NEW SPEECH ENCODING ALGORITHMS FOR ULTRA LOW BIT RATE AT 600/300 BPS

Jian Cong[1] *and Suo Cong*[2]

[1] The 30th Institute of CETC, 610041, China
[2] University of Zurich, Switzerland

ABSTRACT

We utilize the inter-frame redundancy with the larger-size super-frame structure to realize ultra low bit rate speech encoding. A new clustering model of speech characteristics is proposed to process effectively the parameters of large super-frames. Based on the model, we present algorithms for ultra low bit rate speech encoding at 600 bps and 300 bps for applications in acoustically harsh environments. At the decoder, a close-loop excitation signal magnitude estimation model is employed to improve the naturalness of synthesized speech. Two prototypes have been realized and evaluated using the DRT tests based on the national standard of China. Both prototypes are able to synthesize high quality of speech with DRT score 88.85 and 81.78 respectively.

1. INTRODUCTION

Although the bandwidth of wired/wireless communication channels has been increased dramatically, low rate bit speech coding is still required significantly by many applications esp. in acoustically harsh environments for anti-jamming communication. For instance, the objective of ASE (Advance Speech Encoding) program, supported by the DARPA, is to achieve ultra low bit rate (300 bps) speech coding with acceptable intelligibility and quality in acoustically harsh environments within the phase 2 of this program.

A great number of algorithms as well as speech coders, such as code-excited linear prediction (CELP) [1], mixed excitation linear prediction (MELP) [2], [3], multi-band excitation (MBE) [4], enhanced MBE [5], sinusoidal transform coder (STC) [6], waveform interpolation (WI) [7], and enhanced WI [8], have been proposed to optimize the process of determining a concise representation of speech signals for the purpose of transmission, storage, and reconstruction. The state-of-the-art coders manage to produce speech with sufficiently good quality at bit rates as low as 1200 bps. These coders split the speech signal into a series of frames of 20 to 30 ms, compute and extract the parameters characterizing the frames, and transmit the encoded parameters to the corresponding decoders. It is pragmatic to lower the bit rate by coding the set of parameters of a few successive frames structured as a *super-frame*. The NATO STANAG 4470 defines an 800 bps speech coder using the super-frame structure in which a super frame consists of 3 frames.

For the encoding algorithms based on linear prediction, the major bottleneck of reducing bit rate is the quantization of the linear predictive coding (LPC) filter coefficients. The multistage vector quantization (MSVQ) approach presented in [9] has achieved rather good performance at 21-25 bit per 20 ms frame. In addition, the multistage structure gains flexibility in terms of search complexity, codebook storage, and channel error protection. Based on the MSVQ approach, the residual MSVQ (R-MSVQ) approach is proposed in [10] and a 1200 bps coder using R-MSVQ is developed in [11].

We have developed two algorithms for ultra low bit rate (ULBR) speech coding at 600 bps and 300 bps respectively. A large super-frame structure is utilized in both algorithms. To improve the encoding efficiency of parameters, a clustering model of speech characteristics based on acoustic perception has been proposed to effectively reduce the amount of samples in the characteristic space due to large size of super-frames.

This paper is organized as follows. In section 2, we describe the clustering model of speech characteristics for large-size super-frames. In section 3, we present the multistage matrix quantization schemes for 600 bps and 300 bps respectively. The pitch diction and encoding is described in section 4. A closed-loop estimation model of the excitation signal is presented in section 5. Section 6 presents the schemes of bit allocation. The performed DRT experiments and results have been presented in section 7. Finally we conclude in section 8.

2. THE CLUSTERING MODEL

The source speech frames $f_i, i = 1,2,\ldots$, are aggregated and structured as a set of super-frames. A super-frame is then defined as $s_n = \{f_{n\cdot K}, f_{n\cdot K+1}, \ldots, f_{n\cdot K+K-1}\}$ in which K is the size of the super-frame.

We employ the line spectral pair (LSP) vector to represent the LPC filter. Therefore, the super-frame s_n can

be described by the matrix of LSP vectors $Mlsp_n$, the vector of pitch Vp_n, and the vector of energy parameters Ve_n:

$$Mlsp_n = \left\{ lsp_{n\cdot K}{}^t, lsp_{n\cdot K+1}{}^t, \ldots, lsp_{n\cdot K+K-1}{}^t \right\} \quad (1.a)$$

$$Vp_n = \left(pitch_{n\cdot K}, pitch_{n\cdot K+1}, \ldots, pitch_{n\cdot K+K-1} \right) \quad (1.b)$$

$$Ve_n = \left(ene_{n\cdot K}, ene_{n\cdot K+1}, \ldots, ene_{n\cdot K+K-1} \right) \quad (1.c)$$

In most cases, traditional clustering approaches fail to classify super-frames based on actual acoustic perception. Super-frames, which are perceptually similar, may be classified into different subspaces just because they are physically dissimilar. Therefore, the overall quantization performance is degraded.

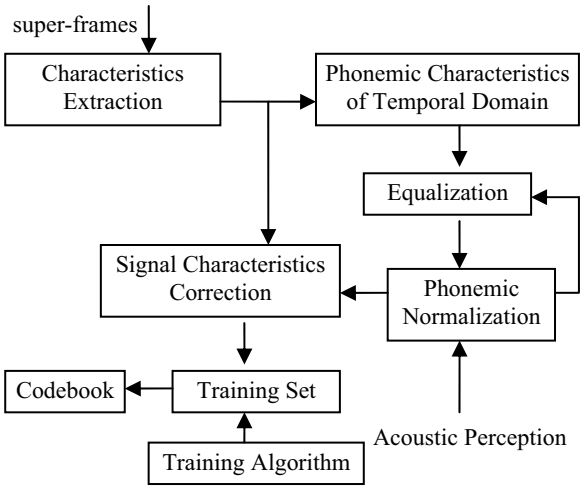

Figure 1: Clustering Model of Speech Characteristics

With the size of super-frame increases, the amount of samples in the characteristic space of the speech signal increases too quickly. Therefore, it is hardly to be handled properly by traditional encoding and clustering approaches. In order to acquire lower bit rate, the quality of the speech has to be sacrificed considerably.

In this paper, we propose a new clustering model of speech characteristics based on acoustic perception (figure 1) to reduce the amount of samples in the signal characteristic space significantly. This model is competent to classify effectively the parameters of larger-size super-frames (e.g. a 12x25ms super frame structure for 300 bps encoding).

We define the acoustic structure of super-frame asc_s as follows: n_s is the number of syllables in the super-frame s_n, uv_s^i, st_s^i, du_s^i represents the unvoiced/voiced characteristics, starting time, and the duration respectively.

$$asc_s = \left\{ n_s; \left\{ uv_s^i, st_s^i, du_s^i \right\}, i = 1,2,\ldots,n_s \right\} \quad (2)$$

A normalized acoustic structure characteristics table (NASCT) is defined to describe the acoustic structure of all samples. Given a super-frame, its *asc* vector is used to search a matched normalized acoustic structure vector $nasc_k = \left\{ n_k; \left\{ uv_k^i, st_k^i, du_k^i \right\}, i = 1,2,\ldots,n_k \right\}$ in the NASCT table based on (3):

$$dist(asc_s, nasc_k) = w_n \left| n_s - n_k \right| + \sum_{i=1}^{n_s} \left(w_{st}^0 \left| uv_s^i - uv_k^i \right| + w_{st}^1 \left(uv_s^i \right) \right) \left| st_s^i - st_k^i \right| + \sum_{i=1}^{n_s} \left(w_{du}^0 \left| uv_s^i - uv_k^i \right| + w_{du}^1 \left(uv_s^i \right) \right) \left| du_s^i - du_k^i \right| \quad (3)$$

$$w_n = 1.0, \quad w_{st}^0 = 0.2, \; w_{du}^0 = 0.25 \quad (4.a)$$

$$w_{st}^1(uv) = \begin{cases} uv = unvoice & 0.03 \\ uv = voice & 0.06 \end{cases}, \; w_{du}^1(uv) = \begin{cases} uv = unvoice & 0.05 \\ uv = voice & 0.10 \end{cases} \quad (4.b)$$

The matched *nasc* vector is used to correct the *asc* vector of the super-frame. And the corrected *asc* vector is used for generating the training set or encoding the speech signal.

In order to prevent the accumulative error caused by the acoustic correction, a feedback control algorithm is designed to compensate the correction based on the principle of minimize the impacts of acoustic perception.

3. MULTISTAGE MATRIX QUANTIZATION

In order to achieve bit rate lower than 600 bps, we utilize the redundancy between successive frames by introducing larger size super-frame structure. A 6x25ms super-frame structure is used in the 600 bps algorithm and a 12x25ms super-frame structure is used in the 300 bps algorithm.

3.1 600 bps: Residual 4-stage Matrix Quantization

The speech frames are classified as *basic frame* (B-frame) and *interpolation frame* (I-frame). Each super-frame consists of 2 B-frames and 4 I-frames. Because the voiced speech frames (V-frame) are more significant to acoustic perception and the LPC paramters of successive V-frames are highly correlated, V-frames, esp. the V-frames at the boundary to unvoiced frames (U-frame), have higher priorites during the selection of B-frames. The LPC parameters of B-frames are quantized usying a 4-stage MSVQ and different coding tables are used accroding to the voiced/unvoiced feature. The LSP vectors of I-frames are estimated based on neighboring B-frames using optimal 1st order linear prediction.

We denote two successive B-frames as f_n, f_{n+m+1}, and the frames between them $f_{n+1}, \ldots, f_{n+m}$ are I-frames. The corresponding LSP vectors are denoted as $\ldots lsp_n, lsp_{n+1}, \ldots, lsp_{n+m}, lsp_{n+m+1}, \ldots$. Then the estimation of the LSP vector of I-frames is obtained using (5), in which the *i=1,2, ... L* and *L* is the order of the LPC filter.

$$\underset{a_{k,m+1-k}^i, b_{k,m+1-k}^i}{\arg\min} \; E\left[\left(lsp_{n+k}[i] - \hat{lsp}_{n+k}[i] \right)^2 \right] \quad (5.a)$$

$$\hat{lsp}_{n+k}[i] = a_{k,m+1-k}^i \cdot lsp_n[i] + b_{k,m+1-k}^i \cdot lsp_{n+m+1}[i] \quad (5.b)$$

The coefficients $a^i_{k,m+1-k}$ and $b^i_{k,m+1-k}$ are given by (6), in which $r^i_k = E[lsp_{n+k}[i] lsp_n[i]]$

$$a^i_{k,m+1-k} = \left(r^i_0 r^i_k - r^i_{m+1-k} r^i_{m+1}\right) \Big/ \left(r^{i\,2}_0 - r^{i\,2}_{m+1}\right) \quad (6.a)$$

$$b^i_{k,m+1-k} = \left(r^i_0 r^i_{m+1-k} - r^i_k r^i_{m+1}\right) \Big/ \left(r^{i\,2}_0 - r^{i\,2}_{m+1}\right) \quad (6.b)$$

r^i_k, $a^i_{k,m+1-k}$, and $b^i_{k,m+1-k}$ are calculated based on training sets. Some estimation coefficients are given in table 1.

The residual vector is defined as $\Delta_{n+k} = lsp_{n+k} - \hat{lsp}_{n+k}$. A 4-stage matrix quantization of 7-7-6-6 bit is applied to the L x 4 matrix of the residual vectors of the 4 I-frames in one super-frame. The Generalized Lloyd Algorithm (GLA) is applied to this sequence to generate the codebook [12].

Table 1: Estimation Coefficients of LSP vectors of I-frames

i	1	2	3	4	5	6	7	8	9	10
$a^i_{1,2}$	0.65	0.66	0.73	0.60	0.70	0.76	0.70	0.64	0.61	0.63
$b^i_{1,2}$	0.35	0.34	0.27	0.40	0.30	0.24	0.30	0.36	0.39	0.37

3.2 300 bps: 7-stage Matrix Quantization

The speech signal characteristics $Mlsp_n$ is divided into 3 sub-spaces $Mlsp_n = \{Mlsp^0_n, Mlsp^1_n, Mlsp^2_n\}$ as defined in (7):

$$Mlsp^i_n = \{lsp_{n\cdot K+i}{}^t, lsp_{n\cdot K+3+i}{}^t, lsp_{n\cdot K+6+i}{}^t, \ldots\}, i = 0,1,2 \quad (7)$$

At stage 1, the sample space is classified according to the distribution of voiced/unvoiced frames in a super-frame and $Mlsp^2_n$ is matrix quantized. At stage 2 and 3, $Mlsp^0_n$ and $Mlsp^0_n$ are estimated using the quantization results at stage 1 based on interpolation and the voiced/unvoiced features. And the residual matrix of the estimation is quantized as in (8) in which Q^i denotes i-stage quantization and M_q denotes matrix quantization.

$$Q^1(Mlsp_n) = Mq\left(Mlsp^2_n\right) \quad (7.a)$$

$$Q^2(Mlsp_n) = Mq\left(Mlsp^0_n / Q^1(Mlsp_n)\right) \quad (7.b)$$

$$Q^3(Mlsp_n) = Mq\left(Mlsp^1_n / \left(Q^1(Mlsp_n), Q^2(Mlsp_n)\right)\right) \quad (7.c)$$

At stage 4, 5, 6, and 7, the residual matrix quantized in previous stage is quantized.

4. PITCH DETECTION AND ENCODING

The estimation of pitch value is an important component in a variety of speech processing systems such as speech analysis synthesis systems, and speech coding systems. A number of algorithms for pitch estimation have been developed, which may utilize the properties of speech signals in either time-domain or frequency-domain.

We apply a 2-stage vector quantization scheme for the pitch value and u/v flag of the frames in super-frames. The whole sample space is classified into several types according to the amount of V-frame in a super-frame and each typed sample space is classified further into a number of subspaces according to the distribution of u/v-frames. For each typed subspaces, a specific sub-codebook is produced and trained with corresponding training set.

The quantization rate at stage 1 is 10 bit /6 frames (for 600 bps) and 12 bit / 12 frames (for 300 bps). At stage 2, it is 4 bit/6 frames (600 bps) and 8 bit /12 frames (300 bps).

5. THE SPEECH SYNTHESIS MODEL

We develop a closed-loop estimation approach (figure 2) of the impulse magnitude at the decoder to improve the naturalness of speech and the articulation of nasals.

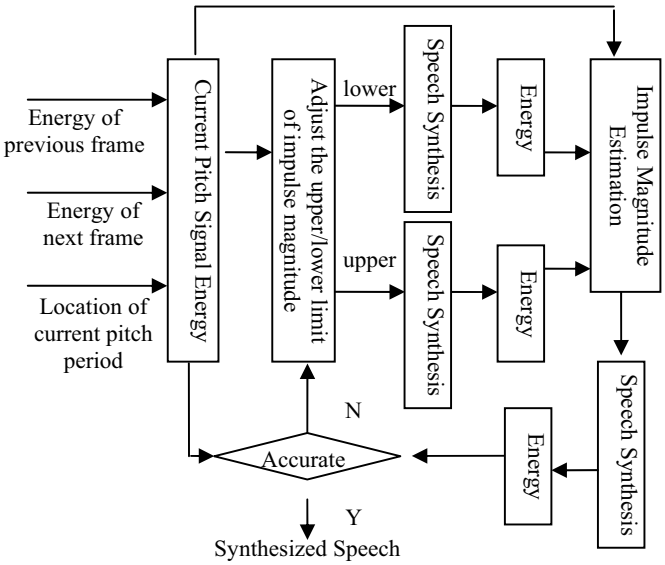

Figure 2: Closed-loop estimation approach of the impulse magnitude

The magnitude of the synthesized speech signal may be varied differently from the magnitude of the excitation impulses. And it may cause the perceptual distortion of the synthesized speech. For some voiced phonemes and nasal phonemes, the magnitude of the speech is very significant to the perception and intelligibility. The proposed approach of closed-loop estimation of excitation impulse magnitude is used in the decoder to improve the quality of the synthesized speech.

6. BIT ALLOCATION

The bit allocation for the 600 bps speech coder is given in table 2 and for the 300 bps coder is given in table 3.

Table 2: Bit allocation for 600 bps coder

LSP (6x25ms)	Pitch and U/V flag	Energy

		(6x25ms)	(6x25ms)
B-frame	I-frame	10＋4 bit	8 bit
21＋21 bit	7＋7＋6＋6 bit		
140＋140 bps	173.33 bps	93.33 bps	53.33 bps
90 bit / 6x25ms =140×2＋173.33＋93.33+53.33 ＝ 600 bps			

Table 3: Bit allocation for 300 bps coder

LSP(10x12)	Pitch and U/V flag (12x25ms)	Energy (12x25ms)
9＋9＋9＋9＋9＋9＋8 bit	12＋8 bit	8 bit
206.67 bps	66.67 bps	26.67 bps
90 bit / 12x25ms =206.67＋66.67＋26.67 ＝ 300 bps		

7. EXPERIMENTS AND RESULTS

We have performed a number of DRT tests using speech materials chosen from the speech signal database based on national standard of China GB/T 16532-1996. The speech is recorded in recording studio and the sample rate is 16 KHz and the quantization level is 16 bit.

Eleven listeners participated in the DRT tests of the 300 bps coder and 10 listeners participated in the DRT tests of the 600 bps coder. The test results are presented in table 4 and table 5.

Table 4a: DRT result of female speakers: 600 bps

Term	Voiced	Nasal	Aspirated	Grave-ness	Compact-ness	Sustention
Avg.	100.00	67.81	95.56	86.67	85.60	90.75
Std.	0.00	12.80	2.09	5.33	7.47	2.57
Total Score: **87.73**						

Table 4b: DRT result of male speakers: 600 bps

Term	Voiced	Nasal	Aspirated	Grave-ness	Compact-ness	Sustention
Avg.	100.00	81.48	97.04	76.54	84.82	96.30
Std.	0.00	5.24	3.26	5.88	5.91	2.79
Total Score: **89.36**						

Table 5a: DRT result of female speakers: 300 bps

Term	Voiced	Nasal	Aspirated	Grave-ness	Compact-ness	Sustention
Avg.	100.00	62.63	93.94	72.05	72.05	89.63
Std.	0.00	9.00	3.00	10.35	5.50	1.68
Total Score: **81.72**						

Table 5b: DRT result of male speakers: 300 bps

Term	Voiced	Nasal	Aspirated	Grave-ness	Compact-ness	Sustention
Avg.	99.26	68.89	95.96	66.33	72.39	88.22
Std.	1.11	6.14	2.60	5.51	6.72	4.56
Total Score: **81.84**						

The average DRT score of the 300 bps coder is (81.72+81.84)/2=81.78 and the average DRT score of the 600 bps coder is (89.36+87.73)/2=88.55. From the test results, the speech encoding algorithms proposed in this paper achieve better quality of processing the male speech.

8. SUMMARY

We have presented two algorithms for ultra low bit rate speech encoding and realized prototype coders to evaluate the performance and efficiency of the proposed approaches. A new clustering model of speech characteristics based on acoustic perception is developed to process the large size super-frame structure used in the algorithms. In order to improve the naturalness of synthesized speech and the articulation of nasals, we have proposed a closed-loop estimation approach to synthesizing speech signal based on the estimation of the magnitude of the excitation impulses.

Our future work will focus on enhancing the performance of the 300 bps encoding algorithm, improving the quality of synthesized speech (DRT score > 85), and developing pragmatic industrial implementations.

11. REFERENCES

[1] M. R. Schroeder and B. S. Atal, “Code-excited linear prediction (CELP): High-quality speech at very low bit rates”, In Proc. ICASSP, pp. 937-940, 1985.

[2] V. McCree and T. P. Barnwell, “A new mixed excitation LPC vocoder”, In Proc. ICASSP, pp. 593-596, 1991.

[3] T. Wang, K. Koishida, V. Cuperman, and et al. , “A 1200/2400 bps coding suite based on MELP”, Proc. of IEEE Workshop on Speech Coding, 2002.

[4] D. W. Griffin and J. S. Lim, “Multiband excitation vocoder”, IEEE Trans. Acoustics, Speech, Signal Processing, 36(8): 1223-1235, 1988.

[5] A. Das and A. Gersho, “Low rate multimode multiband spectral coding of speech”, International Journal of Speech Technology, 2(4): 317-327, 1999.

[6] R. J. McAulay and T. F. Quatieri, “Sinusoidal coding”, In Speech Coding Synthesis, W. B. Kleijn and K. K. Paliwal, Eds. The Netherlands: Elsevier, 1995, chap. 4, pp. 121-173.

[7] W. B. Kleijn, “Speech coding below 4 kb/s using waveform interpolation”, In Proc. GLOBECOM, vol. 3, pp. 1879-1883, 1991.

[8] O. Gottesman and A. Gersho, “Enhanced Waveform Interpolative Coding at Low Bit-Rate”, IEEE Trans. Speech Audio Processing, 9(8), 2001.

[9] W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud, and V. Cuperman, “Efficient search and design procedures for robust multi stage vector quantization of LPC parameters for 4 kbps speech coding”, IEEE Trans. Speech Audio Processing., 1(4), pp. 373-385, 1993.

[10] S. Nandkumar, K. Swaminathan, and U. Bhaskar, “Robust speech mode based LSF vector quantization for low bit rate speech coders”, In Proc. ICASSP, 1998.

[11] S.Ozaydin, B. Baykal, “A 1200 bps speech coder with LSP matrix quantization”, IEEE Int. Conf. on ASSP, 2001.

[12] Ch. Tsao and R. M. Gray, “Matrix quantizer design for LPC speech using the generalized Lloyd algorithm”, IEEE Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-33, No. 3, June, 1985.