NATURAL QUALITY VARIABLE-RATE SPECTRAL SPEECH CODING BELOW 3.0 KBPS

Engin Erzin[1†], *Arun Kumar*[2], *and Allen Gersho*[2]

[1]Lucent Technologies,
600 Mountain Avenue,
Murray Hill, NJ, 07974 USA

[2]Dept. of Electrical & Computer Eng.
University of California,
Santa Barbara, CA, 93106 USA

ABSTRACT

We propose new techniques for natural quality variable rate spectral speech coding at an average rate of 2.2 kbps for dialog speech and 2.8 kbps for monolog speech. The coder models the Fourier spectrum of each frame and it builds on recent enhancements to the classical multiband excitation (MBE) approach. New techniques for robust pitch estimation and tracking, for efficient quantization of voiced and unvoiced spectra and encoding of partial phase information are the key features that result in improved quality over earlier spectral vocoders. Subjective performance results are reported which show that the coder is very close in quality to the ITU-T G.723.1 algorithm at 5.3 kbps.

1. INTRODUCTION

In this paper, we propose several new techniques for efficient low bit rate spectral speech coding. We improve upon recent enhancements of the classical multiband excitation (MBE) paradigm of speech coding [1], in terms of naturalness of decoded speech quality, smoothness of pitch period tracking, rate-distortion performance of spectral shape vector (SSV) quantization, computational complexity and storage requirements. The new techniques are used to design a variable rate spectral speech coder which operates at an average bit rate of approximately 2.8 kbps for monolog speech and 2.2 kbps for dialog (conversational) speech. It requires a maximum bit rate of 3.9 kbps for encoding voiced speech frames and a minimum of 1.25 kbps for encoding silence frames. Listening tests indicate that this coder is nearly equivalent to the ITU-T G.723.1 coder operating at the much higher rate of 5.3 kbps.

Many speech coders in the 2.0-4.0 kbps rate range have evolved from the MBE coding paradigm in recent years. These include coders due to Nishiguchi et al [2], Hassanien et al. [3], Yeldener et al. [4], Das and Gersho [5], [6], and Wang et al. [7]. An improved multiband excitation (IMBE) voice coder became the satellite communication standard in 1991 as the Inmarsat-M voice codec with a source coding rate of 4.15 kbps [8]. Our proposed spectral coder draws several features from the enhanced multiband excitation (EMBE) voice coder of Das and Gersho [9] which has improved upon the Inmarsat-M standard to provide comparable or better speech quality in the range of 1.8-2.9 kbps. We adopt a multimodal analysis scheme as in the EMBE coder. New features of our coder include (1) a robust pitch estimation and tracking algorithm, (2) use of a proposed reduced dimension VQ (RDVQ) technique for spectral shape VQ for voiced frames and a cepstral representation scheme for unvoiced and silence frames, and (3) encoding of a partial phase information term. These features of the coder and its performance comparison results are discussed in the following sections.

* This work was supported in part by the University of California MICRO program, Stratacom, DSP Group, Intel, Moseley Associates, National Semiconductor, Qualcomm, Rockwell International, and Texas Instruments.

† This work was performed when the author was at the University of California, Santa Barbara.

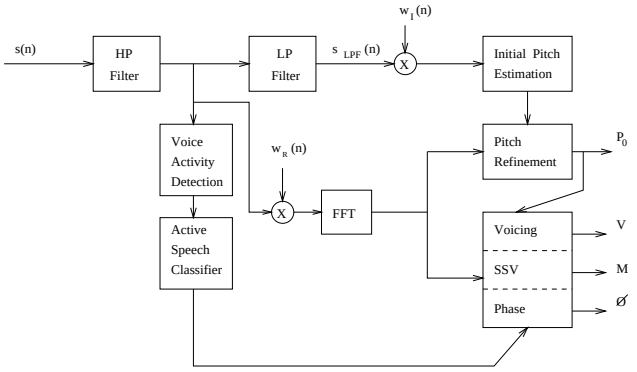

Figure 1: The overall analysis structure of the coder. Pitch, P_0, voicing vector, $\mathbf{V}$, spectral shape vector, $\mathbf{M}$, and phase vector, $\mathbf{\Phi}$, are the parameters that are encoded.

2. PREPROCESSING AND CLASSIFICATION

Figure 1 shows the analysis algorithm of the proposed variable rate spectral coder. The model parameters which are encoded once every frame-length of 20 ms are the pitch period, the voiced character across spectral bands, the spectral or cepstral magnitudes and phase at *lower* harmonics of the estimated pitch frequency in voiced bands for certain frames.

The speech signal is first high pass filtered to remove very low frequency artifacts which are not perceptually relevant and may impede the performance of the pitch estimation algorithm and the spectral shape vector quantizer.

The voice activity detector (VAD) makes a binary decision, silence or speech, based on both the time and frequency energy distributions of the current speech frame and is based on the VAD of Srinivasan and Gersho [10]. If the VAD detects active speech for the current frame, the active speech classifier block classifies the frame further as either voiced or unvoiced. This analysis preclassifies the current speech frame into silence (including a background noise condition), voiced, or unvoiced and serves two important purposes. First, it prevents misclassification of unvoiced and voiced segments for the analysis of the spectral coder, and second, it introduces a convenient multimode structure over which to build a variable bit rate coder.

3. PITCH PERIOD ESTIMATION AND ITS PERFORMANCE EVALUATIONS

Accurate pitch period estimation is a critical task in spectral or sinusoidal coders. Most pitch related errors are caused by incorrectly estimating a multiple or submultiple of the true pitch period, instead of the true one. We propose a new pitch tracking technique that introduces strong smoothness constraints around an average pitch value which is updated adaptively over voiced frames, and uses only one frame look ahead.

The pitch estimation, tracking and refinement is performed with an autocorrelation based error function $E(p)$, derived by Griffin and Lim ([1], pp. 1227, equation (11)). The pitch, P_k, for the current frame is determined in the range of 21 to 114 samples with half-sample accuracy, and uses finalized pitch values P_{k-1} and P_{k-2} of the previous two frames and the initial pitch estimate, P^i_{k+1}, of the future frame. The pitch estimation and tracking steps are as follows:

1. Determination of the future frame's initial pitch estimate, P^i_{k+1}, using the corresponding pitch error function, $E_{k+1}(p)$,
2. Estimation of the current frame's initial pitch value, P'_k, using smoothness constraints (note that $P'_k \neq P^i_k$),
3. Corrections in P'_k by checking for submultiple pitch candidates, to obtain final pitch estimate P_k for current frame.

The first step determines a pitch candidate for the future frame in each of three distinct sets: $\{21, 21.5, \ldots, 39.5\}$, $\{40, 40.5, \ldots, 79.5\}$, $\{80, 80.5, \ldots, 114\}$ to avoid pitch multiple estimates. Then, the initial pitch value P^i_{k+1} of the future frame is chosen from the three candidates by picking the lowest possible pitch value for which $E_{k+1}(p)$ is in an *acceptable* range of the global minimum error value with a bias for choosing lower pitch values.

Next, we obtain the initial pitch value P'_k of the current frame among the initial pitch values P^i_k and P^i_{k+1}, and a backward pitch value P_b that minimizes the current frame's error function $E_k(p)$ in the neighborhood of the previous pitch value P_{k-1}.

The last step is to check for submultiples of the initial pitch value P'_k to avoid pitch doubling and tripling type of errors. The pitch error function values at the submultiples are compared to $E_k(P'_k)$ using a novel frame adaptive threshold function to decide whether the nth submultiple of P'_k is the true pitch period. Finally, the lowest submultiple which satisfies the thresholding is selected; otherwise, if none of the submultiples satisfy the thresholding, then P'_k, is chosen as the final pitch value P_k.

We performed a statistical test to compare the accuracy of the pitch estimates made by the Inmarsat-M codec and the proposed new pitch estimation and tracking algorithm. In the performance evaluations, the pitch contours were hand labeled for 5500 voiced frames (of 20 ms each). We found that the percent of pitch estimation errors (e.g., selecting a multiple or submultiple of the true value) using the proposed pitch tracker instead of the IMBE pitch tracker was reduced from 0.9% to 0.5%. Also the lookahead delay is one frame instead of two. Thus, the new pitch tracking technique is both more robust and computationally more efficient than the Inmarsat-M pitch tracker.

The pitch estimation and tracking algorithm is disabled for frames declared unvoiced by the active speech classifier. Otherwise, the final pitch estimate is used to compute the number of harmonics in the spectrum and boundaries between adjacent harmonic bands as in IMBE. The V/UV decision for harmonic bands in voiced frames is made as in the EMBE codec [5].

4. TECHNIQUES FOR SPECTRAL SHAPE QUANTIZATION

The efficient quantization of spectral shape vector (SSV) which consists of spectral magnitudes for each harmonic interval, is one of the more difficult problems in spectral coding. This is mainly because the SSV is a variable dimension vector whose dimension can vary from 9 to 51. At low bit rates, the overall quality of coded speech significantly depends on how efficiently the variable dimension spectral vectors are quantized. The practical solutions to SSV quantization can be broadly classified into two categories: dimension conversion VQ (DCVQ) and variable dimension VQ (VDVQ). In DCVQ, the variable dimension SSV is converted to a fixed dimension vector prior to applying VQ. Dimension conversion methods include spectral interpolation methods [2], and non-square transform VQ (NSTVQ) [11]. In VDVQ, the variable dimension SSV is encoded using a universal codebook without introducing any modeling distortion [5], [12].

We propose a reduced dimension vector quantization (RDVQ) technique which provides efficient rate-distortion performance and also overcomes the high-dimensionality and high-storage requirements of other methods for possible real-time realization of the coder. Also we propose a quantization structure for the unvoiced spectra using the cepstral sequence of the short-time spectra. The key ideas of these techniques are summarized here.

In RDVQ, the current dimension L_k of the SSV is reduced to a fixed lower dimension L_R by removing some harmonic magnitudes if $L_k > L_R$. Otherwise, we map the smaller dimension vector to an L_R dimension vector, as is done in VDVQ [5], using a universal codebook with dimension L_R. The dimension of the universal codebook as used here is much smaller than in the cited prior work on VDVQ.

Thus, for $L_k > L_R$ we can map any variable-dimension

spectral shape vector to the reduced fixed-dimension vector (of dimension L_R) that enables us to perform a regular L_R dimensional vector quantization. In the decoding process, the missing magnitudes of the SSV are linearly interpolated from neighboring quantized magnitudes in the L_R dimension vector. Reduced dimension VQ approach, thus introduces a modeling distortion only for spectral shape vectors with dimension $L_k > L_R$. On the other hand, spectral shape vectors having dimension $L_k \leq L_R$ can be represented without any modeling distortion.

For frames declared unvoiced or silence by the active speech classifier, the speech quality of the EMBE coder improves upon IMBE by using a fixed number of harmonic bands. However, this representation requires a high dimensional SSV representation for a reasonable modeling distortion of the short-time spectra. Alternatively, the short-time spectra can be parametrically modeled. We introduce a cepstral representation, extracted from Fourier transform domain, to model the short-time spectra. This method exploits the fact that the low-indexed coefficients of the cepstrum sequence carry most of the information for voiced speech. The proposed cepstral representation gives an objective and subjective performance that is similar to the r.m.s. energy representation scheme, but with a *much lower* dimensional representation.

5. PERFORMANCE COMPARISONS OF SPECTRAL SHAPE QUANTIZATION

In the encoding process, the RDVQ scheme is used for voiced and mixed-voiced speech frames, while the cepstral representation is used for the unvoiced and silence frames. In our coder implementation, we use $L_R = 26$. To reduce the computational complexity, we use a multi-survivor trellis-coded and gain-removed multistage VQ structure [13] with 4 survivors. In the multistage VQ structure, we assign 6 bits to each of the first two stages, and 5 bits to each of the other stages. The gain term is quantized using separate 9 bit codebooks, one for voiced frames and another for frames classified as unvoiced (including silence). The voiced and mixed-voiced classes are quantized using 11 stages MSVQ structure. The performance evaluations of RDVQ are done with respect to VDVQ based schemes, where the spectral distortions between two short-time spectra from synthetic speech waveforms with unquantized and quantized harmonic magnitudes are found to be 4.14 dB, 5.12 dB, and 4.10 dB for VDVQ, transform VDVQ, and RDVQ schemes, respectively. As seen, the reduced dimension vector quantization technique performs slightly better than VDVQ based techniques. Also, the storage requirements are reduced to one half compared to transform VDVQ technique which is a reduced-storage (and reduced complexity) version of VDVQ [6].

The evaluation of cepstral domain representation is performed considering the modeling and quantization distortions, and is presented in Table 1. The first row shows the traditional method which divides the spectra into fixed number of bands and represents each band with its r.m.s. value. The second row shows the performance of 16 dimensional cepstral representation. The last row shows the performance of an LP spectral model. The evaluations of the modeling schemes are presented in terms of spectral distortion between the original and model short-time spectra. The modeling performance of the 16-th order cepstral representation is found to be similar to that of a 26-dimensional r.m.s. representation in terms of spectral distortion. The spectral distortion between original and quantized spectra is presented in Table 1. The performance of 4 stages MSVQ for unvoiced spectra is shown in the second-last column, and that of 2 stages MSVQ for silence spectra is shown in the last column. In the final form of the coder we chose 4 and 2 stages of MSVQ structure for unvoiced and silence classes respectively.

Methods	Model SD (dB)		SD after Quantization (dB)	
	UV	Sil.	UV	Sil.
			9+22 bits	9+12 bits
RMS(1+26)	4.86	5.04	5.60	5.70
Cep.(1+15)	4.84	5.01	5.24	5.51
LP(1+15)	5.54	5.68	-	-

Table 1: Quantization performance of unvoiced and silence spectra.

6. SYNTHESIS WITH PARTIAL PHASE INFORMATION

In spectral domain coding, the synthetic voiced speech is represented as a sum of sinusoidal oscillators:

$$\hat{s}_v(t) = \sum_m \alpha_m(t) \cos(\Theta_m(t)) \qquad (1)$$

where m spans all harmonic bands, $\alpha_m(t)$ are linearly interpolated harmonic magnitudes between frames, and $\Theta_m(t)$ is the synthetic phase track with initial phase ϕ_m:

$$\Theta_m(t) = \int_0^t \omega_m(\tau) d\tau + \phi_m. \qquad (2)$$

If we use the *same* initial phase, $\phi_m = \phi_0$, for all m, as in IMBE type coders, then the lack of relative phase information among harmonic frequencies causes the synthesized time domain waveform to look artificially *impulsive* at pitch period intervals.

We introduce partial phase information for the nine lowest harmonics, which for each UV/V transition is sent three frames after the last unvoiced frame. This relative phase information is tracked through the subsequent voiced frames. Since we have one voicing decision for three consecutive harmonic intervals, we need to send three phase terms three frames after each UV/V transition. The phase terms $\phi_{3l-2}, \phi_{3l-1}, \phi_{3l}$ corresponding to the l-th voicing band are grouped as a vector Φ_l. We represent Φ_l, $l = 1, 2, 3$ with a 7-bit VQ codebook. It is observed that the bit rate overhead is 50-100 bps for this amount of partial phase information. The synthetic waveform reconstructed with the partial phase information is perceptually more natural than the waveform reconstructed without the phase information.

Parameters	Modes (bits)			
	S	UV	V	UV/V
Voicing	4	4	4	-
Pitch	-	-	8	-
SSV-gain	9	9	9	-
SSV-shape	12	22	57	-
Phase	-	-	-	7-21
Total	25	35	78	21
Rate	1.25	1.75	3.90	-

Table 2: Bit allocation.

Also, the artificially impulsive character of the synthesized time domain waveform is significantly reduced.

7. BIT ALLOCATION AND SUBJECTIVE PERFORMANCE RESULTS

The bit allocation for voiced, unvoiced, and silence frames (of 20 ms each) are 78, 35, and 25 bits/frame respectively. Table 2 gives the break-up for the different modes. The SSV gain term represents the mean value of the SSV for voiced frames and the first cepstral parameter for unvoiced and silence frames. The overall bit rate of the coder can be derived according to the target application. For telephone communication application, the overall bit rate can be estimated to be around 2.2 kbps assuming 50% activity with 20% unvoiced and 30% voiced frames. However, for storage applications, such as answering machines, the overall bit rate can be estimated as 2.8 kbps assuming 10% silence, 40% unvoiced, and 50% voiced frames.

Natural quality synthesis is one of the important targets of our coder. To evaluate the subjective quality of the proposed coder, we performed ACR and CCR tests with 20 subjects. The procedure was very similar to the subjective qualification test plan of the ITU-T 4.0 kbps speech coding study group. The MOS score for the proposed coder is found to be 3.4 for ACR test and 3.6 for the ITU-T G.723.1 coder at 5.3 kbps. However, the standard deviation was relatively high – approximately 0.9, indicating that the MOS difference 0.2 between two coders is not significant to distinguish the proposed coder from the G.723.1 coder. Comparative category rating (CCR) MOS scores were also obtained for office type babble noise at 30 dB SNR. The uncoded and coded noisy speech pairs are compared and evaluated in a -3 to 3 numerical scale. The final comparative MOS scores are obtained as the absolute degradation of coded speech from uncoded speech, and they are found to be 0.8 for the proposed coder and 0.6 for ITU-T G.723.1 coder at 5.3 kbps.

8. CONCLUSIONS

Several techniques have been found which contribute significantly to achieving high performance natural-sounding speech with a variable-rate spectral modeling coder with average rate of 2.8 kbps. The test results as well as informal listening confirm that the quality is very similar to that of the "nearly" toll-quality ITU-T G.723.1 coder operating at the much higher (fixed) rate of 5.3 kbps.

9. REFERENCES

[1] D. Griffin and J. Lim, "Multi-band excitation vocoder," *IEEE Trans. ASSP*, vol. 36, no. 8, pp. 1223–1235, 1988.

[2] M. Nishiguchi, J. Matsumoto, R. Wakatsuki, and S. Ono, "Vector quantized MBE with simplified v/uv decision at 3.0 kbps," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process.*, pp. 151–154, 1993.

[3] H. Hassanein, A. BrindAmour, and K. Bryden, "A hybrid multiband excitation coder for low bit rates," in *Proc. Int. Conf. on Wireless Communications*, pp. 184–187, 1992.

[4] S. Yeldener, A. Kondoz, and B. Evans, "High quality multi-band LPC coding of speech at 2.4 kb/s," *IEE Electronics Letters*, vol. 27, no. 14, pp. 1287–1289, 1991.

[5] A. Das and A. Gersho, "Variable dimensional spectral coding of speech at 2.4 kb/s and below with phonetic classification," in *Proc. IEEE Int. Conf. Acoust. Speech., Sig. Process.*, vol. 1, pp. 492–495, 1995.

[6] A. Das and A. Gersho, "Multimode spectral coding of speech for satellite communications," in *Proc. Eusipco 96*, (Trieste, Italy), Sept. 1996.

[7] K. T. T. Wang and C. Feng, "A high quality mbe-lpc-fe speech coder at 2.4 kbps and 1.2 kbps," *Proc. IEEE Intl. Conf. Accoust. Sig. Speech Process.*, pp. 208–211, 1996.

[8] Digital Voice Systems, "Inmarsat-M voice codec specifications, Version 2," tech. rep., 1991.

[9] A. Das, A.V. Rao, and A. Gersho, "Enhanced multiband excitation coding of speech at 2.4 kbits/s with discrete all-pole spectral modeling," in *Proc. IEEE Globecom Conference*, pp. 863–866, 1994.

[10] K. Srinivasan and A. Gersho, "Voice activity detection for digital cellular networks," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, pp. 85–86, 1993.

[11] P. Lupini and V. Cuperman, "Vector quantization of harmonic magnitudes for low rate speech coders," in *Proc. IEEE Globecom Conf.*, vol. 2, pp. 858–862, 1994.

[12] A. Das, A.V. Rao, and A. Gersho, "Variable dimension vector quantization," *IEEE Signal Proc. Letters*, vol. 3, no. 7, pp. 200-202, July 1996.

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