MULTIMODE SPECTRAL CODING OF SPEECH FOR SATELLITE COMMUNICATIONS

Amitava Das* and Allen Gersho**

** Qualcomm Inc. 6455 Lusk Boulevard, San Diego, CA 92121.*
Tel/Fax: 619 651-4006/ 658-1562, E-Mail: adas@qualcomm.com

*** Dept. of Electrical & Computer Engineering. University of California.*
Santa Barbara, CA 93106. Tel/Fax: 805 893-2037 / 893-3262, E-Mail: gersho@ece.ucsb.edu

ABSTRACT

We present a multimode spectral coding algorithm which employs the *enhanced MBE* (EMBE) spectral model and a new spectral quantization technique called *transformed variable dimension vector quantization* (TVDVQ) offering good speech quality at low rate. The EMBE model represents the short-term speech spectrum in a mode-specific way. TVDVQ encodes the variable-dimension spectral components efficiently at low complexity. The resulting 2.9 kb/s source coder offers good speech quality comparable to the 4.8 kb/s CELP 1016 and the 4.15 kb/s IMBE coder. An additional 1.1 kb/s of channel coding preserves the speech quality and intelligibility quite well with up to 2% random bit errors.

1. INTRODUCTION

Satellite-based global communication systems are about to revolutionize the telecommunication industry. Major industrial initiatives such as the Iridium and the Globalstar low earth orbit (LEO) satellite based communication schemes promise to provide person-to-person communication between almost any two points on the globe. System designs typically target a bit rate (including channel error protection) of 4 kb/s. We present a low bit rate multimode speech coding algorithm suitable for such applications.

Our algorithm is based on the *enhanced multiband excitation* (EMBE) model [1, 2, 3, 4] and a new spectral quantization technique called *transformed variable dimension vector quantization* (TVDVQ). EMBE is a multimode spectral model which represents the short-term spectrum of speech in a mode-specific way. The new spectral quantization method, TVDVQ, delivers high performance while reducing complexity and storage by significant margin compared to the VDVQ method we introduced earlier [5]. The resulting 2.9 kb/s source coder offers speech quality comparable to the 4.15 kb/s IMBE [6] and the 4.8 kb/s FS 1016 [7] standard coders. An additional 1.1 kb/s channel coding is added to create a robust 4 kb/s multimode spectral coder which offers good, intelligible speech quality up to 2% random bit error. An overview of the multimode spectral coder is presented in Figure 1.

This work was supported by Fujitsu Laboratories Ltd., National Science Foundation, the UC Micro program, Rockwell International Corp., Texas Instruments, Hughes Aircraft Co., Lockheed Missile and Space Co., DSP Group, Moseley Associates, QUALCOMM Inc., National Semiconductor Corp, & Speech Technology Labs.

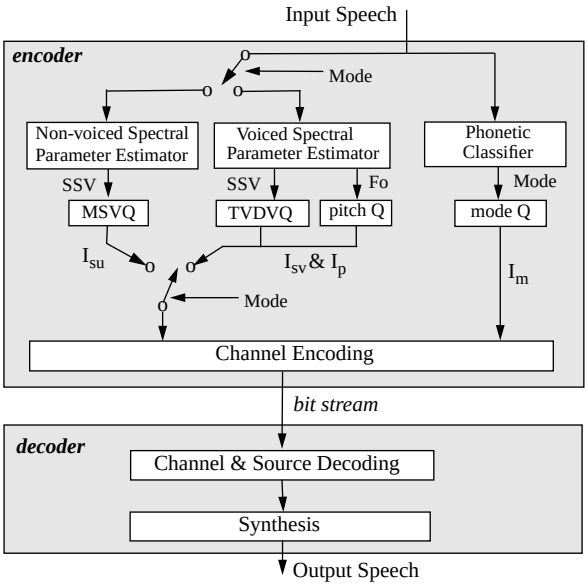

Figure 1. *System overview of the multimode spectral coder*

2. MULTIMODE SPECTRAL CODING

In spectral coding of speech, the spectrum of each speech frame is represented by a model with a set of parameters and then these spectral parameters are encoded by some quantization scheme. In multimode spectral coding (Figure 1) both spectral modeling and quantization are performed in a mode-specific way. A phonetic classifier (Figure 2) labels each 22.5 ms frame as one of 3 non-voiced modes (silence, noise, unvoiced) or 13 voiced modes.

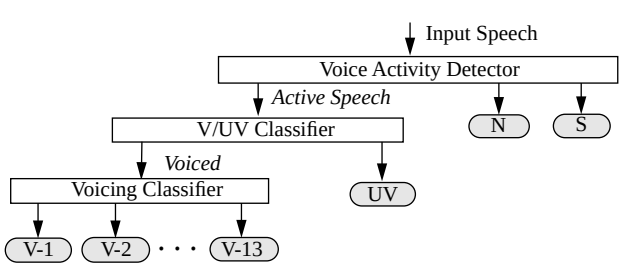

Figure 2. *Mode selection by phonetic classification*

The spectra of non-voiced frames are represented by the mode and a vector of uniformly-sampled spectral amplitudes, called the *spectral shape vector* (SSV). The spectra of voiced frames are represented by the pitch F_o, a variable dimension SSV having components at pitch harmonics (Figure 3) and the *degree of voicing*. Note that unlike IMBE [6], the frequency domain binary voicing information is not transmitted. Instead, it is approximated by a step function as shown in Figure 3 and the transition point or *degree of voicing* is encoded by the mode parameter. As a result, more bits can be allocated to spectral quantization which enhances the overall quality. The synthesis is similar to IMBE [6] but it is performed in a mode-specific way.

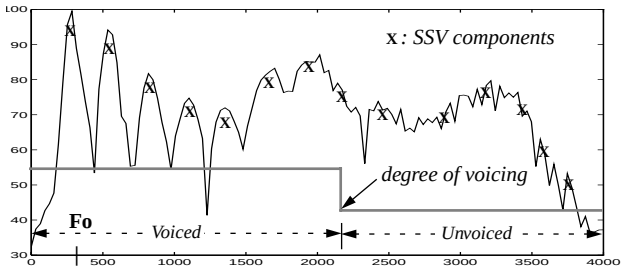

Figure 3. *Voiced spectral parameter estimation*

The fixed dimension SSVs of the non-voiced modes are quantized with multiple-survivor trellis-coded multistage VQ [9]. For the voiced modes, a new technique called TVDVQ directly and efficiently quantizes the variable dimension SSVs. During quantization, a 90 ms *superframe* is formed by combining the parameters of four 22.5 ms frames. Table 1 presents the bit allocation..

	non-voiced	voiced
Mode	4x4	4x4
Pitch	-	8x4
SSV	62x4	54x4
Total	66x4	66x4
Rate	2.9 kb/s	2.9 kb/s

Table 1. *Source coding bit allocation (for the 90 ms superframe)*

3. TRANSFORM VARIABLE DIMENSION VECTOR QUANTIZATION (TVDVQ)

The majority of existing methods for quantizing variable dimension vectors, such as [8, 9], follow a *model-based VQ* (MVQ) approach, where the variable (L) dimension spectral vector is approximated by a model having a fixed (M) number of parameters. The M-dimension parameter vector is then encoded with VQ. The main problem of MVQ is that on top of the VQ distortion, the modeling itself introduces an additional *modeling* distortion. For example (see [9] for details), the 10th order (M=10) all-pole MVQ method [8] exhibits a modeling distortion of 3.8 dB, whereas NSTVQ [9], another MVQ method, exhibits a modeling distortion of 1.1 dB for M=20.

The *transform variable dimension vector quantization* (TVDVQ) method efficiently encodes these variable-dimension spectral vectors without any prior modeling. Therefore, TVDVQ has no modeling distortion and it offers a low-complexity, low-memory and high performance solution by exploiting the benefits of transform coding [10] and structured VQ [10], as explained below.

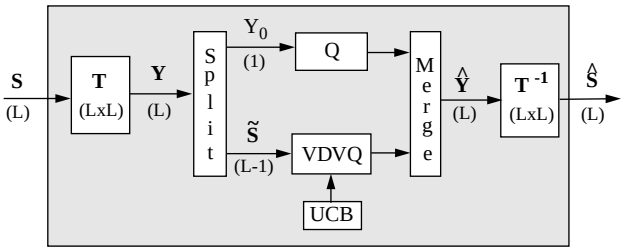

Figure 4. *Transform variable dimension vector quantization*

In TVDVQ, the incoming variable (L) dimension SSV $\mathbf{S}$ is mapped into an L-dimension transform vector $\mathbf{Y}$ by a suitable orthogonal linear transform $\mathbf{T}$. The L-dimension estimate $\hat{\mathbf{X}}$ is computed by applying the inverse transform $\mathbf{T^{-1}}$ to the quantized transform vector $\hat{\mathbf{Y}}$. Unlike NSTVQ [8], here the dimension of the transformed vector $\mathbf{Y}$ is not fixed, but variable. Since $\mathbf{T}$ is orthogonal, the overall distortion $\|\mathbf{X}-\hat{\mathbf{X}}\|^2$ equals the quantization distortion $\|\mathbf{Y}-\hat{\mathbf{Y}}\|^2$. Therefore, the transformation process in TVDVQ does not incur any additional modeling distortion. The *discrete cosine transform* is used as $\mathbf{T}$.

The transform vector $\mathbf{Y}$ is encoded by a gain-shape coding scheme which ensures robustness against signal level variations. The Y[0] component, which is related to the signal energy level, is separately encoded. The variable (L-1) dimension shape vector, $\tilde{\mathbf{S}}$, formed by the remaining components of $\mathbf{Y}$, is encoded as follows.

3.1 TVDVQ encoding and decoding algorithms

The shape vector $\tilde{\mathbf{S}}_k$ of dimension $(L_k\text{-}1)$ is compared to each code-vector $\mathbf{C_j}$ of a universal codebook, $U = \{\mathbf{C_j}\}$, i= 1, 2, ..., N, to find the best match that minimizes the distortion:

$$D(\mathbf{S_k}, \mathbf{C_j}) = \sum_{i=1}^{L_{max}-1} (\mathbf{S_k}[i] - \mathbf{C_j}[i])^2 \times \mathbf{W_k}[i]$$

where

$$\mathbf{W_k}[i] = \begin{cases} 1/(L_k - 1) & \text{if } (i < L_k) \\ 0 & \text{otherwise} \end{cases}$$

where L_{max} is the maximum possible dimension of the input SSV $\mathbf{S}$ and $\mathbf{W_k}$ is the weight vector. The index j*, for which $D(\mathbf{S_k}, \mathbf{C_j})$ is minimum over all j =1, 2,...,N, is selected.

The decoder is a simple table look-up operation where the $(L_k\text{-}1)$ dimension $\hat{\mathbf{S}}_\mathbf{k}$ is formed by the first $(L_k\text{-}1)$ components of $\mathbf{C_{j*}}$.

3.2 TVDVQ training algorithm

Given a large training set of pairs $\{(\mathbf{S_k}, L_k)\}$, and an initial codebook of size N and dimension (L_{max}-1), the universal codebook is designed in an iterative manner similar to the *Generalized Lloyd Algorithm* (GLA) [10]. Let C_j, j=1,2,...,N, be the codevector prior to the current iteration. The two steps of each training iteration are:

1. Nearest neighbor partitioning.

For each $(\mathbf{S_k}, L_k)$, form $\mathbf{W_k}$ and assign $(\mathbf{S_k}, L_k)$ to partition P_m if $D(S_k, C_m) \le D(S_k, \mathbf{C_j})$ for j=1,2,...,N.

2. Centroid computation.

For each partition P_m. m=1,2,...,N, find a new codevector, Y'_m, the centroid, that minimizes $D_m = \sum_{S_j \in P_m} D(S_j, C)$, over all $C \in \Re^{L_{max}}$ as:

$$C'_m[n] = \frac{\sum_{S_k \in P_m} S_k[n] \times W_k[n]}{\sum_{S_k \in P_m} W_k[n]} \quad \text{for n = 1,2,...,}(L_{max}\text{-1})$$

3.3 Implementation and performance of TVDVQ

In order to attain high performance at low complexity and low memory usage, a multiple-survivor multistage trellis VQ structure [10] is employed in TVDVQ (Figure 5). This imparts a great deal of flexibility to TVDVQ. By selecting the right TVDVQ design parameters, M (number of stages), N (number of codevectors/ stage) and K (number of best codevectors retained/ stage), one can easily meet the requirements (distortion, complexity, memory usage) of the target application.

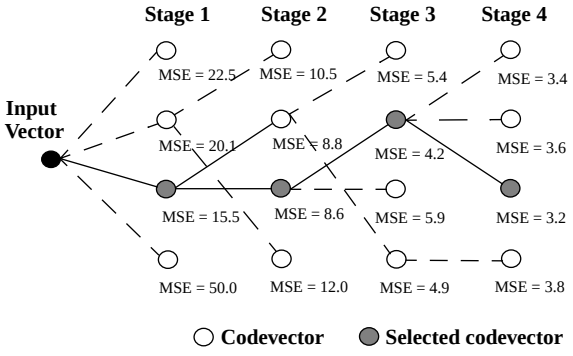

Figure 5. *Multiple survivor multistage trellis VQ implementation of TVDVQ (M=4, N=4, K=1)*

The performance of TVDVQ (N=64, K=1, the number of stage M is varied for different rates) is compared in terms of the spectral distortion (SD) measure [2] with VDVQ [5], the LP-MVQ [8] and the quantization method IMBE [6]. As evident from Table 2, TVDVQ offers similar SD as VDVQ and lower SD than LP-MVQ. Compared to the IMBE quantization method, TVDVQ offers similar SD but with less bits/frame, saving 12 bits/frame.

Method	Rate (bits/frame)	SD (dB)
LP-MVQ10	30	4.24
VDVQ	30	2.84
TVDVQ	30	2.87
VDVQ	54	1.62
TVDVQ	54	1.57
IMBE	66	1.61

Table 2. *Spectral distortion comparison*

Method	Rate (bits/frame)	Peak Complexity (multiply/encoding)	Memory (words)
VDVQ	54	229 x 1024	229 x 1024
TVDVQ	54	27.5 x 1024	27.5 x 1024

Table 3. *Comparison of complexity and storage*

As evident from Tables 2 and 3, TVDVQ offers the same high performance as VDVQ but at significantly less complexity as well as with significantly less storage.

4. CHANNEL CODING

In the proposed channel coding scheme (Figure 6), the source bits are classified into 3 groups according to their perceptual significance or the effect of error in these bits on the overall perceptual quality. Class 0 represents the perceptually most significant bits

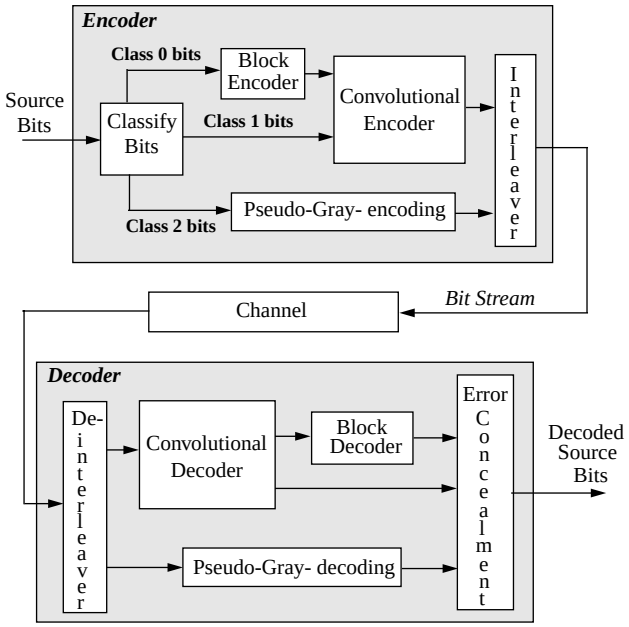

Figure 6. *Channel coding in the multimode spectral coder*

and they enjoy the maximum protection offered by (24, 12) block coding and a 3/5 rate convolutional coder. Class 1 bits are protected by convolutional coding alone. The remaining class-3 bits are protected by only pseudo-Gray coding [11]. Interleaving is used to decorrelate the "bursty" type errors typical of satellite channels. At the decoder, after the proper block and convolutional decoding, an estimate is made to determine whether any catastrophic error has occurred and in such an event soft error concealment is applied.

4.1 Performance of the channel error protection scheme

The performance of the channel error protection scheme is evaluated with three objective measures by comparing *test* speech signals with the corresponding *reference* signals. The test signals are speech processed by the multimode spectral coder under various random bit error conditions. The reference signal is speech processed by the 2.9 kb/s source coder when there was no bit error. Three bit error rates (BER), 0.5%, 1% and 2%, are considered. A 12000 frame speech database is used for this experiment.

The objective measures are: 1) the percentage of frames (F) for which the test signal is identical to the reference signal indicating a full recovery from channel errors, 2) log spectral distortion (SD) defined in [2], and logSNR (SNR). A large value of SD or a low value of SNR indicates distortion in the test signal. The outliers of SD and SNR, O_s and O_n are also computed. O_s is defined as the percentage of times SD is greater than 3 dB and O_n is defined as the percentage of time SNR is less than 0 dB. There outliers indicate large deviation in spectrum or waveform and hence they are indicative of significant artifacts.

Coder	BER (%)	F (%)	SD (dB)	O_{SD} (%)	SNR (dB)	O_{SNR} (%)
EMBE 4 kb/s	0.5	57.3	0.5	0.4	38.1	0.2
	1.0	37.4	0.9	1.9	29.4	2.7
	2.0	15.1	1.5	4.3	28.4	5.3
EMBE 2.9 kb/s	0.5	12.9	3.0	47.2	8.5	50.7
	1.0	1.8	4.0	63.1	0.1	72.0
	2.0	0.2	4.9	70.1	-3.0	76.3

Table 4. *Performance of the channel coding scheme*

The results are presented in Table 4. The bottom 3 rows exhibit the effect of different extents of bit error on the coder in the absence of any protection, whereas the top 3 rows indicate the performance of channel coding. By comparing these results, we see that a significantly large percentage of frames are fully recovered by the channel coding scheme. SD, SNR and their outliers are also significantly reduced by channel coding, indicating the absence of any significant distortion in the decoded speech.

Informal listening comparisons indicate that up to 1% BER condition, the speech quality of the 4 kb/s robust coder is virtually identical to the reference signal. Under 2% BER condition, there are occasional artifacts, but the speech quality is fairly close to the reference signal. In all cases (up to 2% BER), speech intelligibility is fully preserved.

5. CONCLUSIONS

We presented a multimode spectral coding algorithm which employs the EMBE multimode spectral model and a new spectral quantization technique called *transform variable dimension vector quantization* (TVDVQ) to deliver good speech quality at low rate. Informal listening tests indicate that the speech quality of the 2.9 kb/s source coder is comparable to the 4.15 kb/s IMBE coder [6] and the 4.8 kb/s CELP 1016 coder[7]. We also presented a 1.1 kb/s channel coding scheme comprising of both hard and soft error protection mechanisms. The source bits are selectively provided different levels of error protection depending on their sensitivity to bit errors. The resulting 4 kb/s robust multimode spectral coder demonstrated good speech quality and high intelligibility under various channel error conditions up to 2% random bit error rates.

6. REFERENCES

[1] A. Das, "Multimode Spectral Coding of Speech at Low bit Rates", *PhD Thesis*, University of California, Santa Barbara, June 1996.

[2] A. Das and A. Gersho, "Multimode Spectral Coding of Speech at 2400 bps and Below",, *Proc. IEEE Speech Coding Workshop-1995*, Annapolis, USA, pp. 107-108, September 1995.

[3] A. Das and A. Gersho, "Variable Dimension Spectral Coding of Speech at 2400 bps and Below With Phonetic Classification", *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 492-495, May 1995.

[4] 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.

[5] 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.

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

[7] 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 Pub., 1991, pp. 121-133.

[8] M. S. Brandstein, "A 1.5 Kbps Multi-Band Excitation Speech Coder", *S.M. Thesis*, EECS Department, MIT, 1990.

[9] P. Lupini and V. Cuperman, "Vector quantization of harmonic magnitudes for low rate speech coders", *Proc. IEEE Globecom conf*, pp. 858-862, November 1994.

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

[11] K. Zeger and A. Gersho, "Pseudo-Gray Coding", IEEE Trans. on Comm., vol. 38, pp. 2147-2158, 1990.