VARIABLE DIMENSION SPECTRAL CODING OF SPEECH AT 2400 BPS AND BELOW WITH PHONETIC CLASSIFICATION

Amitava Das and Allen Gersho

Dept. of Electrical & Computer Engineering
University of California
Santa Barbara, CA 93106 U.S.A.
das@kane.ece.ucsb.edu; gersho@ece.ucsb.edu

ABSTRACT

The low bit rate *enhanced multiband excitation* or EMBE speech coder adds several important new features including phonetic classification and a novel spectral quantization technique called *variable dimension vector quantization* (VDVQ) to the basic multiband excitation vocoder. Phonetic classification allows the adaptation of spectral modeling and quantization to the local acoustic-phonetic character of the speech signal, enhancing quality and robustness. The VDVQ scheme quantizes the log-spectrum with relatively few bits while preserving perceptually important features. Both the fixed rate (2.4 kb/s) and the variable rate (1.44 kb/s average) implementations of EMBE deliver speech quality comparable to the 4.8 kb/s Federal Standard 1016 CELP coder and the 4.15 kb/s Inmarsat-M standard IMBE coder.

1. INTRODUCTION

There is a surge of research and commercial interest aimed at the development of high quality speech coders at bit rates of 2400 b/s and below. The application area is quite large, ranging from wireless telephony, satellite communicationand and storage to multimedia and other PC applications. The U.S. Government is also seeking a new 2400 b/s standard speech coder [11] for secure voice communications.

An effective technique for efficient low rate speech coding is to classify speech segments into phonetic categories and perform class-based coding. By adapting to the particular needs of each class, better performance can be achieved than with a fixed scheme which applies the same coding method to different speech categories, including background noise. Two distinct benefits of phonetic class-based processing are: 1) individual phonetic classes can enjoy "customized" enhanced processing including class-specific bit allocations, leading to better quality, and 2) a different coding resolution (number of bits) can be given to different speech segments, delivering the same high quality at a lower average rate.

This work was supported by Fujitsu Laboratories Ltd., the National Science Foundation, the UC Micro program, Rockwell International Corporation, Echo Speech Corporation, Signal Technology Inc., Lockheed Missile and Space Company, Qualcomm, Inc., and Hughes Aircraft Company.

We propose a parametric speech coding algorithm called *enhanced multiband excitation* or EMBE, which delivers coded speech with reasonably high quality at a low bit rate by exploiting the phonetic character of the input signal. EMBE integrates phonetic classification into the MBE spectral vocoder [1, 2] enhancing both modeling and quantization. A novel technique called *variable dimension vector quantization* (VDVQ) efficiently performs spectral quantization by preserving perceptually important features with relatively few bits. Both a fixed-rate (2.4 kb/s) and a variable-rate (1.44 kb/s average rate) implementation of the EMBE algorithm deliver speech quality comparable to the 4.8 kb/s Federal Standard 1016 (FS 1016) CELP coder [10] and the 4.15 kb/s Inmarsat-M standard IMBE coder [2].

As seen in Figure 1, in the EMBE coder, each frame is first phonetically classified. Based on the class information, a spectral model is chosen and the spectral parameters are quantized by class-based quantization schemes. Finally, the output speech frame is synthesized using the class information and the decoded spectral parameters. In this way the coding process is adapted, frame by frame, to the acoustic-phonetic character of the incoming speech signal.

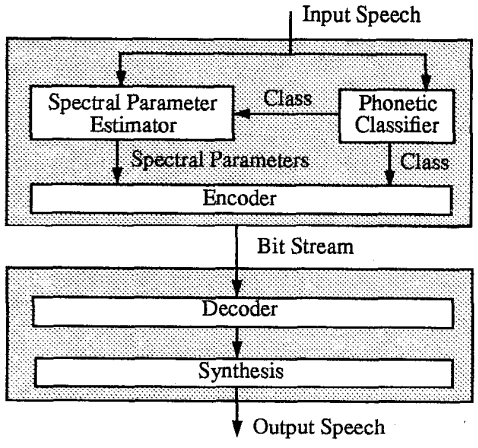

Figure 1. System overview of the EMBE coder

2. THE EMBE SPECTRAL MODEL

The EMBE model classifies each 20 ms input frame into one of the four phonetic catagories: 1) *noise* (N), 2) *unvoiced* (UV), 3) *fully voiced* (FV) and 4) *mixed voiced* (MV). The variable-rate implementation uses another class called silence (S) which is a *noise* type frame with low energy.

The phonetic classification scheme is shown in Figure 2. First, a voice activity detector [12] classifies the input frame either as *active speech* or *noise* (N). All non-*noise* frames are then subdivided by the V/UV classifier into *unvoiced* (UV) or *voiced* classes by evaluating a number of parameters including energy, spectral tilt, low-band energy, and periodicity. The spectral classifier further divides the *voiced* frames into *fully voiced* (FV) or *mixed voiced* (MV) classes using the estimated pitch and the short term spectrum.

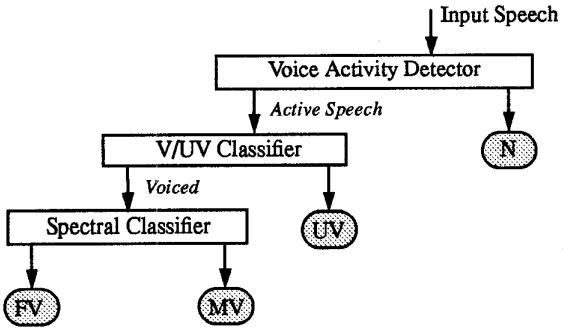

Figure 2: Phonetic classification in the EMBE model

Different parameter sets and different bit allocations are used to represent the spectra of the different phonetic classes. For the UV and N classes, a coarse spectral shape representation is sufficient and therefore the spectrum is modeled by a fixed dimension *spectral shape vector* or SSV, formed by spectral magnitude samples taken at fixed locations. For an FV type frame, the spectrum exhibits a regular harmonic structure and it is represented by the pitch F_0 and an SSV with elements estimated at the harmonics of F_0. For an MV type frame, some parts of the spectra exhibit a regular harmonic structure, while some parts have a noise-like character. In addition to F_0 and the SSV, a frequency domain multiband binary voiced/unvoiced decision vector, VUV, is also used to represent the spectrum. The EMBE synthesis is similar to MBE [1, 2] except for minor changes added to implement class-based processing. The coding delay of EMBE is the same as in the IMBE coder [2].

The phonetic class based spectral modeling in EMBE more accurately represents the high energy unvoiced/noise segments of speech as seen in Figure 3. Here, comparison is made with the IMBE coder [2], which adds undue periodicity (leading to annoying audible artifacts) during such segments. External phonetic classification and a subsequent class-based modeling mitigate these problems. The fixed-rate spectral sampling during unvoiced frames guarantees an adequate representation of spectral shape leading to better quality. It also allows the use of standard VQ methods for unvoiced SSVs.

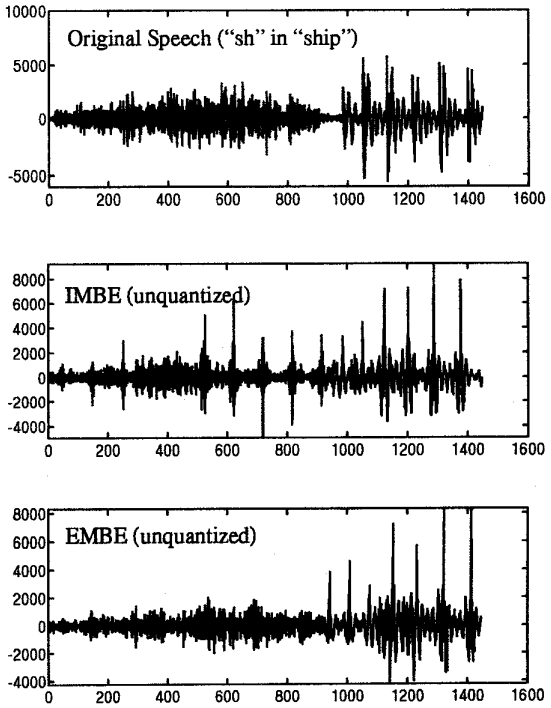

Figure 3. Processing of high energy unvoiced segments

3. QUANTIZATION OF SPECTRAL PARAMETERS

The formidable obstacle in quantizing the spectral parameters is the encoding of the variable dimension *spectral shape vector* or SSV for the *voiced* classes (FV and MV). Since the spectral samples are estimated at pitch harmonics, the dimension of the SSV varies as pitch varies from frame to frame and from speaker to speaker.

Several indirect methods have been studied in the past [2-6] the common theme being the conversion of the variable dimension vector into a fixed dimension parameter set. Such a dimension conversion leads to additional distortion. Since the overall perceptual quality of the coder depends heavily on the quality of the spectral quantization, any added spectral distortion degrades the performance of the coder. Our algorithm uses an innovative method called *variable dimension vector quantization* or VDVQ [7, 9] which does not require any such dimension conversion and instead directly quantizes the variable dimension SSVs with relatively few bits while preserving perceptually important features.

3.1 Variable Dimension Vector Quantization (VDVQ)

The main principle behind VDVQ is the fact that by encoding these SSVs we are essentially encoding a set of samples of the "spectral shape". The spectral shape (largely determined by the vocal tract) is sampled at harmonics of the current pitch, Fo, to generate the SSV. Any particular phoneme will exhibit roughly the same spectral shape but will generate different SSVs when sampled with different F_o, as seen in Figure 4. Female speech will produce lower dimensional SSVs and male speech will generate higher dimensional SSVs.

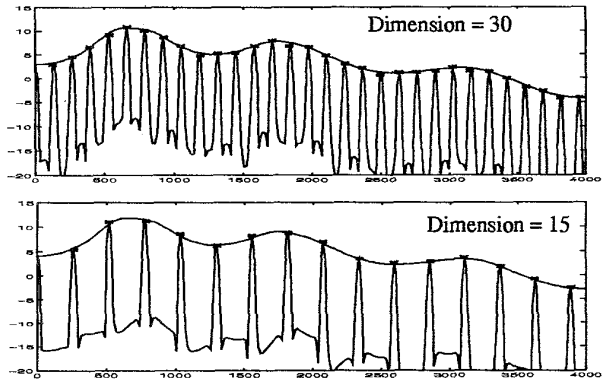

Figure 4: Spectral shape vector generation with various F_o

Therefore, even though SSVs of varying dimension are generated due to varying pitch, the key problem is the quantization of the underlying spectral shape. Consequently, with a *universal spectral shape codebook* C containing K-dimensional code vectors $\mathbf{Y_j}$, where K is the number of DFT points, any SSV, denoted by S, can be well approximated (regardless of its dimensionality) by suitably sub-sampling a matching shape code vector $\mathbf{Y_{j^*}}$ as seen in Figure 5.

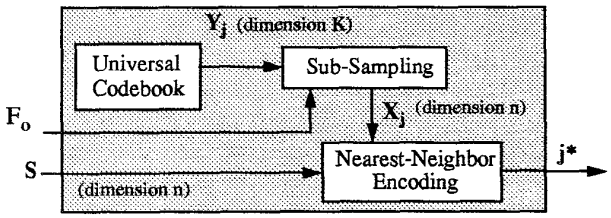

Figure 5: VDVQ encoding

3.2 Class-based Quantization and Bit Allocations

Pitch is losslessly encoded with 8 bits. The frequency domain binary VUV vector is represented by 8 and 12 bits in the fixed and the variable rate implementations respectively. As seen in Figure 6, the SSV S is quantized in the log domain using *mean removed vector quantization* (MRVQ). The mean is scalar quantized and the residual vector is quantized using VDVQ (for FV and MV) or regular VQ (for UV and N).

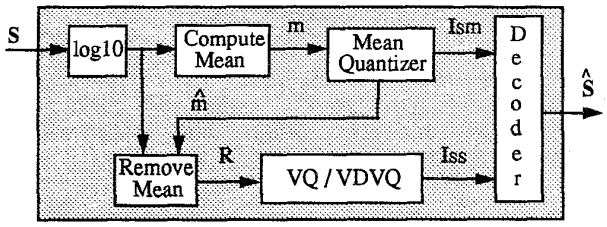

Figure 6. Spectral shape vector encoding/decoding

Table 1 and 2 presents the bit allocations of the fixed rate (2.4 kb/s) and the variable rate implementations of our algorithm. The last row in Table 2 represents estimated class probabilities assuming 50% voice activity.

	N	UV	FV	MV
Class	2	2	2	2
Pitch	-	-	8	8
VUV	-	-	-	8
SSV	46	46	38	30
Total Bits	48	48	48	48

Table 1: Bit allocation: 2.4 kb/s fixed rate implementation

	N	UV	FV	MV
Class	2	2	2	2
Pitch	-	-	8	8
VUV	-	-	-	12
SSV	13	26	30	30
Total	15	28	40	52
Rate	0.75	1.40	2.00	2.60
Class_Prob.	0.50	0.15	0.15	0.20
Average Rate = 1.44 kb/s				

Table 2: Bit allocation: variable rate implementation

4. PERFORMANCE

The performance of our algorithm is compared with other algorithms using 1) subjective quality measures and 2) using spectral distortion (SD). Table 3 compares the SD of the 2.4 kb/s EMBE coder with IMBE [2] and a 2.4 kb/s coder using 10th order all-pole modeling (LP method) similar to [4].

	IMBE	LP	VDVQ
Bits/SSV	63	30	30
Male	2.2	5.6	2.8
Female	1.8	5.0	2.4
Overall	2.0	5.0	2.6

$$SD = \sqrt{\left(\frac{1}{L-1}\cdot\sum_{n=1}^{L-1}\left(20\log S(n\omega_o) - 20\log \hat{S}(n\omega_o)\right)^2\right)}$$

S: original SSV of dimension L; $\hat{S}$: quantized SSV

Table 3: Spectral distortion (SD) comparison

The results show that VDVQ is superior to the all-pole method. VDVQ also enabled the 2.4 kb/s EMBE coder to approach the spectral distortion of the IMBE coder with less than half the bits. However, to obtain the benefit of VDVQ, additional complexity for VQ codebook searching is needed.

For subjective quality evaluation, informal A-B comparison tests (with 20 listeners and 64 sentence pairs including clean and noisy speech) were performed using the 4.15 kb/s Inmarsat-M IMBE coder [2] and the 4.8 kb/s FS 1016 CELP coder[10], as reference coders.

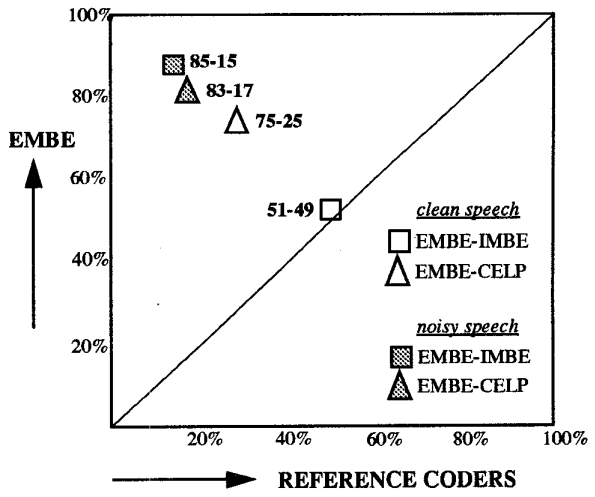

Figure 7: Percentage of preference of the 2.4 kb/s EMBE coder

As seen in Figure 7, the 2.4 kb/s EMBE coder was judged comparable to both the IMBE and the FS 1016 CELP coders for clean speech and was distinctly preferred for noisy speech. The variable rate implementation delivers identical performance at a much lower average bit rate of 1.44 kb/s. Compared to the rigorous subjective tests conducted in speech coding standard competitions, these tests were fairly limited in scope. Even then, the indication is quite clear that phonetic classification based multimode coding enabled our coders to achieve significant rate-distortion tradeoff over the higher rate IMBE and FS 1016 CELP standard coders.

5. CONCLUSIONS

We have presented a low bit rate parametric speech coding algorithm which exploits phonetic classification and a novel spectral quantization technique to deliver high speech quality. Integration of phonetic classification enables efficient distribution of coding resolutions among the various phonetic classes enhancing the quality and robustness of the model. A novel direct quantization scheme (VDVQ) effectively encodes the variable dimensional spectral shape vectors with relatively few bits while preserving the perceptually important features. Two implementations of the algorithm, a fixed-rate (2.4 kb/s) coder and a variable rate (1.4 kb/s average rate) coder, deliver high speech quality, comparable to the higher rate IMBE and FS 1016 CELP coders. Variations of VDVQ and other improvements to IMBE are being explored to further enhance quality and robustness.

6. REFERENCES

[1] D. W. Griffin, J. S. Lim, "Multiband Excitation Vocoder", IEEE Trans. Acoust., Speech, Sig. Process., vol. 36, pp. 1223-1235, August 1988.

[2] Digital Voice Systems, "Inmarsat-M Voice Codec, Version 2", Inmarsat-M specs, Inmarsat, February 1991.

[3] P. C. Meuse, "A 2400 bps Multi-Band Excitation Vocoder", Proc. IEEE Conf. Acoust., Speech, Sig. Process., pp. 9-12, April 1990.

[4] D. Rowe, W. Cowley, A. Perkis, "A Multiband Excitation Linear Predictive Speech Coder", Proc. of Eurospeech, 1991.

[5] A. Das, A. Rao, A. Gersho, "Enhanced Multiband Excitation Coding of Speech at 2.4 Kb/s with Discrete All-Pole Spectral Modeling", Proc. IEEE Globecom Conf., pp. 863-866, November 1994.

[6] M. Nishiguchi et al, "Vector Quantized MBE With Simplified V/UV Decision at 3.0 kbps", Proc. IEEE Conf. Acoust., Speech, Sig. Proc., Minneapolis, pp. 151-154, April 1993.

[7] A. Das, A. Rao, A. Gersho, "Variable-Dimension Vector Quantization of Speech Spectra for Low-Rate Vocoders", Proc. IEEE Data Compression Conf., pp. 420-429, April 1994.

[8] A. Das and A. Gersho, "A Variable-Rate Natural-Quality Parametric Speech Coder", Proc. Inter. Commun. Conf., New Orleans, vol. 1, pp. 216-220, May 1994.

[9] A. Das and A. Gersho, "Enhanced Multiband Excitation Coding of Speech at 2.4 kb/s with Phonetic Classification and Variable Dimension VQ", Proc. EUSIPCO-94, Edinburgh, vol. 2, pp. 943-946, September 1994.

[10] J. P. Campbell Jr., T. E. Tremain, V. C. Welch, "The DoD 4.8 kbps Standard (Proposed Federal Standard 1016)", in B.S. Atal, V. Cuperman, and A. Gersho, editors, Advances in Speech Coding, Kluwer Academic Publ., 1991, pp. 121-133.

[11] V. Welch , T. E. Tremain, "A New Government Standard 2400 bps Speech Coder", Proc. IEEE Speech Coding Workshop, pp. 41-42, October 1993.

[12] K. Srinivasan, A. Gersho, "Voice Activity Detection for Cellular Networks", Proc. IEEE Speech Coding Workshop, pp. 85-86 , October 1993.