to the receiver. The quantized prediction error is also added to $\tilde{s}'(n)$ to reconstruct the speech sample $s'(n)$. In the absence of channel errors $s'(n) = \hat{s}(n)$. In the simplest case $A(z)$ is a first order polynomial.

Two other types of scalar coders are the Deltamodulation (DM) and the Adaptive DPCM (ADPCM) [55][105][106][324] coders. DM is a sub-class of DPCM where the difference (prediction error) is encoded only with one bit. DM typically operates at sampling rates much higher than the rates commonly used with DPCM. The step-size in DM may also be adaptive (ADM). The DM and DPCM are low to medium complexity coders and perform better than ordinary PCM for rates at and below 32 kbits/s. In ADPCM the step-size and the predictor are allowed to adapt and track the time-varying statistics of speech. The predictor can be forward adaptive or backward adaptive. In forward adaptation the prediction parameters are estimated from the current speech data which are not available at the receiver. Therefore the prediction parameters must be encoded and transmitted separately in order to reconstruct the signal at the receiver.

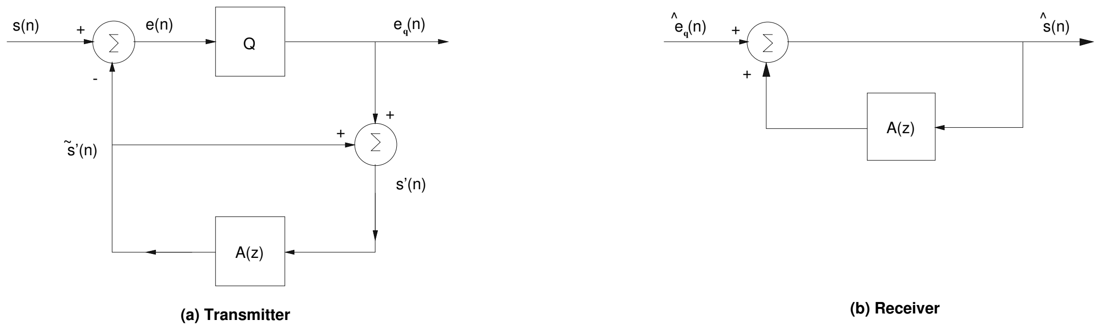


Fig. 4. A DPCM System.

In backward adaptation the parameters are estimated from past speech data, $(e_q(n))$ or $(\hat{s}(n))$, which is also available at the receiver. Therefore the prediction parameters can be estimated locally at the receiver. Backward predictor adaptation is amenable to low-delay coding [38][107]. ADPCM encoders with pole-zero decoder filters have proved to be particularly versatile in speech applications. In fact, the ADPCM 32 kbits/s algorithm adopted for the G.721 CCITT standard (1984) [32] uses a pole-zero adaptive predictor. The block diagram of the G.721 CCITT encoder [22] is shown in Fig. 5.

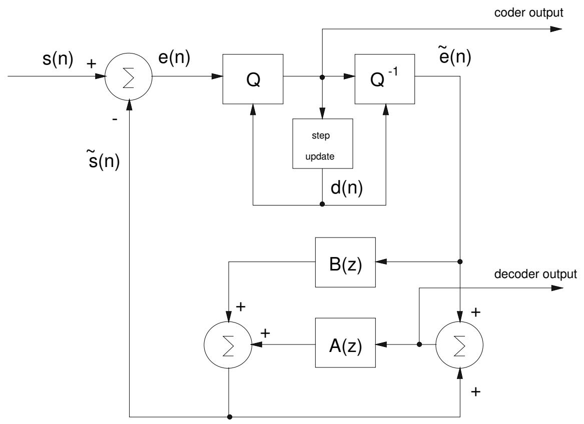
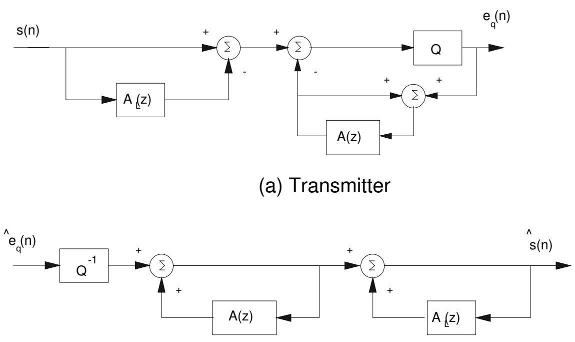


Fig. 5. The ADPCM G.721 encoder [22].

The algorithm consists of an adaptive quantizer and an adaptive pole-zero predictor. As it is seen in Fig. 5 the decoder is embedded in the encoder. The pole-zero predictor (2 poles, 6 zeros) estimates the input signal and hence it reduces the variance of *e(n)*. The quantizer encodes the sequence *e(n)* into a sequence of 4-bit words. The prediction coefficients are estimated using a gradient algorithm and the stability of the decoder is checked by testing the two roots of $A(z)$. The performance of the coder, in terms of the MOS scale, is above 4 but it degrades as the number of asynchronous tandem codings increases. The G.721 ADPCM algorithm was also modified to accommodate 24 and 40 kbits/s in the G.723 standard [34]. More details on the algorithm and its performance are given in [22] and the references therein. The performance of ADPCM degrades quickly for rates below 24 kbits/s.

The use of adaptive long-term prediction in addition to short-term prediction (Fig. 6) provides additional coding gain (at the expense of higher complexity) and high quality speech at 16 kbits/s [7]. The long-term (long delay) predictor $\left(A_L(z) = \sum_{i=-j}^{j} a_i z^{-i-\tau} \right)$



(b) Receiver

Fig. 6. The Adaptive Predictive Coder with short- and long- term predictors.

provides for the pitch (fine) structure of the short-time voiced spectrum. The index τ is the pitch period in samples and j is a small integer. The long-term predictor (ideally) removes the periodicity and thereby redundancy. At the receiver the synthesis filter $(1/(1-A_L(z)))$ introduces periodicity while the synthesis filter associated with the short-term prediction polynomial represents the vocal tract. The parameters of the short-term predictors are computed for every frame (typically 10 to 30 ms). The long-term prediction parameters are computed more often. Although the adaptive predictive coder is now of limited interest, the work on adaptive predictive coders by Atal [7] in 1978 provided several ideas such as the application of short- and long- term prediction and perceptual error weighting in analysis-by-synthesis predictive coders. The issues associated with adaptive short- and long- term prediction algorithms will be discussed in more detail in Section 5.

Our description of scalar quantizers was brief, and space precluded in-depth analysis of this family of coders. The book by Jayant and Noll [152] presents a comprehensive treatment of the subject with an in-depth analysis of performance and complexity. In addition, an excellent tutorial review of adaptive prediction as applied to DPCM systems is given by Gibson [104].

2.1.2 Vector Quantization Methods

Data compression via Vector Quantization (VQ) is achieved by encoding a data set in block or vector form. Although Shannon's rate distortion theory [273] essentially stated that better performance can be achieved by coding vectors instead of scalars, significant results in VQ-based speech coding were not reported until the late seventies. This is mainly due to the inherent complexity of VQ. Recently, however, with the emergence of new and efficient methods for encoding high dimensionality data blocks, VQ became associated with high-quality speech coding at low-rates. In the following, we describe the basic methodologies for VQ.

Figure 7a shows a VQ scheme which consists of an N-dimensional quantizer and a codebook. The incoming vectors are formed from consecutive data samples or from model parameters. The quantizer maps the *i-th* incoming Nx1 vector, $\mathbf{s}_i = [s_i(0)s_i(1)\ldots s_i(N-1)]^T$, to a channel symbol $\{u_n,\ n = 1, 2, \ldots, L\}$. For practical reasons we assume that the channel is noiseless, that is $(u_n = \hat{u}_n)$. The codebook consists of L codevectors, $\{\hat{\mathbf{s}}_n = [\hat{s}_n(0)\hat{s}_n(1)\ldots\hat{s}_n(N-1)]^T,\ n = 1, 2, ..L\}$, which reside in the memory of the transmitter and the receiver.

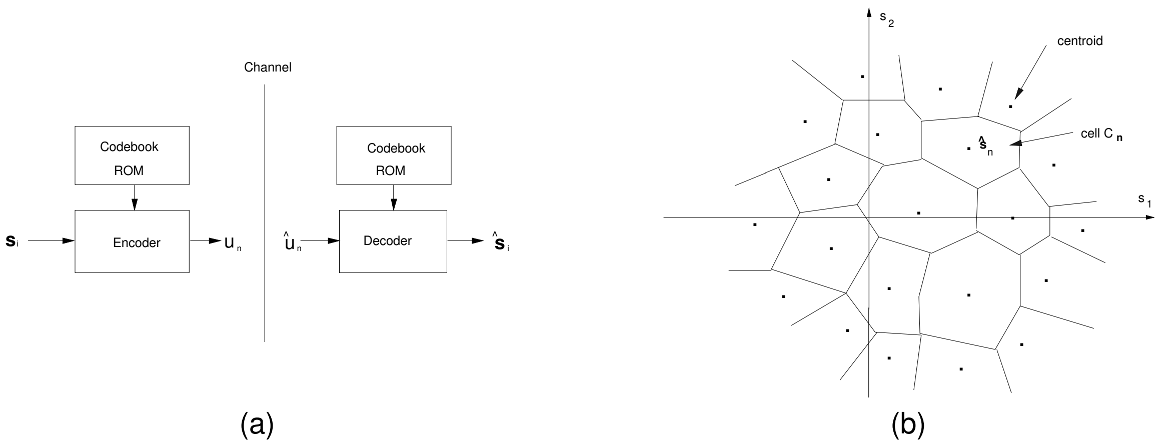


Fig. 7. Vector Quantization Scheme, (a) Block Diagram, (b) Cells for two-dimensional VQ.

A memoryless full search vector quantizer works as follows: incoming vectors, $\mathbf{s}_i$, are compared to each codeword and the address of the closest codeword, w.r.t. a distortion measure or fidelity criterion $\varepsilon(\mathbf{s}_i, \hat{\mathbf{s}}_n)$, determines the channel symbol to be transmitted. The simplest and most commonly used distortion measure is the sum of squared errors which is given by

$$\varepsilon(\mathbf{s}_i, \hat{\mathbf{s}}_n) = \sum_{k=o}^{N-1} (s(k) - \hat{s}(k))^2 \tag{5}$$

The L entries of the codebook, i.e., the L Nx1 real valued vectors, are designed by dividing the vector space into L non-overlapping cells, C_n, Fig. 7b. Each cell, C_n, is associated with a template vector $\hat{\mathbf{s}}_n$. The quantizer assigns the channel symbol u_n to the vector $\mathbf{s}_i$, if $\mathbf{s}_i$ belongs to C_n. This means that if $\mathbf{s}_i$ belongs to C_n it will be represented by $\hat{\mathbf{s}}_n$ which is the centroid of the cell. The channel symbol u_n is usually the binary representation of the index or address of $\hat{\mathbf{s}}_n$.

The simplest form of a vector quantizer can be considered as a generalization of the scalar PCM and in [98] is called Vector PCM (VPCM). In VPCM, the codebook is fully searched (full search VQ or F-VQ) for each incoming vector. The number of bits per sample in VPCM is given by

$$B = \frac{\log_2 L}{N} \tag{6}$$

and the signal-to-noise ratio for VPCM is given by

$$SNR_N = 6B + K_N \qquad \text{(dB)} \tag{7}$$

Note that for $N = 1$ VPCM defaults to scalar PCM and therefore (3) is a special case of (7). Although the two equations are quite similar, VPCM yields improved SNR (reflected in K_N) since it exploits the correlation within the vectors. In the case of speech coding, Gersho and Cuperman [98] reported that K_2 is larger than K_1 by more than 3 dB while K_8 is larger than K_1 by more than 8 dB.

Even though VQ offers significant coding gain by increasing N and L, its memory and computational complexity grows exponentially with N for a given rate. More specifically, the number of computations required for F-VQ is of the order of 2^{BN} while the number of memory locations required is $N\ 2^{BN}$. In general the benefits of VQ are realized at rates of 1 bit per sample or less.

The codebook design process, also known as the training or populating process, can be fixed or adaptive. Fixed codebooks are designed a priori and the basic design procedure involves an initial guess for the codebook and then iterative improvement by using a large number of training vectors. An iterative codebook design algorithm that works for a large class of distortion measures was given by Linde, Buzo, and Gray [183]. This is essentially an extension of Lloyd's [185] scalar quantizer design and is often referred to as the LBG algorithm. In general, the number of training vectors per code vector must be at least ten and preferably fifty [192]. The computational complexity for the design process is quite high. In particular the design complexity grows exponentially with the dimension of the codevectors and linearly with the number of training vectors. The robustness of the codebook, the efficiency in the codebook search process, and the choice of distortion measure [143][318] are central issues in codebook design.

The complexity in high-dimensionality VQ can be reduced significantly with the use of structured codebooks which allow for efficient search. Tree-structured [28] and multistep [160] vector quantizers are associated with lower encoding complexity at the expense of loss of performance and in some cases increased memory requirements. Gray and Abut [116] compared the performance of F-VQ and binary tree search VQ for speech coding and reported a degradation of 1 dB in the SNR for tree-structured VQ. Multistep vector quantizers consist of a cascade of two or more quantizers each one encoding the error or residual of the previous quantizer. Gersho and Cuperman [98] compared the performance of full search (dimension 4) and multistep vector quantizers (dimension 12) for encoding speech waveforms at 1 bit per sample and reported a gain of 1 dB in the SNR in the case of multistep VQ.

The complexity of VQ can also be reduced by normalizing the vectors of the codebook and encoding the gain separately. The technique is called Gain/Shape VQ (GS-VQ) and has been introduced by Buzo et al [28] and later studied by Sabin and Gray [263]. The waveform shape is represented by a codevector from the shape codebook while the gain can be encoded from the gain codebook, Fig. 8. Comparisons of GS-VQ with F-VQ in the case of speech coding at one bit per sample revealed that GS-VQ yields about 0.7 dB improvement at the same level of complexity.

The idea of encoding the gain separately allows for the encoding of vectors of high dimensionality with manageable complexity and is being widely used in encoding the excitation signal in code excited linear predictive coders [9][100].

An alternative method for building highly structured codebooks consists of forming the codevectors by linearly combining a small set of basis vectors. This approach was proposed by Gerson and Jasiuk [100] and is called vector sum quantization. This method not only leads to

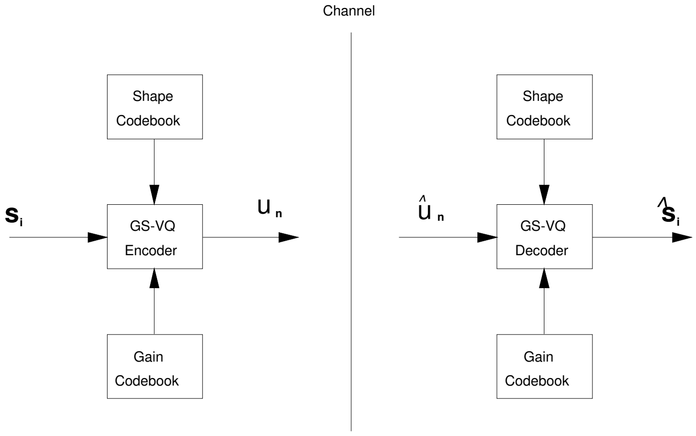


Fig. 8. GS-VQ Encoder and Decoder.

efficient full search but is also associated with robustness to channel errors. Vector sum quantization will be discussed in more detail in a subsequent section.

The VQ methods discussed thus far are associated with time-invariant (fixed) codebooks. Since speech is a non-stationary process, one would like to adapt the codebooks ("codebook design on the fly") to its changing statistics. VQ with adaptive codebooks is called adaptive VQ (A-VQ) and applications to speech coding have been reported in [56] and [238]. There are two types of A-VQ, namely, forward adaptive and backward adaptive. In backward adaptive VQ, codebook updating is based on past data which is also available at the decoder. Forward A-VQ updates the codebooks based on current (or sometimes future) data and as such additional information must be encoded. The principles of forward and backward A-VQ are similar to those of scalar adaptive quantization [104][152]. Practical A-VQ systems are backward adaptive and they can be classified into vector predictive quantizers [57] and finite state quantizers [88]. Vector predictive coders are essentially an extension of scalar predictive DPCM coders. A class of coders that are related to predictive VQ are the trellis tree coders [196][311].

Under certain conditions [171] the long-term predictor in CELP coders can be interpreted as a special case of a backward adaptive vector quantization process. For example in VSELP [100] past excitation forms and updates the codebook, and the long-term prediction lag is viewed as the codebook index. More details on the use of adaptive codebooks in CELP coders will be given in Section 5.

In concluding our discussion we note that the advances in structured high-dimensionality VQ have been one of the main reasons for the dramatic progress in low-rate speech coding. Adaptive, and highly structured stochastic codebooks are used to encode the excitation in the hybrid analysis-by-synthesis speech coder which is currently regarded by many as one of the leading approaches to low-rate coding. For further reading on VQ we refer the reader to the excellent tutorial papers of Gersho et al [98], Gray [115], and Makhoul et al [192]. There is also a collection of papers edited by Abut [1], and a new textbook on VQ by Gersho and Gray [99].

2.2 Sub-band and Transform Coders

The sub-band coders [21][52][63][73][257][265][281] and transform coders [20][36][277][303] exploit the redundancy of the signal in the transform domain. Most of these coders rely on frequency-domain representations of the signal obtained through filter bank representations (sub-band coders) or by using discrete-time unitary transforms (transform coders). The opportunity to reduce the bit-rate in both cases lies both in the structure of the short-time power spectrum of speech and also in the perceptual properties of the human ear.

2.2.1 Sub-band Coders

In the sub-band coder [49][51][52][53][257], Fig. 9, the signal band is divided into frequency sub-bands using a bank of bandpass filters. The output of each filter is then sampled (or down-sampled) and encoded. At the receiver the signals are demultiplexed, decoded, demodulated, and then summed to reconstruct the signal. This system lends itself to both analog and digital implementations depending upon the way the filter bank is implemented. The encoding process introduces quantization noise while the sampling/demodulation processes introduce aliasing distortion due to the overlapping nature of the sub-bands. The sub-band coder (SBC) exploits the statistics of the signal and/or perceptual criteria to encode the signal in each band using a different number of bits. For example, in speech the lower frequency bands are usually allotted more bits than higher bands in order to preserve critical pitch and formant information. In one of the earliest papers on sub-band coding, Crochiere et al [52] presented a four-band SBC operating at 16, 9.6, and 7.2 kbits/s. The four sub-bands are associated with the following frequency ranges: 200-700, 700-1310, 1310-2020, and 2020-3200 Hz. Low-pass translation and decimation is done prior to coding. The 16 kbits/s SBC compared favorably against 16 kbits/s ADPCM, and the 9.6 kbits/s SBC compared favorably against 10.3 and 12.9 kbits/s ADM [149].

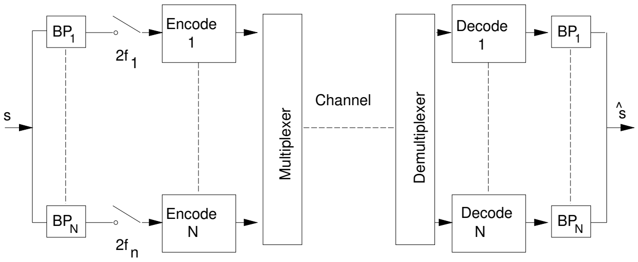


Fig. 9. A typical Sub-band Coder.

The design of the filter bank is a very important consideration in the design of a SBC. The filter bandwidth may be equal or unequal. For example, the low-band filters in speech specific implementations are usually associated with narrower widths so that they can resolve more accurately the low-frequency narrowband formants. In the absence of quantization noise, perfect reconstruction can be achieved using Quadrature-Mirror Filter (QMF) banks [72][281][305][308]. Information on the filter bank design can be found in the book by Crochiere and Rabiner [51]. We also note that the filter bank design has recently been associated with the theory of wavelet series expansions [258].

In the following, we describe briefly two coding standards that are based on sub-band coding algorithms, namely, the AT&T voice store and forward standard [159], and the CCITT G.722 standard [33][155]. The AT&T SBC was used for voice storage at 16 or 24 kbits/s and consists of a five-band non-uniform tree-structured QMF bank in conjunction with APCM coders, Fig. 10.

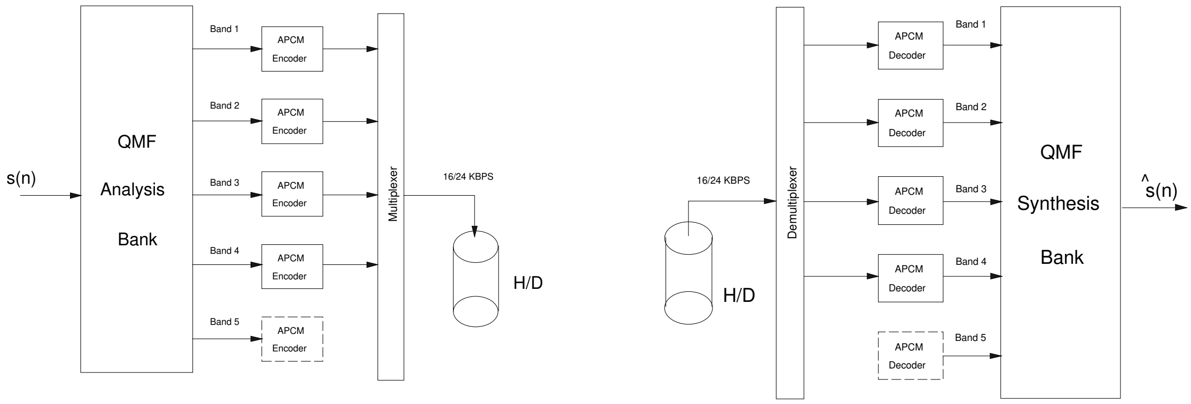


Fig. 10. The AT&T Sub-band Coder.

A silence compression algorithm is also part of the standard. The frequency ranges for each band are: 0-0.5 kHz, 0.5-1 kHz, 1-2 kHz, 2-3 kHz, and 3-4 kHz. For the 16 kbits/s implementation the bit allocations are {4/4/2/2/0} and for the 24 kbits/s the bit assignments are {5/5/4/3/0}. The one-way delay of this coder is less than 18 ms. It must be noted that although this coder was the workhorse for the older AT&T voice store and forward machines, the most recent AT&T audix machines use the new 16 kbits/s Low Delay CELP algorithm which will be described in Section 5.

The CCITT standard (G.722) for 7kHz audio at 64 kbits/s for ISDN teleconferencing is based on a two-band sub-band/ADPCM coder, Fig. 11. The low-frequency sub-band is quantized at 48 kbits/s while the high-frequency sub-band is coded at 16 kbits/s. The G.722 coder includes an adaptive bit allocation scheme and an auxiliary data channel. Provisions for lower rates have been made by quantizing the low-frequency sub-band at 40 kbits/s or at 32 kbits/s. The MOS at 64 kbits/s is greater than four for speech and slightly less than four for music signals [155], and the analysis-synthesis QMF banks introduce a delay of less than 3 ms. Details on the real-time implementation of this coder are given in [297].

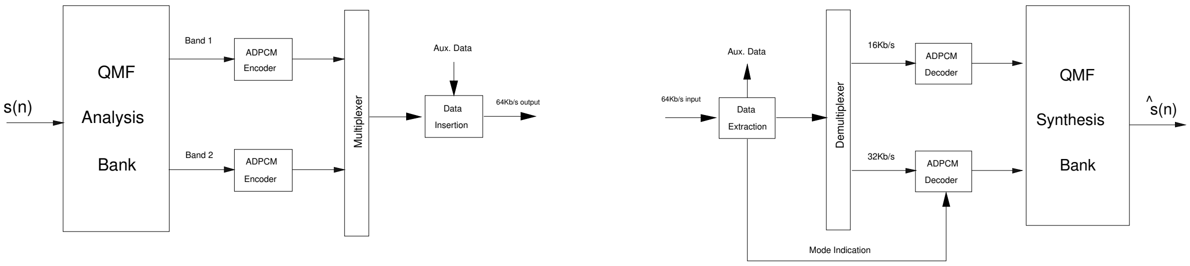


Fig. 11. The CCITT G.722 Sub-band Coder.

Before we close our discussion on SBC, we note that a low-delay 16 kbits/s sub-band coder which employs gain-shape VQ for the coding of sub-band signals was proposed by Cox et al [49]. The VQ-SBC was implemented real-time on a floating-point processor and achieved a SEGSNR of about 20 dB with a coding delay of about 15 ms.

2.2.2 Transform Coders

In transform coders [20][36][48][76][188][285][303] , Fig. 12, the transform components of a unitary transform [3] are quantized at the transmitter and decoded and inverse-transformed at the receiver. The potential for bit rate reduction in transform coding (TC) lies in the fact that unitary transforms tend to generate near-uncorrelated transform components which can be coded independently.

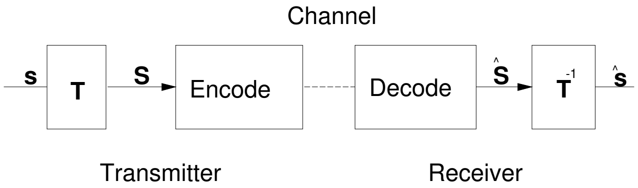


Fig. 12. The Transform Coder.

Furthermore, the variances of these components often exhibit consistent or slowly time-varying patterns which can be exploited for redundancy removal using fixed or adaptive bit-allocation rules. The signal in TC is processed frame-by-frame and each frame is transformed using a discrete unitary transform. This operation can be represented by a matrix multiplication, i.e.,

$$\begin{bmatrix} S(0) \\ S(1) \\ S(2) \\ . \\ . \\ . \\ S(N-1) \end{bmatrix} = \begin{bmatrix} t_{1,1} & t_{1,2} & t_{1,3} & \cdots & t_{1,N} \\ t_{2,1} & t_{2,2} & t_{2,3} & \cdots & t_{2,N} \\ t_{3,1} & t_{3,2} & t_{3,3} & \cdots & t_{3,N} \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ t_{N,1} & t_{N,2} & t_{N,3} & \cdots & t_{N,N} \end{bmatrix} \begin{bmatrix} s(0) \\ s(1) \\ s(2) \\ . \\ . \\ . \\ s(N-1) \end{bmatrix} \tag{8}$$

or in a more compact form

$$\mathbf{S} = \mathbf{T}\ \mathbf{s} \tag{9}$$

The inverse transform is applied for signal synthesis

$$\mathbf{s} = \mathbf{T}^{-1}\ \mathbf{S} \tag{10}$$

Equation (8) is known as the analysis expression while equation (10) is the synthesis formula. For unitary transforms $\mathbf{T}^{-1} = \mathbf{T}^{H}$, where ^{H} in the superscript denotes the hermitian (complex conjugate transpose) operator. The column vectors of $\mathbf{T}^{H}$ are known as the basis vectors of the transformation and the signal vector in (10) is essentially reconstructed by a linear combination of the basis vectors. In the absence of quantization and channel errors the synthesis expression (10) yields exact reconstruction. When the transform components are quantized, however, reconstruction is no longer exact and depends on the quantization noise and the type of the transform. There are several discrete transforms [3][71] that can be used for TC, e.g., the Discrete Cosine Transform (DCT), the Discrete Fourier Transform (DFT), the Walsh Hadamard Transform (WHT), the Karhunen Loeve Transform (KLT), etc. The KLT is the optimal [29][326] unitary transform in the sense that the transform components are "maximally" decorrelated for any given signal. An analytical proof for the optimality of the KLT in TC is given in [326]. The basis vectors of the KLT are the normalized eigenvectors of the autocorrelation matrix of the signal. Therefore the KLT is data-dependent and in many cases impractical because of the large number of computations required to determine the eigenvectors. The DFT and the DCT are associated with sinusoidal basis vectors and can be computed efficiently using the Fast Fourier Transform (FFT) [46][193]. For a normalized DFT, the transform matrix consists of sampled complex exponentials, i.e., $t_{i,j} = e^{-j2\pi(i-1)(j-1)/N}/\sqrt{N}$. The DCT on the other hand is given by

$$S(k) = \sum_{n=0}^{N-1} s(n)\lambda(k)\cos[(2n+1)\pi k/2N] \tag{11}$$

where $\lambda(k)$ is equal to one for $k = 0$ and equal to $\sqrt{2}$ for $k = 1, 2, .., N-1.$ The inverse DCT is defined as

$$s(n) = \frac{1}{N} \sum_{k=0}^{N-1} S(k) \lambda(k) \cos[(2n+1)\pi k / 2N] \tag{12}$$

The transform matrix for the WHT consists of ones and minus ones, i.e., the basis vectors for the WHT are essentially sampled pulse functions which take the value of one or minus one. These pulse functions are characterized by their sequency, i.e., the number of zero crossings per unit time. An N-point WHT can be computed efficiently using an algorithm which requires arithmetic operations of order $\mathbf{O}(N \log_2 N)$ [3].

The performance of the DCT, DFT, and WHT is suboptimal. The performance of the DCT, however, is near optimal. This is because the DCT vectors are eigenvectors of a large class of Toeplitz matrices [3]. The performance of the DFT approaches that of the DCT and KLT for very large block lengths [326]. Early results on TC reported by Campanella and Robinson [29], Gethoffer [92], and Zelinski and Noll [326] claimed gains over PCM of the order of 9-10 dB for the KLT, 5 dB for the DFT, and 3 dB for the WHT.

A transform coder, that employs the DCT and encodes the transform components using adaptive quantization and bit assignment rules, was proposed by Zelinski and Noll [326]. The bit assignment in this adaptive transform coder (ATC) depends on the short-term spectrum which is determined from the transform components, Fig. 13. In particular, a coarse description of the short-term spectrum is formed at the encoder and transmitted (as side information) to the decoder on a frame-by-frame basis. An estimate of the short-term spectrum is then formed using linear interpolation in the log domain, and optimal bit assignments are determined using the estimated spectrum. It was shown in [326], that for rates between 16-32 kbits/s the ATC outperformed the log-PCM by about 17-23 dB (SNR). In addition, the ATC outperformed ADPCM by about 6 dB at 16 kbits/s. Perceptual distortion tends to be small for rates down to 12 kbits/s. At lower rates, however, the ATC suffers from the "low-pass filtering" effect [327] which is a consequence of the fact that only high-energy components are coded. Solutions to the low-pass filtering effect and complexity reductions for the ATC were proposed in [327]. Also a "speech-specific" or "vocoder-driven" ATC was proposed by Tribolet and Crochiere [303]. This coder derives its adaptation strategy from the pitch and the formant structure of speech and uses an LPC model for the side information. A speech-specific ATC, which uses a homomorphic model for the side information, was proposed by Cox and Crochiere [48]. This algorithm is quite efficient, in terms of pipelining, and its performance was shown to be at least as good as the LPC-driven ATC.

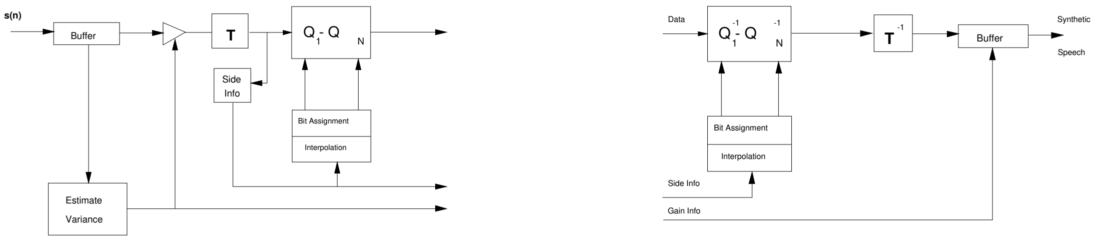


Fig. 13. The Adaptive Transform Coder (ATC) [326].

The DCT not only is near optimal, but it also has a transform spectrum which exposes the formant and pitch structure of speech [303], much like the DFT spectrum. In addition, a filter bank interpretation for the DCT is available and provides additional insight and possibilities for redundancy removal. For a unified treatment of the transform and sub-band coders we refer the reader to the excellent paper on frequency-domain coding by Tribolet and Crochiere [303].

Before we close this section, we note that a transform coder that is based on the WHT was developed in the early seventies by Shum et al. Additionally, analysis-synthesis of speech using mixed Walsh and Fourier transforms was reported by Mikhael and Spanias in [220]. The mixed Walsh and Fourier analysis-synthesis system is based on the notion that narrowband components of speech are represented by a set of narrowband Fourier functions, while broadband components are represented by broadband Walsh functions. Mixed Fourier/Walsh transform coding at 16, 9.6, and 4 kbits/s is given in [284] and [288].

3. Speech Coding Using Sinusoidal Analysis-synthesis Models

In this section we present another class of speech coders which relies on sinusoidal representations of the speech waveform. Our presentation starts with a description of some of the early digital speech analysis-synthesis methods which employed the Short-time Fourier Transform (STFT) [84][85][240][245][268]. We then continue with descriptions of the sinusoidal transform coders (STCs) [205]-[212] and the multiband excitation (MBE) coders [117][118][121] which were proposed in the eighties. Although sinusoidal and MBE coders rely heavily on speech properties (i.e., they are speech specific much like vocoders), they tend to be more robust than the traditional two-state voiced/unvoiced linear predictive vocoders in the sense that they work well with a wider class of signals.

3.1 Speech Analysis-synthesis using the Short-Time Fourier Transform

The basic assumption in transform-based speech analysis-synthesis is that the signal is slowly time-varying (quasi-stationary) and can be modeled by its short-time spectrum. The spectral analysis of speech can not be accommodated by the classical Fourier transform since the signal has time-varying characteristics. Instead, a time-frequency transformation is required. Time-varying spectral analysis [5][240][278] can be performed using the short-time Fourier transform (STFT). The analysis expression for the STFT is given by

$$S(n,\Omega) = \sum_{m=-\infty}^{\infty} s(m)h(n-m)e^{-j\Omega m} = h(n)*s(n)e^{-j\Omega n} \tag{13}$$

where $\Omega = \omega T = 2\pi f T$ is the normalized frequency in radians, and $h(n)$ is the sliding analysis window. The synthesis expression (inverse transform) is given by

$$h(n-m)s(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(n,\Omega)\ e^{j\Omega m}\ d\Omega \tag{14}$$

Note that if $n = m$ and $h(0) = 1$ [242][251], then $s(n)$ can be obtained from (14). The temporal and spectral resolution of the STFT are controlled by the length and shape of the analysis window. For speech, the length of the window is often constrained to be about 5-20 ms and hence spectral resolution is sacrificed. The sequence $h(n)$ can also be viewed as the impulse response of a linear time-invariant filter which is excited by a frequency-shifted signal (see equation 13). The latter leads to the filter-bank interpretation of the STFT, i.e., for a discrete frequency variable

$\Omega_k = \Delta\Omega\ k\ \{k = 0, 1, .., N-1\}$ and $\Delta\Omega$ and N chosen such that the speech band is covered, then the analysis expression is written as

$$S(n, \Omega_k) = \sum_{m=-\infty}^{\infty} s(m)h(n-m)e^{-j\Omega_k m} = h(n)*s(n)e^{-j\Omega_k n} \tag{15}$$

and the synthesis expression is

$$\tilde{s}_{STFT}(n) = \sum_{k=0}^{N-1} S(n, \Omega_k)\ e^{j\Omega_k n} \tag{16}$$

where $\tilde{s}_{STFT}(n)$ is the signal reconstructed within the band of interest. If $h(n)$, $\Delta\Omega$ and N are chosen carefully (see [266]) the reconstruction by (16) can be exact, i.e., $(\tilde{s}_{STFT}(n) = s(n))$. The k-th channel analysis-synthesis scheme is depicted in Fig. 14a, where $h_k(n) = h(n)e^{j\Omega_k n}$, $t(kn) = e^{j\Omega_k n}$, and $t(-kn) = e^{-j\Omega_k n}$. Figure 14b shows the analysis-synthesis including coding for a single channel.

The phase vocoder proposed by Flanagan and Golden [87] was perhaps the first attempt to represent speech explicitly in terms of its short-time magnitude and phase spectra. In the actual vocoder the derivative of the phase is coded and at the receiver the actual phase is retrieved by integration. Synthesis at the k-th channel is performed according to

$$\tilde{s}^k_{Ph}(n) = |S(n, \Omega_k)| \cos\left(\Omega_k n + \sum_{i=0}^{m} \Delta\phi(\Omega_k, i)\right) \tag{17}$$

where $\left\{\sum_{i=0}^{m} \Delta\phi(\Omega_k, i)\right\}$ represents the numerical integration of the phase derivatives.

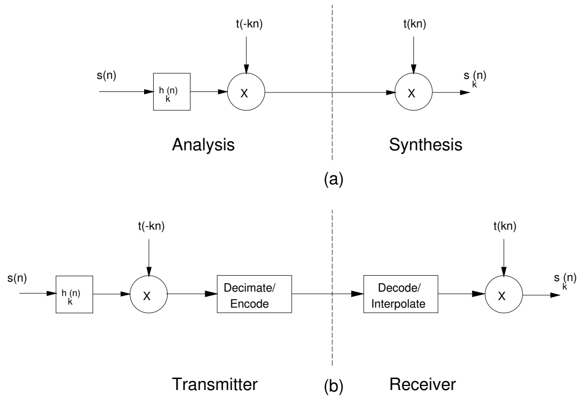


Fig. 14. The k-th channel of the analysis-synthesis filter bank.[251]

The phase vocoder was simulated [87] using a sixth-order Bessel sequence, $h(n)$, and 30 channels $(k = 1, 2, \ldots, 30)$ uniformly covering a 50 to 3050Hz bandwidth. One of the shortcomings of the approach is that (17) assumes a zero initial phase which often gives rise to reverberant speech quality. In a paper that appeared in 1973, Rabiner and Schafer [268] presented theoretical, as well as design considerations for STFT-based speech analysis-synthesis systems. In addition, the paper described an FFT-based method for computing the channel signals. Coding results were given for bit rates down to 14 kbits/s. Refinements of this STFT system were later proposed by Portnoff [243] and a weighted overlap-add method for analysis-synthesis was presented by Crochiere [50]. Speech coders using the STFT in conjunction with VQ were reported by Chang et al [36] and Loizou and Spanias in [187][188]. An in-depth treatment of the STFT and its application to time-frequency representation of speech is given in the papers of Portnoff [240]-[243].

3.2 Sinusoidal Analysis-Synthesis Models for Speech Coding

Speech modeling using unitary sinusoidal transforms implies that the speech waveform is represented by a set of narrowband functions. The phase vocoder [87] was perhaps the first attempt to represent speech as a combination of narrowband components. Sinusoidal representations are also implied in a number of transform-based analysis-synthesis systems such as [84][85][195][225][240][243]. Explicit sinusoidal models for speech were used in [6][124][202][205][203][245]. Hedelin proposed [124] a pitch-independent sinusoidal model for coding the baseband signal for speech coding. The amplitudes and phases in Hedelin's coder are determined using a Kalman estimator. Almeida and Silva proposed [6] a pitch-dependent speech compression system that encodes a set of harmonically related sine waves along with a residual waveform. George and Smith [94][95] also used the sinusoidal model for coding at 4.8 kbits/s and proposed a perceptually weighted closed-loop analysis-by-synthesis method for estimating the sinusoidal parameters. Another approach that uses a harmonic sinusoidal model to represent the narrowband speech components and a sub-set of Walsh functions to reproduce the broadband components was proposed by Spanias and Loizou [288]. The magnitudes and phases of the harmonic components and the amplitudes of the Walsh components were vector quantized at 4 kbits/s.

The most comprehensive study on sinusoidal modeling of speech, which also included successful low-rate representations, was presented in a series of papers by McAulay and Quatieri [205]-[212]. The sinusoidal model proposed by McAulay and Quatieri represents speech by a linear combination of L sinusoids with time-varying amplitudes, phases, and frequencies, i.e.,

$$\hat{s}_{SR}(n) = \sum_{k=1}^{L} A_k \cos(\Omega_k n + \phi_k) \qquad (18)$$

where the subscript $_{SR}$ stands for sinusoidal representation. Note that the number of sinusoids, L, is time-varying. The opportunity to reduce the bit rate using this model stems from the fact that voiced speech is typically highly periodic and hence it can be represented by a constrained set of sinusoids. In addition, the statistical structure (short-time spectrum) of unvoiced speech can be preserved by a sinusoidal model with appropriately defined random phases [212]. The sinusoidal model can be related to the source-system model (Fig. 3) by replacing the simplified two-state (impulse-train and random) excitation by a more general excitation model which consists of L sinusoidal components of arbitrary amplitudes, frequencies, and phases. The output of the vocal tract filter (synthetic speech) at steady-state can then be written as per (18). The basic assumption here is that the parameters of the sinusoidal model are slowly time-varying relative to the duration of the vocal tract impulse response.

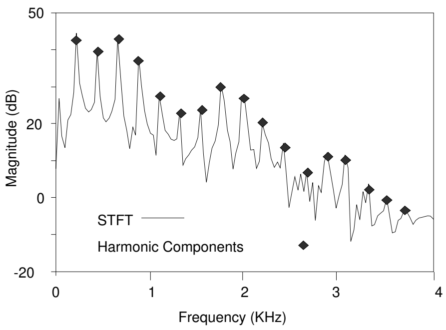
McAulay and Quatieri have shown that high quality reconstruction can be achieved by using sinusoids with amplitudes, frequencies, and phases corresponding to the peaks of the STFT. A Hamming window width that is 2.5 times the average pitch is adequate and ensures that the sine-waves are well resolved. Moreover the sinusoidal model is weakly dependent on pitch and voicing since the average pitch is used only to determine the length of the analysis window.

The sinusoidal model was presented above in its most general form. The main contributions in the work of McAulay and Quatieri lie in the analysis of a minimal parameter sinusoidal model and also in the development of algorithms for tracking the sinusoidal parameters from frame to frame. Firstly, since the number of sinusoids changes with the pitch, the concept of "birth" and "death" of sinusoidal components was established to allow for dynamic parameter matching. In addition, novel amplitude and phase interpolation algorithms were developed to match these parameters from one frame to the next. These algorithms are described in [205]. Experiments with the sinusoidal model revealed that as many as eighty sinusoids can be used for synthesis. These experiments were performed using an adaptive Hamming window, having width 2.5 times the average pitch, and a 1024-point FFT which was updated every 10ms. The sinusoidal analysis-synthesis system performed very well with a large class of signals (multiple speakers, music, biological sounds, etc), as well as with speech in the presence of background noise.

For low-rate coding applications the frequencies of the sine waves can be restricted to be integer multiples of the fundamental (pitch) frequency, i.e.,

$$\hat{s}_{HR}(n) = \sum_{k=1}^{L(\Omega_o)} A_k \cos(k\Omega_o n + \phi_k) \tag{19}$$

where the subscript $_{HR}$ stands for harmonic representation, Ω_o is the fundamental frequency, and $L(\Omega_o)$ is the number of harmonics within the speech bandwidth of interest (usually within 4 kHz). The harmonic representation provides an "optimal" set of frequencies only for perfectly voiced segments. The underlying assumption for voiced speech is that the pitch period is constant over the duration of the analysis frame [205]. For unvoiced speech, a set of sinusoids which are equidistant in frequency will generally preserve the statistics (short-time spectral density) of unvoiced segments [306]. The assumption for unvoiced speech is that the frequencies of the sinusoids are close enough [204] such that they follow the changes of the measured short-time spectral density. The fact that an equidistant set of sinusoids is used is advantageous in two ways, namely, a) the sinusoidal parameters can be conveniently estimated by sampling a high resolution DFT at the peaks for voiced speech (Fig. 15a), and b) the frequencies of the sinusoids can be encoded efficiently.



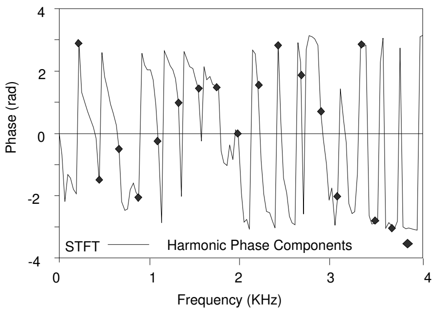


Fig. 15. Magnitudes and Phases of Harmonic components of the DFT for a voiced segment.

A simple example of the reconstruction of a voiced speech segment by a linear combination of harmonically related sinusoids is shown in Fig. 16. The speech segment was formed using a 32ms rectangular window and the amplitudes and phases of the sine waves were estimated from the peaks of the DFT of the segment (Fig. 15).

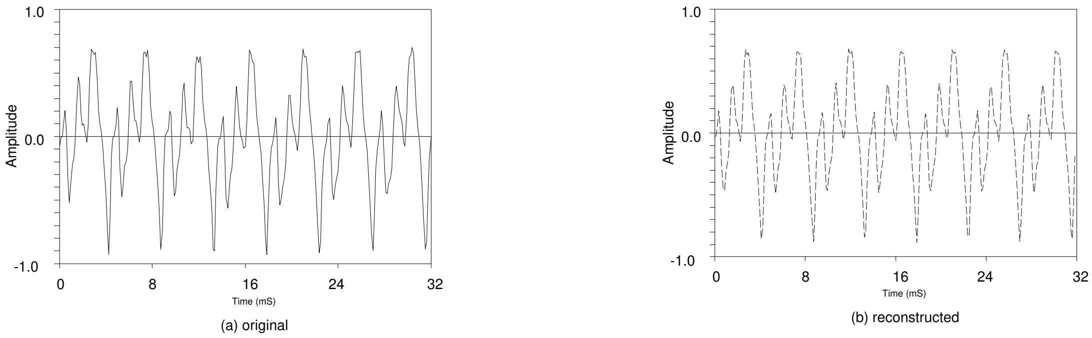


Fig. 16. Signal reconstruction from a harmonic set of sinusoidal components.

Sinusoidal models have been applied successfully to low-rate speech coding. In one of their early attempts to demonstrate low-rate sinusoidal representations, McAulay and Quatieri developed a sinusoidal coder operating at 8 kbits/s [204]. This is based on the basic sinusoidal analysis-synthesis system [205] shown in Fig. 17. The frame rate in the 8 kbits/s sinusoidal coder was 50Hz and 3.6 kbits/s were used to code the amplitudes and 4.4 kbits/s were used to code the phases and the fundamental frequency. For voiced speech the frequencies of the sinusoids were forced to be harmonic of the pitch-fundamental. For unvoiced speech the frequencies were equi-distant but closely spaced (distance less than 100Hz). Pitch and voicing information were derived using a sinusoidal pitch estimator [208]. The amplitudes of the sinusoids were coded using delta-modulation techniques across the frequency [126]. The bit-allocation rules were pitch-adaptive, i.e., for high-pitch speakers more bits per amplitude were used. Phases were coded using 4-5 bits and for low-pitch speakers phase regeneration techniques were used. The sinusoidal coder produced high-quality speech even in the presence of background noise.
In 1987, a multirate sinusoidal coder operating from 1.8 to 8 kbits/s was also developed at MIT Lincoln Labs [206]. The multirate coder employed a coding strategy similar to that of the JSRU vocoder [126]. At rates lower than 4.8 kbits/s phase information was not coded explicitly. Instead, a synthetic phase model was used which consisted of phase-locking all the component phases to the fundamental, adding voicing dependent random phase, and using quadratic pitch-dependent phase dispersion techniques [206].

Although the aforementioned early coding efforts have demonstrated the utility of the sinusoidal model in speech coding, the coded parameter set used was found to be sensitive to both quantization noise and channel errors particularly at the lower rates. A new robust parametric representation of the sinusoidal analysis-synthesis model was developed more recently and was

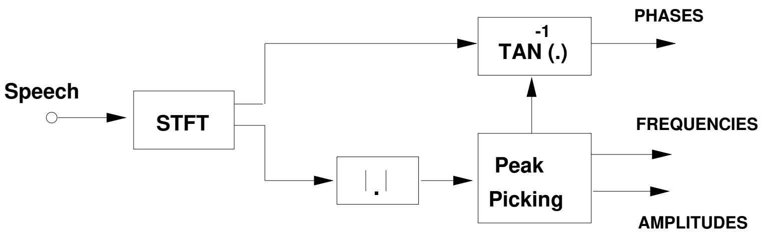


Fig. 17.a. Sinusoidal Analysis [205].

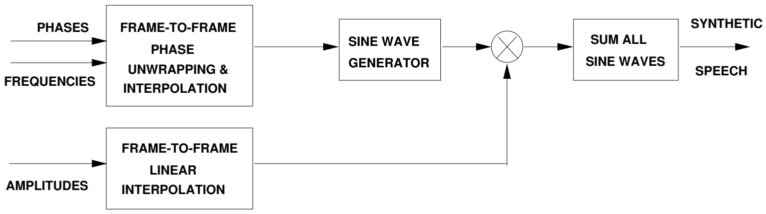


Fig. 17.b. Sinusoidal Synthesis [205].

shown [91] to be amenable to high-quality low-rate speech coding. The parametric model used in the low-rate sinusoidal transform coder (STC) uses a harmonic model for the frequencies, a pitch-adaptive amplitude envelope for the sine wave amplitudes, and a voicing dependent model for the sine wave phases.

The envelope used for the amplitudes is similar to the envelope used in Paul's Spectral Envelope Estimation Vocoder (SEEVOC) [237]. The SEEVOC envelope, $A(\Omega)$, is constructed by applying linear interpolation between the STFT peaks which are at the frequencies $(\Omega_1,\ \Omega_2, \ldots.)$. These peaks are determined as follows. The first peak and its frequency, Ω_1, are determined by searching the interval $[\overline{\Omega}_o/2,\ 3\overline{\Omega}_o/2]$ where $\overline{\Omega}_o$ denotes the average pitch. The rest of the peaks are determined by searching the intervals $[\Omega_{k-1}+\overline{\Omega}_o/2, \Omega_{k-1}+3\overline{\Omega}_o/2]$ until the edge of the band-width. The SEEVOC peak-picking algorithm has two inherent advantages over other peak-picking methods, namely a) it is only weakly dependent on the pitch (since it uses $\overline{\Omega}_o$), and b) it avoids low-level peaks (which are often due to spectral leakage).

A truncated cepstral representation is obtained for the spectral envelope $A(\Omega)$, i.e.,

$$\log(A(\Omega)) = Ce_0 + 2\sum_{m=1}^{\infty} Ce_m \cos(m\Omega) \tag{20}$$

where Ce_m is the cepstral sequence given by

$$Ce_m = \frac{1}{\pi}\int_0^{\pi} \log(A(\Omega))\cos(m\Omega)d\Omega \tag{21}$$

The cepstral representation is used because: a) when used along with frequency warping (MEL scale) it provides a basis for exploiting the perceptual properties of the ear, b) it provides sinusoidal phase information under the assumption that the system function (composite vocal tract and glottal pulse functions) is minimum phase, and c) transformed cepstral coefficients can be encoded efficiently. The perceptual properties of the human ear are exploited by applying a frequency warping function of the amplitude envelope. The warping function is linear in the low-frequency region and exponential in the high-frequency region. Details on the design and utility of the warping function are given in [212].

The phase model used in the STC assumes a source-system representation of speech and considers the phase components due to the excitation, the glottis, and the vocal tract. The phase model includes the system (composite glottal and vocal tract) phase and a linear phase component which ensures that the excitation sinewaves are in-phase with the pitch pulse. The system phase can be estimated, under the assumption of minimum phase, from the cepstral sequence, i.e.,

$$\Phi_s(\Omega) = -2\sum_{m=1}^{\infty} Ce_m \sin(m\Omega) \tag{22}$$

Notice that the same system amplitude and phase is identical for either $s(n)$ or $-s(n)$. Therefore the system phase is generalized, i.e., it is written as $\Phi_s(\Omega)+\beta\pi$, where β can take the value of zero or one. Comparisons of the measured phase with estimates obtained from the phase model revealed that the phase residual is essentially zero for voiced speech (up to a certain frequency) and random on $[-\pi,\pi]$ for unvoiced speech. Based on the above the following voicing-dependent phase model was adopted for the STC:

$$\hat{\phi}_k = -n_o\Omega_k + \Phi_s(\Omega_k) + \beta\pi + e_\phi(\Omega_k) \tag{23}$$

where n_o is the estimate of the pitch onset time and $e_\phi(\Omega_k)$ is the voicing-dependent residual phase introduced. The process for estimating n_o and β are given in [212]. The residual phase is defined such that the sinusoidal excitation model operates as a mixed-excitation model, e.g., of the type proposed by Makhoul et al [194]. The voicing-dependent residual phase is formed by

$$e_\phi(\Omega) = \begin{Bmatrix} 0 & \text{if } \ \Omega \le \Omega_c(P_v) \\ U[-\pi,\pi] & \text{if } \ \Omega > \Omega_c(P_v) \end{Bmatrix} \tag{24}$$

where $U[-\pi, \pi]$ stands for uniformly random phase on $[-\pi, \pi]$, $\Omega_c(P_v)$ is a voicing dependent cutoff frequency, and P_v represents the probability that speech is voiced. The voicing probability P_v depends on the pitch estimator and empirical expressions for P_v are given in [212].

The low-rate STC, Fig. 18, is based on a harmonic representation and the amplitudes of the sinusoids are obtained by sampling the SEEVOC envelope. The harmonic model is "optimal" only for perfectly voiced speech and in order to accommodate unvoiced or mixed voiced-unvoiced speech the sinusoidal frequencies are determined using the following rule:

$$\Omega_k = \begin{Bmatrix} k\Omega_o & \text{for } k\Omega_0 \le \Omega_c(P_v) \\ k^{'}\Omega_o + (k - k^{'})\Omega_u & \text{for } k\Omega_o > \Omega_c(P_v) \end{Bmatrix} \tag{25}$$

where $\Omega_u = 2\pi 100T$, $\Omega_c(P_v)$ is constrained to be no smaller than $(2\pi 1500T)$, and $k^{'}$ is the largest value of k for which $k^{'}\Omega_o <= \Omega_c(P_v)$. Note that if $\Omega_o < \Omega_u$ then $\Omega_k = k\Omega_o$ for all the frequencies. Therefore, the frequencies of the sinusoids are harmonic for perfectly voiced speech and also when the pitch is less than 100 Hz. Otherwise, the frequencies are harmonic up to a cutoff frequency $\Omega_c(P_v)$ and aharmonic (but equidistant) above the cutoff frequency.

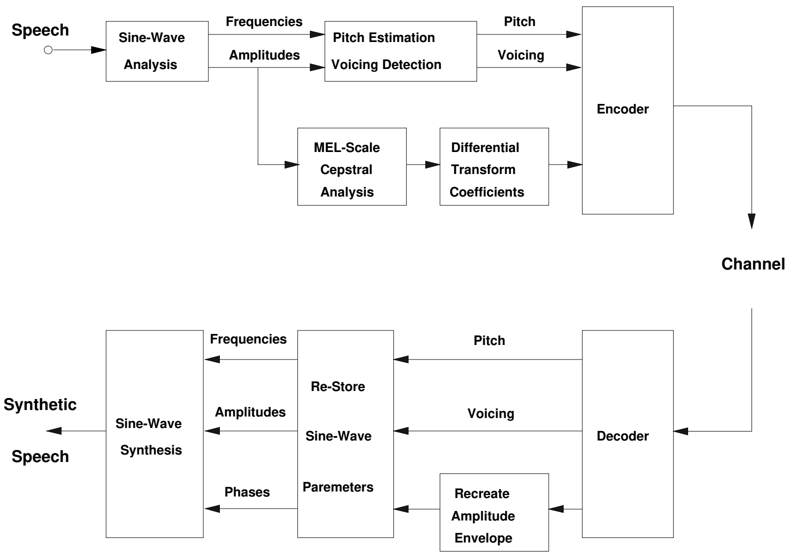


Fig. 18. The STC [210].

The STC, Fig. 18, encodes the pitch, voicing probability, and a transform representation of the cepstral coefficients. Because the cepstral sequency is largely uncorrelated and characterized by a large dynamic range, a DCT representation of the cepstral sequence is coded. The DCT coefficients can be encoded using DPCM techniques similar to those used in the JSRU vocoder [126] and the SEEVOC. In one of the most recent implementations of the STC [210][211], it was shown that robustness to channel errors can be improved by using vector quantization schemes similar to those suggested by Cox et al [49] in sub-band coders. Additional coding gain is realized using postfiltering [212] which reduces the quantization noise in the formant nulls. In addition frame fill [216] techniques are used to reduce further the bit rate. Fixed-point real-time implementations, on Analog Devices DSP chips, of a harmonic zero-phase $(\Phi_s(\Omega)=0)$ STC were described in [212]. More recently full-duplex STCs were implemented on the TI TMS3020C30 chip and the complexity reported was 13 MIPS while the MOS scores were 3.52 (4.8 kbits/s) and 2.9 (2.4 kbits/s) [213]. The STC has been included in Comtech's secure videophone.

Before we close this section we note that sinusoidal transform coding is currently one of the leading approaches to low-rate coding. We also note that the sinusoidal model was used to develop speech processing algorithms for time-scale modification, pitch estimation, and cochannel interference attenuation [208][246][247][248].

3.3 The Multiband Excitation Coder

The Multiband Excitation (MBE) coder, proposed by Griffin and Lim [117], relies on a model that treats the short-time speech spectrum as the product of an excitation spectrum and a vocal tract envelope. Although this type of spectral representation is implied in the classical two-state source-system model (Fig. 3), the difference here is that the excitation spectrum is modeled by a combination of harmonic and random-like contributions (i.e., voicing is frequency dependent). This mixed modeling approach is based on the fact that the spectra of mixed sounds or noisy speech contain both voiced (harmonic) and unvoiced (random-like) regions. Consequently the spectrum is divided into sub-bands and each sub-band is declared voiced or unvoiced. The number of sub-bands is much higher than the traditional sub-band coders [52] and can be chosen to be equal to the number of harmonics. The spectrum model in the MBE is essentially a generalization of the mixed source model suggested by Makhoul et al [194].

The synthetic speech spectrum is the product of a spectral envelope and an excitation magnitude spectrum, i.e.,

$$\hat{S}(\Omega)=H(\Omega)\ |X(\Omega)| \tag{26}$$

where $H(\Omega)$ is the vocal tract envelope and $|X(\Omega)|$ is the magnitude of the excitation spectrum. The vocal tract envelope $H(\Omega)$ is a smoothed version of the short-time speech spectrum and can be obtained by linearly interpolating between the harmonic samples of the spectrum [237]. The short-time spectra of the speech, the vocal tract envelope, and the excitation spectra are shown in Fig. 19.

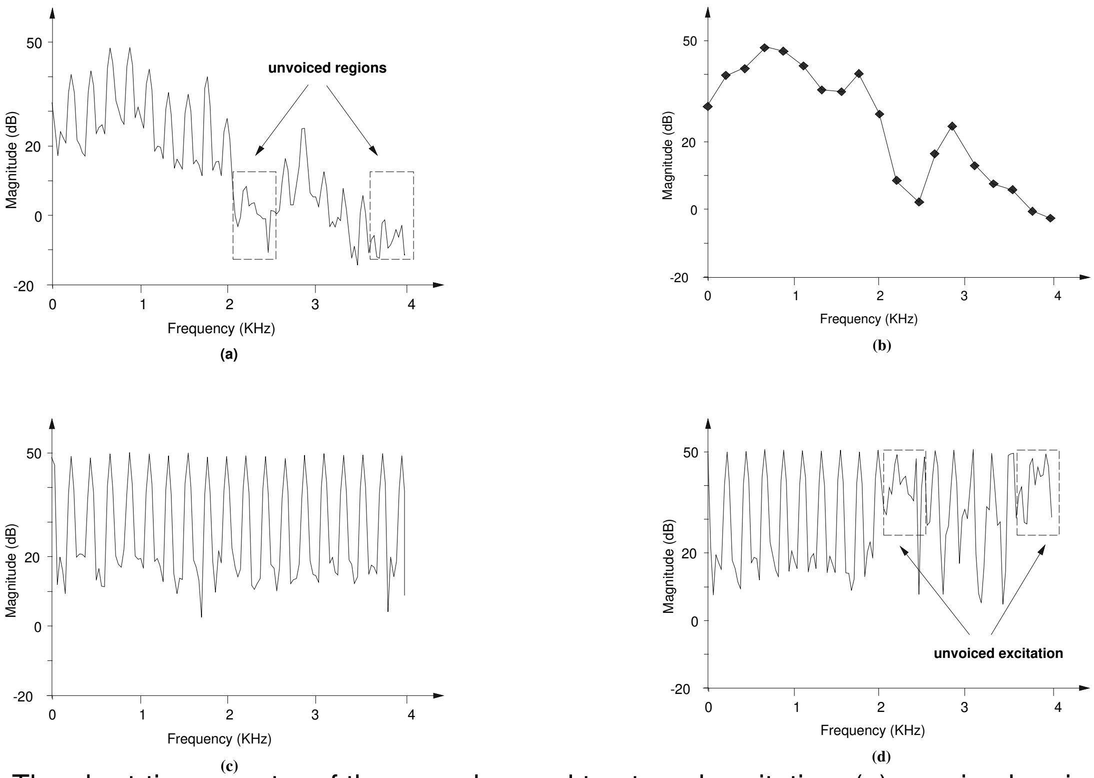


Fig. 19. The short-time spectra of the speech, vocal tract, and excitation: (a) unvoiced regions of the speech spectrum, (b) spectral envelope obtained by interpolation, (c) harmonic voiced excitation spectrum, (d) MBE voiced/unvoiced excitation spectrum.

The synthetic short-time spectrum, $\hat{S}(\Omega)$, is given by the product of the spectral envelope (Fig. 19b) and the MBE excitation spectrum (Fig. 19d). The parameters used in the MBE model are the fundamental frequency, the spectral envelope, voicing information per each harmonic, and phase information only for voiced harmonics. A least mean squares (LMS) analysis-by-synthesis approach is used to estimate $\hat{S}(\Omega)$, i.e., the excitation and envelope parameters are estimated simultaneously by minimizing the error (in the LMS sense) between the original and the synthetic spectra:

$$\varepsilon_{MB} = \sum_{m} \tilde{\varepsilon}_{MB}(m) \tag{27}$$

where the subscript $_{MB}$ stands for MBE and

$$\tilde{\varepsilon}_{MB}(m) = \frac{1}{2\pi} \int_{\phi_{1,m}}^{\phi_{2,m}} | S(\Omega) - \hat{S}(\Omega) |^2 \, d\Omega \tag{28}$$

and $\phi_{1,m}$ and $\phi_{2,m}$ are the boundaries of the m-th sub-band. If only magnitude information is desired then

$$\tilde{\varepsilon}_{MB}(m) = \frac{1}{2\pi} \int_{\phi_{1,m}}^{\phi_{2,m}} \{| S(\Omega) | - | \hat{S}(\Omega) |\}^2 d\Omega \tag{29}$$

The MBE model was described above in most general form. In practice, the voiced portion of the signal is generated in the time domain and the unvoiced signal is generated in the frequency domain, Fig. 20.

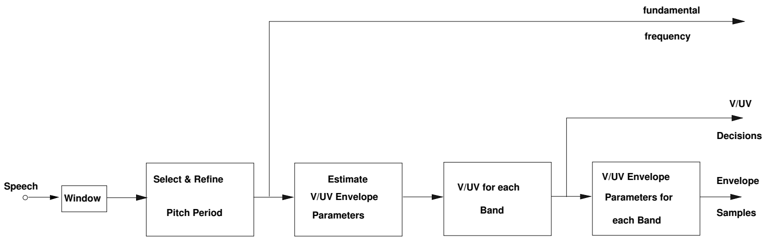


Fig. 20.a. MBE Analysis [117].

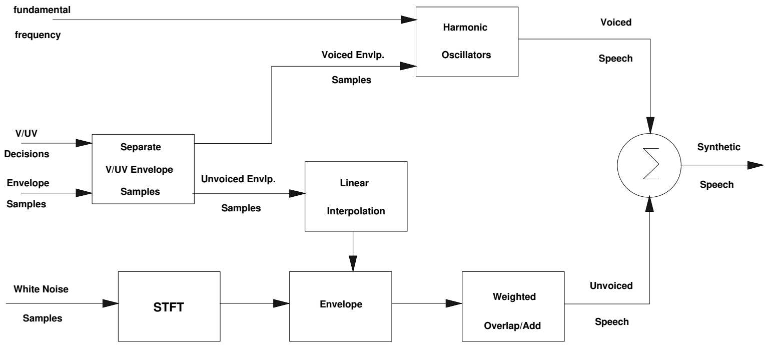


Fig. 20.b. MBE Synthesis [117].

The analysis process consists of determining: a) the pitch period, b) the voiced and unvoiced envelope parameters, c) the voicing information for each sub-band, and d) selecting the voiced and unvoiced envelope parameters for each sub-band. An integer pitch period is first estimated using an autocorrelation-like method and a pitch tracker is used to smooth the estimate for interframe continuity. This is then followed by a frequency-domain pitch refinement process. The details of the pitch estimation process are described in [117]. The spectral envelope is described by samples located at the harmonics of the fundamental. For voiced harmonics the magnitude and phases of the envelope samples are determined using a least-squares process. For unvoiced harmonics only the magnitudes are determined. The sub-bands are centered around the pitch-harmonic frequencies and voicing is determined by comparing the normalized error,

$$\bar{\varepsilon}_{MB}(m) = \frac{\tilde{\varepsilon}_{MB}(m)}{\int_{\phi_{1,m}}^{\phi_{2,m}} |S(\Omega)|^2 \, d\Omega} \tag{30}$$

to a threshold. When $\bar{\varepsilon}_{MB}(m)$ is below a threshold then the associated harmonic is marked voiced. A threshold value of 0.2 was suggested in [117]. It must be noted that in more efficient implementations voicing is typically determined for groups of three harmonics [118].

The synthesis process is shown in Fig. 20.b. The voiced portion of speech is synthesized in the time domain using a bank of harmonic sinusoids. The amplitudes of the sinusoids are obtained from the voiced envelope samples. Linear interpolation techniques, similar to the ones used in the STC, are used for the amplitudes. The amplitudes of the sinusoids associated with unvoiced harmonics are set to zero. The phases of the sinusoids (voiced bands) are determined using a phase prediction algorithm which is similar to the STC phase interpolation algorithm. The unvoiced portion of speech segments is determined by applying the FFT on a windowed segment of white noise. The normalized transform samples are then multiplied by the spectral envelope and unvoiced synthesis is performed using the weighted overlap/add method.

Initial results from coding the MBE parameters at 8 kbits/s are given in [117]. The frame rate in this implementation is 50 Hz and the harmonic magnitudes are coded using DPCM techniques similar to those used in the JSRU vocoder [126]. The difference between the estimated and predicted phases is coded only for voiced harmonics. When all the harmonics are unvoiced then no phase information is coded and the extra bits are allotted to the magnitudes. Voicing information is coded at one bit per sub-band. The average DRT scores for this implementation were: 96.2 for noiseless speech and 58 for speech corrupted by wideband noise [117].

A more efficient 4.8 kbits/s MBE coder was developed by Hardwick and Lim [120]. The MBE parameters, in this coder, were generated and quantized at a frame rate of 50 Hz. The voiced/unvoiced decision was encoded in groups of three harmonics each and a total of 12 voiced/unvoiced decisions per frame were encoded at 1 bit per group. For frames with more than 36 harmonics the rest of the harmonics are declared unvoiced. The pitch is encoded using a variable differential coding scheme with an average of six bits per frame. Phase is quantized only for the first 12 voiced harmonics using a phase prediction algorithm. The phase prediction residual is block quantized (in groups of 3 phases) at 2 bits per phase component. The rest of the phases, associated with voiced harmonics, are not coded and are chosen randomly using a uniform distribution. The phases of unvoiced harmonics are not needed for synthesis and hence they are not coded. The quantization of harmonic magnitudes is based on a hybrid time-frequency framework [120]. The differential magnitudes (across time in a dB scale) are formed and organized into 8-point blocks. Each block is then transformed using an 8-point DCT transform and the DCT coefficients are encoded using uniform quantizers. The 4.8 kbits/s coder was one of the candidates for the DOD FS1016 standard. The DRT/DAM scores reported [169] for the 4.8 kbits/s MBE were 92.7/60.4 and its complexity was estimated at 7 MIPS.

An improved MBE (IMBE) was proposed more recently by Hardwick and Lim [121]. Although the IMBE is based on the MBE analysis-synthesis model, the IMBE employs more efficient methods for quantizing the MBE model parameters. In addition, the IMBE coding scheme is more robust to channel impairments. A real-time implementation of a multirate (8, 4.8, 2.4 kbits/s) IMBE coder on the AT&T DSP32C signal processor was presented in [26]. An IMBE that operates at 6.4 kbits/sec recently became part of the Australian (AUSSAT) mobile satellite standard and the International Mobile Satellite (Inmarsat-M) [121][141][322] standard. The 2,250 bits/s of the Inmarsat-M IMBE are used for forward error correction. The remaining 4,150 bits/s are used for coding the IMBE parameters. The pitch period in the IMBE is quantized with one half sample accuracy at eight bits. A maximum of 12 voiced/unvoiced decisions were encoded; one bit per group of three harmonics. Differential (in time) log-amplitudes of the harmonics are divided into six blocks and each block is transformed using the DCT. The D.C. coefficients from each block are combined into a 6 element Prediction Residual Block Average (PRBA). The mean of the PRBA vector is subtracted from the vector and encoded using a 6-bit non-uniform quantizer. The zero-mean PRBA vector is vector quantized using 10-bit VQ. The rest of the DCT coefficients are quantized using scalar uniform quantizers. The phases of the harmonic components in IMBE are obtained using a phase prediction algorithm. The details of this algorithm, as well as the entire IMBE analysis-synthesis process are given in the Inmarsat-M standard documentation [141]. The 6.4 kbits/s IMBE was implemented on the AT&T DSP32C with an algorithmic delay of 78.75 ms and subjective evaluations revealed a MOS of 3.4.

4. Vocoder Methods

This section deals with speech-specific coders or vocoders. Unlike the waveform coders presented in the previous section, the performance of vocoders generally degrades for non-speech signals. Vocoders rely on speech-specific analysis-synthesis which is mostly based on the source-system model. Before we start our discussion on source-system models, however, we note that articulatory or physiological models have also been proposed for voice analysis-synthesis. Articulatory models represent the human speech production mechanism directly, i.e., distinct human voice-production organs are modeled explicitly. A time-frequency articulatory model for speech analysis-synthesis was proposed by Sondhi and Schroeter [282]. Even though the articulatory system exhibited potential for natural low-rate speech production, its practicality is limited because there are no efficient procedures for the extraction of articulatory parameters from the speech sequence. An excellent review of physiological models as applied to speech coding is given in Chapter 8 of [91].

The source-system representation has been utilized in most of the well known low-rate vocoders. Interestingly, many of the concepts associated with modern source-system vocoders have been around since Dudley's vocoder [66][67] (Fig. 2). In particular, two-state excitation (pulse/noise), voicing and pitch detection, and filter-bank spectral representation were implemented using analog components in Dudley's channel vocoder. Even though such basic concepts were introduced as early as 1939, it was also recognized early that the perceptual quality of speech depended a great deal on the estimation of the parameters associated with the spectral envelope and the excitation model. Reliable envelope estimators, based on linear prediction [190] and homomorphic signal processing [229], were proposed during the late sixties and early seventies. The estimation of the excitation parameters, however, proved to be more challenging and also more critical [270] for speech synthesis. Many of the low-rate channel and linear predictive vocoders employ the simple two-state excitation (pulse/noise). Although this simple excitation model is associated with attractive low information rates, it is also responsible for the synthetic ("buzzy") quality of speech. More sophisticated excitation models, proposed in the seventies and eighties, provided improved quality at the cost of increased complexity.

In the following, we describe four types of vocoders, namely, the channel and formant vocoders, the homomorphic vocoder, and linear predictive coders. Since most of the successful vocoders and hybrid coders make extensive use of linear prediction, the concepts associated with linear prediction will be presented in more detail.

4.1 The Channel and the Formant Vocoder

Dudley's vocoder [66][67] is the oldest method for speech analysis-synthesis. The channel vocoder relies on representing the speech spectrum as the product of vocal tract and excitation spectra. A vocal tract envelope representation is obtained using a bank of band-pass filters, Fig. 2. The number of channels is typically between 16 and 19 and the spectral representation becomes more accurate as the number of channels increases. The bandwidth of the channels is generally chosen to increase with the frequency. The fine structure of the voiced spectrum is represented using pitch-periodic pulse-like waves. Unvoiced speech is reproduced using noise-like excitation. An excellent description of the research on channel vocoders during the sixties is given in the classical paper by Gold and Rader [108].

Improvements in the channel vocoder were reported by Gold et al [110]. In particular, improvements in low-rate channel vocoders were realized by increasing the number of channels, introducing spectral flattening techniques on the excitation signal, and exploiting the correlation of the channel signals in the time and frequency domain using DPCM techniques. Gold et al also introduced a number of new applications of channel vocoders, such as, channel vocoding of the LP residual and noise suppression.

Perhaps one of the most comprehensive studies on channel vocoders was performed by the Joint Speech Research Unit (JSRU) of the U.K. In particular, JSRU produced a 2.4 kbits/s channel vocoder which was also evaluated by the U.S. DOD. The JSRU vocoder used a 19-channel filter bank and employed DPCM techniques across the frequency band to encode the channels. This coding technique was indeed very efficient and was later used in STC and in MBE coders. The 2.4 kbits/s JSRU vocoder scored a DRT of 87. It was also remarkable that the same coder scored a DRT of 81 in the presence of 5% transmission errors. An excellent description of the JSRU channel vocoder is given by Holmes [126].

The main difference between the channel vocoder and early formant vocoders is that the resonant characteristics of the filter-bank in formant vocoders adapt to the trajectories of the formants. More recent implementations of the formant vocoders employ cascade and parallel resonator configurations [127][234][252]. A general purpose cascade configuration of a formant vocoder is shown in Fig. 21.

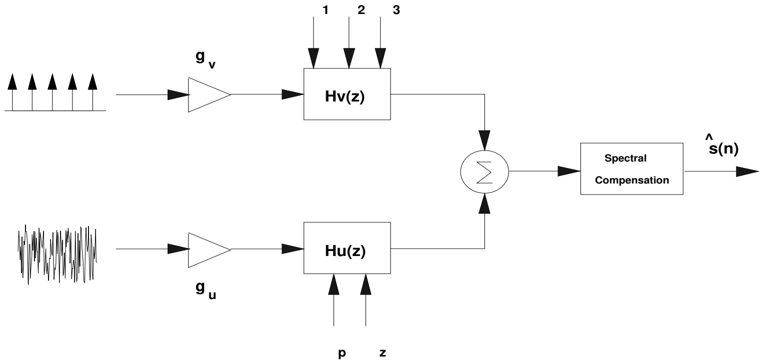


Fig. 21. A typical Formant Vocoder.

The transfer function, $H_v(z)$, for voiced speech synthesis consists of a cascade of three (or generally L) second order all-pole resonators, i.e.,

$$H_v(z) = \prod_{i=1}^{L} H_i(z) \tag{31}$$

where

$$H_i(z) = \frac{1 - 2e^{-\Omega_B(i)}\cos(\Omega_i) + e^{-\Omega_B(i)}}{1 - 2e^{-\Omega_B(i)}\cos(\Omega_i)z^{-1} + e^{-\Omega_B(i)}z^{-2}} \tag{32}$$

and Ω_i and $\Omega_B(i)$ denote the i-th formant frequency and bandwidth respectively. For unvoiced speech $H_u(z)$ consists of a cascade of a second order all-zero function (zero at Ω_z) and a second-order all-pole function (pole at Ω_p). The fixed spectral compensation function accommodates the effects of the glottal pulse and the lip radiation. The major difficulty in formant vocoders lies in the computation of the formants and their bandwidths.

Before we close our discussion on channel and formant vocoders, we note that although research on the channel and formant vocoders in the late eighties and early nineties has been (relatively) minimal, the work done earlier continues to provide valuable insight in the design of modern vocoders.

4.2 Homomorphic Vocoders

Homomorphic signal processing methods [251], such as homomorphic deconvolution, can be used both for vocal tract characterization [230], as well as, for extraction of information relating to excitation [227]. The basic idea in homomorphic vocoders is that the vocal tract and the excitation log-magnitude spectra can be combined additively to produce the speech log-magnitude spectrum.

A speech analysis-synthesis system that uses the cepstrum is shown in Fig. 22. The inverse Fourier transform of the log-magnitude spectrum of speech produces the cepstral sequence $Ce(n)$. It can be shown that the ("quefrency") samples of the cepstrum that are near the origin are associated with the vocal tract [41][229][230]. These coefficients can be extracted using a cepstral window, Fig. 22. The length of the cepstral window must generally be shorter than the shortest possible pitch period. It can be also shown that for voiced speech the cepstral sequence has large samples at the pitch period. Therefore the fundamental frequency can be estimated from the cepstrum [227].

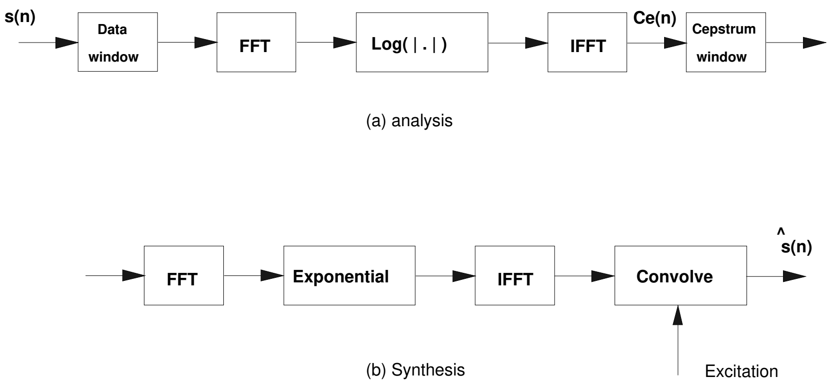


Fig. 22. A Homomorphic Speech Analysis-synthesis System.

The synthesizer takes the FFT of the cepstrum and the resulting frequency components are exponentiated. The IFFT of these components gives the impulse response of the vocal tract which is convolved with the excitation to produce synthetic speech. Although the cepstral vocoder did not find many applications at the time it was proposed, cepstrum-based methods for pitch and vocal tract estimation found many other speech processing applications. In addition, in recent contributions by Chung and Schafer [44][45] it was reported that good quality speech at 4.8 kbits/s can be produced by combining homomorphic deconvolution with analysis-by-synthesis excitation modeling.

4.3 Linear Predictive Vocoders

Linear predictive vocoders have been the most widely researched techniques for speech coding in the last twenty years. This section describes the concepts associated with linear predictive coders. The section starts with a description of the source-system model and continues with a discussion on linear prediction algorithms. Finally, we present the two excitation models for open-loop linear predictive coding, namely, the two-state excitation model and the residual excitation model.

A linear speech production system has been developed and analyzed by Fant [75], Fig. 23a. This system assumes a two-state excitation (impulse-train for voiced and random noise for unvoiced speech). The vocal tract is modeled as an all-pole transfer function which is formed by a cascade of a small number of two-pole resonators representing the formants. The glottal model is represented as a two-pole low-pass filter and the lip-radiation model is represented by $L(z) = 1 - z^{-1}$. Finally, a spectral correction factor is included to compensate for the low-frequency effects of the higher poles. In digital speech representations the spectral correction is omitted [249] and the zero of the lip-radiation function is essentially cancelled by one of the glottal poles [199]. Hence the system can be reduced to an all-pole model, Fig. 23b.

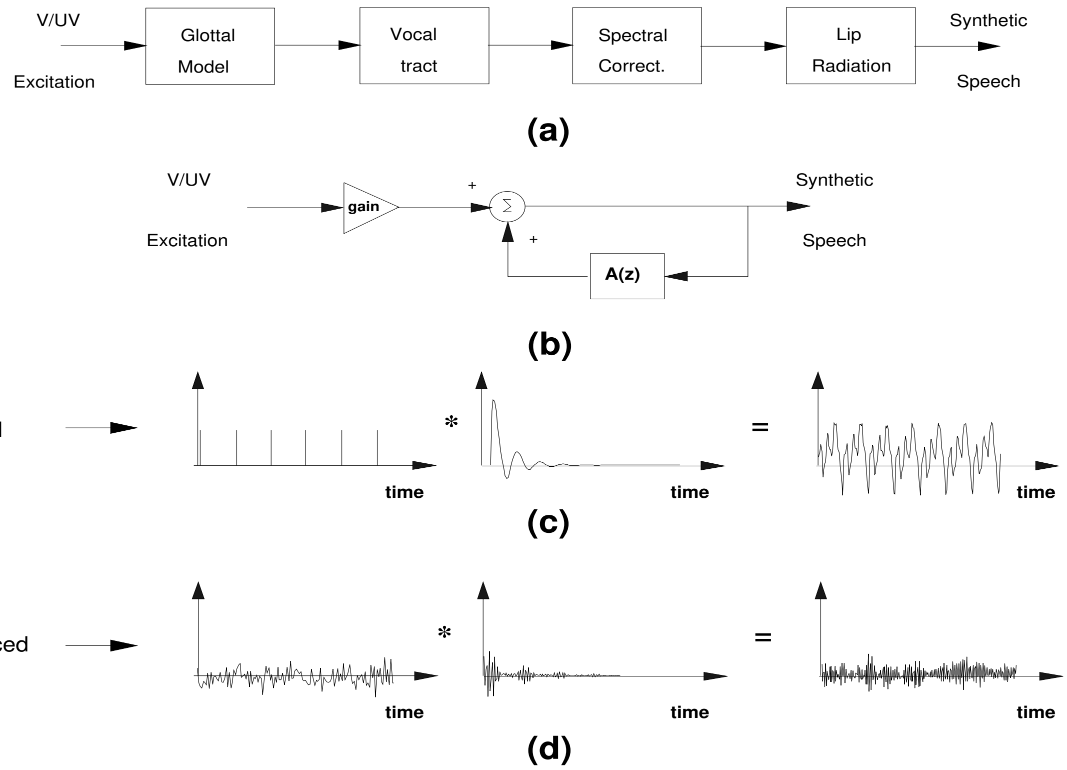


Fig. 23. Linear Speech Models and Voiced/Unvoiced Speech Representations: (a) Fant's Speech Production Model, (b) All-Pole Source-System Model, (c) Graphical Representation of Voiced Speech Production, (d) Graphical Representation of Unvoiced Speech Production.

The source-system synthesis model can be represented using z-domain functions, i.e.,

$$S(z) = \frac{g}{1 - A(z)} X(z) \tag{33}$$

where g represents the gain and $A(z)$ is given in (4). Figures 23 c and d give a graphical time-domain description of voiced and unvoiced speech synthesis using the linear source-system model. The system and excitation parameters in (33) are unknown and must be determined from a finite set of speech samples. The coefficients of $A(z)$ are obtained using Linear Prediction (LP) [189][199]. In a p-th order forward linear predictor the present sample of the speech sequence is predicted from a linear combination of p past samples (Fig. 24a), i.e.,

$$\hat{s}^f(n) = \sum_{i=1}^{p} a_i^f \; s(n-i) \tag{34}$$

where $\hat{s}^f(n)$ is the predicted sample of the process. The superscript ^{f} denotes a forward prediction process. The prediction parameters are obtained by minimizing the mean square of the forward prediction error, ε^f, i.e.,

$$\frac{\partial \varepsilon^f}{\partial a_i^f} = 0 \quad \text{for} \quad i = 1, 2, .., p \tag{35}$$

where

$$\varepsilon^f = E[(e^f(n))^2] = E[(s(n) - \hat{s}^f(n))^2] \tag{36}$$

and $E[.]$ is the statistical expectation operator. This minimization yields a set of Toeplitz equations, that is for $m = 1, 2, .., p$:

$$r_{ss}(m) - \sum_{i=1}^{p} a_i^f \; r_{ss}(m-i) = 0 \tag{37}$$

where

$$r_{ss}(m) = E[s(n+m)s(n)] \tag{38}$$

is the autocorrelation sequence of the speech segment. The autocorrelation sequence $\{r_{ss}(m)\}$ can be estimated from N samples of speech $\{s(n),\; n = 0, 1, 2, \ldots, N-1\}$ using the unbiased estimator,

$$\hat{r}_{ss}(m) = \frac{1}{N - |m|} \sum_{i=0}^{N-|m|-1} s(n + |m|)s(n) \tag{39a}$$

Alternatively one may use a biased estimator

$$\tilde{r}_{ss}(m) = \frac{1}{N} \sum_{i=0}^{N-|m|-1} s(n + |m|)s(n) \tag{39b}$$

Biased estimators are often desirable (particularly for short records) because they tend to yield minimum phase polynomials when used in conjunction with (37).

It must be noted that for real-valued data and coefficients, the minimization of the mean square of the backward prediction error (Fig. 24b):

$$e^b(n) = s(n-p) - \sum_{i=1}^{p} a_i^b \ s(n-p+i) \tag{40}$$

yields a Toeplitz set of equations similar to (37) with backward prediction parameters $\{a_i^b,\ i = 1, 2, .., p\}$. It can be shown (under the assumption of stationarity) that $\varepsilon^f = \varepsilon^b$ and $a_i^b = a_i^f = a_i$. The term "backward" here does not imply backward adaptation (as presented in Section 2.1). It simply means that the oldest sample in the predictor memory is predicted from samples that are indexed later in the predictor memory (40). These samples are in the same analysis frame.

The system of equations given in (37) is Toeplitz and symmetric and can be solved efficiently using an order-recursive algorithm. A general order-recursive Toeplitz matrix inversion algorithm was developed by Levinson [181]. Durbin [69] developed a more efficient algorithm that exploits even further the structure of the equations (37) . The Durbin recursive coefficient update for the m-th order predictor $\{m = 1, 2, .., p\}$ is given by

$$\varepsilon^f(0) = r_{ss}(0) \tag{41a}$$

$$a_m(m) = \frac{r_{ss}(m) - \sum_{i=1}^{m-1} a_i(m-1) r_{ss}(m-i)}{\varepsilon^f(m-1)} \tag{41b}$$

$$a_i(m) = a_i(m-1) - a_m(m) a_{m-i}(m-1) \ , \qquad 1 \le i \le m-1 \tag{41c}$$

$$\varepsilon^f(m) = (1 - (a_m(m))^2) \varepsilon^f(m-1) \tag{41d}$$

where the integer (m) in the parenthesis denotes the order of prediction during the order recursion. A detailed description of the Durbin algorithm is given in [162] and [201]. The coefficients

$$k_m = a_m(m), \quad m = 1, 2, \ldots, p \tag{42}$$

are known as the reflection coefficients due to their interpretation as physical parameters of the acoustical tube model for speech [189].

The negated reflection coefficient, $-k_m$, is also called the partial correlation (PARCOR) coefficient because it represents the normalized correlation between $s(n)$ and $s(n-m+1)$ with the correlation of $s(n-1), \ldots, s(n-m+1)$ removed [24]. The reflection coefficients are used in the

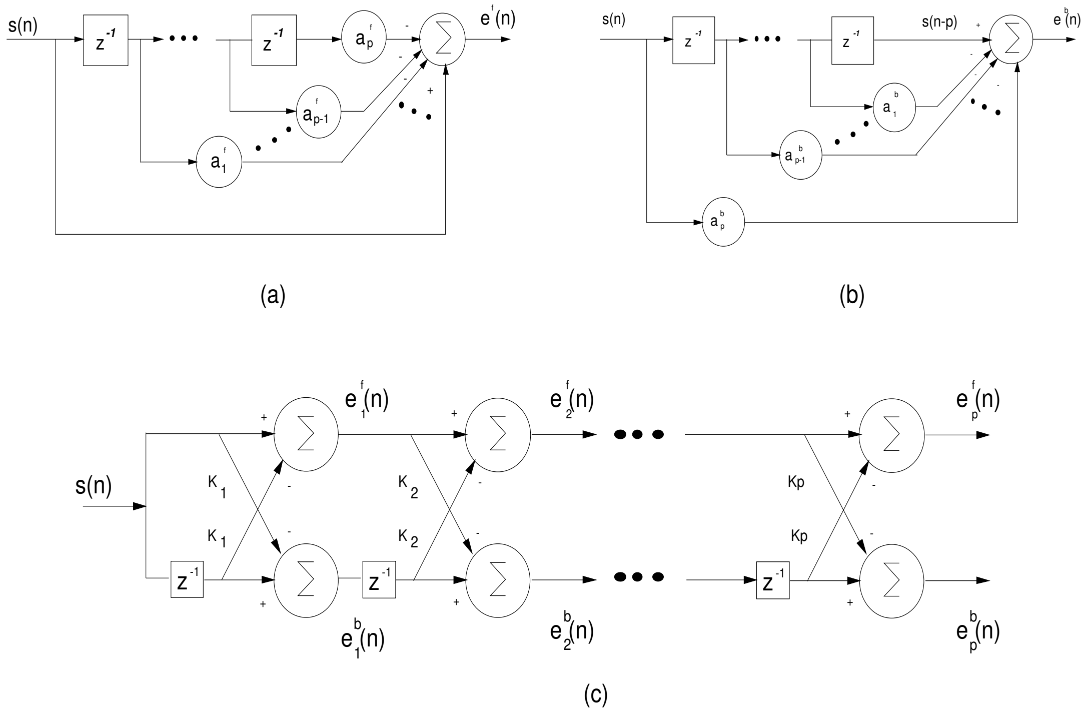


Fig. 24. Linear Prediction Realizations: (a) Direct Forward LP analysis, (b) Backward LP, (c) Lattice Forward-Backward Predictor.

lattice realization (Fig. 24c) of the forward and backward predictor. The lattice filter has implementation advantages over the direct realization in that it is less sensitive to round-off noise and coefficient quantization [144][309]. In addition, the condition $\{|k_m| < 1, \quad m = 1,2,..,p\}$ implies that the polynomial $(1 - A(z))$ is minimum phase (all its zeros are inside the unit circle) and hence stability in the all-pole synthesis filter can be checked by inspecting the reflection coefficients of the lattice predictor. Lattice parameters can be also determined using the harmonic algorithm of Burg [27][114], i.e.,

$$\hat{k}_m = \frac{2 \sum_{n=m+1}^{N} e_{m-1}^{f}(n) e_{m-1}^{b}(n-1)}{\sum_{n=m+1}^{N} | e_{m-1}^{f}(n) |^2 + \sum_{n=m+1}^{N} | e_{m-1}^{b}(n-1) |^2} \tag{43}$$

It can be shown that $\{|\hat{k}_m| \leq 1 \quad m = 1,2,..,p\}$ which is desirable for speech synthesis. Reflection to direct parameter transformations can be performed using (41c) and (42).

Notice that equation (43) computes the reflection coefficients directly from the lattice forward and backward error data. Algorithms that compute direct predictor parameters directly from short data records are also available (e.g., covariance and modified covariance methods [201][286]). In particular, the covariance method has been used in the government LPC-10 standard [301]. The covariance equations are given in the following

$$\begin{bmatrix} \bar{v}_1 \\ \bar{v}_2 \\ \bar{v}_3 \\ . \\ . \\ . \\ \bar{v}_p \end{bmatrix} = \begin{bmatrix} v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,p} \\ v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,p} \\ v_{3,1} & v_{3,2} & v_{3,3} & \cdots & v_{3,p} \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ v_{p,1} & v_{p,2} & v_{p,3} & \cdots & v_{p,p} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \\ . \\ a_p \end{bmatrix} \tag{44}$$

where

$$v_{i,j} = \sum_{n=p+1}^{N} s(n-i)s(n-j) \tag{45}$$

and

$$\bar{v}_i = \sum_{n=p+1}^{N} s(n)s(n-i) \tag{46}$$

Notice that the matrix in (44) is symmetric but not Toeplitz and therefore the Levinson algorithm can not be used for its inversion. The covariance matrix, however, can be inverted using the Cholesky decomposition [294]. The reflection coefficients are a by-product of the decomposition process [301].

One of the major issues in LPC is the quantization of the LP parameters [144][145][162][235][308]. In general 8-14 LP parameters are sufficient for vocal tract representation. Quantization of direct form coefficients is generally avoided since quantization error can lead to instability of the synthesis filter. On the other hand, quantization of the zeros of $(1-A(z))$ may be done such that the stability of the synthesis filter is ensured. The zeros, however, are more difficult to compute and they do not form an ordered set of parameters and hence it is difficult to develop statistical patterns for efficient quantization. The reflection coefficients are an ordered set of parameters and when coded at discrete levels within the range of -1 and 1 stability is ensured. In addition, their ordering can be exploited by encoding the first few reflection coefficients with higher precision. Transformation of the reflection coefficients can also lead to a set of parameters that are less sensitive to quantization. In particular, the log area ratios (LARs), given by

$$LAR(m) = \log\left\{\frac{1+k_m}{1-k_m}\right\} \tag{47}$$

have been used in many coding applications. In addition, the inverse sine transformation, i.e., $Si(m) = arcsin(k_m)$, has also been used in some predictive coding applications. An attractive feature of the inverse sine transformation is that the transformed parameters are bounded between $-\pi/2$ and $\pi/2$.

Another representation of LP parameters that has been widely used is the Line Spectrum Pairs (LSPs) [54][161][164][166][296]. The LSP representation of a typical 10-th order polynomial:

$$A_{10}(z) = 1 + a_1 z^{-1} + \ldots + a_{10} z^{-10} \tag{48}$$

can be explained by considering two polynomials $A_{10_1}(z)$ and $A_{10_2}(z)$ such that

$$A_{10_1}(z) = A_{10}(z) + z^{-11} A_{10}(z^{-1}) \tag{49a}$$

$$A_{10_2}(z) = A_{10}(z) - z^{-11} A_{10}(z^{-1}) \tag{49b}$$

Each of $A_{10_1}(z)$ and $A_{10_2}(z)$ has a set of five complex conjugate pairs of zeros that lie on the unit circle. Hence each polynomial can be represented by the five frequencies of the zeros (the other five frequencies are their negatives). These frequencies are the LSPs. The polynomial $A_{10}(z)$ is minimum phase if the roots of $A_{10_1}(z)$ and $A_{10_2}(z)$ alternate on the unit circle [296]. The polynomial $A_{10}(z)$ can be obtained using the LSPs and the relation

$$A_{10}(z) = \frac{A_{10_1}(z) + A_{10_2}(z)}{2} \tag{50}$$

The major advantage of the LSPs is that they can be coded using perceptual quantization rules. This is due to their relationship to the speech formants. The main drawback of LSPs lies in the complexity associated with their computation.

In linear predictive coding the analysis window is typically 20-30 ms long and parameters are generally updated every 10-30 ms. Since lower frame rates imply a large change in the LP parameters the frame is usually divided into subframes (typically 5ms long) and subframe parameters are obtained by linear interpolation [8] of adjacent frame parameters. It has been shown that the interpolation of prediction coefficients [8] and LARs [307] results in improved speech quality.

Although linear prediction has been by far the dominant method for extracting all-pole models for the vocal tract, there have also been many studies on pole-zero representations [15][218][219][293]. These studies were motivated by the fact that nasalities must be characterized by zeros in addition to poles. Atal and Schroeder [15] compared the pole-zero representation against the all-pole representation and reported improvements in the perceptual quality of speech for nasalized sounds. They also reported, however, that similar improvements were realized by increasing the order of the all-pole model. The latter is in agreement with the fact that spectral zeros can be represented by a large number of poles.

For a comprehensive treatment of linear prediction algorithms and their application to speech we refer the reader to the tutorial paper by Makhoul [189] and the book by Markel and Gray [199].

4.3.1 Open-loop Excitation Modeling

The ideal excitation for LPC synthesis is the prediction residual $x(n) = e^f(n)$. In the classical LPC this excitation is modeled by a pitch-periodic impulse sequence for voiced speech and a random noise sequence for unvoiced speech. Mixed excitation schemes in conjunction with LPC were proposed by Makhoul et al [194] and more recently revisited by McCree and Barnwell [214][215]. A different approach is taken in Residual Excited LP (RELP) where the baseband of the prediction residual $e^f(n)$ is encoded. In the following, we describe these excitation models and we also review some of the well known LPC and RELP algorithms. Analysis-by-synthesis excitation models such as those used in Multi-pulse LP and in Code Excited LP will be described in Section 5.

4.3.1.1 The Classical Two-State Excitation Model and the LPC-10 Algorithm

The two-state excitation parameters consist of the gain, the pitch period, and the binary voicing parameter. The gain of voiced and unvoiced segments is typically determined such that the energy of the synthetic speech segment matches that of the analysis segment. For unvoiced speech the excitation is produced by a random number generator. Since unvoiced segments are associated with small energy and large number of zero crossings, voicing can be determined by energy and zero-crossing measurements. In many cases voicing information is also provided by the pitch detection algorithm. The estimation of the fundamental (pitch) frequency presented a challenge in the early years, because it had to be determined by suppressing the higher harmonics using analog filters. Although this problem was partially solved using digital techniques [65][111][130][186][200][227][250][283], the estimation of the pitch period of noisy and non-perfectly periodic speech segments is still, to some extent, a difficult problem. Most of the well known algorithms for pitch detection appeared in the late sixties and in the seventies. In particular, we note the time-domain pitch tracker by Gold and Rabiner [111], the SIFT algorithm proposed by Markel [200], the cepstrum based algorithm by Noll [227], and the Average Magnitude Difference Pitch Extractor which is used in the LPC-10 algorithm. These algorithms are described in an IEEE collection of papers on speech analysis [269] and also in the book by Hess [125]. In very simple terms, the most straight-forward approach to pitch detection is based on peak-picking the autocorrelation sequence of center-clipped speech. A more expensive but also very robust pitch detector relies on peak-picking (within a range; see section 4.2) the cepstrum. The SIFT algorithm is based on peak-picking the autocorrelation sequence of the prediction residual associated with downsampled speech. Post-processing algorithms [125] for pitch smoothing are also used to provide for frame-to-frame pitch continuity. An excellent comparative study on these and other successful algorithms was presented by Rabiner et al [250]. Current pitch detection algorithms yield high-resolution (sub-sample) estimates for the pitch period and are often specific to the analysis-synthesis system. For example, in the STC and the IMBE coder the pitch estimation process is an integral part of the analysis-by-synthesis process. In addition, in most analysis-by-synthesis linear predictive coders the pitch is measured by a closed-loop process which accounts for the impact of the pitch on the overall quality of the reconstructed speech.

Real-time linear predictive coders were demonstrated in the early seventies. A 4.8 kbits/s LPC was implemented by Itakura and Saito [144] and tests revealed an average DRT score of 87.3. A real-time 3.6 kbits/s LPC was presented by Kang [162] in 1974 with a reported average DRT score of 87.6. In addition, a 600 bits/s linear predictive formant coder was presented by Kang and Coulter [163].

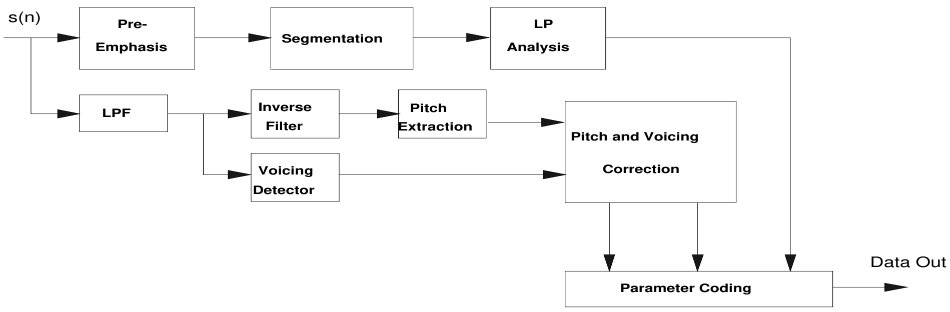
In 1976 a consortium established by the Department of Defense (DoD) recommended an LPC algorithm for secure communications at 2.4 kbits/s. The algorithm, known as the LPC-10, eventually became the Federal Standard FS-1015 [77]. The LPC-10 uses a 10-th order predictor to estimate the vocal-tract parameters. The prediction parameters are estimated by solving the covariance matrix equations (44). A block diagram for the transmitter and the receiver of the LPC-10 is shown in Fig. 25.

In order to reduce the effects of fixed-point arithmetic, speech (100-3600 Hz) is pre-emphasized using a first-order FIR high-pass filter. Segmentation and frame processing depends on voicing. For voiced speech the spacing between successive analysis intervals is a multiple of the pitch period and during unvoiced periods the frame is centered at 22.5 ms intervals. The entries of the covariance matrix are computed recursively prior to the Cholesky decomposition. Voicing and pitch information are estimated by operating on speech which was filtered by a low-pass filter with a cutoff frequency of 800 Hz. A second-order inverse filter is used to enhance the pitch estimator for input signals whose frequency content is low for frequencies below 300 Hz.

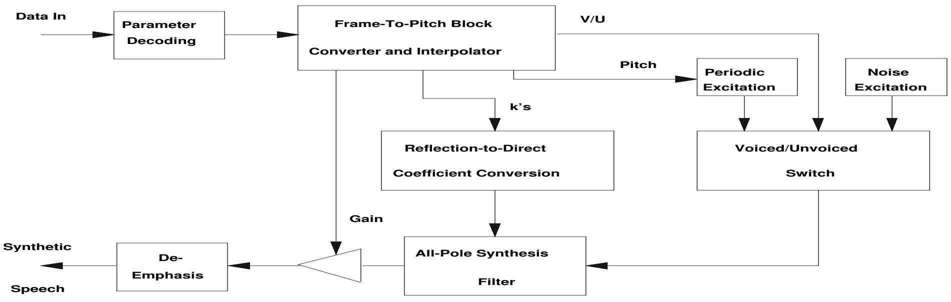
Pitch extraction is based on the average magnitude difference function (AMDF) of the filtered waveform. The AMDF is given by

$$AMDF(\tau) = \sum_{n=1}^{130,4} | s(n) - s(n+\tau) | \tag{51}$$

The pitch resolution is 1 for values of τ between 20 and 39, 2 for $40 \le \tau \le 78$, and 4 for $80 \le \tau \le 156$. Voicing is estimated using energy measurements, zero-crossing measurements, and the maximum to minimum ratio of the AMDF.



(a) Transmitter



(b) Receiver

Fig. 25. The Federal Standard FS1015: (a) LPC-10 Transmitter, (b) LPC-10 receiver.

Pitch and voicing are smoothed using a dynamic programming algorithm and encoded at seven bits. The excitation signal for voiced speech in the LPC-10 consists of a sequence that resembles a sampled glottal pulse. This sequence is defined in the standard [77] and periodicity is created by a pitch-synchronous pulse-repetition process. Gain information is transmitted by encoding a root mean square (RMS) parameter (obtained from $\overline{v}_0$) at five bits per frame. The coding of the reflection coefficients depends on voicing. For voiced speech segments, ten reflection coefficients are encoded while for unvoiced segments only four reflection coefficients are coded. The first two reflection coefficients are encoded as LARs at five bits each. The DRT and DAM score for the

LPC-10 were found to be 90 and 48 respectively for noiseless speech. The DRT score under the presence of typical airborne command post noise was 82.

An enhanced LPC-10 algorithm, the LPC-10e, was proposed in the mid-eighties [31]. This incorporates some of the LPC analysis and synthesis enhancements proposed by Kang et al [167]. The DRT score associated with the LPC-10e is 89.9. Creating better sounding vocoders operating at 2400 bits/s is still a challenging problem and the U.S. Government is seeking a new and improved algorithm [320] to replace the LPC-10. In addition, a great deal of effort is directed towards developing vocoders operating at very-low-rates [42][80][128][184][239][262][302][321]. In particular, current research efforts in military secure communications are focussed upon developing vocoders operating at rates between 400-800 bits/s with speech quality comparable to the FS1015 LPC-10. These vocoders target applications that require increased tolerance to bit errors, low probability of intercept, and narrowband voice/data integration by using very-low-rate algorithms with existing 2.4 kbits/s channels. Excellent overviews on very-low-rate secure military communications are given by Kang and Fransen [166], Jaskie and Fette [148], and Weinstein [319].

4.3.1.2 Mixed Excitation Models

The development of mixed excitation models in LPC was motivated largely by voicing errors in the two-state excitation model which degrade the speech quality and intelligibility and also by the inadequacy of the two-state excitation model in cases of voicing transitions (mixed voiced-unvoiced speech) or weakly voiced speech. The problem of mixed excitation in channel vocoders was examined by Fujimura [89] and later studied by Makhoul et al [194] for the case of linear predictive coders. Makhoul et al proposed a mixed source model, Fig. 26a, where the impulse train (buzz) excites the low-frequency region of the LPC synthesis filter and the noise excites the high-frequency region of the synthesis filter. The excitation filters and gains are chosen such that the overall excitation spectrum is flat. The same time-varying cutoff frequency (fc) is used for both excitation shaping filters. This is estimated using a peak-picking algorithm that determines the region of the spectrum associated with strong periodicity. Makhoul et al reported that mixed excitation lead to reduction of buzziness and raspiness (breathiness) in synthetic speech.

A more elaborate mixed excitation model was proposed by McCree and Barnwell [214][215]. The excitation shaping is done using first order FIR filters ($(H1(z)$ and $H2(z))$ with time-varying parameters. The mixed source model also uses (selectively) pulse position jitter for the synthesis of weakly periodic or aperiodic voiced speech. An adaptive pole-zero spectral enhancer is used to boost the formant frequencies. Finally a dispersion filter is used after the LPC synthesis filter to improve the matching of natural and synthetic speech away from the formants.

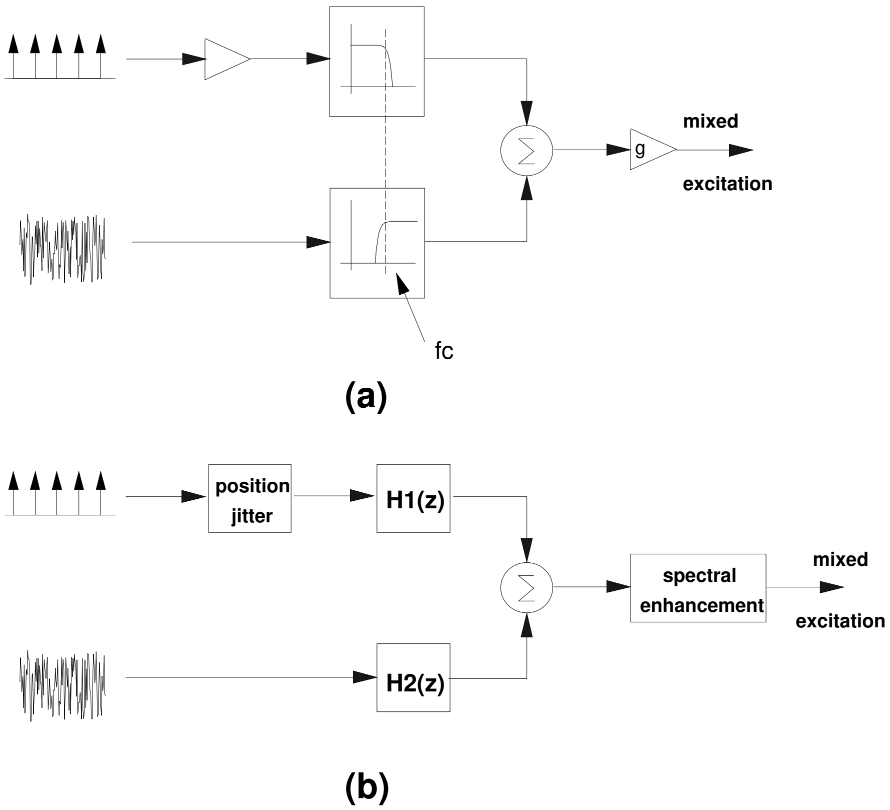


Fig. 26. (a) Mixed excitation model proposed by Makhoul et al [194], (b) (a) Mixed Excitation model proposed by McCree and Barnwell [215].

A 2400 bits/s mixed excitation LPC vocoder was implemented real time on TMS320C30 DSP series chips. The mixed excitation LPC vocoder achieved a DAM score of 58.9 with clean speech and 41 with noisy speech. A 4800 bits/s implementation was also reported in [215] and achieved a DAM score of 61.6 with clean speech and 44 with noisy speech.

4.3.1.3 Residual Excited Linear Prediction

From our previous discussion it is apparent that the linear prediction filter acts as a short-term "decorrelator". Therefore (ideally) the prediction residual has a relatively flat power spectrum. The residual is the perfect excitation for the all-pole synthesis filter, Fig. 27, and a class of linear predictive vocoders [2][43][79][122][304][328] relies on encoding this signal efficiently. The residual excitation signal essentially carries all the information that has not been captured by LP analysis, e.g., phase and pitch information, zeros due to nasal sounds, etc. Although the concept of coding the prediction residual is also utilized in ADPCM and in Adaptive Predictive Coders, RELP is different in that the residual encoding is based on spectral rather than waveform matching. In addition, RELP coders rely on the fact that the low-frequency components of speech are perceptually important.

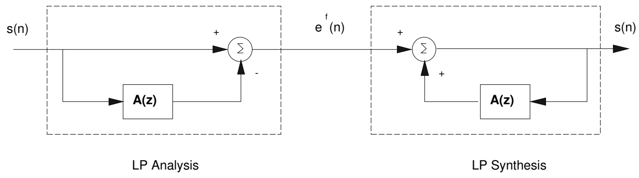


Fig. 27. Analysis-synthesis Using the Prediction Residual.

A RELP vocoder operating between 6 and 9.6 kbits/s was proposed in the mid-seventies by Un and Magill [304], Fig. 28. This RELP vocoder compresses the bandwidth of the residual to 800 Hz thereby coding only the baseband of the residual at 5 kbits/s. The residual is down-sampled and coded using ADM techniques. At the receiver the baseband residual is processed by a non-linear spectral flattener whose function is to regenerate the high-frequency harmonics. The excitation of the synthesis filter is derived by combining the flattened residual with an appropriate amount of white random noise.

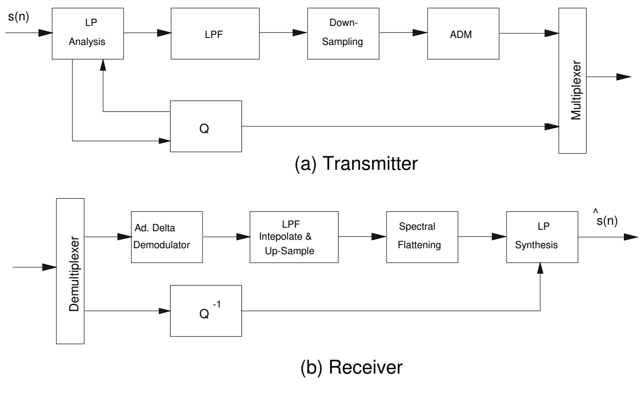


Fig. 28. The RELP Vocoder [304].

High-frequency regeneration can also be achieved by operating directly on the frequency components of the residual. The block diagram of a RELP vocoder which encodes the residual in the frequency domain using the FFT is shown in Fig. 29. In this system, the FFT of the residual is computed and the magnitudes and phases of the frequency components within the baseband (typically below 1 kHz) are encoded and transmitted. At the receiver a pitch-dependent high frequency "copy-up" procedure is performed to generate the high-frequency residual [165]. Notice that the quantization noise in the coded residual is spectrally shaped by the synthesis filter and therefore is masked (to a certain extent) by the speech. A RELP vocoder that employs long-term prediction and adaptive bit allocation was proposed by Fette et al [79]. This coder was also one of the candidates for the 4.8 kbits/s federal standard 1016. The DRT score obtained for this RELP was 91.7 [169]. We also note that a RELP coder that uses VQ for the encoding of the residual and LP parameters was proposed by Adoul et al [2].

In general, the speech quality of the RELP coder at rates above 4.8 kbits/s is higher than the analogous two-state excited LPC mainly because of the emphasis in the coding of the perceptually important residual components. The speech quality of RELP coders, however, is also limited by the information lost in the residual baseband filtering. The analysis-by-synthesis linear predictive coders presented in the next section avoid this problem by using efficient excitation models which can be optimized both for waveform matching and perception.

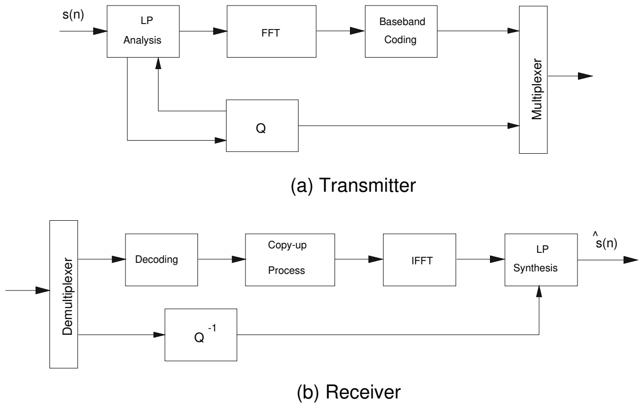


Fig. 29. The FFT-based RELP Vocoder [165].

5. Analysis-by-Synthesis Linear Predictive Coders

This section describes a class of source-system coders in which the system parameters are determined by linear prediction and the excitation sequence is determined by *closed-loop* or *analysis-by-synthesis* optimization. The optimization process determines an excitation sequence which minimizes a measure of the weighted difference between the input speech and the coded speech. The weighting or filtering function is chosen such that the coder is "optimized" for the human ear. The block diagram of a typical analysis-by-synthesis coder is shown in Fig. 30. The system consists of a short-term LP synthesis filter which represents the formant structure of speech, a long-term LP synthesis filter which represents the pitch (fine) structure of speech, a perceptual weighting filter, $W(z)$, which shapes the error such that the quantization noise is masked by the high-energy formants, and the excitation generator which forms or selects an excitation sequence such that the weighted Mean Squared Error (MSE) is minimized. The short-term predictor is typically updated at rates between 30 to 100 times per second while the long-term predictor is updated at higher rates, typically between 60 to 200 times per second [179]. Although the analysis-by-synthesis scheme shown in Fig. 30 is fairly common, some of the proposed coders employ closed-loop configurations without the long-term predictor, or with the short-term predictor preceding the long-term predictor [39]. It must be noted, however, that the use of a closed-loop long-term predictor (originally proposed in [280]) was a major milestone in speech coding and provided a big quality improvement.

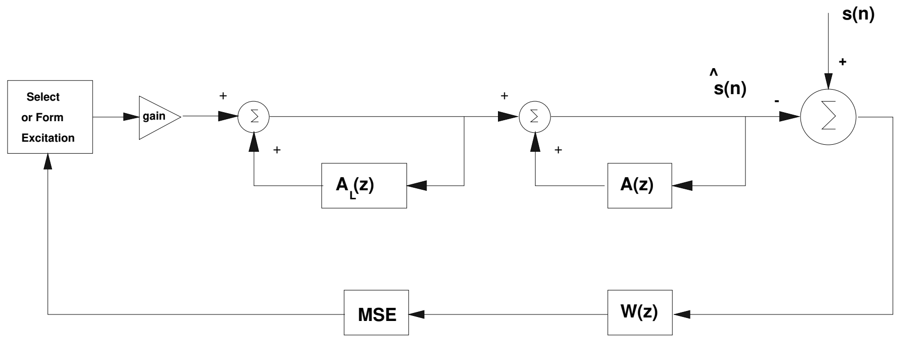


Fig. 30. A typical Analysis-by-Synthesis Linear Predictive Coder.

The three most common excitation models for analysis-by-synthesis linear predictive coding are: the multi-pulse model which was proposed by Atal and Remde [13], the regular pulse excitation model proposed by Kroon et al [178], and the vector or code excitation model which was originally proposed by Atal and Schroeder [272]. The excitation model in the Self Excited Vocoder (SEV), which was proposed by Rose and Barnwell [259][260], also received a lot of attention. In the latter the excitation is derived from a feedback loop by including explicitly the long-term predictor (LTP) in the excitation model. One of the drawbacks of this approach, however, is that it relies totally on feedback to form the excitation and hence it is sensitive to channel errors. Channel error effects may be reduced by resetting the LTP.

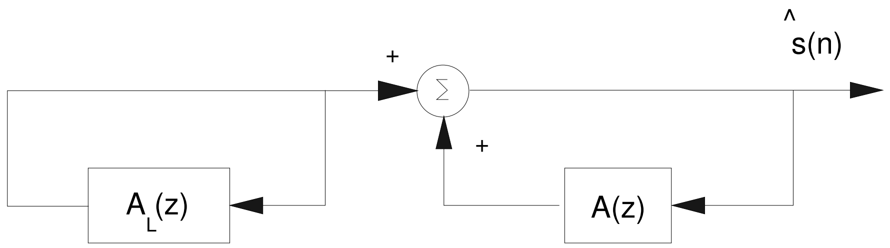


Fig. 31. Simplified synthesis scheme for the Self Excited Vocoder (SEV).

Note that the analysis-by-synthesis linear predictive coder is essentially a hybrid coder in the sense that it combines the features of model-based vocoders, by representing the formant and the pitch structure of speech, and the properties of waveform coders by providing for the matching of the input speech waveform. In addition, the properties of the human auditory system are exploited by incorporating perceptual weighting. In the following, we will describe speech coders based on multi-pulse excited linear prediction (MPLP), regular pulse excitation (RPE), and code excited linear prediction (CELP). In addition, we discuss the hybrid algorithms associated with the recent communications standards.

5.1 Multi-pulse Excited Linear Prediction

The MPLP algorithm forms an excitation sequence which consists of multiple non-uniformly spaced pulses, Fig. 32. During analysis both the amplitude and locations of the pulses are determined (sequentially) one pulse at a time such that the weighted mean squared error is minimized. The MPLP algorithm typically uses 4-6 pulses every 5 ms. The weighted error can be written as

$$e_M(n) = (s(n) - \hat{s}_M(n)) * w(n) \tag{52}$$

where the subscript $_M$ stands for multi-pulse, and $w(n)$ is the impulse response corresponding to the transfer function

$$W(z) = \frac{1 - A(z)}{1 - A(z/\gamma)} = \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} \gamma^i a_i z^{-i}} \tag{53}$$

The role of the γ is to de-emphasize the error energy in the formant regions. This deemphasis strategy is based on the fact that in the formant regions quantization noise is partially masked by speech [16][271].

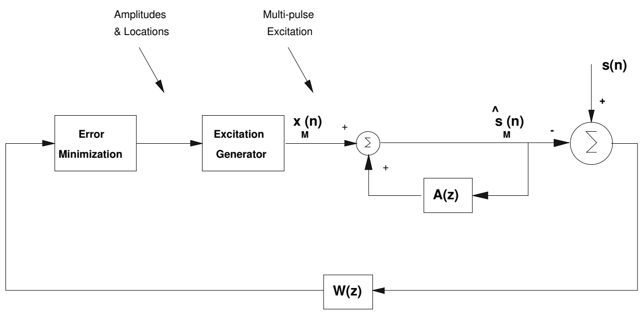


Fig. 32. MPLP Analysis.

The parameter $\{\gamma,\ 0 \le \gamma \le 1\}$ affects the bandwidth, Δf, of the peaks (poles) of $W(z)$ as follows

$$\Delta f = -\frac{1}{\pi T} \ln(\gamma) \qquad \text{(Hz)} \tag{54}$$

A typical value for γ is 0.8. The original MPLP proposed by Atal and Remde [13] considers only the short-term predictor and each pulse is obtained by minimizing

$$\varepsilon_M(k) = \sum_{n=0}^{N-1} (\bar{s}(n) - x_M(k) h_\gamma(n-k))^2 \tag{55}$$

where $x_M(k)$ is the amplitude of the pulse at location k, $h_\gamma(n)$ is the impulse response associated with the transfer function

$$H(z/\gamma) = \frac{1}{1 - A(z/\gamma)} \tag{56}$$

N is the number of samples per frame, and $\bar{s}(n)$ is the output of the filter with transfer function $H(z/\gamma)$ when the residual is the input. By setting

$$\frac{\partial \varepsilon_M(k)}{\partial x_M(k)} = 0 \tag{57}$$

we get

$$x_M(k) = \frac{\hat{r}_{\bar{s}h}(k)}{\hat{r}_{hh}(0)} \tag{58}$$

where $r_{\bar{s}h}(k)$ is the cross-correlation estimate between $\bar{s}(n)$ and $h_\gamma(n)$, and $r_{hh}(0)$ is an energy estimate of $h_\gamma(n)$. Substituting (58) in (55) we get

$$\varepsilon_M(k) = \left(\sum_{n=0}^{N-1} \bar{s}^2(n)\right) - \frac{(\hat{r}_{\bar{s}h}(k))^2}{\hat{r}_{hh}(0)} \tag{59}$$

The location k is obtained, in a suboptimal but efficient manner, by maximizing $((\hat{r}_{\bar{s}h}(k))^2/\hat{r}_{hh}(0))$. One pulse is computed per iteration and at each iteration $\bar{s}(n)$ is updated by subtracting the pulse determined in the previous cycle from the residual. Note that the initial residual is $e^f(n)$ and every time a new pulse is determined the residual is updated.

Excitation coding in the MPLP algorithm is more expensive than in the classical linear predictive vocoder because MPLP codes both the amplitudes and the locations of the pulses. The MPLP algorithm produced good quality speech at rates as low as 10 kbits/s [13][147]. One of the problems reported with MPLP, however, is that its performance generally deteriorates with high-pitch speakers. This problem was later addressed [280] using pitch prediction $(A_L(z))$ which yielded SNR improvements of about 6 to 10 dB. In particular, an SNR of 17 dB was reported at 10 kbits/s. Finally, we note that a 9.6 kbits/s MPLP algorithm is used in Skyphone airline applications [25].

5.1.1 The 9.6 kbits/s MPLP for the Skyphone Service

British Telecom International (BTI) introduced an aeronautical telephone service, known as skyphone, to accommodate passenger conversations and also provide additional communication capabilities between airlines and their aircraft. The service relies on digital transmission between aircraft and the BTI international exchanges and requires a speech coder in order to utilize the limited satellite power efficiently. The speech coder selected for this application is based on a 9.6 kbits/s MPLP algorithm developed by British Telecom Research Laboratories.

The MPLP algorithm for the Skyphone includes both short- and long-term prediction, Fig. 33. The parameters for the 10th-order short-term predictor are obtained using the autocorrelation method and the Levinson-Durbin recursion. The analysis window is formed using a 32ms Hamming window and the prediction parameters are updated every 20 ms. The short-term prediction parameters are coded in an ordered manner as inverse sine coefficients.

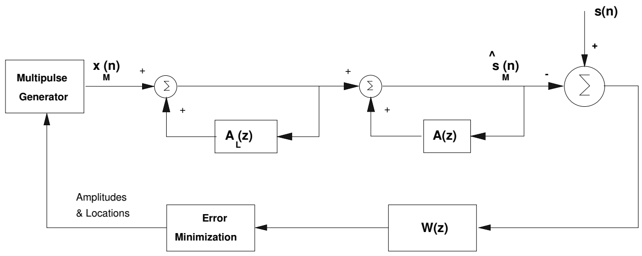


Fig. 33. MPLP for the 9.6 kbits/s Skyphone.

The long-term prediction parameters are determined from the short-term prediction residual using an open-loop configuration. The prediction residual, $e^f(n)$, is formed by filtering speech using the inverse filter $(1 - A(z))$. The long-term prediction (LTP) synthesis equation is given by

$$x_\tau(n) = x_M(n) + a_\tau x_\tau(n - \tau) \tag{60}$$

The LTP lag, τ, is determined by finding the maximum of the autocorrelation sequence $(\hat{r}_{ee}(m))$ of the prediction residual. The autocorrelation sequence is searched for 64 possible values (6 bits). The LTP gain, a_τ, is determined using the expression

$$a_\tau = \frac{\hat{r}_{ee}(\tau)}{\hat{r}_{ee}(0)} \tag{61}$$

and is coded using a non-linear quantizer. The pulse locations and amplitudes of the MPLP are determined using equations (58) and (59) and the minimization procedure described in Section 5.1. The impulse response, in this case, is the combined impulse response of the short- and the long-term predictors with $\gamma = 1$ (i.e., no perceptual weighting, $W(z) = 1$).

The Skyphone MPLP algorithm was implemented on an AT&T WE DSP32C signal processor and consumed about 75% of the available processing. Listening tests revealed an average MOS (for noiseless speech) of about 3.4 and a one-way delay that is less than 40ms. Additional tests have shown that the coder is capable of passing DTMF signalling tones and low-rate modem data.

5.2 Regular Pulse Excitation Coders

RPE coders [178] also employ an excitation sequence which consists of multiple pulses. The basic difference of the RPE algorithm from the MPLP algorithm is that the pulses in the RPE coder are uniformly spaced and therefore their positions are determined by specifying the location, k, of the first pulse within the frame and the spacing between non-zero pulses. A typical spacing factor for RPE is 3 to 4 and the location of the first pulse is generally updated every 5 ms. The number of pulses in RPE is typically between 10 and 13 per 5 ms segment and their amplitudes are determined by solving a set of linear equations. Note that pulse spacing does not need to be coded since the number of pulses per frame is usually fixed.

The analysis-by-synthesis optimization in RPE algorithms considers an inverse filtering scheme where the residual is formed by exciting the short-term prediction filter, $(1-A(z))$, with the input speech. The residual is then represented by a regular pulse sequence which is determined by weighted error minimization, Fig. 34.

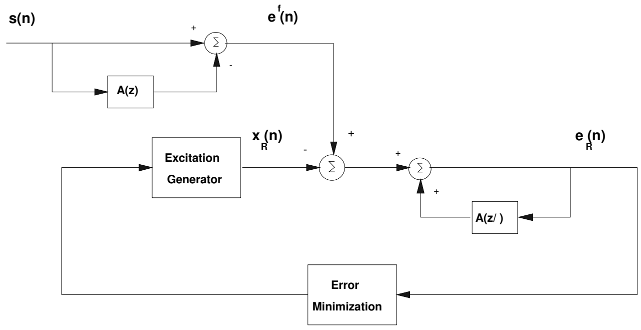


Fig. 34. RPE Analysis.

The minimization process can be described compactly using vector notation. For an N-sample speech frame the excitation vector corresponding to the $k-th$ Nx1 excitation vector, $\mathbf{x}_R(k)$, is written as

$$\mathbf{x}_R(k) = \mathbf{M}(k)\bar{\mathbf{x}}_R(k) \tag{62}$$

where the subscript $_R$ stands for regular pulse, $\bar{\mathbf{x}}_R(k)$ is a Jx1 vector containing the amplitudes of the J (J<N) non-zero pulses of the $k-th$ excitation vector, and $\mathbf{M}(k)$ is an NxJ *location* matrix which consists of ones or zeros and signifies the location of the J non-zero pulses in the k-th excitation vector. The Nx1 vector of excitation errors can be written as

$$\mathbf{e}_R(k) = \mathbf{e}_0 + \mathbf{H}\mathbf{e}^f - \mathbf{H}\mathbf{M}(k)\bar{\mathbf{x}}_R(k) \tag{63}$$

where $\mathbf{e}_0$ is the Nx1 vector containing the output of $H(z/\gamma)$ due to the initial filter state, $\mathbf{e}^f$ is the Nx1 vector containing the original prediction residual, and $\mathbf{H}$ is the NxN matrix that contains the impulse response samples corresponding to $H(z/\gamma)$, i.e.,

$$\mathbf{H} = \begin{bmatrix} h_\gamma(0) & 0 & 0 & \dots & 0 \\ h_\gamma(1) & h_\gamma(0) & 0 & \dots & 0 \\ h_\gamma(2) & h_\gamma(1) & h_\gamma(0) & \dots & 0 \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ h_\gamma(N-1) & h_\gamma(N-2) & h_\gamma(N-3) & \dots & h_\gamma(0) \end{bmatrix} \tag{64}$$

The minimization of

$$\varepsilon_R(k) = \mathbf{e}_R^T(k)\mathbf{e}_R(k) \tag{65}$$

w.r.t. the vector $\bar{\mathbf{x}}_R(k)$ gives

$$\bar{\mathbf{x}}_R(k) = (\mathbf{H}^T(k)\mathbf{H}(k))^{-1}\mathbf{H}^T(k)\ \bar{\mathbf{e}}^f \qquad k = 1,2,..,\frac{N}{J} \tag{66}$$

where $\mathbf{H}(k) = \mathbf{H}\,\mathbf{M}(k)$ and $\bar{\mathbf{e}}^f = \mathbf{e}_0 + \mathbf{H}\,\mathbf{e}^f$. Substituting (66) in (65) we obtain an expression for $\varepsilon_R(k)$ which is a function of $\bar{\mathbf{x}}_R(k)$. The $\bar{\mathbf{x}}_R(k)$ that minimizes $\varepsilon_R(k)$ over all *k* is the one selected.

Although the number of computations required for the selection of an excitation sequence seem to be fairly large, the matrices involved are quite structured and therefore computational complexity can be reduced considerably. In addition, under certain assumptions matrix inversion can be avoided [178] and the minimization problem can be written as

$$\min\{\varepsilon_R(k)\} \cong \max\{\bar{\mathbf{x}}_R^T(k)\bar{\mathbf{x}}_R(k)\} \tag{67}$$

Equation (67) simplifies the minimization problem a great deal and it has been shown [178] that the loss of performance is perceptually insignificant. The RPE selection process presented above included only the short-term predictor. The RPE coder that uses a one-tap pitch predictor was also described in [178] and was shown to improve significantly the performance of the RPE coder, particularly for female (high-pitch) speakers. The pitch synthesis filter is driven by the input $x_R(n)$ and its output, $x_\tau(n)$, is given by

$$x_\tau(n) = x_R(n) + a_\tau x_\tau(n-\tau) \tag{68}$$

For zero input $\{x_R(n),\ n \geq 0\}$ the prediction gain a_τ and the pitch period τ are determined in a closed-loop manner by minimizing

$$\varepsilon_L(\tau, a_\tau) = \sum_n (\bar{e}^f(n) - a_\tau \bar{x}_\tau(n))^2 \tag{69}$$

where the subscript $_L$ stands for long-term prediction, and $\bar{x}_\tau(n)$ represents the output of $H(z/\gamma)$ due to the input $x_\tau(n)$. The minimization approach consists of finding a_τ for all possible $\{\tau, 16 \leq \tau \leq 80\}$ and then selecting the prediction gain and the pitch period for which $\varepsilon_L(\tau, a_\tau)$ is minimal. Searches for τ between the range 20 to 147 (7 bits) have also been suggested [179].

Modified RPE coders have been proposed by Kang and Fischer [168]. In particular, Kang and Fischer proposed two new coding schemes for the excitation sequence in RPE coders, namely a Pyramid Vector Quantizer (PVQ) [81] and a Trellis Coded Quantizer (TCQ) [197]. Improvements in terms of SEGSNR of the order of 1 dB were reported for the PVQ and 1.3 dB for the TCQ.

5.2.1 The GSM RPE-LTP

A sophisticated 13 kbits/s coding scheme that uses RPE and LTP was adopted for the full-rate GSM Pan-European digital mobile standard. The RPE-LTP algorithm [307], Fig. 35, is briefly described in the following. Speech is sampled at 8 kHz and quantized at 13 bits per sample. The RPE-LTP processes speech in 20 ms frames (160 samples) which are in turn segmented into sub-frames of 5ms. The preprocessing stage in RPE-LTP involves pre-emphasis and DC offset compensation. The second stage involves an eighth order short-term prediction analysis which is performed every 20 ms. In order to allow for efficient quantization, the prediction parameters are encoded as LARs. The eight LARs are quantized in an ordered manner, i.e., the number of bits allotted from *LAR(1)* to *LAR(8)* are {6/6/5/5/4/4/3/3}. After decoding, sub-frame LARs are generated by interpolating (linearly) LARs from one frame to the next. The sub-frame LARs are converted back to reflection coefficients which are used to generate the prediction residual. The lag and gain parameters of the LTP are computed in every sub-frame and encoded at 7 and 2 bits respectively. The LTP operates on the LP residual and generates the LTP residual which is then filtered by a block FIR weighting filter. The filtered LTP residual is then down-sampled by a ratio of three. Four 13-sample sub-sequences are formed per sub-frame and the sub-sequence with the maximum energy is quantized using block adaptive PCM.

The RPE-LTP algorithm described above can also be viewed as a generalization of the decimation process in baseband RELP coders. The performance of the GSM codec in terms of MOSs was reported to be between 3.47 (min) and 3.9 (max) and its complexity is 5 to 6 MIPS. More details on the algorithm can be found in the ETSI standard [119] and information on the testing procedures for the selection of the algorithm is given in [307]. In addition, some of the issues associated with the real-time implementation of the speech and the channel coder for the GSM are examined in [290] and [289].

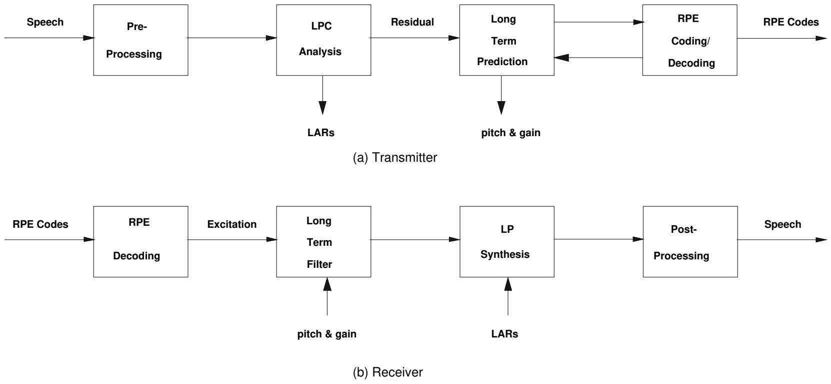


Fig. 35. The RPE-LTP scheme of the GSM standard.

5.3 Code Excited Linear Prediction

The two analysis-by-synthesis multi-pulse excitation schemes described in the previous Sections (5.1 and 5.2) achieve high-quality speech at medium rates. For low-rate high-quality speech coding a more efficient representation of the excitation sequence is required. This problem was addressed in a paper by Atal [7] in which it was suggested that high quality speech at low rates may be produced by using non-instantaneous (delayed decision) coding of Gaussian excitation sequences in conjunction with analysis-by-synthesis linear prediction and perceptual weighting. A novel vector excitation scheme for analysis-by-synthesis linear predictive coding was proposed in the mid-eighties by Atal and Schroeder [17][272]. The vector or code excited linear prediction (CELP) algorithm, Fig. 36, produced low-rate coded speech comparable to that of medium-rate waveform coders thereby bridging the gap between waveform coders and vocoders. The CELP algorithm encodes the excitation using a codebook of Gaussian sequences. The codebook used in [272] contains 1024 vectors and each vector is 40-samples (5ms) long. A gain factor scales the excitation vector and the excitation samples are filtered by the long- and short-term synthesis filters. The "optimum" vector is selected such that the perceptually weighted MSE is minimized.

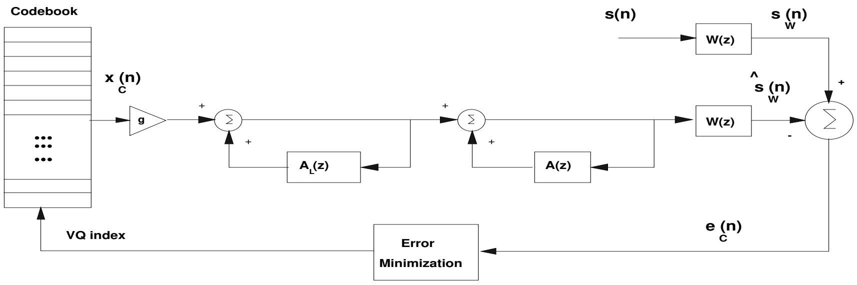


Fig. 36. Analysis-by-Synthesis CELP [272].

The search process in CELP can be explained by considering the analysis-by-synthesis scheme shown in Fig. 36. Notice that in order to facilitate the analysis that follows, the perceptual weighting is applied directly on the input and synthetic speech. The Nx1 error vector in this case is written as

$$\mathbf{e}_C(k) = \mathbf{s}_w - \hat{\mathbf{s}}_w^0 - g_k \hat{\mathbf{s}}_w(k) \tag{70}$$

where the subscript $_C$ stands for CELP, $\mathbf{s}_w$ is the Nx1 vector that contains the filtered speech samples, $\hat{\mathbf{s}}_w^0$ is the vector that contains the output due to the initial filter state, $\hat{\mathbf{s}}_w(k)$ is the filtered synthetic speech vector, and g_k is the gain factor. Minimizing $\varepsilon_C(k) = \mathbf{e}_C^T(k)\mathbf{e}_C(k)$ w.r.t. g_k we get

$$g_k = \frac{\bar{\mathbf{s}}_w^T \hat{\mathbf{s}}_w(k)}{\hat{\mathbf{s}}_w^T(k)\hat{\mathbf{s}}_w(k)} \tag{71}$$

where $\bar{\mathbf{s}}_w = \mathbf{s}_w - \mathbf{s}_w^0$. Considering (71), $\varepsilon_C(k)$ can be written as

$$\varepsilon_C(k) = \bar{\mathbf{s}}_w^T \bar{\mathbf{s}}_w - \frac{(\bar{\mathbf{s}}_w^T \hat{\mathbf{s}}_w(k))^2}{\hat{\mathbf{s}}_w^T(k)\hat{\mathbf{s}}_w(k)} \tag{72}$$

The $k-th$ excitation vector, $\mathbf{x}_C(k)$, that minimizes (72) is selected and its gain is obtained from (71).

Although long-term prediction parameters can be determined using an open-loop configuration, the closed-loop approach yields by far superior speech quality. In the closed-loop lag (τ) search the LTP parameters, a_τ and τ, are determined before the stochastic excitation parameters. The range of values for τ are typically within the integers 20 to 147. The search for an integer lag in CELP is generally similar to that described in Section 5.2 (see [179] for details). Note that if the pitch period is larger than the subframe length, then LTP filtering can be interpreted as an adaptive codebook (a terminology introduced in [171]). Other issues associated with

LTP were examined by Ramachandran and Kabal [254][255], Kroon and Atal [176], and Shoham [276]. Ramachandran and Kabal examined the LTP parameter estimation problem [255] and the issues associated with the stability [254] of the pitch prediction filter. Kroon and Atal proposed an efficient pitch predictor with sub-sample resolution (i.e., non-integer lags) which was shown to achieve similar or better performance than higher order integer-delay predictors. In addition, Shoham examined the efficiency of the LTP and the relationship between the prediction gain and the gain of the stochastic excitation. Shoham argued that the gain of the stochastic excitation must be constrained because it often contributes noise components that can not be removed by the synthesis filters. The constrained excitation algorithm proposed by Shoham was shown to perform better than the conventional CELP algorithm at the same rate.

Speech quality in CELP coders can be enhanced by applying post-filtering on the speech sequence. Post-filters [256] are used to emphasize the formant and the pitch structure of speech and a typical post-filtering configuration consists of a cascade of long- and short-term weighting filters with appropriate gains and bandwidth expansion parameters (γ).

One of the disadvantages of the original CELP algorithm is the large computational effort required for the codebook search. In fact, many of the CELP algorithms require processors capable of executing more than 20 MIPS and codebook storage of the order of 40 kbytes . This problem motivated a great deal of work focussed upon developing structured codebooks [4][60][100][170] and fast search procedures [300]. In particular, Davidson and Gersho [60] proposed sparse codebooks and Kleijn et al [170] proposed a fast algorithm for searching stochastic codebooks with overlapping vectors. In addition, Gerson and Jasiuk [100][101] proposed a vector sum excited linear predictive coder which is associated with fast codebook search and robustness to channel errors. The vector sum scheme will be discussed in more detail in Section 5.3.2. Other implementation issues associated with CELP include the quantization of the CELP parameters, the effects of channel errors on CELP coders, and the operation of the algorithm on finite-precision and fixed-point machines. A study on the effects of parameter quantization on the performance of CELP was presented in [176], and the issues associated with the channel coding of the CELP parameters were discussed by Kleijn in [172]. Some of the problems associated with the fixed-point implementation of CELP algorithms were presented in [287].

CELP coding achieves communications quality at 8 kbits/s and significant progress has been made towards achieving high quality at 4 kbits/s for the half-rate North American Cellular Telephone Standard. Several papers have appeared in the literature suggesting new techniques and strategies aimed at producing high-quality CELP coders operating below 4.8 kbits/s. Many approaches concentrated on dynamic bit allocation among the excitation and the LP parameters. Taniguchi et al [298] proposed a multi-mode coding scheme in which the speech parameters are coded using several coders with different bit allocations and the optimum coder is selected by minimizing a fidelity criterion. Similar ideas were proposed independently by Yong and Gersho [325] and Jayant and Chen [151]. Taniguchi et al [298] reported SEGSNR improvements of the order of 2 dB. A novel method for dynamic coding of the CELP parameters, based on phonetic segmentation, was proposed by Wang and Gersho [315]. This is called phonetically segmented vector excitation coder (PS-VXC) and classifies speech into three general phonetic categories, namely voiced, unvoiced, and onset (rapid transitions from unvoiced to voiced). The voiced category is further divided into four sub-categories, namely full-band steady-state, full-band transient, low-pass steady-state, and low-pass transient. The PS-VXC partitions input speech into contiguous segments and each segment contains an integer number of frames. Each segment is phonetically classified into one of the six categories and each frame is coded by a coder which has been tailored for a specific phonetic category. The coding strategy in PS-VXC not only controls the bit allocation for each category but it also determines the frame size, LP order, the LTP (on/off), etc. Speech coded with a 3.6 kbits/s PS-VXC was reported to be comparable to that of the conventional 4.8 kbits/s CELP coder.

Kleijn et al promoted [173] interpolative coding strategies as a means for reducing further the bit rate in analysis-by-synthesis linear predictive coders. The basic idea is to allow interpolation not only of short-term prediction parameters, which is customary in CELP coders, but also in excitation parameters such as the LTP parameters. For example, in CELP the LTP lag (τ) is updated every subframe. Since the pitch period in natural speech varies slowly it would seem that an intelligent interpolation scheme on the LTP lag may lead to coding efficiencies. Although interpolative coding approaches lead to coding gains they are associated with some loss of waveform matching. The notion of relaxing the strict requirements for matching the original speech waveform was exploited in an approach called generalized analysis-by-synthesis [174]. In generalized analysis-by-synthesis instead of using the original speech signal as a reference for waveform matching, the performance of the speech coder is evaluated over a multitude of modified speech waveforms that are perceptually similar to the original signal. The modified speech waveform that yields the best coding gain is selected and the parameters corresponding to this signal are transmitted. The potential for improved coding gain lies in using a set of model parameters, such as interpolated LTP parameters, to match the modified speech waveform. In that sense, generalized analysis-by-synthesis bridges the gap between waveform matching and interpolative coding. A 5.85 kbits/s relaxation CELP (RCELP) algorithm that is based on the concept of generalized analysis-by-synthesis was proposed in [175]. RCELP was implemented real-time on a single AT&T DSP32 chip and MOS tests revealed that RCELP coded speech was very similar to that of the 13 kbits/s GSM algorithm.

To summarize, CELP essentially broke the 9600 bits/s "barrier" which was considered for years as the lower boundary for communications quality speech. There are at least four CELP type of algorithms that are part of national and international communications standards. In the following, we describe briefly three CELP algorithms that have been standardized.

5.3.1 The Federal Standard 1016 CELP

A 4.8 kbits/s CELP algorithm [9][30] has been adopted by the Department of Defense for possible use in the third-generation secure telephone unit (STU-III). This algorithm is described in the Federal Standard 1016 [78] (FS1016) and was jointly developed by the DOD and AT&T Bell Labs. The candidate algorithms and the selection process for the standard are described in [169]. The synthesis configuration for the FS1016 CELP is shown in Fig. 37. Speech in the FS1016 CELP is sampled at 8 kHz and segmented in frames of 30ms. Each frame is segmented in sub-frames of 7.5 ms. The excitation in this CELP is formed by combining vectors from an adaptive and a stochastic codebook with gains g_a and g_s respectively (gain-shape VQ). The excitation vectors are selected in every sub-frame by minimizing the perceptually weighted error measure. The codebooks are searched sequentially starting with the adaptive codebook. The term "adaptive codebook" is used because the LTP lag search can be viewed as an adaptive codebook search where the codebook is defined by previous excitation sequences (LTP state) and the lag, τ, determines the specific vector. The adaptive codebook contains the history of past excitation signals and the LTP lag search is carried over 128 integer (20 to 147) and 128 non-integer delays. A sub-set of lags is searched in even sub-frames to reduce the computational complexity. The stochastic codebook contains 512 sparse and overlapping codevectors [170]. Each codevector consists of sixty samples and each sample is ternary valued (1,0,-1) [182] to allow for fast convolution.

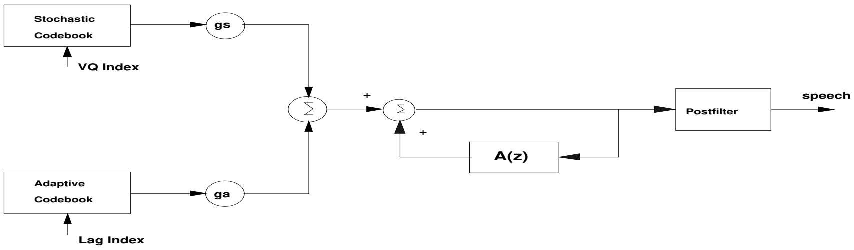


Fig. 37. FS1016 CELP synthesis [9].

Ten short-term prediction parameters are encoded as LSPs on a frame-by-frame basis. Sub-frame LSPs are obtained by applying linear interpolation of frame LSPs. A short-term pole-zero postfilter (similar to that proposed in [40]) is also part of the standard.

The details on the bit allocations are given in the standard. The computational complexity of the FS1016 CELP was estimated at 16 MIPS (for partially searched codebooks) and the DRT and MOS scores were reported to be 91.5 and 3.2 respectively.

5.3.2 Vector Sum Excited Linear Prediction (VSELP)

The Vector Sum Excited Linear Prediction (VSELP) algorithm was proposed by Gerson and Jasiuk [100] for use in digital cellular and mobile communications. An 8 kbits/s VSELP algorithm was adopted for the North American Digital Cellular System. The 8 kbits/s VSELP algorithm uses highly structured codebooks which are tailored for reduced computational complexity and increased robustness to channel errors. The VSELP excitation is derived by combining excitation vectors from three codebooks, namely, a pitch-adaptive codebook and two highly structured stochastic codebooks, Fig. 38.

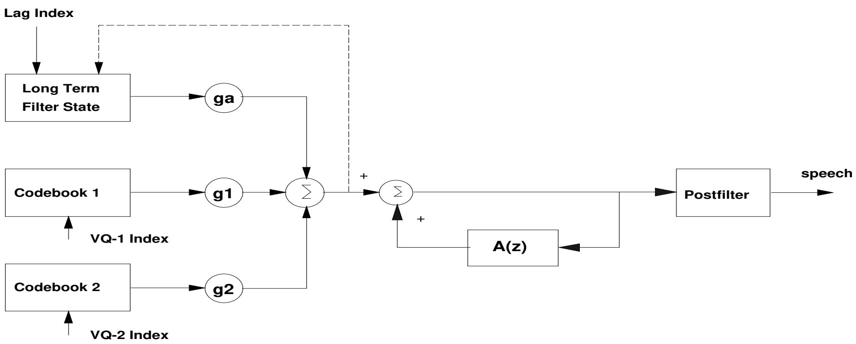


Fig. 38. VSELP synthesis [100].

Speech sampled at 8 kHz is first pre-processed using a fourth-order high-pass Chebyshev filter. The frame in the VSELP algorithm is 20 ms long and each frame is divided into four 5ms sub-frames. A 10-th order short-term synthesis filter is used and its coefficients are encoded as reflection coefficients once per frame with bit allocation {6/5/5/4/4/3/3/3/3/2}. Sub-frame LPC parameters are obtained through linear interpolation. The excitation parameters are updated every 5ms. The excitation is coded using gain-shape vector quantizers. The codebooks are searched sequentially and the codevectors are determined using closed-loop perceptually weighted MSE minimization. The long-term prediction lag (adaptive codebook) is searched first; assuming no input from the stochastic codebooks. The adaptive codebook is fully searched for lags of 20 to 146 (127 codes) and the 128th code for τ is used to indicate that the LTP is not used. The 128 40-sample vectors in each stochastic codebook are formed by linearly combining seven basis vectors $\mathbf{b}_1(k,m)$, i.e.,

$$\mathbf{x}_{V_1}(k,i) = \sum_{m=1}^{7} \theta_{i,m} \ \mathbf{b}_1(k,m) \tag{73}$$

The coefficients $\theta_{i,m}$ are allowed to take the values of one or minus one. The value of $\theta_{i,m}$ is one if the m-th bit of the codeword is 1, or -1 if the m-th bit of the codeword is 0. Note also that for every codevector its negative is also a codevector. Clearly the effect of changing one bit in the codeword, possibly due to a channel error, is not catastrophic since the codevectors corresponding to adjacent (gray code-wise) codewords are different only by one basis vector. The search of the codebook is also greatly simplified since the response of the short-term synthesis filter, to codevectors from the stochastic codebook, can be formed by combining filtered basis vectors. In addition, the codebook structure lends itself to an efficient recursive search process.

The sequential search of the codebooks also incorporates an orthogonalization process in which the filtered basis vectors of the first stochastic codebook are made orthogonal to the filtered LTP vector. Similarly the filtered basis vectors of the second stochastic codebook are made orthogonal to the previous set of filtered vectors. The orthogonalization essentially relies on a Gram-Schmidt approach [294]. The codebook gains, which are also determined from this process, are parameterized and jointly quantized using a vector quantizer. The implementation details of the algorithm, including codebook information, post-filtering, etc, are presented in the IS-54 standard [70]. Fixed-point implementations of VSELP were reported in [287] and [295]. The complexity of the 8 kbits/s VSELP was reported to be more than 13.5 MIPS (typical 20 MIPS) and the MOSs reported were 3.45 (low) and 3.9 (high).

Improvements on VSELP, in terms of complexity, were reported in [101]. In particular, it was shown that the computational effort in the lag-search can be significantly reduced by using a two-step estimation process. In the first step an initial estimate of the lag is obtained by searching the autocorrelation sequence. The final estimate is obtained by carrying a constrained closed-loop search around the neighborhood of the initial estimate.

VSELP algorithms operating at lower rates have also been proposed. In fact, a 6.7 kbits/s VSELP algorithm [102][103] was adopted for the Japanese digital cellular standard and VSELP algorithms are candidates for the half-rate North American and the GSM cellular standards.

5.3.3 The 16 kbits/s Low-Delay CELP

One of the problems in speech compression methods which utilize delayed-decision coders is that coding gain is achieved at the expense of coding delay. The one-way delay is basically the time elapsed from the instant a speech sample arrived at the encoder to the instant that this sample appears at the output of the decoder [38]. This definition of one-way delay does not include channel- or modem-related delays. The delay is basically attributed to data buffering, processing, and generation of coding bits (channel symbols). Roughly speaking, the one-way delay is generally between two and four frames. For example, a typical CELP algorithm with 20 ms frames is associated with a delay of about 60 ms [38]. Fast processing and encoders that transmit coding bits as they become available (on the fly) can reduce this delay.

In 1988 CCITT launched an investigation on low-delay speech coding algorithms operating at 16 kbits/s for possible standardization as a G. series recommendation for universal applications. This standardization effort motivated a great deal of research [38][58][64][107][146][198][299] on low-delay coding. Of all the available coders the low-delay CELP (LD-CELP) coder, proposed by Chen et al [38], was selected [35], Fig. 39.

The low-delay CELP coder achieves low one-way delay by: a) using a backward-adaptive predictor, and b) short excitation vectors (5 samples). In backward-adaptive prediction, the LP parameters are determined by operating on previously quantized speech samples which are also available at the decoder. The LD-CELP algorithm does not utilize LTP. Instead, the order of the short-term predictor is increased to fifty $(p = 50)$ to compensate for the lack of a pitch loop. The autocorrelation analysis for LP is based on a novel hybrid window which consists of recursive and non-recursive portions. The hybrid window is described in [38] and allows for efficient computation of the autocorrelation sequence using single-precision integer arithmetic. The frame-size in LD-CELP is 2.5 ms and the sub-frames are 0.625ms long. The parameters of the 50-th order predictor are updated every 2.5 ms. The LD-CELP uses gain-shape VQ for the excitation. The codebook consists of a 3-bit gain and 7-bit shape codebooks. A backward-adaptive excitation gain is also used. The gain information is obtained from previously quantized excitation using a 10th order predictor which operates on logarithmic gains.

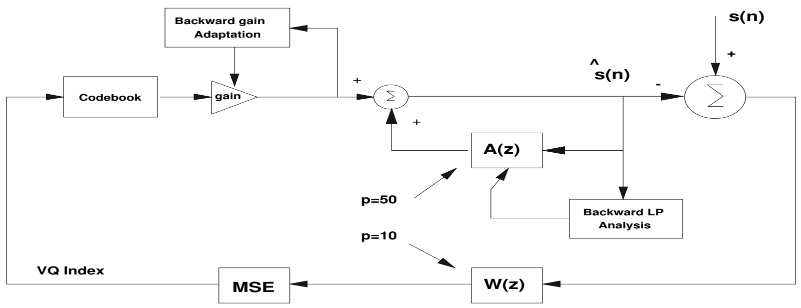


Fig. 39a. Low-Delay CELP (G.728) Encoder [38].

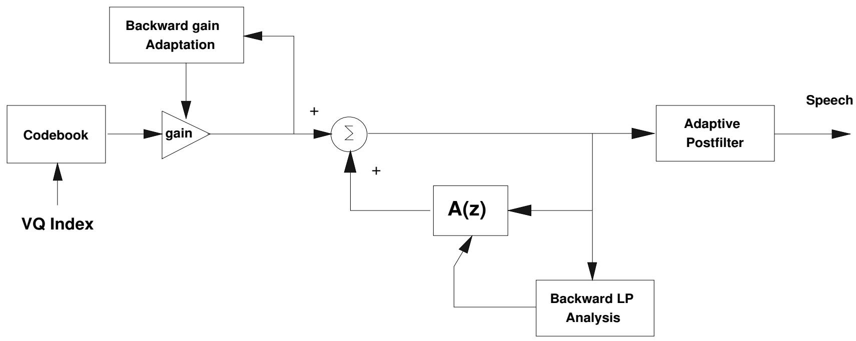


Fig. 39b. Low-Delay CELP (G.728) Decoder [38].

The gain is updated for every vector (0.625 ms) and the coefficients of the log-gain predictor are updated every 2.5 ms. The perceptual weighting filter is based on 10th order LP operating directly on unquantized speech and is updated every 2.5 ms. The transfer function of the weighting filter is more general than the one introduced (see (53)) for conventional analysis-by-synthesis linear predictive coders, i.e.,

$$W(z) = \frac{1 - A(z/\gamma_1)}{1 - A(z/\gamma_2)} \quad , \quad 0 < \gamma_2 < \gamma_1 \leq 1 \tag{74}$$

Improvements in the perceptual quality were realized for $\gamma_1 = 0.9$ and $\gamma_2 = 0.6$. In order to limit the buffering delay in LD-CELP only 0.625 ms of speech data are buffered at a time. Finally, the

LD-CELP utilizes adaptive short- and long-term postfilters to emphasize the pitch and formant structures of speech. The single-tap long-term postfilter requires knowledge of the pitch which is estimated at the decoder from the decoded speech. The short-term postfilter is of the form

$$H_p(z) = \frac{1 - A(z/\gamma_1)}{1 - A(z/\gamma_2)}(1 + b_1 z^{-1}) \tag{75}$$

The 10-th order rational polynomial emphasizes the formant structure of speech while the first order all-zero term compensates for the spectral tilt. The parameters of the rational polynomial are obtained as a by-product of the 50-th order recursive analysis process. These parameters are updated every 2.5 ms. The parameters γ_1, γ_2, μ as well as the LTP parameters were tuned to satisfy the requirements of the standard and their values are given in [38].

A two-phase testing procedure for the standardization of the LD-CELP was completed in 1991. CCITT required that the selected coder has a coding delay of less than 5 ms and speech quality that is not worse than the G.721 32 kbits/s ADPCM coder. In addition, it was required that the selected coder is capable of transmitting signalling information such as DTMF tones. The one-way delay of the LD-CELP is less than 2 ms and MOSs as high as 3.93 and 4.1 were obtained. The speech quality of the LD-CELP was judged to be equivalent or better than the G.721 standard even after three asynchronous tandem encodings. The coder was also shown to be capable of handling voiceband modem signals at rates as high as 2400 baud (provided that perceptual weighting is not used). The coder was implemented on the AT&T DSP32C processor and the complexity and memory requirements were found to be: 10.6 MIPS and 12.4 kbytes for the encoder and 8.06 MIPS and 13.8 kbytes for the decoder. More details on the performance of the LD-CELP and the standardization process are given in [38] which is the source of the information presented in this section.

6. Concluding Remarks

A. Summary

Speech coding has enjoyed a very successful period during the past decade. Advances in analysis-by-synthesis linear predictive coding and in transform coding provided a new generation of coders capable of producing communications quality speech at rates at and below 8 kbits/s. Several types of medium- and low-rate analysis-by-synthesis linear predictive coders have been adopted in national and international telephony standards. Similarly, low-rate speech-specific transform-based coders, such as the IMBE coder and the sinusoidal transform coder, were standardized or are strong candidates for standards under development. Table 2 summarizes the performance and complexity of several well known algorithms. We note that the performance is given in terms of Mean Opinion, DRT, and DAM scores, which were obtained from individual subjective tests (reported in the references cited in the table) and not from exhaustive comparative tests. Given that MOS tests may vary by as much as 0.4 from one test to another (DAM scores have a similar problem), the differences in these scores do not always constitute an absolute measure for comparison. Similarly, the complexity figures, which are expressed in terms of the number of MIPS, were obtained from different sources and are processor-dependent. Therefore the number of MIPS reported for an implementation on a floating-point signal processor, could increase by a factor of two or three when implementation is attempted on a fixed-point processor. In addition, in some of the real-time CELP implementations codebook searches are restricted at the expense of speech quality.

B. Current and Future Research

Digital cellular standards that use 13, 8, and 6.7 kbits/s speech coding algorithms were established and significant progress has been made in getting high quality speech coders for the future half-rate standards. In addition, high-quality low-delay coding at 16 kbits/s has also been achieved but low-delay coding at lower rates is still a challenging problem. Current research is now focussed upon achieving high-quality speech at and below 4.0 kbits/s. Improving the performance of low-rate coders operating in noisy channels is also an open problem. Additionally there is a demand for robust low-rate coders that will accommodate signals other than speech such as music. Finally, military-oriented coding research is geared towards obtaining improved 2.4 kbits/s coders, and also very-low-rate coders for applications that require increased tolerance to channel errors and low probability of intercept.

Although some opportunities to reduce the bit rate lie in efficient parameter quantization

Table. 2. Performance and Complexity of Selected Algorithms.

Algorithm	Bit Rate (bits/sec)	MOS/DRT/DAM	MIPS*	References
PCM (G.711)	64k	4.3/95/73	0.01	[150][152]
ADPCM (G.721)	32k	4.1/94/68	~2	[22][32][150]
LD-CELP (G.728)	16k	4.0+/-/-	~19	[35][38]
RPE-LTP (GSM)	13k	3.47+/-/-	6	[119][307]
Skyphone-MPLP	9.6k	3.4/-/-	11	[25]
VSELP (IS-54)	8k	3.45+/-/-	13.5	[70][100]
CELP (FS1016)	4.8k	3.2/93.7/62.2	16	[30][78]
STC-1	4.8k	3.52/92.7/63.	13	[210][212][213]
IMBE	4.15k	3.4/-/-	3	[26][121][141]
STC-2	2.4k	2.9/90.1/56	13	[210][212][213]
LPC-10e (FS 1015)	2.4k	2.3/89.9/52.3	~7	[77][301]
LPC-LSP	800	-/91.2/-	~20	[166]
~ estimated, + low score reported, * processor dependent				
The above complexity and performance figures were obtained from different sources and correspond to different implementation platforms and test environments. Therefore the performance and complexity figures do not always constitute an absolute measure for comparison.				

techniques [233][323] and new signal analysis methods, such as non-linear prediction [97][317], multiresolution time-frequency representation (wavelets) [308], and higher order statistics [158][226], most of the work still concentrates on exploiting further the perceptual properties of the ear. Improvements in low-rate speech coding are likely to come by optimizing further the speech coder to the "human receiver". This will require better understanding of the properties of the human auditory system and also development of automatic subjective tests. An excellent article on opportunities for future research in speech coding, by exploiting human perception, was published recently by Jayant [157].

From the implementation standpoint, high quality speech at lower information rates will come at the expense of increased algorithmic complexity. Although the development of faster and highly integrated signal processing devices may partially solve this problem, low complexity and low sensitivity to machine precision will still be important for low-power single-chip implementations.

7. Acknowledgements

The author would like to thank the following friends and colleagues for their comments on an earlier version of the paper: J. Aberle, L. Afxentiou, C. Balanis, C. Chakrabarti, D. Cochran, M. Deisher, P Loizou, A. Rodriguez, K. Tsakalis, of ASU, B. Fette of Motorola GSTG, D. Griffin of DVSI, B. Mears and G. Nair of Intel Corporation, P. Papamichalis of TI, T. Quatieri of MIT Lincoln Labs, and T. Taniguchi of Fujitsu. Special thanks to B. McAulay of MIT Lincoln Labs and B. Kleijn of AT&T Bell Labs for detailed comments on sinusoidal transform coding and code excited linear prediction respectively. Also the author would like to thank the anonymous reviewers for their constructive comments and corrections.

Portions of this work have been supported by Intel Corporation under contract agreements 92PJM120 and 91PJM107.

8. References

[1] H. Abut, Editor, "Vector Quantization," *IEEE Press,* 1990.

[2] J. Adoul et al, "4800 bits/s RELP vocoder using vector quantization for both filter and residual representation," *Proc. ICASSP-82, p.601, 1982*.

[3] N. Ahmed and K. Rao, Orthogonal Transforms for Digital Signal Processing, *Springer-Verlag*, New York 1975.

[4] M. Akamine and K. Miseki, "CELP Coding with an Adaptive Density Pulse Excitation Model," *Proc. ICASSP-90,* p. 29, New Mexico, April 1990.

[5] J.B. Allen and L.R. Rabiner, "A Unified Theory of Short-Time Spectrum Analysis and Synthesis," *Proc. IEEE,* Vol. 65, No. 11, pp. 1558-1564, Nov. 1977.

[6] L. Almeida and F. Silva, "Variable-frequency synthesis: An improved harmonic coding scheme," *IEEE ICASSP-84,* 27.5.1, San Diego, 1984.

[7] B. Atal, "Predictive Coding of Speech at Low Bit Rates," *IEEE Trans. COM-30*, No. 4, p. 600, April 1982.

[8] B. Atal, R. Cox, and P. Kroon, "Spectral Quantization and Interpolation for CELP coders," *Proc. ICASSP-89,* pp. 69-72, 1989.

[9] B. Atal, V. Cuperman, and A. Gersho, Advances in Speech Coding, Kluwer Ac. Publ., 1990

[10] B. Atal, V. Cuperman, and A. Gersho, Speech and Audio Coding for Wireless and Network Applications, Kluwer Ac. Publ., 1993.

[11] B. Atal and S. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Waves," *J. Acoust. Soc. Am.*, Vol. 50, p. 637, 1971.

[12] B. Atal and L. Rabiner, "Speech Research Directions," *AT&T Technical Journal*, Vol. 65(5), Sept.-Oct. 1986

[13] B. Atal and J. Remde, "A new model for LPC excitation for producing natural sounding speech at low bit rates," *Proc. ICASSP-82,* pp. 614-617, April 1982.

[14] B. Atal and M.R. Schroeder, "Adaptive Predictive Coding of Speech Signals," *Conf. Comm. and Proc.*, pp. 360-361, 1967

[15] B. Atal and M.R. Schroeder, "Linear Prediction Analysis of Speech based on a pole-zero representation," *J. Acoust. Soc. Am.*, Vol. 64(5), p. 1310, 1978.

[16] B. Atal and M.R. Schroeder, "Predictive Coding of Speech and Subjective Error Criteria," *IEEE Trans. ASSP-27*, No. 23, pp. 247-254, June 1979.

[17] B. Atal and M.R. Schroeder, "Stochastic Coding of Speech Signals At Very Low Bit Rates," *Proc. Int. Conf. Comm.*, pp. 1610-1613, May 1984.

[18] AT&T Technical Journal, Special Issue on "Speech Processing Technologies," Vol. 65(5), Sept.-Oct. 1986

[19] AT&T Technical Journal, Special Issue on "Speech Technologies," Vol. 69(5), Sept.-Oct. 1990

[20] B.N.S. Babu, "Performance of an FFT-Based Voice Coding System in Quiet and Noisy Environments," *IEEE Trans. ASSP-31*, No. 5, p. 1323, Oct. 1983.

[21] H. Bellanger et al, "Digital Filtering by Polyphase Network: Application to Sample-Rate Alteration and Filter Banks," *IEEE Trans. ASSP*, pp. 252-259, 1976.

[22] N. Benevuto et al, "The 32Kb/s coding standard," AT&T Technical Journal, Vol. 65(5), pp. 12-22, Sept.-Oct. 1986

[23] P. Blankenship, "A Review of Narrowband Speech Coding Techniques," Int. Specialist Seminar on Case Studies in Adv. Sig. Proc., Sept. 1979.

[24] G. Box and G. Jenkins, Time Series Analysis Forecasting and Control, HoldenDay, Inc., San Francisco, 1970.

[25] I. Boyd and C. Southcott, "A Speech Codec for the Skyphone Service," Br. Telecom Techn. J., Vol. 6(2), pp. 51-55, April 1988.

[26] M. Brandstein, P. Monta, J. Hardwick, J. Lim, "A Real-Time Implementation of the Improved MBE Coder," *Proc. ICASSP-90,* pp. 5-8, New Mexico, April 1990.

[27] J. Burg, "Maximum Entropy Spectral Analysis," *Proc. 37th Meeting of the Society of Exploration Geophysicists*, 1967.

[28] A. Buzo et al, " Speech Coding Based Upon Vector Quantization," *IEEE Trans. ASSP-28*, No. 5, pp. 562-574, Oct. 1980.

[29] S. Campanella et al, "A Comparison of Orthogonal Transformations for Digital Speech Processing," *IEEE Trans. COM-19(6)*, p. 1045, 1971.

[30] J. Campbell, T.E. Tremain, and V. Welch, "The Proposed Federal Standard 1016 4800 bps Voice Coder: CELP," *Speech Technology*, pp. 58-64, April 1990.

[31] J. Campbell and T.E. Tremain, "Voiced/Unvoiced Classification of Speech with Applications o the U.S. Government LPC-10e algorithm," *Proc. ICASSP-86, pp. 473-476, Tokyo 1986*.

[32] CCITT Recommendation G.721,"32kb/s Adaptive Differential Pulse Code Modulation (ADPCM)," Blue Book, Vol. III, Fascicle III.3, Oct. 1988.

[33] CCITT Recommendation G.722,"7 KHz Audio Coding within 64 kbits/s," Blue Book, Vol. III, Fascicle III, Oct. 1988.

[34] CCITT Recommendation G.723," Extensions of Recommendation G.721 ADPCM to 24 and 40 kbits/s for DCME Application," Blue Book, Vol. III, Fascicle III, Oct. 1988.

[35] CCITT Draft Recommendation G.728,"Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction (LD-CELP)," 1992.

[36] P. Chang, R. Gray, J. May, "Fourier Transform Vector Quantization for Speech Coding," *IEEE Trans. COM-35(10)*, p. 1059, Oct. 1987.

[37] J. Chen and R. Cox, "A Fixed-Point LD-CELP Algorithm and its Real-Time Implementation," *Proc. ICASSP-91*, pp. 21-24, Toronto, May 1991.

[38] J. Chen, R. Cox, Y. Lin, N. Jayant, and M. Melchner, "A Low-Delay CELP Coder for the CCITT 16 KB/s Speech Coding Standard," *IEEE Trans. on Selected Areas in Communications*, Special Issue on Speech and Image Coding, Ed. N. Hubing, pp. 830-849, June 1992.

[39] J. Chen and A. Gersho, "Vector adaptive predictive coding of speech at 9.6 kbps," *Proc. ICASSP-86*, pp. 1693-1696, 1986.

[40] J. Chen and A. Gersho, "Real Time Vector APC speech coding at 4800 bps with adaptive postfiltering," *Proc. ICASSP-87*, pp. 2185-2188, 1987.

[41] D.G. Childers, "The Cepstrum," *Proc. IEEE,* No. 10, pp. 1428-1443, 1977.

[42] Y. Cheng and D. O'Shaughnessy, "A 450 bits/s Vocoder with Natural Sounding Speech," *Proc. ICASSP-90,* pp. 649-652, New Mexico, April 1990.

[43] P. Ching et al, "Walsh Transform Coding of the Speech Residual in RELP coders," *IEE Proceedings,* Vol. 131-1, p. 29, 1984.

[44] J. Chung and R. Schafer, "A 4.8 Kbps Homomorphic Vocoder using analysis-by-synthesis excitation analysis," *Proc. ICASSP-89,* p. 144, 1989.

[45] J. Chung and R. Schafer, "Excitation Modeling in a Homomorphic Vocoder," *Proc. ICASSP-90,* p. 25, New Mexico, April 1990.

[46] Cooley et al, "The Fast Fourier Transform and its Application," *IBM Res. Paper, RC 1743*, IBM Watson Research Center, New York, 1968.

[47] M. Copperi and D. Sereno, "Vector Quantization and Perceptual Criteria for low-rate Coding of Speech," *Proc. ICASSP-85,* pp. 252-255, Tampa, March 1985.

[48] R.V. Cox and R. Crochiere, "Real-Time Simulation of Adaptive Transform Coding," *IEEE Trans. ASSP-29*, No. 2, p. 147, Apr. 1981.

[49] R.V. Cox et al, "New Directions in Sub-band Coding," *IEEE Trans. on Selected Areas in Communications*, Special Issue on Voice Coding for Communications, Vol. 6, No. 2, pp. 391-409, Feb. 1988.

[50] R. Crochiere, "A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis," *IEEE Trans. ASSP-28*, No. 1, p. 99, Feb. 1980.

[51] R. Crochiere and L. Rabiner, Multirate Digital Signal Processing, Prentice Hall, 1983.

[52] R. Crochiere, S. Webber, and J. Flanagan, "Digital Coding of Speech in Sub-bands," *The Bell Tech. J.*, Vol. 55(8), p. 1069, Oct. 1976.

[53] R. Crochiere et al, "Real-time Speech Coding," *IEEE Trans. COM-30(4)*, p. 621, Apr. 1982.

[54] J. Crosmer and T.P. Barnwell, "A Low Bit Rate Segment Vocoder Based on Line Spectrum Pairs," *Proc. ICASSP-85,* pp. 240-243, Tampa, March 1985.

[55] P. Cummiskey et al, "Adaptive Quantization in Differential PCM Coding of Speech," *Bell Sys. Tech. J.*, Vol. 52, No. 7, p. 1105, Sept, 1973.

[56] V. Cuperman and A. Gersho, "Adaptive differential vector coding of speech," *Proc. Globecom 1982*, p. E6.6.1, Miami, Dec. 1982.

[57] V. Cuperman and A. Gersho, "Vector predictive coding of speech at 16 kb/s," *IEEE Trans. COM-33*, p. 685, July 1985.

[58] V. Cuperman, A. Gersho, R. Pettigrew, J. Shynk, and J. Yao, "Backward Adaptive Configurations for Low-Delay Vector Excitation Coding," *Advances in Speech Coding*, Eds: B. Atal, V. Cuperman, and A. Gersho, pp. 13-23, Kluwer Ac. Publ., 1990.

[59] W. Daumer, "Subjective Evaluation of Several Coders," *IEEE Trans. COM-30(4)*, p. 655, Apr. 1982.

[60] G. Davidson and A. Gersho, "Complexity Reduction Methods for Vector Excitation Coding," *Proc. ICASSP-86,* p. 3055, Tokyo 1986.

[61] L. Davidson and R. Gray, Eds., Data Compression, Stroudsburg, PA: Dowden Hutchinson & Ross 1976.

[62] J. Deller, J. Proakis, and J. Hansen, Discrete-time Processing of Speech Signals, Macmillan, 1990

[63] J. Derby and C. Galand, "Multirate Sub-band Coding Applied to Digital Speech Interpolation," *IEEE Trans. ASSP-35*(12), p. 1684, Dec. 1987.

[64] S. Dimolitsas et al, "Use of Low-Delay CELP Technology in Circuit Multiplexed Networks," *Proc. ICASSP-93,* p. II-608, Minneapolis, April 1993.

[65] J.J. Dubnowski et al, "Real-time Digital Hardware Pitch Detector," *IEEE Trans. ASSP-24*, pp. 2-8, Feb. 1976.

[66] H. Dudley, "Remaking Speech," *J. Acoust. Soc. Am.*, Vol. 11, p. 169, 1939.

[67] H. Dudley, "The Vocoder," *Bell Labs. Record.,* 17, p. 122, 1939.

[68] H. Dudley, "Phonetic Pattern Recognition Vocoder for narrow-band Speech Transmission," *J. Acoust. Soc. Am.*, Vol. 30, pp. 733-739, 1958.

[69] J. Durbin, "The Fitting of Time-Series Models," *Rev. Inst. Int. Stat.*, Vol. 28, pp. 233-244, 1960.

[70] EIA/TIA-PN2398 (IS-54), "The 8 kbits/s VSELP algorithm," 1989.

[71] D. Elliott and K. Rao, Fast Transforms - Algorithms, Analyses, and Applications, *Academic Press*, New York 1982.

[72] D. Esteban and C. Galand, "Application of QMF to Split Band Voice Coding Schemes," *Proc. ICASSP-77,* p. 191, May 1977.

[73] D. Esteban and C. Galand, "16 kb/s Sub-band Coder Incorporating Variable Overhead Information," *Proc. ICASSP-82,* p. 1684, April 1982.

[74] G. Fairbanks, "Test of Phonemic Differentiation: The Rhyme Test," *J. Acoust. Soc. Am.*, Vol. 30, p. 596-600, 1958.

[75] G. Fant, "Acoustic Theory of Speech Production," *Mounton and Co.,* Gravenhage, The Netherlands, 1960.

[76] N. Farvardin and Y. Hussain, "Adaptive Block Transform Coding of Speech based on a Hidden Markov Model," *Proc. EUSIPCO,* p. 883, May 1988.

[77] Federal Standard 1015, Telecommunications: Analog to Digital Conversion of Radio Voice By 2400 Bit/Second Linear Predictive Coding, National Communication System - Office Technology and Standards, Nov. 1984.

[78] Federal Standard 1016, Telecommunications: Analog to Digital Conversion of Radio Voice By 4800 Bit/Second Code Excited Linear Prediction (CELP), National Communication System - Office Technology and Standards, Feb. 1991.

[79] B. Fette, "Experiments with High-Quality Low Complexity 4800 bps Residual excited LPC (RELP)," *Proc. ICASSP-88,* pp. 263-266, New York, April 1988.

[80] B. Fette and C. Jaskie, "A 600 bits/s LPC Voice Coder," *MILCOM-91*, pp. 1215-1219, 1991.

[81] T. Fischer, "A Pyramid Vector Quantizer," *IEEE Trans. IT-32,* p. 568, Jul. 1986.

[82] J. Flanagan et al., "Speech Coding," *IEEE Trans. COM-27(4)*, p.710, Apr. 1979.

[83] J. Flanagan, "Focal Points in Speech Communications Research," *IEEE Trans. COM-19*, pp. 1006-1015, Dec. 1971.

[84] J. Flanagan, "Parametric Coding of Speech Spectra", *J. Acoust. Soc. Am.,* Vol. 68(2), p. 412, Aug. 1980.

[85] J. Flanagan et al, "Computer Studies on parametric coding of speech spectra," *J. Acoust. Soc. Am.,* Vol. 68(2), p. 420, Aug. 1980.

[86] J. Flanagan, Speech Analysis, Synthesis and Perception, Springer-Verlag, 1972.

[87] J. Flanagan and Golden, "Phase Vocoder," *Bell Tech. J.*, 45, p. 1493, Nov. 1966.

[88] J. Foster and R. Gray, "Finite state vector quantization for waveform coding," *IEEE Trans. IT-31,* p. 348, May 1985.

[89] O. Fujimora, "Speech Coding and the Excitation Signal," *IEEE Int. Comm. Conf.* p. 49, 1966.

[90] S. Furui, Digital Speech Processing, Synthesis, and Recognition, *Marcel Dekker Inc., New York 1989*.

[91] S. Furui and M.M. Sondhi, Advances in Speech Signal Processing, *Marcel Dekker Inc., New York 1992*.

[92] H. Gethoffer, "Transform Coding of speech using Walsh functions," Symp. on Theory and Applications of Walsh Functions," Naval Res. Labs, Washington, D.C., 1973.

[93] J.S. Garofolo et al, "DARPA CD-ROM: An Acoustic Phonetic Continuous Speech Database, Training Set: 420 Talkers, 4200 sentences," Dec. 1988.

[94] E. George and M.J.T. Smith, "A New Speech Model Based on a Least-Squares Sinusoidal Representation," *Proc. ICASSP-87,* pp. 1641-1644, May 1987.

[95] E. George and M.J.T. Smith, "Perceptual Considerations in a Low Bit Rate Sinusoidal Vocoder," *Proc. IPCCC-90*, p. 268, Phoenix, Mar. 1990.

[96] A. Gersho and S. Wang, "Recent Trends and Techniques in Speech Coding," *Proc. 24th. Asilomar Conf.*, Pacific Grove, Nov. 1990.

[97] A. Gersho, "Linear Prediction Techniques in Speech Coding," *Proc. Globecom 1991*, Phoenix.

[98] A. Gersho and V. Cuperman, "Vector Quantization: A Pattern Matching Technique for Speech Coding," *IEEE Com. Mag.*, Vol. 21, p. 15, Dec. 1983.

[99] A. Gersho and R.M. Gray, Vector Quantization and Signal Compression, *Kluwer Academic Publishers, 1992*

[100] I. Gerson and M. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8 kbits/s," *Proc. ICASSP-90,* pp. 461-464, New Mexico, Apr. 1990.

[101] I. Gerson and M. Jasiuk, "Techniques for Improving the Performance of CELP-type speech coders," *Proc. ICASSP-91,* pp. 205-208, Toronto, May 1991.

[102] I. Gerson, M. Jasiuk, M. McLaughlin, and E. Winter, "Combined Speech and Channel Coding at 11.2 kbits/s," *Signal Processing V: Theories and Applications,* Elsevier, pp. 1339-1342, 1990

[103] I. Gerson, "Vector Sum Excited Linear Prediction (VSELP) Speech Coding for Japan Digital Cellular," *Meeting of IEICE, RCS90-26, Nov. 1990*.

[104] J. Gibson, "Adaptive Prediction for Speech Encoding," *ASSP Mag.*, Vol. 1, pp. 12-26, 1984.

[105] J. Gibson, "Adaptive Prediction in Speech Differential Encoding Systems," *Proc. IEEE*, vol. 68 pp. 1789-1797, Nov. 1974.

[106] J. Gibson, "Sequentially Adaptive Backward Prediction in ADPCM Speech Coders," *IEEE Trans. Com.*, pp. 145-150, Jan. 1978.

[107] J. Gibson et al, "Backward Adaptive Prediction in Multi-Tree Speech Coders," *Advances in Speech Coding*, Eds: B. Atal, V. Cuperman, and A. Gersho, pp. 5-12, Kluwer Ac. Publ., 1990.

[108] B. Gold and C. Rader, "The Channel Vocoder," *IEEE Trans. AU-15,* pp. 148-161, Dec. 1967.

[109] B. Gold, "Digital Speech Networks," *Proc. IEEE,* No. 12, pp. 1636-1658, Dec. 1977.

[110] B. Gold et. al. "New Applications of Channel Vocoders," *IEEE Trans. ASSP-29,* No. 1, p. 13, Feb. 1981.

[111] R. Gold and L. Rabiner, "Parallel Processing Techniques for Estimating the Pitch Periods of Speech in the Time Domain," *J. Acoust. Soc. Am.,* Vol. 46, pp. 146-152, 1969.

[112] R. Golden, "Digital Computer simulation of a sampled-datavoice-excited vocoder," *J. Acoust. Soc. Am.,* Vol. 35, pp. 1358-1366, 1963.

[113] P. Grant, C. Cowan, J. Dripps, and B. Mulgrew, "Analogue and Digital Signal Processing and Coding," Chartwell-Bratt, Bromley, U.K., 1989.

[114] A. Gray and D. Wong, "The Burg Algorithm for LPC Speech Analysis/Synthesis," *IEEE Trans. ASSP-28*, No. 6, pp. 609-615, Dec. 1980.

[115] R. Gray, "Vector Quantization," *ASSP Mag.*, Vol. 1, p. 4, Apr. 1984.

[116] R. Gray and H. Abut, "Full-search and tree-searched Vector Quantization for speech waveforms," *Proc. ICASSP-82,* pp. 593-596, Paris 1982.

[117] D. Griffin and J. Lim, "Multiband Excitation Vocoder," *IEEE Trans. ASSP-36*, No. 8, p. 1223, Aug. 1988.

[118] D. Griffin, "Multiband Excitation Vocoder," *Ph.D. Dissertation,* M.I.T, Cambridge, MA 1987.

[119] GSM 06.10, "GSM Full-Rate Transcoding," *Technical Rep. Vers. 3.2, ETSI/GSM,* July 1989.

[120] J. Hardwick and J. Lim, "A 4800 bps Multiband Excitation Speech Coder," *Proc. ICASSP-88*, pp. 374-377, Apr. 1988.

[121] J. Hardwick and J. Lim, "The Application of the IMBE Speech Coder to Mobile Communications," *Proc. ICASSP-91*, pp. 249-252, May. 1991.

[122] P. Hedelin, "RELP Vocoding with Uniform and Non-Uniform Down-Sampling," *Proc. ICASSP-83*, p. 1320, Apr. 1983.

[123] P. Hedelin and A. Bergstrom, "Amplitude Quantization for CELP Excitation Signals," *Proc. ICASSP-91*, pp. 225-228, Toronto, May 1991.

[124] P. Hedelin, "A tone-oriented voice-excited coder," *Proc. ICASSP-84*, p. 27.5.1, San Diego, 1984.

[125] W. Hess, Pitch Determination of Speech Signals, *Springer Verlag, New York 1983*.

[126] J.N. Holmes, "The JSRU Channel Vocoder," *Proc. IEE-Part F,* No. 1, Vol. 127, pp. 53-60, Feb. 1980.

[127] J.N. Holmes, "Formant Synthesizer: Cascade or Parallel?," *Proc. Speech Communication,* North Holland, vol. 2(4), pp. 251-273, Dec. 1983.

[128] M. Honda and Y. Shiraki, "Very Low-Bit-Rate Coding," in Advances in Speech Signal Processing, *Ed:* S. Furui and M.M. Sondhi, *pp. 209-230, Marcel Dekker Inc., New York 1992*.

[129] A. House et al, "Articulation Testing Methods," *J. Acoust. Soc. Am.*, Vol. 37, p. 158-166, 1966.

[130] N. Hubing and K. Yoo, "Exploiting Recursive Parameters Trajectories in Speech," *Proc. ICASSP-92,* p. I-125, San-Francisco, Apr. 1992.

[131] IEEE Communications Magazine, "Speech Processing and Applications," *IEEE Com. Mag.*, Vol. 28(1), Jan. 1990.

[132] IEEE Communications Magazine, "Speech Processing," *IEEE Com. Mag.*, Vol. 31(11), Nov. 1993.

[133] IEEE International Conference on Acoustics Speech and Signal Processing, annual conference proceedings.

[134] IEEE Journal on Selected Areas in Communications, Special Issue on Voice Coding for Communications, Vol. 6(2), Feb. 1988.

[135] IEEE Journal on Selected Areas in Communications, Special Issue on Speech and Image Coding, Vol. 10(5), June 1992.

[136] IEEE Proceedings, Special Issue on "man-machine speech communication," Nov. 1985.

[137] IEEE Trans. on Speech and Audio Processing, A publication of the Signal Processing Society.

[138] IEEE Trans. Communications, Special Issue on Bit Rate Reduction and Speech Interpolation, COM-30(4), Apr. 1982.

[139] IEEE Trans. Information Theory, Special issue on Quantization, Mar. 1982.

[140] IEEE Workshop on Speech Coding for Telecommunications, Proceedings, Sainte-Adele, Quebec, October 1993.

[141] Inmarsat Satellite Communications Services, Inmarsat-M System Definition, Issue 3.0 - Module 1: System Description, Nov. 1991.

[142] H. Irii, N. Kitawaki, "Multilingual Speech Data Base for Speech Quality Evaluation and Statistical Features," *Trans. IEICE*, Vol. E74, No. 1, Jan. 1991.

[143] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," *Proc. 6th International Congress of Acoustics*, pp. C-17-C.20, August 1968.

[144] F. Itakura and S. Saito, "Speech Analysis/Synthesis by Partial Correlation Coefficients," Proc. 4th Electrical Engineers Society Joint Conf., Paper 2835, 1970.

[145] F. Itakura and S. Saito, "On the Optimum Quantization of Feature Parameters in the PARCOR speech synthesizer," Proc. IEEE Conf. on Speech Commun. and Proc., pp. 434-437, April 1972.

[146] V. Iyengar and P. Kabal, "A Low Delay 16 Kb/s Coder," *IEEE Trans. ASSP-39*, No. 5, pp. 1033-1048, May 1991.

[147] V.K. Jain and R. Hangartner, "Efficient Algorithm for Multi-Pulse LPC Analysis of Speech," *Proc. ICASSP-84,* p. 1.4.1, 1984.

[148] C. Jaskie and B. Fette, "A Survey of Low Bit Rate Vocoders," *Speech Technology, 1982*.

[149] N.S. Jayant, "Step-size Transmitting Differential Coders for Mobile Telephony," *Bell Sys. Tech. J.*, Vol. 54, pp. 1557-1581, Nov. 1975.

[150] N.S. Jayant, "Coding Speech at low bit rates," *IEEE Spectrum*, pp. 58-63, Aug. 1986.

[151] N.S. Jayant and J. Chen, "Speech Coding with Time-Varying Bit Allocation to Excitation and LPC parameters," *Proc. ICASSP-89,* pp. 65-68, Glascow, Apr. 1989.

[152] N.S. Jayant and P. Noll, Digital Coding of Waveforms, Prentice-Hall, INC. Englewood Cliffs, NJ 1984.

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

[154] N.S. Jayant, "High-Quality Coding of Telephone, Speech, and Wideband Audio," *IEEE Com. Mag.*, Vol. 28, N0. 1, Jan. 1990.

[155] N.S. Jayant, V. Lawrence, D. Prezas, "Coding of Speech and Wideband Audio," *AT&T Tech. J.*, Vol. 69(5), pp. 25-41, Sept.-Oct. 1990.

[156] N.S. Jayant, Waveform Quantization and Coding, *IEEE Press,* 1976.

[157] N.S. Jayant, "Signal Compression: Technology Targets and Research Directions," *IEEE Trans. on Selected Areas in Communications*, Special Issue on Speech and Image Coding, Ed. N. Hubing, pp. 796-818, June 1992.

[158] I. Jebali, P. Loizou, and A. Spanias, "Speech Processing Using Higher Order Statistics," *Proc. ISCAS-93,* pp. 160-163, May 1993.

[159] J. Josenhans et al, "Speech Processing Applications Standards," *AT&T Technical Journal*, Vol. 65(5), p. 23, Sept.-Oct. 1986

[160] B. Juang and A. Gray, "Multiple stage vector quantization for speech coding," *Proc. ICASSP-82,* pp. 597-600, Paris 1982.

[161] P. Kabal and R. Ramachandran, "The Computation of Line Spectral Frequencies Using Chebyshev Polynomials," *IEEE Trans. ASSP-34*, No. 6, pp. 1419-1426, Dec. 1986.

[162] G. Kang, "Application of linear prediction to a narrowband voice digitizer," *NRL Report 7774,* Oct. 1974.

[163] G. Kang and D. Coulter, "600 Bit-per-second voice digitizer (linear predictive format coder)," *NRL Report 8043,* Nov. 1976.

[164] G. Kang and L. Fransen, "Application of Line-Spectrum Pairs to Low-Bit-Rate Speech Encoders," *Proc. ICASSP-85*, p. 244, Tampa 1985.

[165] G. Kang, L. Fransen, and E. Kline, "MultiRate Processor (MRP)," *NRL Report 8295,* March 1979.

[166] G. Kang and L. Fransen, "High Quality 800 b/s Voice Processing Algorithm," *NRL Report 9301,* Feb. 1991.

[167] G. Kang and S. Everett, "Improvement in the excitation source in the narrow-band linear predictive vocoder," *IEEE Trans. ASSP-33*, No. 2, pp. 377-386, 1985.

[168] S. Kang and T.R. Fischer, "Improved Excitation Sequence Quantization in RPE Speech Coding, *Advances in Speech Coding*, Eds: B. Atal, V. Cuperman, and A. Gersho, pp. 311-320, Kluwer Ac. Publ., 1990.

[169] D. Kemp et al, "An evaluation of 4800 bits/s Voice Coders," *Proc. ICASSP-89,* p. 200, Glascow, Apr. 1989.

[170] W.B. Kleijn et al, "Fast Methods for the CELP Speech Coding Algorithm," *IEEE Trans. ASSP-38*(8), p. 1330, Aug. 1990.

[171] W.B. Kleijn et al, "Improved Speech Quality and Efficient Vector Quantization in SELP," *Proc. ICASSP-88*, pp. 155-158, New York, April 1988.

[172] W.B. Kleijn, "Source-Dependent Channel Coding and its Application to CELP," *Advances in Speech Coding*, Eds: B. Atal, V. Cuperman, and A. Gersho, pp. 257-266, Kluwer Ac. Publ., 1990.

[173] W.B. Kleijn, "Methods for Waveform Interpolation," *Digital Signal Processing*, 1(4), pp. 215-230, 1991.

[174] W.B. Kleijn et al, "Generalized Analysis-by-Synthesis Coding and its Application to Pitch Prediction," *Proc. ICASSP-92*, pp. I-337-I340, 1992.

[175] W.B. Kleijn et al, "A 5.85 Kb/s CELP Algorithm for Cellular Applications," *Proc. ICASSP-93*, pp. II-596-599, 1993.

[176] P. Kroon and B. Atal, "Pitch Predictors with High Temporal Resolution," *Proc. ICASSP-90*, pp. 661-664, New Mexico, April 1990.

[177] P. Kroon and B. Atal, "Quantization Procedures for the Excitation in CELP coders," *Proc. ICASSP-87*, pp. 1649-1652, April 1987.

[178] P. Kroon, E. Deprettere, and R.J. Sluyeter, "Regular-Pulse Excitation-A Novel Approach to Effective and Efficient Multi-pulse Coding of Speech," *IEEE Trans. ASSP-34(5), Oct. 1986*.

[179] P. Kroon and E. Deprettere, "A Class of Analysis-By-Synthesis Predictive Coders for High Quality Speech Coding at Rates Between 4.8 and 16 kbits/s," *IEEE Trans. on Selected Areas in Communications*, Special Issue on Voice Coding for Communications, Vol. 6, No. 2, pp. 353-363, Feb. 1988.

[180] R. Kubichek, "Standards and Technology Issues in Objective Voice Quality Assessment," *Digital Signal Processing: A Review Journal*, pp. 38-44, DSP 1, April 1991.

[181] N. Levinson, "The Wiener RMS Error Criterion in Filter Design and Prediction," *J. Math. Phys.,* vol. 25, pp. 261-278, Jan. 1947.

[182] D. Lin, "New Approaches to Stochastic Coding of Speech Sources at Very Low Bit Rates," *Proc. EUPISCO-86,* p. 445, 1986.

[183] Y. Linde, A. Buzo, and A. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Comm.,* vol. COM 28, pp. 84-95, Jan. 1980.

[184] Y.J. Liu and J. Rothweiler, "A High Quality Speech Coder at 400 bits/s," *Proc. ICASSP-89,* Glascow, Apr. 1989.

[185] S.P. Lloyd, "Least Squares quantization in PCM," originally published in 1957 and republished in *IEEE Trans. IT*-28, part II, pp. 127-135, March 1982.

[186] J. LoCicero and R. Sukkar, "Pitch Detection: A Parallel processing Technique," pp.1237-1242, *Proc. 1986 Nat. Comm. Forum*, Chicago, Sept. 1986.

[187] P. Loizou and A. Spanias, "Vector Quantization of Principal Spectral Components for Speech Coding at 4800 BPS," *Proc. 24th. Asilomar Conf.*, Pacific Grove, Nov. 1990.

[188] P. Loizou and A. Spanias, "Vector Quantization of Transform Components for Speech Coding at 1200 BPS," *Proc. ICASSP-91*, pp. 245-248, Toronto, May 1991.

[189] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, Vol. 63, No. 4, pp. 561-580, April 1975.

[190] J. Makhoul and J. Wolf, "Linear Prediction and Spectral Analysis of Speech," *BBN Inc.*, Rep. 2304, 1972.

[191] J. Makhoul, "Speech Processing and Coding," *in Modern Signal Processing, T. Kailath Ed., Springer-Verlag, pp. 211-248, 1985*.

[192] J. Makhoul et al, "Vector Quantization in Speech Coding," *Proc. IEEE*, Vol. 73, No. 11, pp. 1551-1588, Nov. 1985.

[193] J. Makhoul, "A Fast Cosine Transform in One and Two Dimensions," *IEEE Trans. ASSP-28*, pp. 27-34, Feb. 1980.

[194] J. Makhoul et al, "A mixed-source model for speech compression and Synthesis," *Acoustical Society of America,* pp. vol. 64, pp. 1577-1581, Dec. 1978.

[195] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals," *IEEE Trans. ASSP-27*, pp. 121-133, 1979.

[196] M. Marcellin, T. Fisher, J. Gibson, "Predictive Trellis Coded Quantization of Speech," *IEEE Trans. ASSP-38*(1), p. 46, Jan. 1990.

[197] M. Marcellin and T. Fisher, "Trellis Coded Quantization of memoryless and Gauss-Markov sources," *IEEE Trans. COM-38*(1), pp. 82-93, Jan. 1990.

[198] M. Marcellin and T. Fisher, "A Trellis-Searched Speech Coder with Low Delay," *Advances in Speech Coding*, Eds: B. Atal, V. Cuperman, and A. Gersho, pp. 47-56, Kluwer Ac. Publ., 1990.

[199] J. Markel and A. Gray, Jr., Linear Prediction of Speech, Springer-Verlag, New York 1976.

[200] J.D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation, *IEEE Trans. AU-20*, pp. 367-377, Dec. 1972.

[201] S.L. Marple, Digital Spectral Analysis with Applications, Prentice Hall, 1987.

[202] J. Marques, L. Almeida, J. Tribolet, "Harmonic Coding at 4.8 KB/S ," *Proc. ICASSP-90,* p. 17, New Mexico, April 1990.

[203] R. McAulay and T. Quatieri, "Magnitude-only reconstruction using a sinusoidal speech model," *Proc. ICASSP-84,* Session 27.6.1, San Diego, Mar. 1984.

[204] R. McAulay and T. Quatieri, "Mid-rate Coding based on a Sinusoidal Representation of Speech," *Proc. ICASSP-85,* pp. 945., Tampa, Apr. 1985.

[205] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. ASSP-34*, No. 4, p. 744, Aug. 1986.

[206] R. McAulay and T. Quatieri, "Multirate Sinusoidal Transform Coding at Rates From 2.4 kbits/s to 8 kbits/s," *Proc. ICASSP-87,* Dallas, Apr. 1987.

[207] R. McAulay and T. Quatieri, "Computationally Efficient Sine-Wave Synthesis and its application to Sinusoidal Transform Coding," *Proc. ICASSP-88,* pp. 370-373, New York, March 1988.

[208] R. McAulay and T. Quatieri, "Pitch Estimation and Voicing Detection Based on a Sinusoidal Model," *Proc. ICASSP-90,* pp. 249-252, Albuquerque, March 1990.

[209] R. McAulay and T. Quatieri, "Sine-Wave Phase Coding at Low Data Rates," *Proc. ICASSP-91,* pp. 577-580, Toronto, May 1991.

[210] R. McAulay and T. Quatieri, "The Sinusoidal Transform Coder at 2400 b/s," *MILCOM-92, San Diego,* Oct. 1992.

[211] R. McAulay and T. Quatieri, "The Application of Subband Coding to Improve Quality and Robustness of the Sinusoidal Transform Coder," *Proc. ICASSP-93,* p. II-39, Minneapolis, April 1993.

[212] R. McAulay and T. Quatieri, "Low-Rate Speech Coding Based on the Sinusoidal Model," in Advances in Speech Signal Processing, Ch. 6, pp. 165-207, Ed. S. Furui and M.M. Sondhi, Marcel Dekker Inc., New York 1992.

[213] R. McAulay, private communication, 1992.

[214] A. McCree and T. Barnwell III, "A New Mixed Excitation LPC Vocoder," pp. 593-596, *Proc. ICASSP-91*, Toronto, May 1991.

[215] A. McCree and T. Barnwell III, "Implementation and Evaluation of a 2400 BPS Mixed Excitation LPC Vocoder," p. II-159, *Proc. ICASSP-93*, Minneapolis, April 1993.

[216] E. McLarnon, "A method for reducing the frame rate of a channel vocoder by using frame interpolation," pp. 458-461, *Proc. ICASSP-78*, Washington D.C., 1978.

[217] M. McLaughlin, "Channel Coding for Digital Speech Transmission in the Japanese Digital Cellular System," *Meeting of IEICE, RCS90-27, Nov. 1990*.

[218] N. Mikami and R. Ohba, "Pole-zero Analysis of Voiced Speech Using Group Delay Characteristics," *IEEE Trans. ASSP-32*, No. 5, p. 1095, Oct. 1984.

[219] W. Mikhael, A. Spanias, G. Kang, and L. Fransen, "Pole-zero Prediction," *IEEE Trans. CAS-33*, No. 3, p. 352, Jan. 1986.

[220] W. Mikhael and A. Spanias, "Accurate Representation of Time-Varying Signals using Mixed Transforms with Applications to Speech," *IEEE Trans. CAS-36*, No. 2, p. 329, Feb. 1989.

[221] R.L. Miller, "Improvement in the vocoder," *J. Acoust. Soc. Am.,* Vol. 33, pp. 667-677, 1959.

[222] Moore, An introduction to the Physiology of hearing, Academic Press, 1989.

[223] W. Munson and H. Montgomerry, "A speech analyzer and synthesizer," *J. Acoust. Soc. Am.,* Vol. 22, p. 678, 1950.

[224] M. Nakatsui et al, "Subjective speech-to-noise ratio as a measure of speech quality for digital waveform coders," *J. Acoust. Soc. Am.,* Vol. 72(4), p. 1136, Oct. 1982.

[225] S.H. Nawab et al, "Signal Reconstruction from Short-time Fourier Transform Magnitude," *IEEE Trans. ASSP-31*, No. 4, p. 996-998, Aug. 1983.

[226] C. Nikias and M. Raghuveer, "Bispectrum Estimation: A Digital Signal Processing Framework," *Proc. IEEE,* No. 75, pp. 869-891, July 1987.

[227] A.M. Noll, "Cepstrum Pitch Determination," *J. Acoust. Soc. Am.*, 41, pp. 293-309, Feb 1967.

[228] B.M. Oliver, J. Pierce, and C.E. Shannon, *"The Philosophy of PCM," Proc. IRE*, pp. 1324-1331, Nov. 1948 (also reprinted in [156]).

[229] A. Oppenheim and R. Schafer, "Homomorphic Analysis of Speech," *IEEE Trans. AU-16,* pp. 221-226, June 1968.

[230] A. Oppenheim, "A Speech Analysis-Synthesis System Based on Homomorphic Filtering," *J. Acoust. Soc. Am.*, Vol. 45, p. 458, Feb 1969.

[231] A. Oppenheim and R. Schafer, Digital Signal Processing, Prentice Hall, Englewood Cliffs, 1975.

[232] F. Owens, Signal Processing of Speech, McGraw Hill Inc., New York, 1993.

[233] K. Paliwal and B. Atal, "Efficient Vector Quantization of LPC parameters at 2.4 Kbits/frame," *Proc. ICASSP-91,* pp. 661-663, Toronto 1991.

[234] P.E. Papamichalis, Practical Approaches to Speech Coding, Prentice Hall, Englewood Cliffs, N.J., 1987

[235] P.E. Papamichalis and T.P. Barnwell III, "Variable Rate Speech Compression by encoding subsets of the PARCOR coefficients," *IEEE Trans. ASSP-31*, No. 3, p. 706-713, Jun. 1983.

[236] T. Parsons, Voice and Speech Processing, McGraw-Hill Book Co., 1987.

[237] D. Paul, "The Spectral Envelope Estimation Vocoder," *IEEE Trans. ASSP-29*, No. 4, p. 786, Aug. 1981.

[238] D. Paul, "A 500-800 b/s adaptive vector quantization vocoder using a perceptually motivated distance measure," *Proc. Globecom 1982*, p. E6.3.1, Miami, Dec. 1982.

[239] J. Picone and G. Doddington*,* "A Phonetic Vocoder," *Proc. ICASSP-89,* pp. 580-583, Glascow, Apr. 1989.

[240] M.R. Portnoff, "Short-Time Fourier Analysis of Sampled Speech", *IEEE Trans. ASSP-29,* no.3, pp. 364-373, June 1981.

[241] M.R. Portnoff, "Time-scale modification of speech based on Short-Time Fourier Analysis", *IEEE Trans. ASSP-29,* no.3, pp. 374-390, June 1981.

[242] M.R. Portnoff, "Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis", *IEEE Trans. ASSP-28,* no.1, pp. 55-69, February 1980.

[243] M.R. Portnoff, "Implementation of the Digital Phase Vocoder using the Fast Fourier Transform",*IEEE Trans. ASSP-24,* no.3, p. 243, June 1976.

[244] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements, Objective Measures for Speech Quality, *Prentice Hall,* Englewood Cliffs, NJ 1988.

[245] T. Quatieri et al, "Frequency sampling of the short-time Fourier-transform magnitude for signal reconstruction," *J. Opt. Soc. Am.,* 73, p.1523, Nov. 1983.

[246] T. Quatieri and R. McAulay, "Speech Transformations Based on a Sinusoidal Representation," *Proc. ICASSP-85,* pp. 489., Tampa, Apr. 1985.

[247] T. Quatieri and R. McAulay, "Shape Invariant Time-Scale and Pitch Scale Modification of Speech," *IEEE Trans. ASSP-40,* no.3, March 1992.

[248] T. Quatieri and R. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. ASSP-38*, pp. 56-69, 1981.

[249] L. R. Rabiner, "Digital-Formant Synthesizer for Speech Synthesis," *J. Acoust. Soc. Am.*, 43, pp. 822-828, 1968.

[250] L. R. Rabiner et al, "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Trans. ASSP-24*, pp. 399-417, Oct. 1976.

[251] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.

[252] L. R. Rabiner and B. Gold, Theory and Application of Digital Signal Processing, Prentice-Hall, 1975.

[253] D. Rahikka, "CELP Coding for Land Mobile Radio Applications," *Proc. ICASSP-90,* p. 465, New Mexico, April 1990.

[254] R. Ramachandran and P. Kabal, "Stability and Performance Analysis of Pitch Filters in Speech Coders," *IEEE Trans. ASSP-35*, No. 7, pp. 937-946, Jul. 1987.

[255] R. Ramachandran and P. Kabal, "Pitch Prediction Filters in Speech Coding," *IEEE Trans. ASSP-37*, No. 4, pp. 467-478, April 1989.

[256] V. Ramamoorthy and N.S. Jayant, "Enhancement of ADPCM speech by adaptive post-filtering," *Bell Syst. J.,* pp. 1465-1475, Oct. 1984.

[257] T.A. Ramstad, "Sub-band Coder with a simple adaptive bit allocation algorithm," *Proc. ICASSP-82,* p. 203, Paris, April 1982.

[258] O. Rioul and M. Vetterli, "Wavelet and Signal Processing," *ASSP Mag.*, Vol. 8, No. 4, p. 14, October 1991.

[259] R.C. Rose and T. Barnwell III, "Quality Comparison of Low Complexity 4800 BPS Self Excited and Code Excited Vocoders," *Proc. ICASSP-87,* April 1987.

[260] R.C. Rose and T. Barnwell III, "The Self-Excited Vocoder - An Alternative Approach to toll quality at 4800 BPS," *Proc. ICASSP-86,* pp. 453-457, April 1986.

[261] A. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," *J. Acoust. Soc. Am.*, Vol. 49, pp. 583-590, 1971.

[262] S. Roucos, R. Schwartz, and J. Makhoul, "A Segment Vocoder at 150 b/s," *Proc. ICASSP-83,* pp. 61-64, April 1983.

[263] M. Sabin and R. Gray, "Product code vector quantizers for speech waveform coding," *Proc. Globecom 1982*, p. E6.5.1, Miami, 1982.

[264] S. Saito and F. Itakura, "The theoretical Consideration of Statistically Optimum Methods for Speech Spectral Density," Report No. 3107, Electr. Comm. Lab., N.T.T., Tokyo 1966 (in Japanese).

[265] A. Satt and D. Malah, "Design of Uniform DFT filter Banks Optimized for Sub-band Coding," *IEEE Trans. ASSP-37*(11), p. 1672, Nov. 1989.

[266] R. Schafer and L. Rabiner, "Digital Representations of Speech Signals", *Proc. IEEE*, Vol. 63, pp. 662-677, April 1975.

[267] R. Schafer and L. Rabiner, "Design of Digital Filter Banks for Speech Analysis," *Bell Syst. Tech. J., Vol. 50, pp. 3097-3115, Dec. 1971*.

[268] R. Schafer and L. Rabiner, "Design and Simulation of a speech Analysis-Synthesis System Based on Short-Time Fourier Analysis," *IEEE Trans. AU-21*, No. 3, p. 165, June 1973.

[269] R. Schafer and J. Markel, Speech Analysis, *IEEE Press,* 1979.

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

[271] M.R. Schroeder, B. Atal, and J. Hall, "Optimizing digital Speech Coders by Exploiting the Masking Properties of the Ear," *J. Acoust. Soc. Am.*, Vol. 66, p. 1647, 1979.

[272] M.R. Schroeder and B. Atal, "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates," *Proc. ICASSP-85,* p. 937, Tampa, Apr. 1985.

[273] C.E. Shannon, A mathematical theory of communication," *Bell Syst. Journal*, vol. 27, pp. 379-423, 623-656, 1948.

[274] K.S. Shanmugam, Digital and Analog Communications, John Wiley, New York, 1979

[275] D. O'Shaughnessy, Speech Communication - Human and Machine, Addison-Wesley, 1987

[276] Y. Shoham, "Constrained Stochastic Excitation Coding of Speech at 4.8 kb/s," In *Advances in Speech Coding*, Eds: B. Atal, V. Cuperman, and A. Gersho, Kluwer Ac. Publ., 1990.

[277] F. Shum et al., "Speech Processing with Walsh-Hadamard Transforms," *IEEE Trans. AU-21*, no. 3, p. 174, June 1973.

[278] H.F. Silverman and N.R. Dixon, "A Parametrically Controlled Spectral Analysis System for Speech," *IEEE Trans. ASSP-22*, pp. 362-381, Nov. 1974.

[279] S. Singhal, D Le Gall, and C. Chen, "Source Coding of Speech and Video Signals," *Proc. IEEE*, Vol. 78, pp. 1233-1248, July 1990.

[280] S. Singhal and B. Atal, "Improving the Performance of Multi-pulse Coders at Low Bit Rates," *Proc. ICASSP-84,* p. 1.3.1, 1984.

[281] M.J.T. Smith and T.P. Barnwell, "Exact Reconstruction Techniques for Tree-Structured Sub-band Coders," *IEEE Trans. ASSP-34*(3), p. 434, June 1986.

[282] M. Sondhi and J. Schroeter, "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer," *IEEE Trans. ASSP-35*, No. 7, p. 955, July 1987.

[283] M. Sondhi "New Methods of Pitch Extraction," *IEEE Trans. AU-16*, pp. 153-157, June 1968.

[284] A. Spanias, "A Hybrid Transform Method for Analysis/Synthesis of Speech," *Signal Processing,* pp. 217-229, Elsevier, August 1991.

[285] A. Spanias, "A Hybrid Model for Speech Synthesis," *Proc. ISCAS-90*, p. 1521, New Orleans, May 1990.

[286] A. Spanias et al, "Block Time and Frequency Domain Modified Covariance Algorithms," *Proc. ICASSP-92,* pp. 529-532, San-Francisco, Apr. 1992.

[287] A. Spanias, M. Deisher, P. Loizou and G. Lim, "Fixed-Point Implementation of the VSELP algorithm, *ASU-TRC Technical Report*, TRC-SP-ASP-9201, July 1992.

[288] A. Spanias and P. Loizou, "A Mixed Fourier/Walsh Transform Scheme for Speech Coding at 4 kbits/s," *Proc. IEE - Part I*, pp. 473-481, October 1992.

[289] A. Spanias, P. Loizou, and G. Lim, "Fixed-Point Implementation of the RPE-LTP Speech Coder," *ASU-TRC Technical Report*, TRC-SP-ASP-9204, August 1992.

[290] A. Spanias, W. Ma, and F. Tiong, "Simulation Models for the GSM Channel Coding and Modem Functions," *ASU-TRC Technical Report*, TRC-SP-ASP-9202, August 1992.

[291] Speech Communication, A publication of the European Association for Signal Processing (EURASIP) and of the European Speech Communication Association (ESCA), North Holland

[292] R. Steele and L. Cassel, "Dynamic encoding as applied to a channel vocoder," *J. Acoust. Soc. Am.,* Vol. 35, p. 789, 1963.

[293] K. Steiglitz, "On the Simultaneous Estimation of Poles and Zeros in Speech Analysis," *IEEE Trans. ASSP-25*, No. 3, p. 229, June 1977.

[294] G. Strang, Linear Algebra, Academic Press, 1982.

[295] M. Sunwoo, S. Park, and K. Terry, " A Real-Time Implementation of Key VSELP routines on a 16-bit DSP Chip," *DSP Technologies*, 1991

[296] N. Sugamura and F. Itakura, "Speech Data Compression by LSP Analysis/Synthesis Technique," *Trans. IEICE*, Vol. J64, pp. 599-606, 1981.

[297] M. Taka et al, "DSP Implementations of Sophisticated Speech Codecs," *IEEE Trans. on Selected Areas in Communications*, Special Issue on Voice Coding for Communications, Vol. 6, No. 2, pp. 274-282, Feb. 1988.

[298] T. Taniguchi et al, "Multimode Coding: Application to CELP," *Proc. ICASSP-89,* pp. 156-159, Glascow, Apr. 1989.

[299] T. Taniguchi et al, "A 16 kbits/s ADPCM with multiquantizer (ADPVM-MQ) codec and its implementation by digital signal processor," *Proc. ICASSP-87,* pp. 1340-1343, Apr. 1987.

[300] I. Trancoso and B. Atal, "Efficient Search Procedures for Selecting the Optimum Innovation in Stochastic Coders," *IEEE Trans. ASSP-38*(3), p. 385, March 1990.

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

[302] T.E. Tremain et al, "Evaluation of Low Rate Speech Coders for HF," *Proc. ICASSP-93,* p. II-163, Minneapolis, April 1993.

[303] J. Tribolet and R. Crochiere, "Frequency domain coding of speech," *IEEE Trans. ASSP-27*(5), p. 512, Oct. 1979.

[304] C. K. Un and D.T. Magill, "The Residual-Excited Linear Prediction Vocoder with Transmission Rate below 9.6 kbits/s," *IEEE Trans. COM-23(12)*, p. 1466, Dec. 1975.

[305] P.P. Vaidyanathan, "Quadrature Mirror Filter Banks M-Band Extensions and Perfect Reconstruction Techniques," *ASSP Mag.*, Vol. 4, No. 3, p. 4, July 1987.

[306] H. Van Trees, Detection, Estimation, and Modulation Theory, Part I, Ch. 3, New York, Wiley 1968

[307] P. Vary et al, "Speech Codec for the European Mobile Radio System," *Proc. ICASSP-88,* p. 227, April 1988

[308] M. Vetterli, "Filter Banks Allowing Reconstruction," *Signal Processing*, Vol. 10, No. 3, p. 219, April 1986.

[309] V. Viswanatham and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," *IEEE Trans. ASSP-23*, pp. 309-321, June 1975.

[310] V. Viswanathan et al, "Speech-Quality optimization of 16 kb/s adaptive predictive coders," *Proc. ICASSP-80,* p. 520, Denver, 1980

[311] A. Viterbi and J. Omura, "Trellis encoding of memoryless discrete sources with a fidelity criterion," *IEEE Trans. IT-20,* p. 325, May 1974.

[312] W.D. Voiers, "Diagnostic Acceptability Measure for Speech Communications Systems," *Proc. ICASSP-77,* p. 204, May 1977.

[313] W.D. Voiers, "Evaluating Processed Speech using the Diagnostic Rhyme Test," *Speech Techn.,* Jan./Feb. 1983.

[314] M. Wakao and M. Oguchi, "Activity of the Codec Algorithm Selection for the Japanese Digital Cellular System," *Meeting of IEICE, RCS90-25, Nov. 1990*.

[315] S. Wang and A. Gersho, "Improved Excitation for Phonetically-Segmented VXC Speech Coding Below 4Kb/s," *Proc. 1990 Globecom, p. 946*, San Diego 1990.

[316] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting the Subjective Quality of Speech Coders," *IEEE Trans. on Selected Areas in Communications*, Special Issue on Speech and Image Coding, Ed. N. Hubing, pp. 819-829, June 1992.

[317] S. Wang, E. Paksoy and A. Gersho, "Performance of Nonlinear Prediction of Speech," Report

[318] S. Wang, A. Sekey and A. Gersho, "Auditory Distortion Measure for Speech Coding," *Proc. ICASSP-91*, pp. 493-496, Toronto, 1991.

[319] C.J. Weinstein, "Opportunities for Advanced Speech Processing in Military Computer-Based Systems," *Proc. IEEE,* Vol. 79, No. 11, pp. 1626-1640, Nov. 1991.

[320] V. Welch and T. Tremain," *IEEE Workshop on Speech Coding and Telecommunications,* pp. 41-42, Oct. 1993.

[321] D. Wong, "Vector/Matrix Quantization for Narrow-Bandwidth Digital Speech Compression," *Signal Technology, Tech. Rep. F30602-81-C-0054, 1982*.

[322] Siu-Wah Wong, "An Evaluation of 6.4 kbits/s Speech Codecs for the INMARSAT-M System," *Proc. ICASSP-91*, pp. 629-632, Toronto, May 1991.

[323] C. Xydeas and K. So, "A Long History Quantization Approach to Scalar and Vector Quantization of LSP Coefficients," *Proc. ICASSP-93,* p. II-1, Minneapolis, April 1993.

[324] P. Yatrou and P. Mermelstein, Ensuring Predictor Tracking in ADPCM Speech Coders Under Noisy Transmission Conditions, *IEEE Journal on Selected Areas in Communications, Special Issue on Voice Coding for Communications*, Vol. 6(2), pp. 249-261, Feb. 1988.

[325] M. Yong and A. Gersho, "Vector Excitation Coding with Dynamic Bit Allocation ," *Proc. Globecom 1988*, p. 290, Dec. 1988.

[326] R. Zelinski and P. Noll, "Adaptive Transform Coding of Speech Signals," *IEEE Trans. ASSP-25*, p. 299, Apr. 1977.

[327] R. Zelinski and P. Noll, "Approaches to Adaptive Transform Coding at Low Bit rates," *IEEE Trans. ASSP-27*, p. 89, Feb. 1979.

[328] R.L. Zinser, "An Efficient, Pitch-Aligned High-Frequency Regeneration Technique for RELP Vocoders," *Proc. ICASSP-85,* p. 969, Florida, April 1985.