MULTISENSOR VERY LOW BIT RATE SPEECH CODING USING SEGMENT QUANTIZATION

Alan McCree, Kevin Brady, and Thomas F. Quatieri

MIT Lincoln Laboratory
Lexington, MA 02420
E-mail: [mccree, kbrady, tfq]@ll.mit.edu

ABSTRACT

We present two approaches to noise robust very low bit rate speech coding using wideband MELP analysis/synthesis. Both methods exploit multiple acoustic and non-acoustic input sensors, using our previously-presented dynamic waveform fusion algorithm to simultaneously perform waveform fusion, noise suppression, and cross-channel noise cancellation. One coder uses a 600 bps scalable phonetic vocoder, with a phonetic speech recognizer followed by joint predictive vector quantization of the error in wideband MELP parameters. The second coder operates at 300 bps with fixed 80 ms segments, using novel variable-rate multistage matrix quantization techniques. Formal test results show that both coders achieve equivalent intelligibility to the 2.4 kbps NATO standard MELPe coder in harsh acoustic noise environments, at much lower bit rates, with only modest quality loss.

Index Terms— Nonacoustic sensor, phonetic vocoder, vector quantization, MELP

1. INTRODUCTION

The problem of encoding speech signals at very low bit rates (less than 1000 bps) is a very difficult one, typically addressed by jointly encoding a block of consecutive speech frames. Variable segment sizes allow the quantizer to match typical speech patterns, as in the phonetic vocoder [1, 2, 3, 4]. By contrast, fixed duration segments can be used with efficient matrix quantization techniques [5, 6].

Since speech coding in severe acoustic noise is even more difficult, recent work has explored the use of non-acoustic sensors to supplement the acoustic microphone information [7, 8]. In particular, waveform fusion of non-acoustic sensors, combined with additional highband speech encoding, produced significant intelligibility improvements for the 2.4 kbps NATO MELPe standard in [8], and a dynamic waveform fusion algorithm that combines sensor fusion, noise suppression, and crosschannel noise cancellation into a time-varying Wiener filter provided further improvement [9].

This paper describes two noise robust, multisensor, very low bit rate speech coders. As shown in Figure 1, both use a dynamic waveform fusion front-end to combine multiple acoustic and non-acoustic sensors. Also, both rely on parametric analysis and synthesis using a wideband MELP algorithm [10]. The first coder, at 600 bps, uses variable-length segmentation and quantization based on the Scalable Phonetic Vocoder [4]. An alternative approach, using fixed block-sizes, relies on novel matrix quantization techniques to attain good performance even at 300 bps.

The remainder of this paper is organized as follows. The dynamic waveform fusion algorithm is reviewed in Section 2. Details of the 600 bps coder are provided in Section 3, along with formal test results. Section 4 presents the 300 bps coder algorithm design and performance, and is followed by concluding remarks.

2. MULTISENSOR DYNAMIC WAVEFORM FUSION

By incorporating a multiplicative noise model into a multichannel Wiener filtering approach, we have shown that non-acoustic signals can be optimally exploited using a minimum mean squared error criterion [9]. When combined with a robust estimator of instantaneous signal-to-noise ratio, this dynamic waveform fusion algorithm automatically adjusts the sensor combination coefficients to achieve the benefits of waveform fusion, noise suppression, and cross-channel noise cancellation. Formal testing results have shown that the resulting dynamic waveform fusion algorithm provides significant intelligibility and quality improvement for low-rate coding in difficult acoustic environments.

In this work, we use this multisensor dynamic waveform fusion as a front-end with a set of six sensor signals. These sensors are a dual channel close-talking noise cancelling microphone from Aliph Corporation, two channels of a second-generation microwave radar sensor mounted on the throat (also from Aliph), a piezo-electric vibrometer also on the throat (P-mic), and a bone conduction microphone located on the top of the skull (bone-mic). In some cases, we also use a resident acoustic microphone (Gentex M175A); this is also the microphone used for the MELPe baseline coders.

3. 600 BPS SCALABLE PHONETIC VOCODER

In a phonetic vocoder, the information content of the speech signal is extracted with a phonetic speech recognizer, and the prosody of the particular utterance is encoded with a separate scheme such as pitch contour quantization [1]. In the Scalable Phonetic Vocoder (SPV), additional information is also transmitted to improve the speech quality [4]. The phonetic speech recognition algorithm uses Hidden Markov Models (HMM's), with each of 39 monophones characterized by five states. Instead of synthesizing with the MELP parameters representing each phone state, scalability is achieved by selecting the closest codeword in a state-specific codebook and transmitting this index.

This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

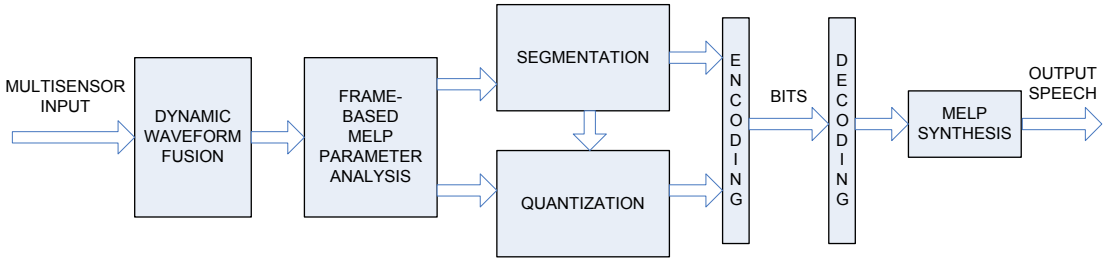

Fig. 1. Multisensor Very Low Bit Rate Coder Structure

Parameter	Bits/unit	Unit	Bit Rate
Phone	5	phone	40
Phone duration	5	phone	40
State path	4	phone	32
Pitch contour	8	100 ms	48
MELP supervector	4.1	10 ms	408
Total			568

Table 1. 600 bps SPV bit allocation.

Coder	BH	APC	M1	HM
Res. mic. 2.4 kbps MELPe	75.0	79.8	76.5	86.4
Multisensor 600 bps SPV	80.8	87.7	83.8	86.8

Table 2. DRT scores for four environmental noise conditions.

Coder	APC	BH	HM
Multisensor 600 bps SPV	30.1	41.8	46.9

Table 3. A/B percent preference scores of coder over 2.4 kbps MELPe for three environmental noise conditions.

3.1. Quantizer Design

In the terminology of quantization theory, the SPV is a classified vector quantizer (VQ) [11]. Each of 195 possible phone states is a separate class, with a corresponding VQ codebook, and the phonetic state sequence is encoded as side information. For each 10 ms frame, the quantizers encode the supervector of MELP parameters consisting of 14 wideband mel-scale Line Spectral Frequencies (LSFs), the frame gain in dB, and the voicing cutoff frequency. An extension of switched-predictive VQ approach is used, where each codeword represents a codevector/predictor pair in a process we call joint predictive vector quantization (also known as selective linear prediction [12]). Since the phonetic vocoder uses variable segment lengths, we also use variable bit rate quantization techniques. In particular, the quantizers are designed using entropy-constrained VQ [11], where the codebook search considers both the distortion and associated bit rate for each variable length codeword index.

The bit allocation for the 600 bps SPV is shown in Table 1. For each phone, the quantized side information includes the phone and the state path. The 39 phones have an uneven distribution with less than 5 bits of entropy. We encode the state path with a product VQ, where the duration of up to 40 frames is encoded exactly with an average of 5 bits, and the fraction of time in each state is quantized with a 4-bit VQ. With a measured average phone duration of 124 ms, these 14 bits of side information require an average of 113 bps. The pitch contour is encoded with a fixed segment duration of 10 frames (100 ms). An 8-bit direct VQ of the logarithm of the pitch is used, The MELP parameter supervector is encoded with entropy-constrained joint predictive VQ at an average of 6 bits per frame. Both pitch and supervector quantizers use a frame weighting function reflecting the perceived loudness of each frame as in [13]. Finally, the average bit rate is further reduced by special handling of non-speech regions; in these cases no pitch is transmitted and the MELP supervector is only sent for every 5th frame and interpolated for the rest.

3.2. Performance

This 600 bps SPV coder was tested by ARCON Corporation as part of the Defense Advanced Research Projects Agency (DARPA) Advanced Speech Encoding program in 2006. Testing consisted of the Diagnostic Rhyme Test (DRT) for intelligibility and an A/B forced choice comparison against 2.4 kbps MELPe for speech quality. Four severe military noise environments were tested: the UH-60 Blackhawk helicopter (BH), the M577 armored personnel carrier (APC), the IPM1 Abrams tank (M1) and the High-Mobility Multipurpose Wheeled Vehicle or HMMWV (HM) traveling over rough terrain. The DRT scores in Table 2 show significant improvement in intelligibility for the first three environments over the baseline 2.4 kbps MELPe. Table 3 shows that this significant bit rate reduction results in only a modest drop in quality.

4. 300 BPS MATRIX QUANTIZATION

An alternative approach to very low bit rate speech coding is matrix quanitzation: joint quantization of a block of consecutive frames [5]. While it might seem that a fixed blocksize is less sophisticated than the variable segment duration of a phonetic vocoder, this method does have a number of advantages. First, a fixed blocksize may be easier to use in potential applications. Also, this approach avoids the significant complexity and look-ahead delay needed for phonetic speech recognition. Finally, matrix quantization can fully exploit dependencies between neighboring frames, even when they are not in the same phone state.

4.1. Joint Predictive Multistage Matrix Quantization

The fundamental problem with matrix quantization is the rapid expansion in codebook size and search complexity with increasing blocksize. Multistage VQ (MSVQ) allows a suboptimal constrained

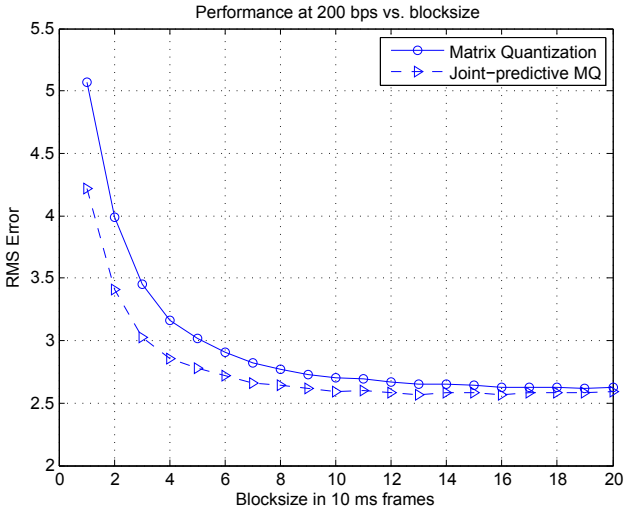

Fig. 2. Performance of 200 bps VQ with increasing blocksize

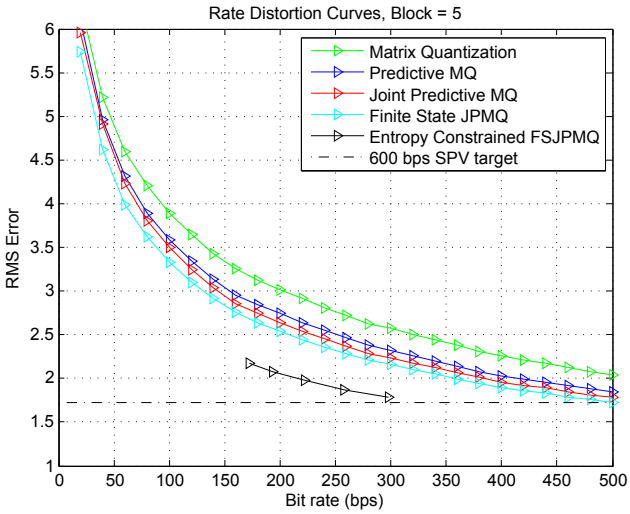

Fig. 3. Rate distortion curves with blocksize of 50 ms.

architecture to simulate very large codebook sizes with reasonable complexity [14]. Therefore, multistage matrix quantization (MSMQ) can be used in very low rate coding [6].

We would like to continue to make use of the prediction gain across blocks, as in the SPV, so we have developed a simple but effective generalization of the single-frame joint predictive approach. Rather than using an entire previous block, the memory term is simply the last *frame* in the previous block. Each frame in the current block has a unique set of predictor coefficients, which typically become weaker for frames away from the beginning of the block. In this way, the prediction gain and smooth time evolution properties of predictive coding are preserved in matrix quantization, while maintaining simple optimal design and search algorithms. In the multistage structure, only the first stage codebook includes predictor coefficients.

To investigate the potential of this approach, we trained MSMQ quantizers of the wideband MELP supervector. For each design, the stages are built using 8 bits for the first stage and up to 6 bits for each remaining stage. The stages are searched and trained jointly using an M-best search with $M = 4$. Figure 2 shows the results for 200 bps quantizers with blocksizes from 10 to 200 ms. With increasing blocksizes, performance improves rapidly at first and then levels off at around 150 ms blocksize. Beyond this point, any further potential gains in coding efficiency are lost due to the multistage constraints. Joint predictive MSMQ reaches its best performance at blocksizes around 100 ms, and surprisingly continues to show a slight performance advantage over the non-predictive approach even with large blocks.

4.2. Classified Matrix Quantization

In the SPV coder, significant coding gains are achieved by classified VQ. One way to attain these gains without the coding overhead of phonetic information is with Finite State VQ (FSVQ). FSVQ performs classification based on the state of the previous frame [11]. We use this approach for matrix quantization, but have extended it in two ways. First, similar to our generalization of predictive coding, we use only the last frame of a block for the finite state memory. This requires a separate classification codebook search for the reconstructed last frame. Second, we combine finite state and prediction by performing classification of the predictively reconstructed frame. While these are often viewed as two different approaches to exploit redundancy, in our experiments, the combination of a small number (16 or less) of finite states with joint predictive quantization performs better than either alone.

4.3. Variable-Rate Quantization

As in the SPV, we use variable-rate design using entropy constraints. For multistage designs, we found significant improvement by using codeword indices for each stage that are conditioned upon the previous stage index as in [15].

4.4. Numerical Performance

We compared the weighted mean squared error of the MELP supervector for various matrix quantizer designs against the quantizer in the 600 bps SPV. Figure 3 shows the rate distortion curves at a fixed blocksize (50 ms) for MSMQ, predictive MSMQ, joint-predictive MSMQ, finite-state joint-predictive MSMQ, and entropy-constrained finite-state joint-predictive MSMQ. As before, the MSMQ structures use 8 bits for the first stage and 6 bits thereafter. From these curves, the performance of 500 bps MSMQ can be achieved at a fixed rate of 340 bps with the addition of finite state and joint prediction, and an average rate of just 200 bps with variable-rate design. Also, this variable rate approach nearly achieves the performance of the variable-rate 600 bps SPV quantization (550 bps for the MELP supervector) at an average rate of only 300 bps.

4.5. 300 bps Coder Design

Based on these encouraging results, we developed a complete 300 bps MSMQ coder. Compared to the preliminary results reported above, this design has a number of optimizations. First, based on listening experiments we cut the bandwidth of our wideband MELP from 8 to 6 kHz. This reduces the information content while preserving the important signal characteristics, and allows the use of

Coder	BH	M2	APC	M1
Res. mic. 2.4 kbps MELPe	84.7	80.0	86.1	79.3
Res. mic. 1.2 kbps MELPe	80.8	77.4	82.2	75.7
Res. mic. 600 bps MELPe	74.2	72.3	75.5	67.7
Multisensor 300 bps MSMQ	83.8	85.1	85.5	79.6

Table 4. DRT scores for four environmental noise conditions.

Coder	BH	M2	APC	M1
Res. mic. 1.2 kbps MELPe	42.6	45.8	44.5	43.5
Res. mic. 600 bps MELPe	16.5	16.2	18.5	16.5
Multisensor 300 bps MSMQ	27.9	38.1	29.2	29.2

Table 5. A/B percent preference scores of coder over 2.4 kbps MELPe for four environmental noise conditions.

linear-scale LPC. Also, we incorporated pitch into the MELP parameter supervector, so that all information is now encoded jointly in one 17-dimensional supervector. This required some effort to design appropriate perceptual weighting, but results in a simple and elegant design: the only transmission for each block is a single codebook index. Two gender-dependent codebooks were used, each with 8 finite states. Finally, we made two changes that increase performance at the price of some additional complexity: increasing the blocksize to 80 ms and the M-best search range to 16.

4.6. Performance Evaluation

Testing of the 300 bps MSMQ coder was conducted at ARCON earlier this year, in four acoustic noise environments: UH-60 Blackhawk helicopter (BH) Bradley Fighting Vehicle (M2). M577 armored personnel carrier (APC), and IPM1 Abrams tank (M1). Table 4 shows that the multisensor 300 bps coder has similar intelligibility to the 2.4 kbps NATO standard, and much higher than the 600 bps MELPe. Note also that the baseline MELPe scores were higher in this testing than in the previous year's; this is due to a change in the acoustic sound field generation during the recordings at ARCON. While quality continues to be more difficult, Table 5 shows that the quality of the 300 bps coder is better than 600 bps MELPe but less than the 2.4 kbps version.

5. CONCLUSION

We have presented two very low bit rate speech coders based on wideband multisensor segmental MELP. By exploiting the additional bandwidth as well as dynamic waveform fusion of acoustic and non-acoustic sensors, both are able to maintain the intelligibility of the much higher bit rate 2.4 kbps NATO MELPe coder, with only a modest drop in quality. The first coder, a 600 bps scalable phonetic vocoder, confirmed the potential of a variable segmentation approach at low rates. However, in our experiments a sophisticated fixed blocksize quantizer performed as well at even lower rates. The inclusion of joint prediction, finite state classification, and entropy-constrained variable rate coding provides a large improvement in rate/distortion performance over traditional matrix quantization, and the M-best search of multistage codebooks allows control of storage and complexity. Formal testing confirmed that our 300 bps coder design based on these techniques provides remarkable performance for such a low bit rate.

6. REFERENCES

[1] J. Picone and G. Doddington, "A phonetic vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1989, pp. 580–583.

[2] C. Ribeiro and I. Trancoso, "Phonetic vocoding with speaker adaptation," in *Proc. EUROSPEECH '97*, 1997, pp. 1291–1294.

[3] T. Masuko, K. Tokuda, and T. Kobayashi, "A very low bit rate speech coder using HMM with speaker adaptation," in *Proc. ICSLP '98*, 1998.

[4] A. McCree, "A scalable phonetic vocoder framework using joint predictive vector quantization of MELP parameters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2006, pp. I 705–708.

[5] D. Y. Wong, B. H. Juang, and D. Y. Cheng, "Very low data rate speech compression with LPC vector and matrix quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1983, pp. 65–68.

[6] S. Ozaydin and B. Baykal, "A 1200 bps speech coder with LSF matrix quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 677–680.

[7] L. C. Ng, G. C. Burnett, J. F. Holzrichter, and T. J. Gable, "Denoising of human speech using combined acoustic and EM sensor signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 229–232.

[8] T. F. Quatieri, K. Brady, D. Messing, J. P. Campbell, W. M. Campbell, M. S. Brandstein, C. J. Weinstein, J. D. Tardelli, and P. D. Gatewood, "Exploiting nonacoustic sensors for speech encoding," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 533–544, Mar. 2006.

[9] A. McCree, K. Brady, and T. F. Quatieri, "Multisensor dynamic waveform fusion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2007, pp. IV 577 – 580.

[10] A. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, July 1995.

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

[12] K. M. Holt and D. L. Neuhoff, "Coding by selective linear prediction: a new scheme for predictive vector quantization," in *Proc. IEEE Int. Conf. Image Processing*, 2002, pp. II657–II660.

[13] E. B. George, A. McCree, and V. R. Viswanathan, "Variable frame rate parameter encoding via adaptive frame selection using dynamic programming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, 1996.

[14] W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud, and V. Cuperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 4, pp. 373–385, Oct. 1993.

[15] F. Kossentini, M. J. T. Smith, and C. F. Barnes, "Necessary conditions for the optimality of variable-rate residual vector quantizers," *IEEE Trans. Information Theory*, vol. 41, pp. 1903–1914, Nov. 1995.