Low Bit Rate Speech Coding

Miguel Arjona Ramírez and Mario Minami
Electronic Systems Eng. Dept. (PSI) - Escola Politécnica
University of São Paulo
05508-900 São Paulo - SP - Brazil

Abstract—
This article is focused on speech coding methods for achieving communication quality speech at bit rates of 4 kbit/s and lower. The speech coding techniques are based on an all-pole model of the vocal tract which may be implemented in the time domain with appropriately selected excitation functions or else may be fit to a spectral analysis of the speech signal. Three main types of coders are described below. Code-excited linear prediction (CELP) coders select their excitation from waveform codebooks using analysis-by-synthesis closed-loop techniques, which need to be supplemented by speech classification and open-loop parametric techniques for keeping up with quality at lower rates. The prototypical sinusoidal coder (SC) has a bank of oscillators for signal synthesis, driven by a model of the magnitude spectrum. However, phase regeneration is important in enhancing speech reconstruction at low rates. Waveform interpolation (WI) coders afford a wider time-frequency footprint for the representation of the excitation, showing a good potential for achieving toll quality at bit rates below 4 kbit/s.

Keywords— **Low bit rate speech coding, vocoder, codec, rate-distortion function, code-excited linear prediction, CELP, algebraic CELP, ACELP, linear prediction, LP, linear predictive coding, LPC, sinusoidal coder, waveform interpolation, WI, complexity, bit rate, fidelity, distortion, speech synthesis.**

Cross-references

speech coding, low bit rate
voice coding, low bit rate

I. Introduction

Speech coders were first used for encrypting the speech signal as they still are today for secure voice communications. But their most important use is bit rate saving to accomodate more users in a communications channel such as a mobile telephone cell or a packet network link. Alternatively, a high resolution coder or a more elaborate coding method may be required to provide for a higher fidelity playback.

Actually, the availability of ever broader-band connection and larger-capacity media has led some to consider speech coding as unnecessary but the increasing population of transmitters and the ever richer content have taken up the "bandwidth" made available by the introduction of broadband services.

Further, coding may be required to counter the noise present in the communication channel, such as a wireless connection, or the decay of the storage media, such as a magnetic or optical disc. In fact, such a coding, called channel coding, will increase the total bit rate and this is usually on a par with encryption. In contrast, the coding mentioned before is called source coding and will be dealt with almost exclusively below.

The speech signal is an analog continuous waveform and any digital representation of it incurs a distortion or lack of fidelity, which is irrelevant for high-fidelity rendering. High-fidelity representations are obtained by filtering the signal within a wide enough frequency band, sampling it at regular intervals and then quantizing each amplitude so obtained with a large number of bits. This kind of direct digital coding is called pulse code modulation (PCM). The sampling operation is reversible if properly done and the large number of bits for quantizer codes makes it possible to have a large number of closely spaced coding levels, reducing quantization distortion.

Since human hearing has a finite sensitivity, a sufficiently fine digital representation may be considered "transparent" or essentially identical to the original signal. In the case of a general audio signal, a bit rate of 706 kbit/s per channel, compact disc (CD) quality, is usually considered transparent while for telephone speech 64 kbit/s is taken as toll quality (Table I). Even though it is rather elusive to impose a range for low bit rate speech coding as it is a moving target, it seems that nowadays it is best bounded by 4 kbit/s from above, given the long lasting effort to settle for a toll quality speech coder at that rate at the ITU-T (1), (2), and it is bounded by about 1 kbit/s from below by considering mainly the expected range of leading coding techniques at the lower low-rate region and the upper very-low-rate region (3). A very good and comprehensive reference to speech coding (4) located low rate between 2.4 kbit/s and 8 kbit/s just some years ago.

II. Speech modeling for low rate speech coding

Speech is a time-varying signal which may be considered stationary during segments of some tens of milliseconds in general. For these segments, usually called frames, an overall characterization is often made by using a spectral model. Complementarily, the energy is imparted to a synthesis filter, which embodies the estimated spectral model, by an excitation signal also carrying more details of the fine structure of the signal spectrum or else the spectral model may be sampled at selected frequencies or integrated over selected frequency bands in order to define a proper reconstructed signal. In addition, the incorporation into the excitation model of the requisite interpolation for the process of synthesis further extends it into the time-frequency domain.

A. Predictive coders

During the first half of the twentieth century, filterbanks were used for synthesizing speech since the first voice coder or "vocoder" developed by Dudley. The major difficulty in vocoding was the separation of vocal source behavior from vocal tract behavior in order to drive a source-filter model for synthesis. A didactic taxonomy of parametric coders is given by (5).

A manageable and accurate acoustical model of speech production was proposed by Fant in 1960 and a good approximation to it is provided by the linear prediction (LP) model. The LP model for speech analysis was originally proposed by Itakura and Saito in 1968 and Atal and Hanauer in 1971 (6) whose spectral models are short-term stationary and nonstationary, respectively. The stationary LP spectral model is the frequency response of

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

whose magnitude may be interpreted as a fit to the envelope of the short-term log spectrum of the signal as shown in Figure 1. The order p of the LP model has to be high enough to enable it to adjust to the overall shape of the spectrum and the gain factor G allows an energy matching between the frequency response of the model and the spectrum of the signal. The LP model is particularly biased toward the peaks of the signal spectrum as opposed to the valleys and is particularly useful as a smooth peak-picking template for estimating the formants, sometimes not at likely places at a first glance like the second formant in Figure 1.

The excitation model proposed by Itakura and Saito combines two signal sources as shown in Figure 2 whose relative intensities may be controlled by the two attenuation factors $U^{1/2}$ and $V^{1/2}$ which are interlocked by the relation

$$U + V = 1. \qquad (2)$$

The pulse source, obtained for $V = 1$ and $U = 0$, is useful for generating voiced speech. In this mode, besides the gain factor G, the pulse repetition rate P has to be controlled. It is obtained in the coder as the pitch period of the speech signal through a pitch detection algorithm. The detected pitch period value may not be appropriate due to a lot of situations which may occur as a consequence of the quasiperiodic nature of voiced speech, the interaction of fundamental frequency (F_0) with the first formant or missing lower harmonics of F_0. On the other hand, for unvoiced speech the gain factor G is enough to match the power level of the pseudorandom source along with $U = 1$ and $V = 0$.

A better mixed excitation is produced by the Mixed Excitation Linear Prediction (MELP) coder which, besides combining pulse and noise excitations, is able to yield periodic and aperiodic pulses by position jitter (7). Further, the composite mixed excitation undergoes adaptive spectral enhancement prior to going through the synthesis filter to produce the synthetic signal which is applied to the pulse dispersion filter.

B. *Sinusoidal coders*

The voiced mode of speech production motivates the sine-wave representation of voiced speech segments by

$$s(n) = \sum_{k=1}^{K} A_k \cos(\omega_k n + \phi_k) \qquad (3)$$

where A_k and ϕ_k are the amplitude and phase of oscillator k, associated with the ω_k frequency track. This model quite makes sense in view of the spectrum of a voiced segment as can be seen in Figure 3. As suggested in this figure, the peak frequencies $\{\omega_k, k = 1, 2, \ldots, K\}$ may be extracted and used as the oscillator frequencies in the equation above. For a strict periodic excitation model, $\omega_k = k\omega_0$, that is, the peak frequencies are equally interspaced and we have the so-called harmonic oscillator model. However, not all sinusoidal coders subscribe to this model because, by distinguishing small deviations from harmony, tonal artifacts may be guarded against. But the harmonic model is more amenable to low-rate implementation and then other techniques have to be resorted to in order to forestall the development of buzzy effects which arise as a consequence of the forced additional periodicity.

The amplitudes may be constrained to lie on an envelope fit to the whole set of amplitudes thereby enabling an efficient vector quantization of the amplitude spectrum. This amplitude model is compatible with the linear prediction filter in Section II-A and the efficient quantization methods available for it may be borrowed just like the sinusoidal transform coder (STC) does (8).

Equation (3) may also be used for synthesizing unvoiced speech as long as the phases are random. In order to reduce the accuracy required of the voicing decision, a uniformly distributed random component is added to the phase of the oscillators with frequency above a voicing-dependent cutoff frequency in the STC as the lower harmonics of F_0 are responsible for the perception of pitch. In the multiband excitation (MBE) coder the band around each frequency track is defined as either voiced or unvoiced and Equation (3) is not used for unvoiced synthesis; instead, filtered white noise is used. The bands are actually obtained after the signal has been windowed and, as the windows have a finite bandwidth, this brings about a similarity of the sinusoidal coder with subband coders.

For low-rate coding, there is not enough rate for coding the phases and phase models have to be used by the synthesizer such as the zero-phase model and the minimum-phase model. When there is a minimum-phase spectral model as in the latter case, the complex amplitude is obtained at no additional cost by sampling its frequency response as

$$H\left(e^{j\omega_k}\right) = A_k^{(r)} e^{j\phi_k^{(r)}} \qquad (4)$$

where $A_k^{(r)}$ and $\phi_k^{(r)}$ are the reconstructed amplitude and phase of frequency track ω_k, respectively.

C. *Waveform-interpolation coders*

Waveform-interpolation coders usually apply linear prediction for estimating a filter whose excitation is made by interpolation of characteristic waveforms. Characteristic waveforms (CWs) are supposed to represent one cycle of excitation for voiced speech. The basic idea for the characteristic waveform stems from the Fourier-series representation of a periodic signal, whose overtones are properly obtained by a Fourier-series expansion. Therefore, the CW encapsulates the whole excitation spectrum provided that the signal be periodic. The rate of extraction of CWs may be as low as 40 Hz for voiced segments as these waveforms are slowly varying in this case. On the other hand, for unvoiced segments the rate of extraction may have to be as high as 500 Hz but each one of them may be represented with lower resolution (9).

The length of sampled characteristic waveforms varies as the pitch period. Therefore, their periods have to be normalized and aligned before coding for proper phase tracking. A continuous-time notation encapsulates a length normalization and the time-domain CW extraction process so that a two-dimensional surface may be built. The normalization of CW length is achieved by stretching or shrinking them so as to fit within a normalized period of 2π radians. This normalized time within a period is referred to as the phase (ϕ). Assuming that linear prediction analysis has been performed and that the prediction residual has been determined for CW extraction and Fourier-series representation, above and below the time-phase plane undulates the characteristic surface

$$u\left(t,\phi\right)=\sum_{k=1}^{K}\alpha_k(t)\cos\left(k\phi\right)+\beta_k(t)\sin\left(k\phi\right). \tag{5}$$

For the sake of coding efficiency, it is convenient to decompose the characteristic surface into a slowly evolving waveform (SEW) and a rapidly evolving waveform (REW). The SEW may be obtained by lowpass filtering $u\left(t,\phi\right)$ along the t axis as shown in Figure 4 and represents the quasiperiodic component of speech excitation whereas the REW may be obtained by highpass filtering $u\left(t,\phi\right)$ along the t axis, representing the random component of speech excitation. Both components must add up to the original surface, that is,

$$u\left(t,\phi\right)=u_{\mathrm{SEW}}\left(t,\phi\right)+u_{\mathrm{REW}}\left(t,\phi\right). \tag{6}$$

Characteristic waveforms may be represented by means other than a Fourier series but in the latter case they may be compared to sinusoidal coders, having smaller interpolation rates due to a more flexible time-frequency representation and to a higher resolution in time. For a common framework that encompasses both sinusoidal coding and waveform interpolation, please refer to (10) where the issue of perfect reconstruction in the absence of quantization errors is brought to bear.

III. Parameter estimation from speech segments

The linear prediction model was introduced in the last section along with the simplest excitation types for time-domain implementation, the frequency-domain parametric models of greater use for low bit rate coders and a harmonic excitation model including waveform interpolation. In this section a more detailed description is provided of the structures used to constrain the excitation and the algorithms used for estimating its parameters. The segmentation of the speech signal for its analysis is complemented by its concatenation in the synthesis phase.

Aimed first at the medium bit rate range from 8 kbit/s to 16 kbit/s, a different approach has come to be used for coding the excitation, called code-excited linear prediction (CELP) (11). The two most important concepts in CELP coding are an excitation quantization by sets of consecutive samples, which is a kind of vector quantization (VQ) of the excitation, and a search criterion based on the reconstruction error instead of the prediction error or differential signal. Figure 5 has been drawn stressing these main distinguishing features.

A CELP coder is provided with a finite set of codevectors to be used for reconstructing each segment or subframe of the original signal. A collection of M codevectors is said to be a codebook of size M. Prior to searching the excitation, a filter is estimated through LP analysis (see Section II-A) to have a frequency response matching the short-term spectral envelope of a block of the original signal called a frame. Each frame typically consists of two to four excitation subframes and the synthesis filter is determined for each subframe by interpolation from the LP filters of neighboring frames. As shown in Figure 5, each codevector $\mathbf{c}_k$ in turn, for $k=1,2,\ldots,M$ is filtered by the synthesis filter

$$H(z)=\frac{1}{1-P(z)} \tag{7}$$

generating all around the encoding loop a reconstruction error vector ε_k. This process of determining the signal to be synthesized within the coder is called the analysis-by-synthesis method. It allows the coder to anticipate the best strategy constrained to the situation that the synthesizer will face. Thus, the minimum square reconstruction error is identified as

$$i=\underset{k=1,2,\ldots,M}{\operatorname{argmin}}\left\{\left\|\varepsilon_k\right\|^2\right\} \tag{8}$$

after an exhaustive search all through the codebook and the actual excitation is delivered as the scaled version

$$\mathbf{e}_r=G\mathbf{c}_i \tag{9}$$

of codevector $\mathbf{c}_i$, where the scale factor $G=G_i$ has been calculated to minimize the square reconstruction error $\left\|\varepsilon_i\right\|^2$ for codevector $\mathbf{c}_i$.

Actually, a CELP coder applies a perceptual spectral weighting to the reconstruction error prior to the minimization by means of the weighting filter, defined by a function of the adaptive synthesis filter as

$$W(z)=\frac{H(z/\gamma_2)}{H(z/\gamma_1)} \tag{10}$$

where $0<\gamma_2<\gamma_1\leq 1$ are bandwidth expansion factors. A very usual combination of values is $\gamma_2=0.8$ and $\gamma_1=1$. Overall, the weighting filter serves the dual purpose of deemphasizing the power spectral density of the reconstruction error around the formant frequencies where the power spectrum of the signal is higher and emphasizing the spectral density of the error in between the formant frequencies where hearing perception is more sensitive to an extraneous error. Both actions come about as consequences of the frequency response of $W(z)$ in Figure 6. In much the same way, in order to achieve a reconstructed signal with a higher perceptual quality an open-loop postfilter is usually applied to the reconstructed signal which is defined as a function of the synthesis filter as well (see Figure 7).

Additionally, toll quality reconstruction can only be achieved if there is a rather precise means of imposing the periodicity of voiced speech segments on the reconstructed signal. This goal can be achieved by using a second adaptive codebook in the CELP coder. This adaptive codebook is fed on a subframe basis the composite coded excitation

$$e(n) = G_a c_a(n) + G_f c_f(n), \qquad (11)$$

where $c_a(n)$ stands for the adaptive codevector with its gain factor G_a and $c_f(n)$ with its gain factor G_f represents the fixed excitation, depicted by the only codebook in Figure 5. The enhanced synthesis model for this CELP coder is illustrated in Figure 7.

Nonetheless, the fixed codebook structure and its search algorithms have been the target for developments leading to the widespread applicability of CELP coders. The fixed codebook in the original CELP coder was stochastically populated from samples of independent and identically Gaussian distributed vectors (11). As the complexity of exhaustive searches through the codebook was overwhelming for the then current signal processors, more efficient search methods were derived, as will be seen in Section IV, which required more structured codebooks such as the center-clipped and overlapped stochastic codebooks. Their searches have lower operational complexity due to the sparse amplitude distribution and the overlapped nature of their codevectors. The latter allows for the use of efficient search techniques originally developed for the adaptive codebook. Even more surprising, they enhance the speech quality as well (12) to a level considered good enough for secure voice and cellular applications at low to medium rates.

Meanwhile, predictive waveform coders borrow the idea of impulse excitation from parametric LP coders (see Section II-A) in order to be able to decrease the bit rate but with a twist to be able to deliver higher quality which involves the increase in the number of pulses per pitch period. A subframe of multipulse excitation is given by

$$e(n) = G \sum_{k=0}^{M-1} \alpha_k \delta(n - m_k), \qquad n = 0, 1, \ldots, L-1, \qquad (12)$$

where M is the number of pulses per excitation subframe, L is the length of the subframe, α_k and m_k represent individual pulse amplitude and position and G is a common excitation vector gain. This new approach was called “multipulse excitation” and is very complex in its most general formulation (13). Moreover, a constrained version of it, known by “regular pulse excitation with long-term predictor” (RPE-LTP), was adopted for the Global System for Mobile Communications (GSM) full rate standard coder for digital telephony and it is notable for its low complexity (14).

This kind of excitation was further structured and inserted into a CELP coder. Pulse positions were constrained to lie in different tracks, which cover in principle all the positions in the excitation subframe whereas pulse amplitudes α_k were restricted to either plus or minus one. The latter feature and its conceptual connection to error-correction codes has established the name “algebraic CELP” for this kind of excitation. These deterministic sparse codebooks made their entrance into standard speech coding with the G.729 conjugate structure, algebraic CELP (CS-ACELP) coder (15). A general ACELP position grid is given in Table II for an M-pulse codebook over an L-sample subframe.

As the bit rate is decreased, further modeling and classification of the signal has to be done at the encoder in order to keep speech quality about the same. For instance, the pitch synchronous innovation CELP (PSI-CELP) coder adapts the fixed random codevectors in voiced frames to have periodicity (16).

Surprisingly, the analysis-by-synthesis operation of CELP is proving capable of delivering toll quality speech at lower rates when generalized to allow for a mixture of open-loop and closed-loop procedures (2) where parameters and excitation are determined in an open-loop fashion for clearly recognizable subframe types such as stationary periodic or voiced segments and closed-loop algorithms are used for unvoiced or transient segments. Due to the scarcity of bits for representing the excitation, it makes sense to predistort the target vector for closed-loop searches when it is clearly voiced since it becomes easier to match a codevector to it. The predistortion has to be perceptually transparent such as the time warping described in (17).

In a different trend, the development of text-to-speech (TTS) systems has been moving away from the rule-based, expert system approach to the new framework of concatenative synthesis, based on model fitting with statistical signal processing (18). In rule-based systems subword speech units are designed as well as rules for concatenating them which take into account the coarticulation between neighboring units as well as their exchange for allophonic variations. On the other hand, concatenative synthesis systems are based on the acquisition of a large database of connected speech from an individual speaker containing instances of coarticulation between all possible units. For the latter systems, the synthesis consists of selecting the largest possible string of original database subunits, thereby borrowing their natural concatenation. The final postprocessing stage of the TTS adjusts the prosody of the synthetic signal, mostly by pitch and time scale modifications. For segment selection, a concatenative synthesizer uses both an acoustic cost within each segment as well as a concatenation cost between consecutive segments (3). If the input feature vector sequence $\mathbf{F} = \mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N$ is to be synthesized by the unit sequence $\mathbf{U} = \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$, the acoustic cost may be defined by

$$J_A(\mathbf{f}_m, \mathbf{u}_m) = \sum_{k=1}^{K} (f_{m,k} - u_{m,k})^2 \qquad (13)$$

for segment m, where k indexes through the K features selected for comparison, normally the spectral representation of the subunits, and the concatenation cost may be calculated by

$$J_C\left(\mathbf{u}_{m-1}, \mathbf{u}_m\right) = \sum_{k=1}^{K} \left(u_{m-1,k} - u_{m,k}\right)^2 . \qquad (14)$$

The best subunit sequence is selected by minimization of the total cost $J\left(\mathbf{F}, \mathbf{U}\right)$ whose simplest definition is

$$J\left(\mathbf{F}, \mathbf{U}\right) = \sum_{m=1}^{N} J_A\left(\mathbf{f}_m, \mathbf{u}_m\right) + \sum_{m=2}^{N} J_C\left(\mathbf{u}_{m-1}, \mathbf{u}_m\right) . \qquad (15)$$

By using these kinds of cost measures in their analysis, concatenative synthesizers are becoming more similar to speech coders.

IV. Low-rate coding approaches

Speech coding allows more users to share a communications channel such as a mobile telephone cell or a packet network link and is concerned with the economical representation of a speech signal with a given distortion for a specified implementation complexity level. Traditionally, a fixed bit rate and an acceptable maximum distortion are specified. More generally, the required maximum bit rate or the acceptable maximum distortion level may be specified. Actually, for modern cellular or packet communications, sometimes the bit rate may be dictated by channel traffic constraints, requiring variable bit rate coders.

Objective fidelity measures such as the segmental signal-to-noise ratio (SNRSEG) are very practical for coder development while more perceptual like objective distortion measures like the perceptual speech quality measure (PSQM) (19) that use to advantage the handicaps of the human ear may be used instead. But still the opinion of human listeners is the best gauge of fidelity and may be assessed by the *mean opinion score* (MOS), obtained in formal listening tests where each listener classifies the speech stimulus on the 5-point scale shown in Table III.

Coder complexity constrains the possibilities of rate-distortion trade-off. Its major component is operational complexity, liable to be measured in million operations per second (MIPS) (20). An artistic conception of the fidelity versus rate behavior of low-rate coders for two levels of complexity is presented in Figure 8, anchored by some real coder test points, listed in Table IV. It should be said that these fidelity curves go through a kind of knee around the 4 kbit/s rate where they evolve at a lower slope, eventually reaching a virtual plateau at high rates (21).

Low bit rate implementations of models tested at higher rates need compensation for the loss of resolution or reduction of parameters whereas very low bit rate implementations admit refinements when upgraded to the low-rate range. In general, low-rate implementations require higher complexity algorithms and incur longer algorithmic delay. But a reduction in complexity may turn the original algorithm useful for a number of applications. This is one reason why a number of efficient search algorithms have been proposed since right after the inception of the CELP coder such as (22) which proposed a residual-based preselection of codevectors and the efficient transform-domain search algorithms elaborated by (23). Another preselection of codevectors was proposed by (24) based on the correlation between the backward-filtered target vector and segments of codevectors. The latter efficient search was called "focused search" and was adopted for the reference ITU-T 8 kbit/s CS-ACELP coder (15) with an open-loop signal-selected pulse amplitude approach. This coder is used for transmitting voice over packet networks among other applications.

In fact, the acceptance of this family of coders is so wide that most of the second-generation digital cellular coders use it, including the Telecommunications Industry Association (TIA) IS-641 enhanced full rate (EFR) coder (25) and the IS-127 enhanced variable rate coder (EVRC) (26) as well as the GSM EFR coder (27). Besides, a general purpose efficient search algorithm for ACELP fixed excitation codebook has been proposed, the joint position and amplitude search (JPAS) (28), which includes a closed-loop sequential pulse amplitude determination and a more efficient search for the EVRC (29) has been advanced as well. Also, a generalization of "algebraic pulses" by "algebraic subvectors" is the basis for the algebraic vector quantized CELP (AVQ-CELP) search, which enhances the IS-127 coder and uses open-loop subvector preselection in order to make it efficient (30).

As the bit rate is decreasead below 6 kbit/s ACELP coder quality degrades due to the uniform pulse density in the pulse position grid (31) and the high level of sparsity in the resulting excitation waveform. In an effort to push down the bit rate for ACELP applications, pulse dispersion techniques have been proposed such as (32) and (33). The former closed-loop technique is incorporated in a partially qualified candidate for the ITU-T 4 kbit/s coder (2). Furthermore, parametric coders such as MELP also implement pulse dispersion but as an open-loop enhancement in the decoder as mentioned in Section II-A. Along with pulse dispersion, the pulse position in the grid should be changed adaptively since it will not be able to cover all the positions (34), (31).

Another technique which holds promise for lower bit rate coding is target vector predistortion. Time-warping predistortions have already been proposed as mentioned in Section III and even used in the IS-127 EVRC.

The segments coded open loop may use enhanced vocoder-like techniques such as those used in the MELP or sinusoidal coders or, alternatively, WI techniques with a partial use of analysis-by-synthesis methods (35).

The judicious application of these enhancement techniques requires the classification of the signal into voice or silence. In the former case, the speech signal is classified into voiced and unvoiced stationary segments at least. Even the identification of transients may be required as a next step. Branching out further, speech classification might get down to subunits such as triphones, diphones and phones. In these cases the segmentation is event-driven as used for very-low-rate coding (36). Anyway, one should bear in mind that irregular segmentation requires time-scale modification as a post-processing stage, which may introduce annoying artifacts into the reconstructed signal. So sometimes it may be wise to maintain regular frame-based segmentation even at very low rates in order to ensure a certain uniform quality level (3).

In conclusion, the CELP framework with some relaxed waveform matching constraints, allowing for perceptual quality preserving signal predistortion and more segments of simple parametric coding, is very likely to be able to achieve toll quality at 4 kbit/s. It is anticipated as well that coders based on codebooks of sequences of speech subunits with properly defined distortion measures will also play an important role in advancing the toll quality frontier into the low bit rate range.

References

[1] S. Dimolitsas, C. Ravishankar, and G. Schröder, "Current objectives in 4-kb/s wireline-quality speech coding standardization," *IEEE Signal Processing Letters*, vol. 1, no. 11, pp. 157–159, Nov. 1994.

[2] J. Thyssen, Yang Gao, A. Benyassine, E. Shlomot, C. Murgia, Huan-yu Su, K. Mano, Y. Y. Hiwasaki, H. Ehara, K K. Yasunaga, C. Lamblin, B. Kovesi, J. Stegmann, and Hong-Goo Kang, "A candidate for the ITU-T 4 kbit/s speech coding standard," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, 2001, vol. 2, pp. 681–684.

[3] Ki-Seung Lee and R. V. Cox, "A very low bit rate speech coder based on a recognition/synthesis paradigm," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 5, pp. 482–491, Jul. 2001.

[4] A. S. Spanias, "Speech coding: A tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct. 1994.

[5] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, chapter 7, pp. 459–487, Macmillan, New Jersey, 1993.

[6] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer, Berlin, 1976.

[7] A. McCree, K. Truong, E. Bryan George, T. P. Barnwell, and V. Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new U. S. Federal Standard," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, 1996, vol. 1, pp. 200–203.

[8] R. J. McAulay and J. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. Bastiaan Kleijn and K. K. Paliwal, Eds., pp. 121–173. Elsevier Science, Amsterdam, 1995.

[9] W. Bastiaan Kleijn and K. K. Paliwal, "An introduction to speech coding," in *Speech Coding and Synthesis*, W. Bastiaan Kleijn and K. K. Paliwal, Eds., pp. 1–47. Elsevier Science, Amsterdam, 1995.

[10] W. Bastiaan Kleijn, "A frame interpretation of sinusoidal coding and waveform interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Istanbul, 2000, vol. 3, pp. 1475–1478.

[11] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, 1985, vol. 2, pp. 437–440.

[12] W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, "Fast methods for the CELP speech coding algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 8, pp. 1330–1342, Aug. 1990.

[13] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, 1982, vol. 1, pp. 614–617.

[14] R. V. Cox, "Speech coding standards," in *Speech Coding and Synthesis*, W. Bastiaan Kleijn and K. K. Paliwal, Eds., pp. 49–78. Elsevier Science, Amsterdam, 1995.

[15] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and description of CS-ACELP, a toll quality 8 kb/s speech coder," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 2, pp. 116–130, Mar. 1998.

[16] K. Mano, T. Moriya, S. Miki, H. Ohmuro, K. Ikeda, and J. Ikedo, "Design of a pitch synchronous innovation CELP coder for mobile communications," *IEEE J. Select. Areas Commun.*, vol. 13, no. 1, pp. 31–40, Jan. 1995.

[17] W. Bastiaan Kleijn, R. P. Ramachandran, and P. Kroon, "Generalized analysis-by-synthesis coding and its application to pitch prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Francisco, 1992, vol. 1, pp. 23–26.

[18] Y. Sagisaka and N. Iwahashi, "Objective optimization in algorithms for text-to-speech synthesis," in *Speech Coding and Synthesis*, W. Bastiaan Kleijn and K. K. Paliwal, Eds., pp. 685–706. Elsevier Science, Amsterdam, 1995.

[19] "*Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*," ITU-T Recommend. P.861, Aug. 1996.

[20] P. Kroon, "Evaluation of speech coders," in *Speech Coding and Synthesis*, W. Bastiaan Kleijn and K. K. Paliwal, Eds., pp. 467–494. Elsevier Science, Amsterdam, 1995.

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

[22] L. A. Hernández-Gómez, F. J. Casajús-Quirós, A. R. Figueiras-Vidal, and R. García-Gómez, "On the behaviour of reduced complexity code-excited linear prediction (CELP)," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 1986, vol. 1, pp. 469–472.

[23] I. M. Trancoso and B. S. Atal, "Efficient procedures for finding the optimum innovation in stochastic coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 1986, vol. 4, pp. 2375–2378.

[24] C. Laflamme, J.-P. Adoul, R. Salami, S. Morisette, and P. Mabilleau, "16 kbps wideband speech coding technique based on algebraic CELP," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Toronto, 1991, vol. 1, pp. 13–16.

[25] T. Honkanen, J. Vainio, K. Järvinen, and P. Haavisto, "Enhanced full rate codec for IS-136 digital cellular system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Munich, 1997, vol. 2, pp. 731–734.

[26] "*Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*," TIA/EIA/IS-127, Jul. 1996.

[27] K. Järvinen, J. Vainio, P. Kapanen, T. Honkanen, P. Haavisto, R. Salami, C. Laflamme, and J.-P. Adoul, "GSM enhanced full rate speech codec," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Munich, 1997, vol. 2, pp. 771–774.

[28] M. Arjona Ramírez and M. Gerken, "Joint position and amplitude search of algebraic multipulses," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 5, pp. 633–637, Sept. 2000.

[29] H. Park, "Efficient codebook search method of EVRC speech codec," *IEEE Signal Processing Letters*, vol. 7, no. 1, pp. 1–2, Jan. 2000.

[30] Fenghua Liu and R. Heidari, "Improving EVRC half rate by the algebraic VQ-CELP," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, 1999, vol. 4, pp. 2299–2302.

[31] V. Cuperman, A. Gersho, J. Lindén, A. Rao, Tung-Chiang Yang, S. Ahmadi, R. Heidari, and Fenghua Liu, "A novel approach to excitation coding in low-bit-rate high-quality CELP coders," in *Proc. IEEE Workshop on Speech Coding*, Delavan, 2000, pp. 14–16.

[32] K. Yasunaga, H. Ehara, K. Yoshida, and T. Morii, "Dispersed-pulse codebook and its application to a 4 kb/s speech coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Istanbul, 2000, vol. 3, pp. 1503–1506.

[33] M. Arjona Ramírez, "Sparsity compensation for speech coders," in *Proc. of IEEE GLOBECOM*, San Antonio, 2001, vol. 4, pp. 2475–2478.

[34] T. Amada, K. Miseki, and M. Akamine, "CELP speech coding based on an adaptive pulse position codebook," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, 1999, vol. 1, pp. 13–16.

[35] O. Gottesman and A. Gersho, "Enhanced waveform interpolative coding at low bit-rate," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 8, pp. 786–798, Nov. 2001.

[36] C. S. Xydeas and T. M. Chapman, "Segmental prototype interpolation coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, 1999, vol. 4, pp. 2311–2314.

[37] M. A. Kohler, "A comparison of the new 2.4 kbps MELP Federal Standard with other standard coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Munich, 1997, vol. 2, pp. 1587–1590.

[38] M. E. Perkins, K. Evans, D. Pascal, and L. A. Thorpe, "Characterizing the subjective performance of the ITU-T 8 kb/s speech coding algorithm - ITU-T G.729," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 74–81, Sept. 1997.

[39] K. Mano, "Design of a toll-quality 4-kbit/s speech coder based on phase-adaptive PSI-CELP," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Munich, 1997, vol. 2, pp. 755–758.

[40] W. Bastiaan Kleijn and J. Haagen, "Waveform interpolation for coding and synthesis," in *Speech Coding and Synthesis*, W. Bastiaan Kleijn and K. K. Paliwal, Eds., pp. 175–207. Elsevier Science, Amsterdam, 1995.

[41] R. V. Cox and P. Kroon, "Low bit-rate speech coders for multimedia communication," *IEEE Commun. Mag.*, vol. 34, no. 12, pp. 34–41, Dec. 1996.

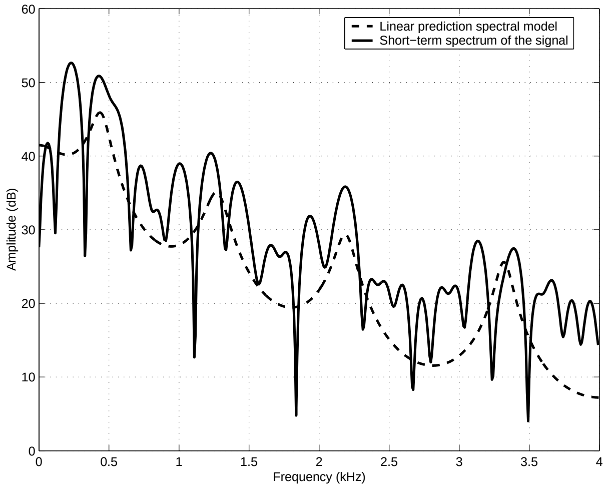

Fig. 1. Linear prediction spectral fit to the envelope of the short-term log spectrum of the signal.

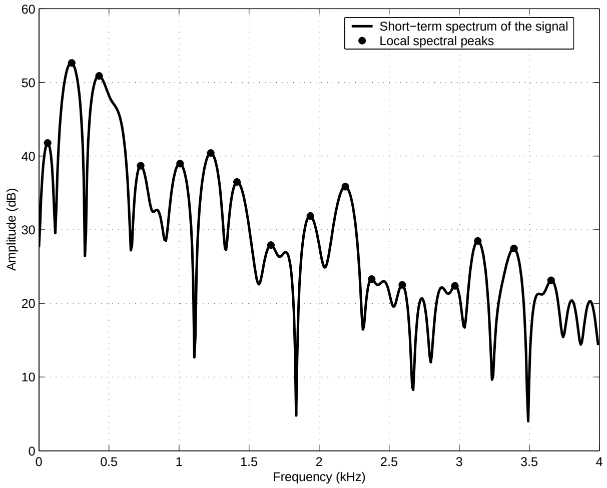

Fig. 3. Short-term log spectrum of the signal with selected local peaks.

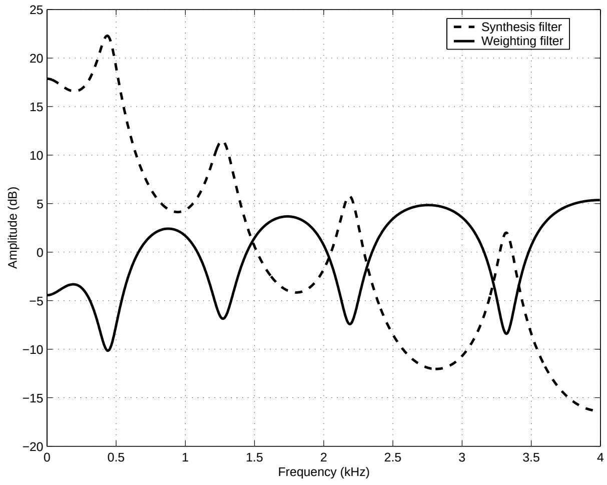

Fig. 6. Frequency responses of synthesis filter and corresponding perceptual weighting filter.

TABLE II

ACELP position grid for M pulse tracks over an L-sample subframe.

Track	Positions				
0	0	M	$2M$	$\cdots$	$L-M$
1	1	$M+1$	$2M+1$	$\cdots$	$L-M+1$
2	2	$M+1$	$2M+2$	$\cdots$	$L-M+2$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$M-1$	$M-1$	$2M-1$	$3M-1$	$\cdots$	$L-1$

TABLE III

Quality scale for subjective listening rating.

Quality	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

TABLE I
Bit rates of typical acoustic signals

	Bandwidth	Sampling frequency	Bits per sample	Bit rate
Narrowband speech	300 Hz - 3.4 kHz	8.0 kHz	8	64 kbit/s
Wideband speech	50 Hz - 7.0 kHz	16.0 kHz	14	224 kbit/s
Wideband audio (DAT format)	10 Hz - 20.0 kHz	48.0 kHz	16	768 kbit/s
Wideband audio (CD format)	10 Hz - 20.0 kHz	44.1 kHz	16	706 kbit/s

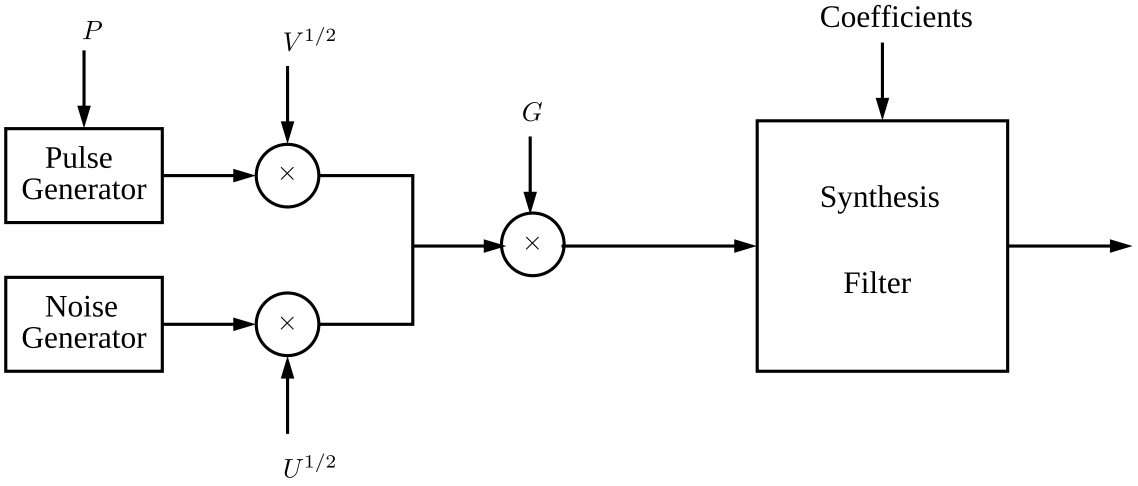

Fig. 2. Mixed source and filter model for speech synthesis.

TABLE IV
Speech quality and operational complexity of some selected coders.

Coder	Bit rate (kbit/s)	Quality (MOS)	Complexity (MIPS)	References
LPC-10e, FS-1015	2.40	2.30	8.7	(37)
MELP, FS-1017	2.40	3.30	20.4	(37)
EWI	2.80	∼ 3.80	∼ 30.0	(35), (38), (33)
PSI-CELP, RCR PDC half-rate	3.45	∼ 3.40	23.0	(16), (39), (14), (38)
IMBE, Inmarsat-M System	4.15	3.40	7.0	(4), (14)
CELP, FS-1016	4.80	3.59	17.0	(40), (37)
STC	4.80	3.53	∼ 25.0	(8)
WI	4.80	3.77	∼ 25.0	(40)
ACELP, G.723.1	5.33	3.55	16.0	(33), (41)
CS-ACELP, G.729	8.00	3.92	20.0	(38), (41)

∼: Estimate

Caution: These performance and complexity figures were obtained under different test and implementation conditions and should be used only as a first guess in comparisons.

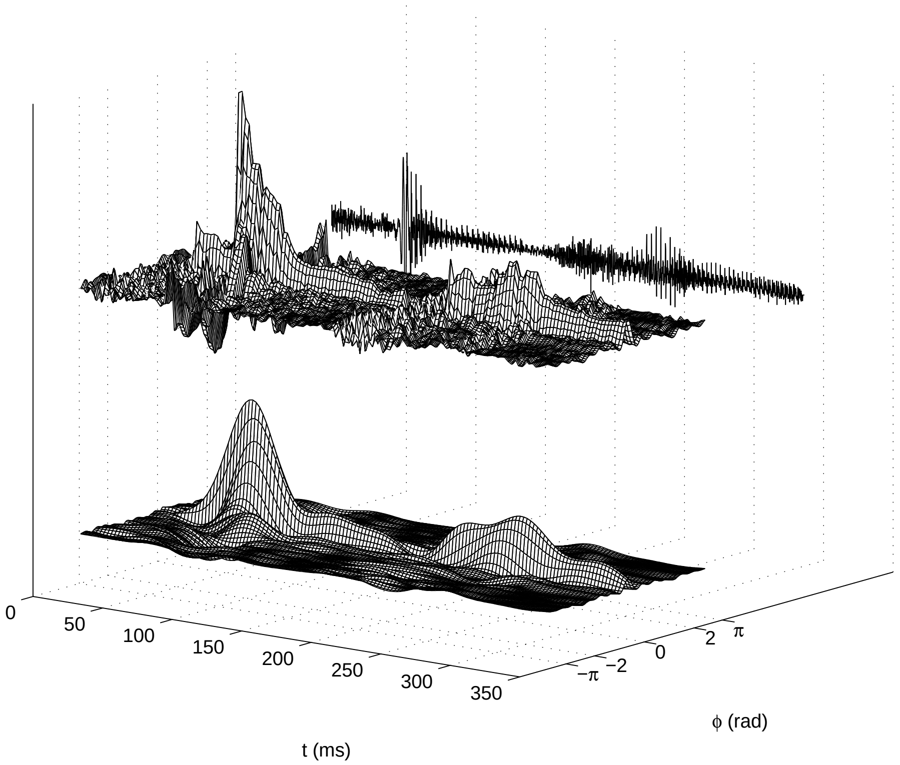

Fig. 4. Characteristic surface for WI coding the residual signal given behind whose underlying CWs have been extracted at a 400 Hz rate. Its SEW component is also shown below which has been obtained by lowpass filtering the characteristic surface along the time axis with a cutoff frequency of 20 Hz.

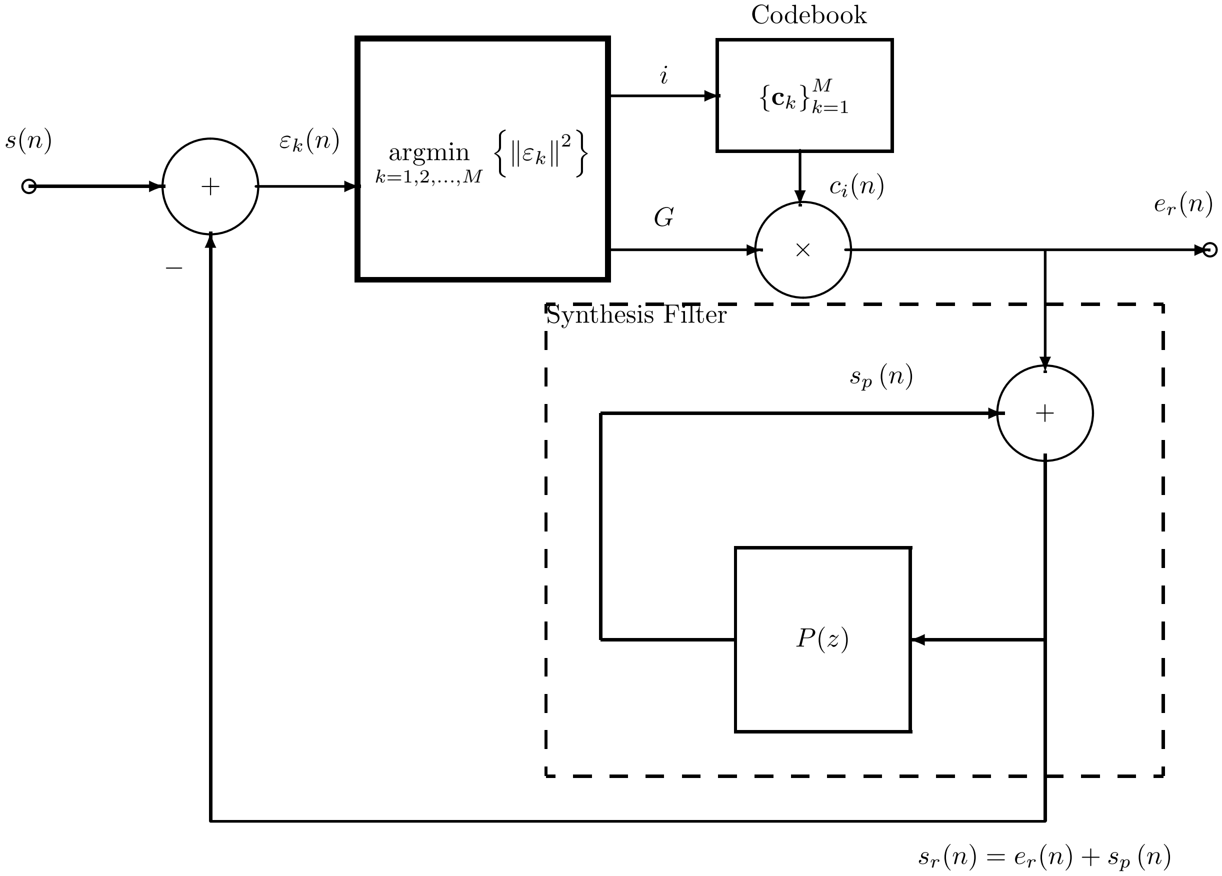
Codebook
$\{\mathbf{c}_k\}_{k=1}^{M}$
i
$s(n)$
$\varepsilon_k(n)$
$\underset{k=1,2,\ldots,M}{\text{argmin}} \left\{\|\varepsilon_k\|^2\right\}$
+
−
$c_i(n)$
G
×
$e_r(n)$
Synthesis Filter
$s_p(n)$
+
$P(z)$
$s_r(n) = e_r(n) + s_p(n)$

Fig. 5. Conceptual block diagram for CELP coding.

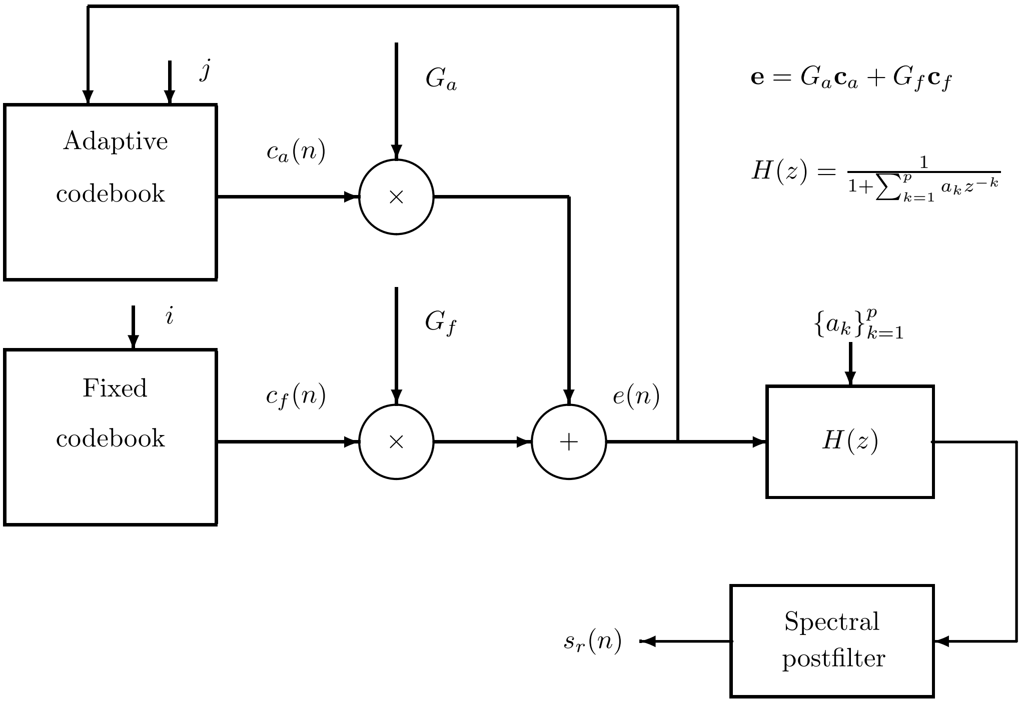

Fig. 7. Two-codebook CELP synthesis model.

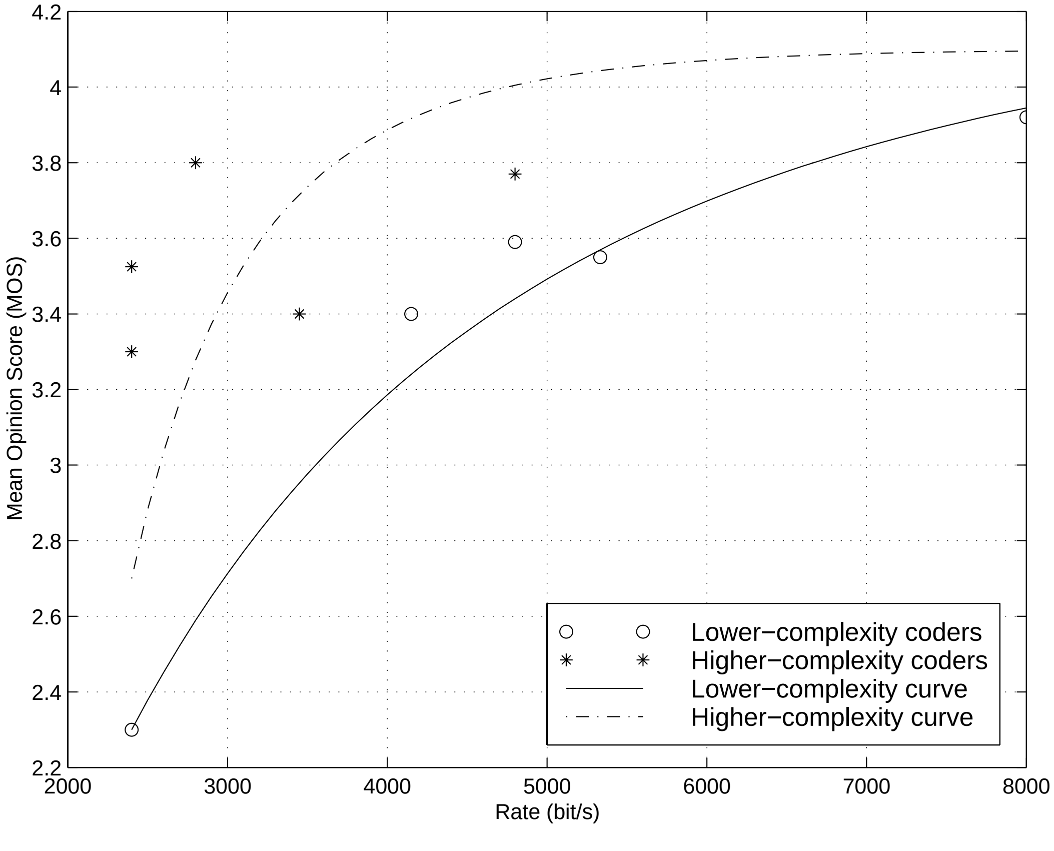

Fig. 8. Conception of the fidelity versus rate behavior of low-rate speech coders for two levels of complexity, anchored by some real coder test points, listed in Table IV.