# SPECTRAL EXCITATION CODING OF SPEECH AT 2.4 KB/S

*V. Cuperman, P. Lupini, and B. Bhattacharya*

School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6
Tel: (604) 291-4371, Fax: (604) 291-4951
email: vladimir@cs.sfu.ca, lupini@cs.sfu.ca, bhaskar@cs.sfu.ca

## ABSTRACT

In this paper we present Spectral Excitation Coding (SEC), a speech codec based on a sinusoidal model applied to the excitation signal. A phase dispersion algorithm allows the same model to be used for voiced as well as unvoiced and transitional sounds. The phase dispersion algorithm significantly improves the perceived quality resulting in more natural reconstructed speech. A new technique for variable-dimension vector quantization called Non-Square Transform Vector Quantization (NSTVQ) is used for quantization of the harmonic magnitudes. The SEC system at 2.45 kb/s achieved an MOS score 0.8 points higher than the 2.4 kb/s LPC-10 standard. A preliminary 1.85 kb/s SEC system which uses zero-bit magnitude quantization is also presented. Informal listening tests indicate that the quality of the 1.85 kb/s system exceeds that of the LPC-10 standard.

## 1. INTRODUCTION

Speech coding systems based on Code Excited Linear Prediction (CELP) may be used to achieve toll speech quality at 16 kb/s, close to toll quality speech at 8 kb/s, and good communications quality for special applications at 4 kb/s. Recently, toll quality has been achieved at 8 kb/s with relatively high complexity CELP systems. However, for rates around 4 kb/s and below, speech coding in the spectral domain has recently shown potential for better quality than the existing CELP based codecs [1]-[3]. Spectral domain coders try to reproduce speech magnitude spectra rather than the precise details of the speech waveform.

Spectral coding of speech is usually based on a sinusoidal speech production model. The sinusoidal model was applied directly to the speech signal in Multi Band Excitation (MBE) [1] and Sinusoidal Transform Coding (STC) [2]. Time Frequency Interpolation (TFI) uses a CELP codec for encoding unvoiced sounds, and the equivalent of a sinusoidal model applied to the excitation signal for encoding voiced sounds [3].

This paper presents Spectral Excitation Coding (SEC), a speech coding technique based on a sinusoidal model applied to the excitation signal. In the encoder, the excitation signal is obtained by passing the speech through a short-term linear prediction filter. The sinusoidal model parameters, including the pitch period, spectral magnitude shape, and signal level (gain) are extracted. The spectral magnitude information is encoded using a new technique called Non-Square Transform Vector Quantization (NSTVQ) and transmitted, while the spectral phases are synthesized in the receiver using phase prediction. A phase dispersion factor is transmitted which can decorrelate the predicted phases, for example during unvoiced sounds. In the receiver the excitation signal is synthesized using the sinusoidal model and passed through the inverse short-term linear prediction filter to obtain the reconstructed speech.

In this paper we present a 2.4 kb/s SEC system as well as a preliminary 1.85 kb/s system. The reduction in rate to 1.85 kb/s is accomplished by using a zero-bit quantization method for the spectral magnitudes.

## 2. SYSTEM OVERVIEW

Figure 1 shows a block diagram of the SEC system. At the transmitter, the speech spectral envelope is estimated using linear prediction techniques. The LPC parameters are transformed into LSP's and quantized by a four stage vector quantizer [4]. The quantized parameters are transformed back into LPCs and used in the short-term filter which computes the excitation signal.

In order to reproduce the excitation signal at the decoder, three parameters are extracted: the fundamental period (pitch) $P$, the harmonic spectral magnitudes $\vec{y}$, and a phase dispersion factor $D_\phi$. For speech segments which are not periodic, the system uses a fixed value for $P$, and the components of $\vec{y}$ are simply samples of the excitation spectrum taken at frequencies $kF_s/P$ where $F_s$ is the sampling frequency. A new variable-length vector quantization method called Non-Square Transform Vector Quantization (NSTVQ) [5] is used to transform $\vec{y}$ into a quantized, fixed length vector $\vec{z_q}$.

During decoding, the spectral magnitudes are reconstructed using the non-square inverse transform. The phase reconstruction is based on a predictive model which is modified using phase dispersion in order to achieve a more natural synthesized speech quality. The excitation signal is reconstructed using a sum-of-sinusoids model and passed through a synthesis filter based on the inverse short-term filter to obtain the reconstructed speech $\hat{s}_n$.

## 3. EXCITATION SYNTHESIS MODEL

The excitation synthesis model used in SEC is shown in Fig. 2. The model is based on a bank of sinusoidal oscil-

This work was supported in part by Rockwell International, Digital Communications Division, Newport Beach, CA.

0-7803-2431-5/95 $4.00 © 1995 IEEE

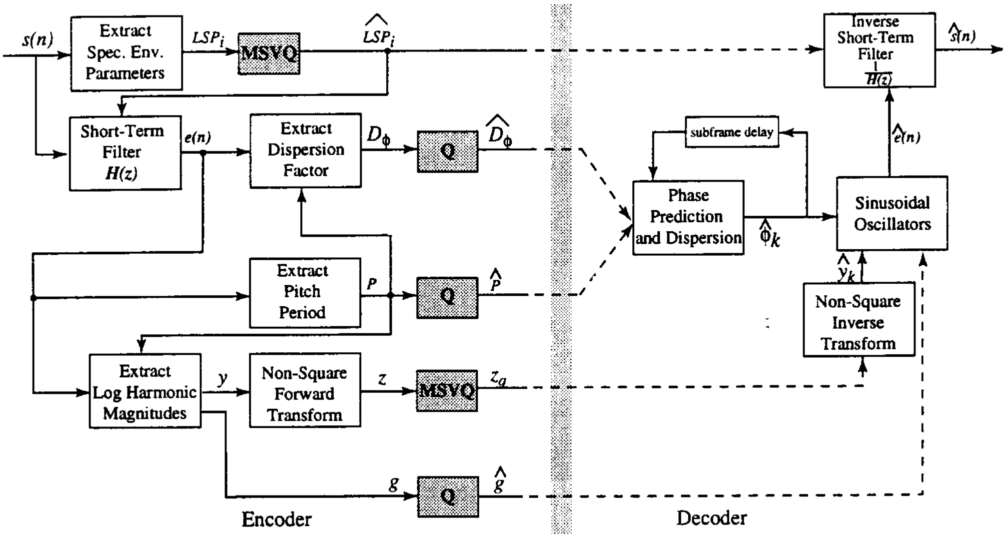

Figure 1: Block Diagram of SEC System

lators having frequencies which are integer multiples of a fundamental (pitch) frequency, $\omega_0(n)$. The index $n$ is a sample index which shows that the fundamental frequency is time varying. The output of each oscillator is scaled using a time varying gain factor $A_k(n)$ where $k$ is the harmonic number. The phases which are applied to each oscillator can come from one of two sources: predicted phases or random noise. Switching between these two sources is controlled by a phase dispersion factor.

At each pitch update instant, the pitch period $\omega_0$ is estimated from the unquantized excitation signal using an autocorrelation-based method. A tracking algorithm is used to estimate the probability that the frame is voiced and the pitch is being correctly tracked. This helps keep the pitch trajectory smooth by avoiding pitch doubling and other short term estimation errors.

Estimation of the excitation spectrum, $\vec{y}$, is performed at regularly spaced time instants. The excitation signal is windowed and the magnitude spectrum is estimated using the Discrete Fourier Transform (DFT). Spectral magnitude estimates for other times are evaluated by linearly interpolating the spectra estimated at these update instants.

In order to avoid transmission of phase information, the SEC phases are synthesized entirely at the receiver using a quadratic interpolation procedure introduced by Almeida [6].

Listening tests show that when predicted phases are used to synthesize the residual signal, the reconstructed speech has a synthetic (unnatural) quality. Part of the problem is due to the periodicity introduced by predicted phases during unvoiced and transition sounds. Voiced sounds are also observed to have a synthetic quality when predicted phases are used in the synthesis model.

Experiments indicate that this undesirable quality is due mainly to the fact that the phase difference between the fundamental harmonic and the upper harmonics of the synthetic residual signal are overly correlated. In fact, this occurs precisely from the definition of the predicted phase which allows for no phase dispersion. Measurements on speech residual signals show that the actual degree of phase correlation varies greatly with utterance and also between speakers.

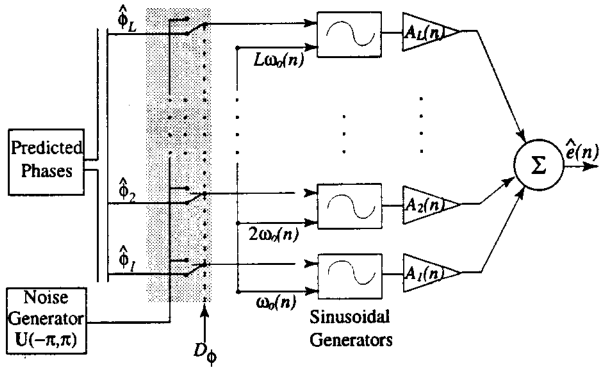

Figure 2: SEC Synthesis Model

To compensate for the excess phase correlation due to phase prediction, we introduced a phase dispersion factor, $D_\phi$, which is used to decorrelate the phases on a subframe by subframe basis. $D_\phi$ is computed using an algorithm based on the normalized autocorrelation coefficient at the pitch lag. Because lower pitched speakers tend to have lower pitch lag correlation, the phase dispersion factor is adapted using the pitch period.

## 4. HARMONIC MAGNITUDE ENCODING

One of the most challenging problems in SEC and other spectral domain coders [1] – [3] involves the encoding of the harmonic magnitudes. Because the pitch period is time-varying, the length of the magnitude vector to be quantized changes from frame to frame resulting in a variable-dimension vector quantization problem. To address this problem we developed a new technique called Non-Square Transform Vector Quantization (NSTVQ) which combines a fixed-dimension vector quantizer with a variable-sized non-square transform. Results were presented which showed that NSTVQ out-performed existing all-pole modeling techniques [5]. Recently, we compared NSTVQ with the hybrid scalar/vector quantization scheme used by IMBE [7] and found that NSTVQ in an IMBE environment can obtain equivalent spectral distortion while reducing the overall rate by 1000-1250 bps [8]. A summary of the NSTVQ technique is presented below.

### 4.1. The Non-Square Transform

Let $\vec{y}$ be a vector of length $N$ where $N$ is variable. For example, if $\vec{y}$ is a vector of harmonic magnitudes for a frame of speech, then $N$ depends on the pitch period for that frame. Given a known $N$x$M$ matrix $\mathbf{A}$ (to be specified below), we want to find a fixed length $M$-dimensional vector $\vec{z}$ which can be used to compute an estimate of $\vec{y}$ using the transformation $\vec{y}_m = \mathbf{A}\vec{z}$. For any given $\mathbf{A}$, our goal is to minimize the mean squared error distortion criterion $D_m$ with respect to $\vec{z}$ where $D_m(\vec{y}, \vec{y}_m) = \frac{1}{N}||\vec{y}_m - \vec{y}||^2$. It can be shown that the vector $\vec{z}_{opt}$ which minimizes $D_m(\vec{y}, \vec{y}_m)$ is obtained as the solution to the following set of linear equations:

$$(\mathbf{A}^T\mathbf{A})\vec{z}_{opt} = \mathbf{A}^T\vec{y} \tag{1}$$

A solution to this equation can always be found regardless of the rank of $\mathbf{A}$ using one of the linear algebra techniques for inverting ill-conditioned matrices, for example Singular Value Decomposition (SVD). If $N \geq M$ and the $M$ columns of $\mathbf{A}$ are linearly independent, the matrix $\mathbf{A}^T\mathbf{A}$ is of full rank, and therefore has an explicit inverse which gives a unique solution. Furthermore, if $N \geq M$ and the columns of $\mathbf{A}$ are orthonormal, the optimal solution vector is simply $\vec{z}_{opt} = \mathbf{A}^T\vec{y}$. For the case of $N < M$, eqn. (1) is under-determined and therefore has no unique solution. It was found experimentally that a zero-padded solution works well when combined with vector quantization. The zero-padded solution is obtained by using $N$ orthonormal vectors for the first $N$ columns of $\mathbf{A}$ and setting the last $M - N$ columns to zeros. Using orthonormal columns and zero-padding, the general solution for $\vec{z}_{opt}$ can be written as:

$$\vec{z}_{opt} = \mathbf{Ap}^T\vec{y} \tag{2}$$

$\mathbf{Ap}$ is defined as:

$$\mathbf{Ap} = \begin{cases} (\vec{a}_1\ \vec{a}_2\ \cdots\ \vec{a}_M) & \text{if } N \geq M \\ (\vec{a}_1\ \vec{a}_2\ \cdots\ \vec{a}_N \mid O) & \text{if } N < M \end{cases} \tag{3}$$

where $\vec{a}_i$ are orthonormal column vectors and $O$ is an $N$ x $(M - N)$ all zero matrix. For SEC we form the columns of $\mathbf{Ap}$ using variable length basis functions derived from the Discrete Cosine Transform (DCT). In this case, the elements $a_i[n]$ for $n = 1 \ldots N$ are given by:

$$a_i[n] = (\frac{2}{N})^{\frac{1}{2}} C_i \cos\left(\frac{(2(n-1)+1)\pi(i-1)}{2N}\right) \tag{4}$$

where $C_i = 1$ when $i \neq 1$, and $C_i = 1/\sqrt{2}$ when $i = 1$.

### 4.2. The Non-Square Transform with Vector Quantization

The non-square transformation derived in section 4.1 transforms a variable length vector $\vec{y}$ into $\vec{z}$ which can be encoded using a fixed-dimension VQ. The quantized fixed-length vector $\vec{z}_q$ is then transformed into the quantized variable length vector $\vec{y}_q$ using $\vec{y}_q = \mathbf{Ap}\vec{z}_q$ . The vector quantizer should be designed to minimize the distortion $D_t(\vec{y}, \vec{y}_q) = \frac{1}{N}||\vec{y} - \vec{y}_q||^2$. It can be shown that

$$D_t(\vec{y}; \vec{y}_q) = \frac{1}{N}||\mathbf{Ap}\vec{z} - \vec{y}||^2 + \frac{1}{N}||\mathbf{Ap}(\vec{z}_q - \vec{z})||^2 \tag{5}$$

The first term of eqn (5) is the modeling distortion due to the non-square transform and the second term is the quantization distortion due to the VQ. The fact that these distortions can be separated shows that once we have chosen an orthonormal transformation matrix $\mathbf{Ap}$ we need not consider it during training. The distortion measure for vector quantizer error minimization is given by

$$D_q(\vec{z}, \vec{z}_q) = \frac{1}{N}\sum_{i=1}^{\min(M,N)} (z[i] - z_q[i])^2 \tag{6}$$

where $z[i]$ is the $i$th element of the vector $\vec{z}$.

We trained our vector quantizers using the generalized Lloyd algorithm (GLA). A training set of size $L$ consists of fixed length vectors $\vec{z}_l$ and corresponding vector lengths $N_l$, where $l = 1 \ldots L$. Given an initial codebook of size $K$ with entries $\vec{c}_k, k = 1 \ldots K$, we encode the training set by assigning each vector, $\vec{z}_l$ to partition $S^i$ if $D_q(\vec{z}_l, \vec{c}_i)$ is minimum over all codebook entries. The centroid rule for computing the new $k$th codebook entry $\vec{c'}_k$ is given by

$$c'_k[n] = \frac{\sum_{l \in S^k} \frac{1}{N_l} z_l[n] p_l[n]}{\sum_{l \in S^k} \frac{1}{N_l} p_l[n]} \quad \text{for } n = 1 \ldots M \tag{7}$$

where $z_l[n]$ is the $n$th element of the $l$th training vector. $p_l[n]$ are the components of a vector which eliminates zero-padded elements from the distortion calculation, and are defined as:

$$p_l[n] = \begin{cases} 1 & \text{if } n \leq \min(N, M) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

## 5. PARAMETER QUANTIZATION FOR 2.4 KB/S SEC

The short-term filter coefficients are converted to Line Spectral Pair (LSP) values and quantized once each 320 sample frame using a 4 stage, 6 bit/stage multi-stage vector quantizer [4]. The pitch period is quantized every 160 samples using a 7 bit scalar quantizer. The phase dispersion factor and excitation gains are both scalar quantized using 4 bits and 5 bits respectively every 80 samples. The spectral magnitude vector is quantized every 80 samples using the NSTVQ technique discussed in section 4 applied to the log of the harmonic magnitude vector. A 6 bit vector quantizer of fixed dimension $M = 30$ is used.

Table 1 shows a summary of the bit allocations for the 2.45 kb/s SEC codec.

| PARAMETER | Bits | Updates | Rates (bps) |
|---|---|---|---|
| Envelope LSPs | 24 | 1 | 600 |
| Pitch Period | 7 | 2 | 350 |
| Phase Disp. Factor | 4 | 4 | 400 |
| Exc. Gain | 5 | 4 | 500 |
| Spectral Mags | 6 | 4 | 600 |
| Total | | | 2450 |

Table 1: Bit Allocations for the 2.45 kb/s SEC Codec

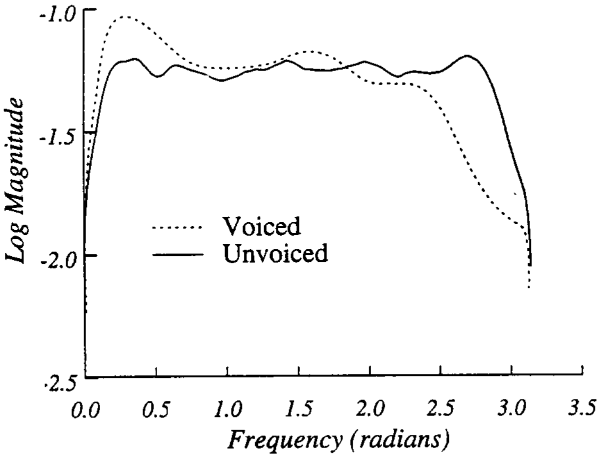

Figure 3: Log magnitude spectrum templates

## 6. ZERO-BIT MAGNITUDE QUANTIZATION FOR RATES BELOW 2 KB/S

In order to reduce the bit rate to less than 2 kb/s, we experimented with zero-bit quantization of the spectral magnitude vector using a V/UV classifier. Since the range of quantized pitch values ($20 \leq P \leq 140$) requires less than the allocated 7 bits, there are unused codes that can be utilized to send class information whenever pitch information is not required. The voiced/unvoiced decision is used to select one of two NSTVQ codebooks, each containing a single entry. The voiced and unvoiced NSTVQ codebooks are obtained from the training set by computing a single NSTVQ centroid for all voiced and unvoiced speech respectively.

When the encoder classifies a subframe as unvoiced, a pitch of zero is transmitted and replaced in the receiver by a fixed a priori pitch value. The unvoiced NSTVQ entry is then selected and transformed to an appropriate variable length magnitude vector. For voiced subframes, the calculated pitch is transmitted and the entry from the voiced NSTVQ codebook is used. To avoid excessive spectral similarity between successive frames, a random number with a small variance is added to each quantized magnitude. Fig. 3 shows an example of the voiced and unvoiced spectral magnitude shapes obtained after inverse transformation of each of the NSTVQ codebook vectors.

If all other bit assignments are kept the same as in Table 1, this technique results in a system with a rate of 1850 bps.

## 7. PERFORMANCE

The performance of the 2.4 kb/s SEC system was evaluated using an informal Mean Opinion Score (MOS) test with 22 participants listening to 12 sentences spoken by male and female speakers. The existing 2.4 kb/s LPC-10e standard and the 4.15 kb/s IMBE standard codecs were included as reference systems. The SEC system obtained an MOS score of 3.0, exceeding the LPC-10e score by 0.8 MOS points. The 4.15 kb/s IMBE system obtained an MOS score of 3.4.

Preliminary tests on the 1.85 kb/s SEC system with 0-bit magnitude quantization have been undertaken and the results indicate that the quality is also better than the LPC-10e standard. The largest degradation in the 1.85 kb/s SEC codec occurs during nasal speech indicating that a separate "nasal" spectral magnitude codebook should be used for these sounds.

Current research efforts are directed towards further reducing the bit-rate through the use of inter-subframe prediction and vector-quantization of other codec parameters. In particular, recent tests show that a significant number of bits can be saved by using vector quantization of the phase dispersion factor. Improvements to the phase model are also being studied.

## 8. REFERENCES

[1] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1223–1235, August 1988.

[2] R. McAulay, T. Parks, T. Quatieri, and M. Sabin, "Sine-wave amplitude coding at low data rates," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, (Vancouver, B.C.), 1989.

[3] Y. Shoham, "High-quality speech coding at 2.4 to 4.0 kbps based on time-frequency interpolation," in *Proc. ICASSP*, (Minneapolis), 1993.

[4] B. Bhattacharya, W. LeBlanc, S. Mahmoud, and V. Cuperman, "Tree searched multi-stage vector quantization of LPC parameters for 4 kb/s speech coding," in *Proc. ICASSP*, pp. 105–108, 1992.

[5] P. Lupini and V. Cuperman, "Vector quantization of harmonic magnitudes for low-rate speech coders," in *Proc. IEEE Globecomm*, (San Francisco), 1994.

[6] L. Almeida and F. Silva, "Variable-frequency synthesis: An improved harmonic coding scheme," in *Proc. ICASSP*, (San Diego), 1984.

[7] Digital Voice Systems, *Inmarsat-M Voice Codec, Version 2.* Inmarsat, February 1991.

[8] P. Lupini and V. Cuperman, "Non-square transform vector quantization," *Submitted to Signal Processing Letters*, 1994.