A 1200/2400 BPS CODING SUITE BASED ON MELP

Tian Wang, Kazuhito Koishida

Microsoft Corp., One Microsoft Way, Redmond, WA 98052
tianwang@microsoft.com, kazukoi@microsoft.com

Vladimir Cuperman

Niftybox LLC, 5635 Cielo Ave., Goleta, CA 93117
vladimir@dsp-consult.com

Allen Gersho

ECE Dept., University of California, Santa Barbara, CA 93106
gersho@ece.ucsb.edu

John S. Collura

National Security Agency, 9800 Savage Rd, STE 6516, Ft. Meade, MD 20755-6516
jscollu@alpha.ncsc.mil

ABSTRACT

This paper presents key algorithm features of the future NATO Narrow Band Voice Coder (NBVC), a 1.2/2.4 kbps speech coder with noise preprocessor based on the MELP analysis algorithm. At 1.2 kbps, the MELP parameters for three consecutive frames are grouped into a superframe and jointly quantized to obtain high coding efficiency. The inter-frame redundancy is exploited with distinct quantization schemes for different unvoiced/voiced (U/V) frame combinations in the superframe. Novel techniques used at 1.2 kbps include pitch vector quantization using pitch differentials, joint quantization of pitch and U/V decisions and LSF quantization with a forward-backward interpolation method. A new harmonic synthesizer is introduced for both rates which improves the reproduction quality. Subjective test results indicate that the 1.2 kbps speech coder achieves quality close to the existing federal standard 2.4 kbps MELP coder.

1. INTRODUCTION AND CODER OVERVIEW

In March 1996, the US government's Digital Voice Processing Consortium (DDVPC) selected the 2.4 kbps Mixed Excitation Linear Prediction (MELP) [1] speech-coding algorithm to be a new standard for narrow band secure voice coding products and applications. The MELP algorithm leads to a significant improvement in both speech quality and intelligibility compared to the LPC10 type algorithms. However, in some difficult satellite and HF radio channels, robust 2.4 kbps transmission is not always possible. To enable high performance interoperable secure speech communications in harsh acoustic and channel error environments, a dual-rate codec operating at rates of 1.2 kbps and 2.4 kbps is needed.

In this paper, we describe a MELP 1.2 kbps speech coder and a new harmonic synthesizer for both 1.2 and 2.4 kbps developed at SignalCom, Inc. and Microsoft under a DoD contract. The proposed 1.2 kbps coder, called Multi-Frame MELP or MF-MELP, shares the core analysis algorithm with the 2.4 kbps MELP standard, and its transmitted parameters are the same as those of the 2.4 kbps MELP coder. In the 1.2 kbps coder, the MELP parameters of three consecutive 22.5 ms frames are grouped together into a superframe and jointly quantized to obtain high coding efficiency. A pitch smoother is incorporated to avoid large pitch errors, and this results in an increase in the look-ahead by 129 samples. The total algorithmic delay for MF-MELP is 103.75 ms.

The subjective test results of the codec integrated with a noise pre-processor [3] show that it achieves quality close to the 2.4 kbps MELP at half the bit-rate. The enhanced dual-rate MELP algorithm, called MELPe has been selected as the new NATO STANAG 4591 standard. The testing was comprehensive and indicated an overwhelming advantage for the MELPe coder. In fact, the combined scores for both MELPe rates outscored the 8 kbps CELP FS 1616 coder.

The Multi-frame MELP (MF-MELP) quantization schemes are designed to efficiently exploit the superframe structure by using vector quantization (VQ) and interpolation, taking into account the statistical properties of voiced (V) and unvoiced (U) speech. Each superframe is categorized into one of several coding states with a different bit allocation for each state. State selection is done according to the U/V pattern of the superframe. The MF-MELP utilizes several techniques for reducing the effect of state mismatch between encoder and decoder due to channel errors.

2. QUANTIZATION OF PITCH AND U/V DECISIONS AT 1.2 KB/S

2.1 Pitch Quantization

Different pitch quantization schemes are used for different U/V combinations in the superframe. Within those superframes where the voicing pattern contains either two or three voiced frames, the

0-7803-7549-1/02/$17.00 ©2002 IEEE.

pitch parameters are vector-quantized. For voicing patterns containing only one voiced frame, the scalar quantizer used in the MELP standard is applied for the pitch of the voiced frame. For the UUU voicing pattern, no pitch information is transmitted.

The pitch values, P_i (i=1,2,3), obtained from the pitch analysis are transformed into logarithmic values, $p_i = \log P_i$, prior to quantization. For each superframe, a pitch vector is constructed with components equal to the log pitch value for each voiced frame and a zero value for each unvoiced frame.

The pitch VQ algorithm has three steps for obtaining the best index. In the first step M-best candidates are selected using a weighted squared Euclidean distance measure [6]. In the second step, we calculate the differentials of the unquantized and quantized log pitch values. Finally, in the last step, we select from the M-best candidates the optimum index that minimizes a weighted distortion measures that combines pitch errors and the pitch differential errors over three frames [6]. This novel distortion measure incorporates pitch differentials to account for the perceptual importance of adequately tracking the pitch trajectory.

2.2 Joint Quantization of Pitch and U/V Decisions

The U/V decisions and pitch parameters for each superframe are jointly quantized using 12 bits consisting of 3 mode bits (representing the 8 possible combinations of U/V decisions for the 3 frames in a superframe) and 9 bits for pitch values (Table 1). The scheme employs six separate pitch codebooks, five having 9 bits (i.e. 512 entries each) and one being the MELP scalar quantizer. The specific codebook is determined according to the 3 mode bits. The pitch vector is quantized employing the procedure described in the previous Section with the selected codebook to generate a 9-bit codeword. Four codebooks are assigned to the superframes in the VVV mode, i.e., the pitch vector is quantized by one of 2048 codewords. If the number of voiced frames in the superframe is not larger than one, the 3-bit codeword is set to 000 and the distinction between different modes is determined within the 9-bit codebook. The latter case consists of the 4 modes UUU, VUU, UVU, and UUV whereby the 9 available bits are sufficient to represent the mode information and the pitch value.

Table 1. Joint quantization of pitch and U/V decisions.

U/V patterns	3-bit CB	9-bit CB
UUU	000	The pitch value is quantized with the same 99-level uniform quantizer as the 2.4kb/s standard. The pitch value and U/V pattern are then mapped to this 9-bit codebook.
UUV		
UVU		
VUU		
VVU	001	These U/V patterns share the same codebook containing 512 codevectors of the pitch triple.
VUV	010	
UVV	100	
VVV	011	512-level codebook A
	101	512-level codebook B
	110	512-level codebook C
	111	512-level codebook D

3. LSF QUANTIZATION AT 1.2 KB/S

For the UUU, UUV, UVU and VUU modes, the LSF vectors of unvoiced frames are quantized using a 9-bit codebook, while the LSF vector of the voiced frame is quantized with the same 25-bit multi-stage VQ (MSVQ) quantizer as in the MELP standard.

The LSF vectors for the other U/V patterns are encoded using a forward-backward interpolation scheme. This scheme works as follows. First the LSFs of the last frame in the current superframe, l_3, are quantized to $\hat{l}_3$ using a 9-bit codebook for the unvoiced case or the same 25-bit MSVQ codebook as in the MELP coder for the voiced case. Predicted values of l_1 and l_2 are then obtained by interpolating $\hat{l}_p$ and $\hat{l}_3$ as follows ($\hat{l}_p$ is the quantized LSFs of the last frame of the previous superframe):

$$\tilde{l}_1(j) = a_1(j)\hat{l}_p(j) + [1 - a_1(j)]\hat{l}_3(j)$$
$$\tilde{l}_2(j) = a_2(j)\hat{l}_p(j) + [1 - a_2(j)]\hat{l}_3(j) \qquad j = 1,\cdots,10$$

where $a_1(j)$ and $a_2(j)$ are the interpolation coefficients, and $\hat{l}_i(j)$ is the j-th component of $\hat{l}_i$. The interpolation coefficients are stored in a codebook of size 16, and the best set of the coefficients are selected by minimizing the distortion measure:

$$E = \sum_{j=1}^{10} w_1(j)\left|l_1(j) - \tilde{l}_1(j)\right|^2 + \sum_{j=1}^{10} w_2(j)\left|l_2(j) - \tilde{l}_2(j)\right|^2$$

where $w_i(j)$ are the weighting coefficients obtained with the same procedure as in the 2.4 kbps MELP standard. After obtaining the best interpolation coefficients, the residual LSF vector for frames 1 and 2 are computed by

$$r_1(j) = l_1(j) - \tilde{l}_1(j)$$
$$r_2(j) = l_2(j) - \tilde{l}_2(j) \qquad j = 1,\cdots,10.$$

The two residual vectors are concatenated and the resulting 20-dimension residual vector is encoded with a MSVQ quantizer having 14 bits (8+6) if the last frame was voiced, and 26 bits if l_3 corresponds to a unvoiced frame .

4. BIT ALLOCATION AT 1.2 KB/S

The bit allocation of the 1.2 kb/s coder is summarized in Table 2. Two gain parameters are calculated per frame, with 6 gains per superframe. The 6 gain parameters are vector-quantized in logarithmic domain using a 10-bit codebook.

The binary voicing decisions for 5 bands are obtained per frame. The bandpass information for the lowest band is determined from the U/V decision. The bandpass decisions of the remaining 4 bands are employed only for voiced frames and quantized with a 2-bit codebook.

Table 2. Bit allocation of 1.2 kbps MF-MELP coder for a superframe of 67.5 ms.

Parameters	U/V patterns of superframe				
	VVV	UVV VUV	VVU	UUV UVU VUU	UUU
Pitch & UV	12	12	12	12	12
LSFs	43	43	39	43	27
Gains	10	10	10	10	10
Bandpass Voicing	6	4	4	2	0
Fourier Magnitudes	8	8	8	8	0
Aperiodic Flag	1	1	1	1	0
Synchronization	1	1	1	1	1
Error Protection	0	2	6	4	31
Total	81	81	81	81	81

The Fourier magnitude vector is computed only for voiced frames. The vector of the last voiced frame in the current superframe is quantized with the same 8-bit quantizer as the MELP standard. The Fourier magnitude vectors for the other voiced frames are reconstructed using the quantized vectors of the current and previous superframes. The reconstruction procedure uses either an interpolation or a repetition method according to the U/V decisions.

The aperiodic flag is computed only from voiced frames and quantized using 1-bit per superframe with codebooks selected by the U/V pattern.

5. A HARMONIC SYNTHESIZER FOR 1.2/2.4 KB/S MELPe

The periodic and the noise excitations are combined in frequency domain and then converted to a time-domain signal of one pitch period in length using inverse Discrete Fourier Transform. The excitation spectrum is generated based on two parameters, the cutoff frequency F and the Fourier magnitude vector $M(k)$, $k=1,2..,L$. The cutoff frequency F is obtained from the quantized bandpass voicing strengths $Vbp_i, i = 2,3,4,5$, and then interpolated for each pitch cycle. F is set to be zero if the overall voicing Vbp_1 of the frame is set to be unvoiced. Otherwise the quantized voicing strengths $Vbp_i, i = 2,3,4,5$ are mapped into a cutoff frequency F according to the Table 3.

A jitter value is computed by multiplying the output of a uniform random number generator with range [-1, 1] by the interpolated jitter strength. A pitch period T, is computed as the interpolated pitch value plus the jitter. This pitch period is rounded to the nearest integer and clamped between 20 and 160.

Two transition frequencies F_H and F_L are then determined according to the cutoff frequency F employing an empirically derived algorithm. The transition frequencies are in the range $F_L \in [0.85F, 0.98F]$, $F_H \in [F, 1.05F]$.

Table 3. Cut-off frequency mapping.

F, cutoff frequency (Hz)	500	1000	2000
Voicing patterns $Vbp_i, i = 2,3,4,5$	0000	1000	1100
	0001	1001	
	0010	1010	
	0011		
	0100		
	0101		
	0110		

The transition frequencies correspond to two DFT frequency component indices V_L and V_H. A voiced model is used for all the frequency samples below V_L, a mixed model is used for frequency samples between V_L and V_H, and an unvoiced model is used for frequency samples above V_H. For the mixed mode, a gain factor g is selected with the value depending on the cutoff frequency (the higher is the cutoff frequency F, the smaller is the gain factor). The frequency components of the excitation are determined as follows,

$$|X(k)| = \begin{cases} M(k) & k < V_L \\ \frac{k - V_L}{V_H - V_L} \cdot g \cdot M(k) + \frac{V_H - k}{V_H - V_L} \cdot M(k) & V_L \le k \le V_H \\ g \cdot M(k) & k > V_H \end{cases}$$

with phases initialized by combining the linear phase with a random component [6]. The frequency samples of the mixed excitation are then converted to time domain using an IDFT.

6. REFERENCES

[1] A.V. McCree, T.P. and Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding", IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 4, pp. 242-250, July 1995.

[2] L.M. Supplee, R.P. Chon, J.S. Collura and A.V. McCree, "MELP: The New Federal Standard at 2400 bps," in Proc. ICASSP-97, vol. 2, pp. 1591-1594, 1997.

[3] R. Martin and R.V. Cox, "New Speech Enhancement Techniques for Low Bit Rate Speech Coding," in Proc. Speech Coding Workshop-99, pp. 165-167, 1999.

[4] D.P. Kemp, J.S. Collura, T.E. Tremain, "Multi-frame coding of LPC parameters at 600-800 bps" Proc. IEEE Inter. Conf. Acoustics, Speech and Signal Processing, vol.1, pp. 609-612, 1991.

[5] A.V. McCree and J.C. De Martin, "A 1.7 kb/s MELP coder with improved analysis and quantization" Proc. IEEE Inter. Conf. Acoustics, Speech and Signal Processing, pp. 593-596, 1998.

[6] T. Wang, K. Koishida, V. Cuperman, A. Gersho, J. Collura, "A 1.2 kbps Coder Based on MELP", Proc. IEEE Inter. Conf. Acoustics, Speech and Signal Processing, 2000.