SPEECH CODING AT 4 KB/S AND LOWER USING SINGLE-PULSE AND STOCHASTIC MODELS OF LPC EXCITATION

W. Granzow, B.S. Atal, K.K. Paliwal, and J. Schroeter

AT&T Bell Laboratories, Murray Hill, NJ 07974-2070, U.S.A.

ABSTRACT

Accurate representation of periodic speech segments is essential for synthesizing high-quality digital speech. For bit rates at and below 4 kb/s, conventional code-excited linear predictive coding (CELP) does not provide the appropriate degree of periodicity. Small codebook size and coarse quantization of gain factors result in large spectral fluctuations between pitch periods. At low bit rates, smoothness of spectral changes can be achieved by an excitation function which contains one excitation pulse of fixed or slowly time-varying shape for each pitch period. Earlier work has shown that single-pulse excitation can achieve reasonably good speech quality. However, this method was based on an optimization procedure that caused a very large coding delay. In this paper, we present a linear predictive coder that classifies speech into periodic and nonperiodic intervals. Nonperiodic speech is synthesized as in CELP. Periodic speech is synthesized using single-pulse excitation. The coder is based on a new algorithm for determining pitch markers within short blocks of periodic speech. This algorithm requires a small coding delay and is implemented efficiently by dynamic programming.

I. INTRODUCTION

In code-excited linear predictive coding (CELP), the excitation for the all-pole synthesis filter is modeled as a sum of two gain-scaled vectors [1]. The first vector is obtained from an adaptive codebook which contains the past excitation [2]. The second vector is obtained from a fixed stochastic codebook. The two excitation vectors are selected by minimizing the perceptually-weighted error between original and reconstructed speech [1]. In CELP, the concept of repeating the past excitation is essential for obtaining an appropriately periodic excitation.

At low bit rates, the stochastic codebook is restricted to a small size and the gain factors are coarsely quantized. This significantly reduces the ability to produce a periodic excitation. The stochastic codebook vectors of a fixed block size cause large fluctuations in the spectrum of the reconstructed speech. As a result, the reconstructed speech has a noisy character.

An obvious method for achieving highly periodic speech is to represent each period of excitation by a pulse-like signal with a fixed or slowly time-varying shape. The pulse shape can be defined as a cluster of delta-impulses or, in its simplest form, as a single delta-impulse. We will refer to this representation of the excitation as single-pulse excitation (SPE). The single-pulse excitation is described by the time location, shape and the gain of each pulse. Such a parametric representation also enables interpolation of the excitation parameters for efficient encoding.

In principle, a pulse-like excitation could be generated with CELP (without an adaptive codebook) by using a fixed codebook that contains individual pulse shapes as well as sequences of such shapes for vector dimensions larger than a possible period. However, the selection of pulse vectors from the codebook, based on individually encoding relatively short speech blocks, usually does not result in an excitation with smoothly evolving pulse intervals. For obtaining such a consistent periodic excitation, large optimization frames are required.

Previous work [3] has shown that good speech quality can be achieved if periodic speech is synthesized by exciting an all-pole LPC filter with an optimized single-pulse excitation consisting of one delta-impulse per period.

In this paper, we present an LPC speech coder that classifies speech into periodic and nonperiodic intervals. The coder uses a stochastic codebook, as in CELP, to synthesize nonperiodic speech, and single-pulse excitation to synthesize periodic speech. We refer to this coder as SPE-CELP. The optimization of the single-pulse excitation is based on a new algorithm that determines the time instants of pitch periods within a short interval of periodic speech of approximately 32 ms. It thus causes a coding delay that is acceptable for many applications. (Earlier methods did not attempt to constrain the delay.) We report on results using a fixed delta-impulse shape and time-varying pulse shapes obtained from codebooks.

II. DETERMINATION OF PITCH MARKERS

We assume that the speech signal is encoded in frames, with a fixed number of bits per frame. These frames will be referred to as coding frames. A typical length of a coding frame is 200 samples (25 ms at a sampling frequency of 8 kHz). A coding frame is subdivided into four subframes of 50 samples each. Each subframe is classified as either periodic or nonperiodic using a modified autocorrelation algorithm [4, 5] that is based on computing the long-term autocorrelation of the preprocessed speech signal within a window around a subframe. For each periodic subframe, an average pitch period $\bar{p}$ is determined. The periodic/nonperiodic classification is smoothed in order to allow, within a coding frame, at most one transition from periodic to nonperiodic or vice-versa. A sequence of up to five periodic subframes following a nonperiodic-to-periodic transition is created to form an optimization frame as shown in Fig. 1. The optimization frames extend beyond the coding frames by at least one subframe if this overlapping interval is also periodic. The pitch markers are successively determined for each optimization frame, but only those markers are retained that fall into the actual coding frame.

The pitch markers are determined under the assumption that they define the optimal locations for single-pulse excitation of a delta-impulse shape based on a subjectively meaningful error criterion.

In our experiments, we found that it is generally not possible to determine a consistent sequence of excitation pulses if an error criterion is used that is based only on minimizing a weighted mean-squared error as in CELP. Therefore, we apply an error criterion that combines different perceptually meaningful optimization criteria in the cost function defined below. The optimization procedure, inspired by the work of Talkin [6], is efficiently implemented using dynamic programming.

CH2977-7/91/0000-0217 $1.00 © 1991 IEEE

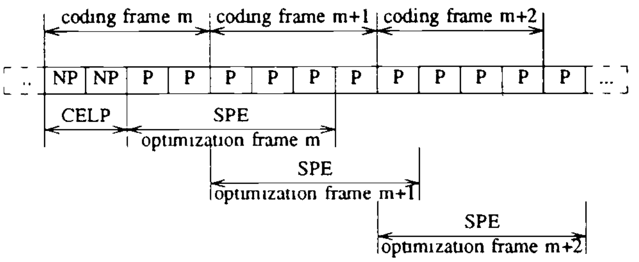

Fig 1 Definition of optimization frames relative to coding frames

We first compute the maximum signal-to-noise ratio, measured over short intervals, that can be obtained by exciting an LPC synthesis filter with a single pulse of optimal amplitude α at time location n. Let us denote the impulse response of the LPC synthesis filter by a vector $\mathbf{h}_0 = [h(0), \ldots, h(2N-1)]^T$. This impulse response vector is updated for each periodic subframe of length N. We define an error vector $\mathbf{e}_n$ as the difference between the speech vector $\mathbf{x}$ and the delayed impulse response vector $\mathbf{h}_n$ multiplied by the pulse amplitude α:

$$\mathbf{e}_n = \mathbf{x} - \alpha \mathbf{h}_n , \tag{1}$$

where

$$\mathbf{x} = [x(0), x(1), \ldots, x(2N-1)]^T ,$$

and

$$\mathbf{h}_n = [0, \ldots 0, h(0), \ldots, h(2N-1-n)]^T , \; n = 0, \ldots, N-1 .$$

Minimization of the error energy with respect to the pulse amplitude α leads to

$$\min_{\alpha} \mathbf{e}_n^T \mathbf{e}_n = \mathbf{x}^T \mathbf{x} - \frac{(\mathbf{x}^T \mathbf{h}_n)^2}{\mathbf{h}_n^T \mathbf{h}_n} , \tag{2}$$

$$\alpha_{opt}(n) = \frac{\mathbf{x}^T \mathbf{h}_n}{\mathbf{h}_n^T \mathbf{h}_n} , \tag{3}$$

and

$$S_{opt}(n) = \max_{\alpha} SNR(n) = 10 \log_{10} \left[\frac{\mathbf{x}^T \mathbf{x}}{\min\limits_{\alpha} \mathbf{e}_n^T \mathbf{e}_n} \right] . \tag{4}$$

We compute $S_{opt}(n)$ and $\alpha_{opt}(n)$ for each time instant n in an optimization frame. Both these functions reveal the quasi-periodicity of the speech signal $x(n)$. Note that $S_{opt}(n)$ does not depend on the actual speech energy.

Next we identify locations at which $S_{opt}(n)$ has a local maximum exceeding a certain threshold, with the constraint that the corresponding $\alpha_{opt}(n)$ is of constant sign. These local maxima, and neighboring locations, form a preliminary set of candidate locations n_i for the pitch markers. The final selection is made as follows:

For each candidate pitch marker, n_i, define the 3-tuple $z_i = (n_i, \alpha_{opt}(n_i), S_{opt}(n_i))$ and let $Z = \{z_i, i = 1, \ldots, M\}$ be the set of z_i found. Let s be a K-tuple ($K \geq 2$) of the z_i's such that the intervals between two candidates n_i lie in a prescribed range, and let S be the set of all such valid K-tuples. Each s defines a sequence of K ($K \geq 2$) pitch markers. The optimum sequence s_{opt} is the one that minimizes an accumulated cost C. This accumulated cost is determined by a cost function f that consists of four summation terms. The first term penalizes candidates with low $S_{opt}(n_i)$; the second term penalizes inconsistency in the amplitude of two successive pulse candidates; the third term penalizes inconsistency in two successive pulse intervals $n_i - n_j$ and $n_j - n_k$; the fourth term penalizes a deviation in the pulse interval $n_i - n_j$ from the initial estimate of the average pitch period $\bar{p}(n_i)$. Thus

$$s_{opt} = \{ z_{q_1}, \cdots, z_{q_K} \}, \; K \geq 2 \tag{5}$$

$$\textit{with indices } Q = \{ q_1, \cdots, q_K \mid q_k \in [1, M], \; n_{q_k} > n_{q_{k-1}} \}$$

is the K-tuple that minimizes the accumulated cost

$$C = \min_{S} \frac{1}{K} \left[f_{ini}(i = q_1) + \sum_{l=2}^{K} f(i = q_l, \, j = q_{l-1}, \, k = q_{l-2}) \right] \tag{6}$$

where

$$\begin{aligned} f(i,j,k) = {} & \frac{a}{S_{opt}(n_i)} + b \left| \ln \frac{\alpha_{opt}(n_i)}{\alpha_{opt}(n_j)} \right| \\ & + c \left| \ln \frac{n_i - n_j}{n_j - n_k} \right| + d \left| \ln \frac{n_i - n_j}{\bar{p}(n_i)} \right| , \; n_i > n_j > n_k . \end{aligned} \tag{7}$$

The initial cost $f_{ini}(i = q_1)$ in (6) is computed as

$$f_{ini}(i) = \begin{cases} \dfrac{a}{S_{opt}(n_i)} + d \ln \dfrac{n_i}{\bar{p}(n_i)} + f_{fix}, & n_i > \bar{p}(n_i) \\[2ex] \dfrac{a}{S_{opt}(n_i)} + f_{fix}, & n_i \leq \bar{p}(n_i) \end{cases} \tag{8}$$

in the first optimization frame following a nonperiodic-to-periodic transition. The term f_{fix} is a constant. In all subsequent optimization frames, the initial cost is computed as $f_{ini}(i = q_1) = f(i = q_1, \, j = q_0, \, k = q_{-1}) + f_{fix}$ where $f(i,j,k)$ is obtained from (7) with n_j and n_k defined as the last two pitch marker locations within the preceding coding frame. This link of the cost computation to the previous coding frame assures continuity of pulse intervals at frame boundaries. The third term in (7) is dropped if there is no predecessor at time location n_k.

The four factors a, b, c, d were determined experimentally in order to get a proper weighting of all summation terms in the cost function. The indices Q of the best subset s_{opt} define the locations of the pitch markers within the current optimization frame. Locations outside of the current coding frame are dropped.

An example illustrating the pitch marker determination is given in [7].

III. HYBRID SPEECH CODER BASED ON SINGLE-PULSE EXCITATION AND CELP

Coder structure

Figure 2 shows the analyzer structure of SPE-CELP. In principle, the structure is identical to a conventional analysis-by-synthesis scheme [1]. We have added a periodic/nonperiodic (P/NP) classification that is used to switch between different modes of generating the excitation. Nonperiodic speech subframes are

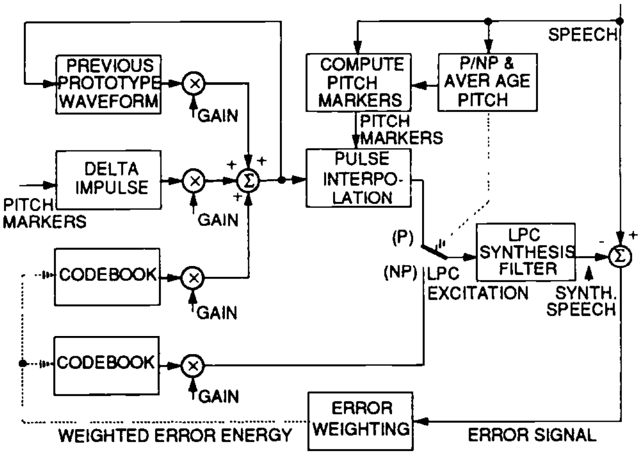

Fig. 2. Analyzer structure of SPE-CELP

individually encoded by selecting an excitation vector from a stochastic codebook as in CELP. In this mode, an adaptive codebook is not used. Periodic speech subframes are synthesized with single-pulse excitation.

For periodic speech intervals, we compute the pitch markers as described in the previous section. Each period of speech is synthesized with a pulse-like excitation that is derived as follows. Let us number the last pitch marker in the previous coding frame as M_0 and the pitch markers within the current coding frame as $M_1, \ldots, M_K$, as shown in Fig. 3. We define an excitation frame $u_k(n)$, $k = 1, \ldots, K$ as the interval around pitch marker M_k that is limited by the midpoints between the pitch markers M_{k-1} and M_k, and M_k and M_{k+1}. We denote the interval between pitch markers M_k and M_{k+1} as p_k. Let us define the pitch marker locations in each excitation frame as $n = 0$ and the right and left boundaries as $-l_k$ and m_k, respectively.

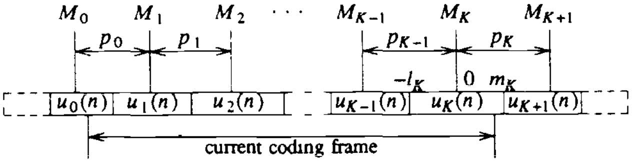

Fig. 3. Definition of excitation frames

For each coding frame, the pulse waveform for the K^{th} excitation frame is encoded and transmitted. These pulse waveforms are referred to as *prototype waveforms* as proposed in [8]. The intermediate excitation frames $u_k(n)$, $k = 1, \ldots, K-1$ are obtained by linear interpolation of the prototype waveforms of the previous and present coding frames $u_0(n)$ and $u_K(n)$ [8]:

$$u_k(n) = \left(u_0(n) \sum_{i=k}^{K-1} p_i + u_K(n) \sum_{i=0}^{k-1} p_i \right) \Big/ \sum_{i=0}^{K-1} p_i, \; -l_k \le n \le m_k. \quad (9)$$

In the simplest case, e.g., for a low-bit-rate implementation, we represent each prototype waveform by one delta-impulse at a pitch marker location,

$$u_K(n) = \alpha_1 \, \delta(n), \; -l_K \le n \le m_K. \quad (10)$$

For this case, the interpolation (9) results in an excitation consisting of one delta-impulse at each pitch marker location with linearly interpolated amplitude.

If additional bits are available, we represent each prototype waveform as the sum [8],

$$u_K(n) = \alpha_1 \, \delta(n) + \alpha_2 \, u_0(n) + \sum_{i=1}^{I} \beta_i \, c_i(n), \; -l_K \le n \le m_K. (11)$$

where the first component is a delta-impulse, the second component is the gain-scaled previous prototype and the third component is a sum of gain-scaled vectors $c_i(n)$ that are selected from I codebooks. The gain factors are jointly optimized by minimizing a perceptually weighted mean-squared error criterion as in CELP [1]. Note that at nonperiodic-to-periodic transitions, the prototype $u_0(n)$ is always defined as a delta-impulse with an amplitude derived from the energy of the previous nonperiodic subframe.

A 10^{th} order LPC analysis is carried out for every coding frame. The filter parameters are vector-quantized in the line-spectral frequency domain [9]. We linearly interpolate the quantized filter parameters and reset them to a new value once every excitation frame during periodic speech intervals and once every subframe during nonperiodic speech intervals. The interpolation smoothes transitions of the LPC spectrum and reduces distortions caused by block adaptation of the synthesis filter.

Bit allocation

For a coding-frame length of $L = 200$ samples and a minimum pulse interval of $p_{\min} = 20$ samples, the maximum number of pulses within a periodic coding frame is $k_{\max} = 10$. We assume that a coding frame contains at least one pulse. The total number N_p of pitch marker patterns can then be computed as

$$N_p = \sum_{k=1}^{k_{\max}} \binom{L-(k-1)(p_{\min}-1)}{k}. \quad (12)$$

We define a certain numbering of the pitch marker patterns and compute a binary code word for each such occurring pattern. In completely periodic coding frames, we require 34 bits for encoding pitch markers.

We allow at most one transition between periodic and nonperiodic subframes within a coding frame. Since the coding frame consists of four subframes, eight different periodic/nonperiodic subframe patterns can occur. For each coding frame, the frame classification is encoded with 3 bits. The filter parameters are encoded with 24 bits per coding frame.

At an overall bit rate of 3 kb/s, the total number of available bits per 25-ms coding frame is 75, leaving 48 bits for encoding of the excitation. Nonperiodic subframes are encoded using an 8-bit stochastic codebook and a 4-bit gain. In periodic speech intervals, the prototype waveform is represented according to (10) using a fixed delta-impulse shape. In completely periodic coding frames, we use a 34-bit block code to encode the pitch markers and 10 bits to represent the gain α_1 (the remaining 4 bits are not used). In transitional coding frames not all possible pitch marker locations are encoded and α_1 is quantized more coarsely.

At an overall bit rate of 4 kb/s, the total number of available bits per 25-ms coding frame is 100, leaving 73 bits for encoding of the excitation. This bit number allows a two-stage quantization of nonperiodic subframes using two stochastic codebooks. Alternatively, an adaptive codebook excitation could be added to the stochastic component as in conventional CELP. The prototype waveform is represented according to (11) using a time-varying pulse shape. The pitch markers are encoded as in the 3 kb/s-example. In completely periodic coding frames, we spend 8 bits for α_1, 5 bits each for α_2, β_1, and β_2, and we use two 8-bit codebooks.

Results

Figure 4 compares a periodic interval of the original and the reconstructed speech signals obtained from SPE-CELP at total bit rates of 4 and 3 kb/s. The single-pulse excitation of time-varying shape for 4 kb/s, and of delta-impulse shape for 3 kb/s are shown in Fig. 4 (c) and (e), respectively. In SPE-CELP, the reconstructed speech is synchronous with the original speech. At 4 kb/s, the synthetic speech achieves a good match to the original speech. Typically, a signal-to-noise ratio around 9 dB is achieved.

Figure 5 compares the narrow-band log-magnitude spectra of the original and single-pulse synthesized speech of a female speaker at 3 kb/s. The spectra were computed by applying a discrete Fourier transform to a 36 ms-speech segment (approximately 8 pitch periods). Both signals are bandlimited to the 0.2 - 3.8 kHz frequency range. The envelope and the harmonic structure of both spectra coincide reasonably well. However, occasionally the SPE-CELP reconstructed speech becomes more periodic in certain frequency intervals than the original speech. This can be seen in Fig. 5 for frequencies above 3 kHz.

At 3 kb/s, SPE-CELP achieves a good speech quality. Unlike low bit-rate CELP, the speech reconstructed with SPE-CELP is not distorted by background noise. However, sometimes a slight buzziness is perceivable that depends on the speaker and the recording conditions. By using a time-varying pulse shape, the buzziness can be completely removed. However, a comparison of the log-magnitude spectra of SPE-CELP reconstructed speech at 3 and 4 kb/s does not reveal significant differences. Thus, the improvement in speech quality at 4 kb/s appears to be primarily due to a better match of the phase characteristics.

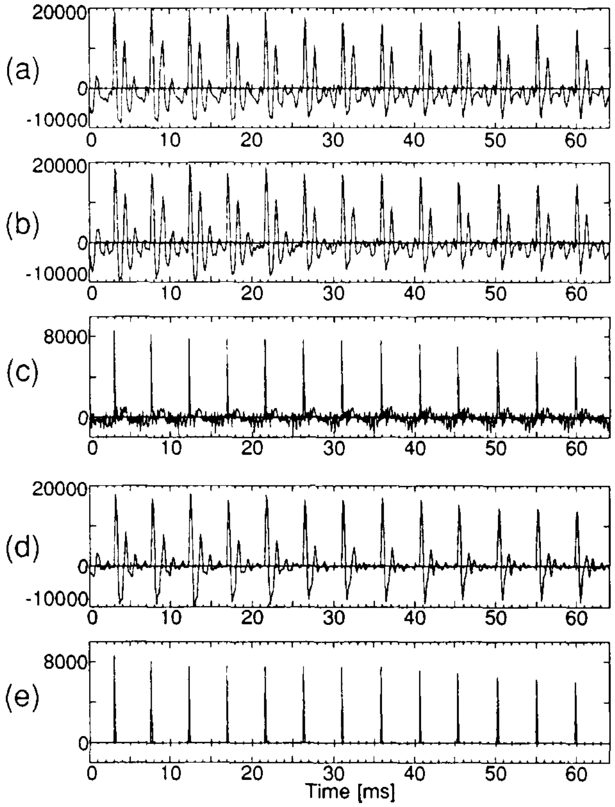

Fig 4. Waveforms of (a) original speech, (b) reconstructed speech at 4 kb/s, (c) single-pulse excitation of time-varying shape, (d) reconstructed speech at 3 kb/s, and (e) single-pulse excitation of delta-impulse shape

IV. SUMMARY

In this paper, we have introduced a speech coder based on a single-pulse and a CELP representation of the excitation. We proposed a new method for determining pitch markers which define the optimum locations of excitation pulses. An important feature of this method is that different perceptually meaningful optimization criteria are combined in a single cost function. These criteria are the mean-squared error between the original and the reconstructed speech and the consistency of successive pulse intervals and pulse amplitudes. The pitch marker determination is implemented efficiently using dynamic programming

The pitch markers have been used to detect individual periods of periodic speech within coding frames. Single-pulse excitation defined as one delta-impulse at a pitch marker location is the simplest form of representing a quasi-periodic LPC excitation. With such an excitation, SPE-CELP produces, at overall bit rates around 3 kb/s, significantly better speech quality than LPC10E, though the synthesized speech still sounds slightly buzzy for certain speakers.

We have also proposed a method to improve the speech quality, if additional bits are available. In this method, one pitch period per coding frame is encoded in an analysis-by-synthesis procedure to obtain a more appropriate time-varying excitation pulse shape. Pulse shapes are linearly interpolated for intermediate periods. The interpolation ensures smooth spectral transitions between pitch periods. By increasing the bit rate from 3 to 4 kb/s, the buzziness of SPE-CELP with fixed pulse shape is removed

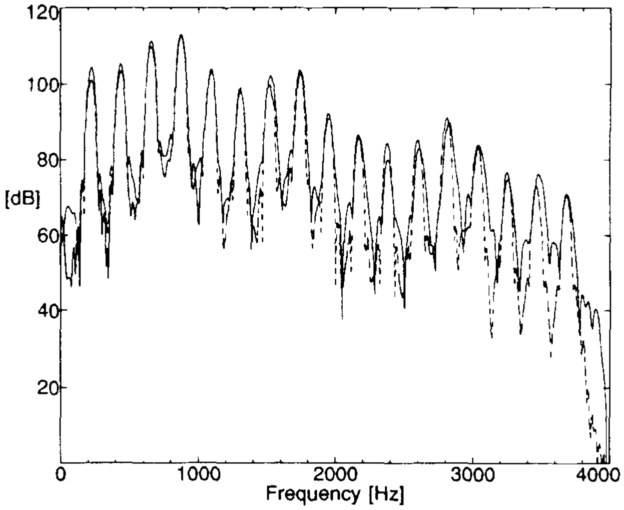

Fig 5. Comparison of spectra of the original speech (solid) and the synthesized speech using SPE-CELP at 3 kb/s (dashed)

ACKNOWLEDGMENT

We acknowledge helpful discussions with W. B. Kleijn, whose work [8] suggested encoding and interpolation of prototype waveforms. We also thank M. M. Sondhi for helpful comments.

REFERENCES

[1] M. R. Schroeder, B. S. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates," *Proc. IEEE Int. Conf. Acoust Speech Sig. Proc.*, pp. 937-940, 1985

[2] W.B. Kleijn, D.J. Krasinski, R.H. Ketchum, "Improved Speech Quality and Efficient Vector Quantization in SELP," *Proc IEEE Int Conf. Acoust. Speech Sig. Proc.*, pp. 155-158, 1988.

[3] B S. Atal, B.E. Caspers, "Beyond Multipulse and CELP Towards High Quality Speech at 4 Kb/s," *Advances in Speech Coding*, Kluwer Academic Publishers, 1990

[4] M M. Sondhi, "New Methods of Pitch Extraction," *IEEE Trans on Audio and Electroacoustics*, **AU-16**, pp 262-266, 1968.

[5] J. J. Dubnowski, R. W. Schafer, L. R. Rabiner, "Real-Time Digital Hardware Pitch Detector," *IEEE Trans Acoust Speech and Sig. Proc.*, **ASSP-24**, pp. 2-8, 1976.

[6] D. Talkin, "Voicing Epoch Determination with Dynamic Programming," *J. Acoust Soc Am.*, **Suppl. 1**, **Vol. 85**, pp. S149, 1989

[7] W Granzow, B.S Atal,"High-Quality Digital Speech at 4 kb/s," *Proc IEEE Global Telecommunications Conf.*, pp 941-945, 1990.

[8] W.B Kleijn, "Continuous Representations in Linear Predictive Coding," *Proc IEEE Int. Conf. Acoust Speech Sig Proc.*, 1991.

[9] K. K. Paliwal, B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," *J. Acoust. Soc Am.*, **Suppl. 1**, **Vol. 87**, pp. S39, 1990 (see also paper in this conference record).