Subjective Performance of Spectral Excitation Coding of Speech at 2.4 kb/s

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

ABSTRACT

This paper presents a low rate speech codec (2.4 kb/s) based on a sinusoidal model applied to the excitation signal. A frame classifier in combination with 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 for all frame classes resulting in more "natural" reconstructed speech. Informal MOS testing indicates that the 2.4 kb/s SEC system achieves MOS scores close to the existing 4 kb/s standards (differences up to 0.2 on the MOS scale) and significantly better than the existing 2.4 kb/s LPC-10 standard (difference of 1.5 on the MOS scale).

1 Introduction

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 speech quality than the existing CELP based codecs [1, 2, 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. The model is used for voiced as well as unvoiced and transitional sounds. The spectral excitation magnitudes are quantized and transmitted, while the corresponding phases are synthesized at the receiver based on a small number of transmitted parameters. The excitation synthesized by the sinusoidal model is applied to a synthesis filter based on short-term linear prediction of the speech signal. The parameters of the synthesis filter are transmitted as quantized Line Spectral Pairs (LSP) using a multi-stage vector quantizer.

2 System Description

Figure 1 shows a block diagram of the 2.4 kb/s SEC system. Once each 30 ms frame, the speech spectral envelope is estimated using tenth order LPC analysis. The coefficients are converted to line spectral pairs and quantized once per frame using the tree-searched multi-stage vector quantization scheme presented in [4]. The 10 LSP coefficients are encoded using 24 bits each frame. We have found that using a 4-stage, 6 bits/stage, MSVQ with 8 candidates results in a robust VQ with low spectral distortion. The quantized coefficients are then transformed back into LPCs and used in the short-term filter which computes the excitation signal. The filter coefficients are updated using LSP interpolation every 2 ms

The excitation signal is analyzed over 5 ms subframes (giving 6 subframes per 30 ms frame) in order to compute estimates for three parameters once per subframe: the fundamental period (pitch) P, the phase dispersion factor D_ϕ, and the harmonic spectral magnitudes $\vec{y}$. Details of the estimation and the definition of the dispersion factor will be presented below.

The excitation signal is reconstructed at the decoder using the sinusoidal model

$$\hat{e}(n) = \sum_{k=0}^{L} \hat{y_k}(n) cos \hat{\phi_k}(n) \tag{1}$$

where $\hat{y_k}$ are the reconstructed harmonic magnitudes and $\hat{\phi_k}$ are the reconstructed phases. The reconstructed excitation is applied to the inverse short-term filter to obtain the reconstructed speech signal $\hat{s}(n)$.

3 Parameter Estimation and Quantization

Every subframe, the pitch period P is estimated from the unquantized excitation signal, $e(n)$, using an autocorrelation-based method. The optimal pitch pe-

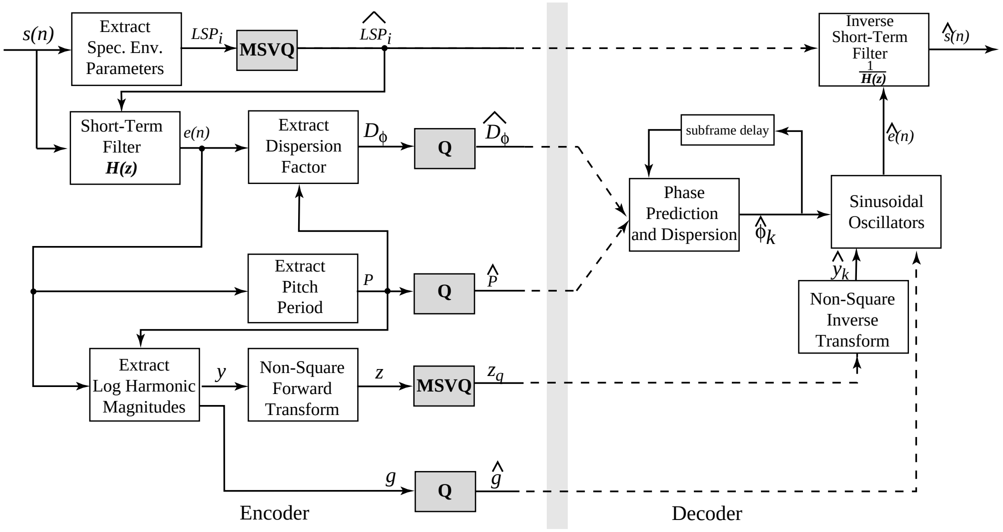


Figure 1: Block Diagram of SEC System

riod, P is given by

$$P = \max_{p_l \le p \le p_h} [\rho(p)] \tag{2}$$

where

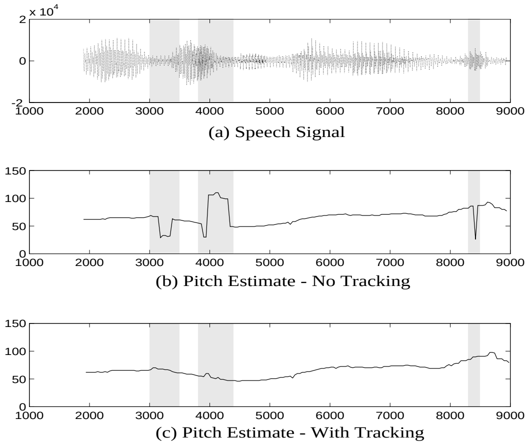
$$\rho(p) = \frac{\sum_{n=-L_p/2}^{L_p/2-1} e(n)e(n-p)}{\sqrt{\sum_{n=-L_p/2}^{L_p/2-1} e^2(n-p)}} \tag{3}$$

and p_l and p_h are the minimum and maximum possible pitch periods respectively. For 8 kHz sampled speech, $p_l = 20$ and $p_h = 147$ are used. Although the pitch must be computed every 5 ms because it is required for estimation of the phase dispersion factor (see below), it is quantized only once every 15 ms using a 7-bit scalar quantizer. Values for unencoded subframes are obtained by linearly interpolating between quantized pitch values.

Sinusoidal coders are particularly sensitive to errors in pitch estimation and produce large artifacts when the pitch contour is discontinuous during segments with steady voicing. To improve subjective performance of SEC, a pitch tracking algorithm was developed which attempts to identify voiced regions during which the pitch is changing slowly. When the algorithm determines that the pitch is being tracked during steady voicing, any large deviations in the estimated pitch period are assumed to be pitch errors, and the open loop pitch estimate is modified to be within close range of the previous pitch values.

Figure 2 shows an example of how the pitch tracker smooths the pitch contour during voiced speech. In fig.2(a), a speech segment containing only voiced speech is plotted. The pitch estimates without and with the pitch tracker are plotted in fig.2(b) and (c) respectively. The areas where the pitch tracker has corrected bad pitch estimates are shown in grey.

The SEC phases, $\hat{\phi_k}$, are synthesized at the receiver



(a) Speech Signal

(b) Pitch Estimate - No Tracking

(c) Pitch Estimate - With Tracking

Figure 2: Example of pitch contour obtained from voiced speech segment with and without the pitch tracking algorithm

using

$$\hat{\phi_k} = \phi'_k + \Delta\phi_k \tag{4}$$

where ϕ'_k are the predicted phases, $\Delta\phi_k$ are the phase residuals, and k is the harmonic number. In order to avoid transmission of phase information, the predicted phases, ϕ'_k, are synthesized at the receiver using a quadratic interpolation procedure introduced by Almeida [5]. Listening tests show that if interpolation is used for the predicted phases, setting the phase residuals, $\Delta\phi_k$, to zero results in encoded speech which sounds buzzy during unvoiced segments and sometimes sounds robotic or unnatural during voiced segments. If, on the other hand, the receiver randomly assigns values to the phase residuals using a uniform distribution between $-\pi$ and π, the unvoiced speech sounds natural while the voiced speech sounds whispered or breathy.

Based on these two extremes, a model was developed in which direct quantization of the phase residuals is replaced by

$$\Delta\phi_k = \begin{cases} 0 & 1 \le k < h_c \\ \mathbf{U}[-\beta\pi, \beta\pi] & h_c < k \le K(\omega_0) \end{cases} \tag{5}$$

where h_c is the cutoff harmonic (defined below), $K(\omega_0)$ is the number of harmonics for the current subframe, $\mathbf{U}[-a, a]$ is a uniform random variable defined over the interval $-a \ldots a$, and β is a parameter which modifies the range of the randomized phase residual.

Experimentally, it was found that a good approach for obtaining the cutoff frequency uses the fundamental frequency, ω_0, and is given by

$$h_c = \begin{cases} 0 & \text{if } D_\phi \le D_l \\ K(\omega_0) & \text{if } D_\phi \ge D_h \\ \left[\frac{(D_\phi - D_l)}{(D_h - D_l)}\right] K(\omega_0) & \text{otherwise} \end{cases} \tag{6}$$

where D_ϕ is defined as the phase dispersion factor, and D_l and D_h are heuristically determined parameters. In SEC, D_ϕ is computed based on the normalized autocorrelation at the pitch lag given by eqn. (3); during strongly voiced frames, D_ϕ is close to one, and during strongly unvoiced frames, D_ϕ is close to zero. Note that when quantization of the phase residuals is replaced with the model defined above, D_ϕ must be quantized and transmitted to the receiver.

During subjective testing with phase dispersion as defined above, the reconstructed speech for male speakers was often described as being too breathy indicating that the phase vector for male speakers contained too large a random component. Analysis showed that for male speakers, $\rho(p)$ tends to be lower, probably due to the longer pitch periods. To improve subjective quality, the upper distortion limit D_h was made to be dependent on the pitch period according to

$$D_h(p) = \begin{cases} 0.85 & 20 \le p < 40 \\ -0.0125p + 1.35 & 40 \le p < 80 \\ 0.35 & 80 \le p \le 147 \end{cases} \tag{7}$$

where all constants were determined through experimentation. By substituting eqn. (7) into eqn. (6), it can be seen that the cutoff frequency is higher for lower pitch speakers resulting in fewer harmonics being randomized. Figure 3 illustrates the effect of the adaptive dispersion equation by plotting the fraction of randomized harmonics versus the dispersion factor for two different pitch values of 20 and 80. It can be seen, for example, that when the phase dispersion factor is 0.5, 70% of the phases are randomized for a high-pitched speaker with $p = 20$, but only 40% of the phases are randomized for a low-pitched speaker with $p = 80$.

Estimation of the excitation spectrum, $\vec{y}$, is performed every 3 subframes (15 ms). The excitation signal is windowed using a pitch-sized rectangular window and the magnitude spectrum is estimated using the Discrete Fourier Transform (DFT). Spectral estimates for intermediate subframes are evaluated by linearly interpolating between quantized log spectral magnitudes. For speech segments which are not periodic, the system uses a fixed value of $P = 100$, 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) [6] is used to transform $\vec{y}$ into a quantized, fixed length vector $\vec{z_q}$. Before quantization, the mean of the variable-length log spectra is removed and quantized separately using 6 bits every 15 ms. The remaining normalized vectors are quantized every 15 ms using NSTVQ with 8 bits. The NSTVQ configuration in SEC uses the DCT-II transform and a fixed-dimension of $M = 30$.

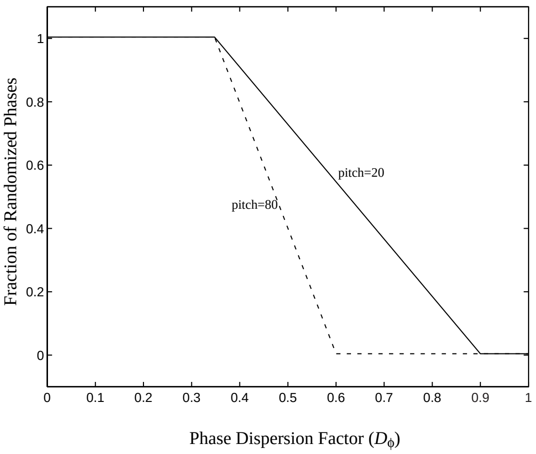


Figure 3: Effect of Pitch Period on Number of Randomized Harmonics using Adaptive Phase Dispersion

PARAMETER	Bits	Updates	Rates (bps)
Envelope LSPs	24	1	800
Pitch Period	7	2	467
Phase Disp. Factor	1 (VQ)	6	200
Exc. Gain	6	2	400
Spectral Mags	8	2	533
Total			2400

Table 1: Bit Allocations for the 2.4 kb/s SEC Codec using a frame length of 240 samples with 40 samples per subframe (SEC-v2)

Table 1 shows a summary of the bit allocations for the 2.4 kb/s SEC codec.

4 Performance Evaluation

The performance of the 2.4 kb/s SEC system was evaluated using an informal Mean Opinion Score (MOS) test in which several speech codecs were used to encode 10 sentences, 5 from male speakers and 5 from female speakers. Fourteen participants took part in the infor-

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 2: Bit Allocations for SEC-v1 using a frame length of 320 samples with 80 samples per subframe.

		Mean Opinion Score		
System	Rate (bps)	All	Male	Female
IMBE	4150	3.4	3.3	3.5
FS 1016	4600	3.3	3.2	3.4
SEC-v2	2400	3.2	3.1	3.3
SEC-v1	2450	3.0	3.0	3.0
LPC-10e	2400	1.8	1.7	1.8

Table 3: Mean opinion scores (MOS) results

mal MOS test and were asked to rate the quality of each speech sample using a scale from 1 to 5 representing a subjective quality of bad, poor, fair, good, and excellent. This provides a total of 140 ratings for each system. Included in the test was the existing 2.4 kb/s LPC-10e standard, the 4.15 kb/s IMBE standard, and the FS 1016 4.6 kb/s CELP standard.

Two SEC systems were included in the test. The first, SEC-v1, is a previous baseline system operating at 2.45 kb/s [7]. The bit allocation for SEC-v1 is given in table 2. The second system, SEC-v2, is an improved version which was summarized in table 1. As can be seen from the two tables, the improved version uses a VQ for quantization of the phase dispersion factor. The use of a VQ makes it possible to encode the dispersion factor every 40 samples rather than every 80 samples, while using only half the rate. Further improvements were made by reducing the frame length from 320 samples to 240 samples. As a result, the pitch period can be encoded at a lower rate, leaving more bits for spectral magnitude encoding.

The results of the test are shown in table 3. The 2.4 kb/s SEC-v2 system scored within 0.1 MOS points of the FS 1016 CELP system operating at 4.6 kb/s, and within 0.2 MOS points of the 4.15 kb/s IMBE standard. The MOS differences were consistent for both male and female speakers. The existing LPC-10e standard performed poorly on these sentences, obtaining an MOS of 1.8. The results also indicate that the quality of the SEC algorithm was improved from the older baseline (SEC-v1) through the use of vector quantization of the phase dispersion factor, a shorter frame length, and a higher encoding rate for the spectral magnitudes.

In an attempt to determine how SEC might be improved, several people were asked to judge the quality of the SEC-v2 system in informal interviews. The most common comment made was that the unvoiced sounds were often unnatural and sometimes annoying. It is possible that this problem may be alleviated through the development of a more sophisticated phase dispersion algorithm. A successful approach to unvoiced sound synthesis was recently reported in [8] in which separate spectral quantization codebooks were used for unvoiced sounds in combination with rapid updates of the RMS gain.

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

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

[7] V. Cuperman, P. Lupini, and B. Bhattacharya, “Spectral excitation coding of speech at 2.4 kb/s,” in *Proc. ICASSP*, pp. 496–499, 1995.

[8] M. Nishiguchi and J. Matsumoto, “Harmonic and noise coding of lpc residuals with classified vector quantization,” in *Proc. ICASSP*, pp. 484–487, 1995.