World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:3, No:11, 2009

A High Quality Speech Coder at 600 bps

Yong Zhang, Ruimin Hu

***Abstract*—**This paper presents a vocoder to obtain high quality synthetic speech at 600 bps. To reduce the bit rate, the algorithm is based on a sinusoidally excited linear prediction model which extracts few coding parameters, and three consecutive frames are grouped into a superframe and jointly vector quantization is used to obtain high coding efficiency. The inter-frame redundancy is exploited with distinct quantization schemes for different unvoiced/voiced frame combinations in the superframe. Experimental results show that the quality of the proposed coder is better than that of 2.4kbps LPC10e and achieves approximately the same as that of 2.4kbps MELP and with high robustness.

***Keywords*—**Speech coding, Vector quantization, linear predicition, Mixed sinusoidal excitation

I. INTRODUCTION

HIGH quality speech coding at a very low bit rate, such as 300~1000bps, is one of the most important research areas[1,2]. It is applied widely in radio communications, secure voice, satellite communications and IP phone etc. With the development of digital communication, frequency resources are more and more deficient, especially in some radio channels, speech coding at low bit rate is required urgently.

For very low bit rate speech coding, the coder must compress the signal at the transmitter, and the decoder synthesize the signal having as closely as possible the original speech at the receiver, according to the coding parameters, certain speech production model, and perceptual rules. In the past decade, the dominant speech production model has been linear predictive coding (LPC), which is related to a simplified version of a basic filter-excitation concept of speech production. Typical 1200-2400bps speech coders are needed for direct LPC implementation, such as mixed excitation linear prediction (MELP) [3], waveform interpolation [4], sinusoidal coder (SC) [5]. However, for 300~1000 bps vocoder [6,7], there is still a challenge currently and it is the key to code speech parameters using a very limited bit rate.

In this paper, a 600 bps coder is proposed based on a sinusoidally excited linear prediction model. To reduce the bit rate of coder, the inter-frame redundancies of the parameters are exploited with a superframe structure, which is composed of three successive frames, and multi-frame joint vector quantization methods are used in the superframes. 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. Recently, a new method of coding LSFs was introduced in [8] which involves the use of a Gaussian mixture model (GMM) to parameterize the probability density function of the source and design the optimum block quantizers. In this paper, we propose a GMM-based block quantizer that operates on superframe using the weighted minimum mean squared error as distortion criterion, which leads to a significant improvement in the performance of the block quantizer.

The proposed coding algorithm is described in this paper and the results of subject tests are reported. These results show that the proposed speech coder still preserves acceptable speech quality with significant savings in bit rate.

Yong Zhang is with the National Engineering Research Center for multimedia software, Wuhan University, Wuhan, Hubei Province, China (phone: +86-27-87648233; e-mail: zhangyhb@vip.sohu.com).

Ruimin Hu is with the National Engineering Research Center for multimedia software, Wuhan University, Wuhan, Hubei Province, China (e-mail: hrm1964@public.hb.cn).

II. CODER DESCRIPTION

A block diagram of this coder is shown in Fig.1. In the Proposed coder, the transmitted parameters are extracted every 22.5 ms frame (or 180 samples of speech at a sampling rate of 8kHz). A superframe structure of length 67.5ms comprising three consecutive frames is adopted in the proposed coder.

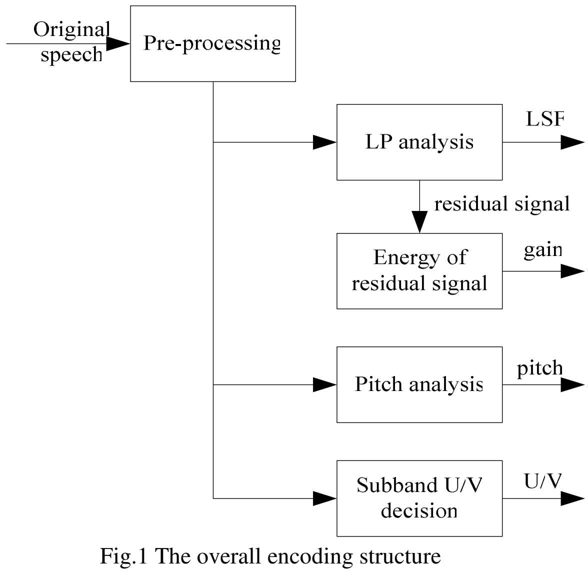

Fig.1 The overall encoding structure

The parameters for each frame in the superframe are jointly quantized to obtain high coding efficiency. The quantization schemes are designed so that the superframe structure is efficiently exploited by jointly vector quantization to reduce the inter-frame redundancy. The statistical properties of voiced (V) and unvoiced (U) speech are also taken into account. The bit allocation of the proposed coder is shown in Table 1, where a total of 45 bits is used per superframe.

A. Multi-Sub-band U/V Decision

The scheme of multi-sub-band is adopted for the vocoder in a

TABLE I
THE BIT ALLOCATION OF PROPOSED CODER

Parameters	Bits
Pitch	9
Gain	9
Band-pass Voicing	4
LSF	22
Synchronization	1
Total	45

frame. There are 5 sub-bands of 0~500Hz, 500~1000Hz, 1000~2000Hz, 2000~3000Hz, 3000~4000Hz in a frame just like in the MELP. The U/V state of each sub-band is estimated by minimizing the error between original speech spectra and reconstructed speech spectra in a frame.

Sub-band U/V patterns with larger probability distributions are listed in Table 2 according to static results using a database of 96.4MB in English and 93.6MB in Chinese, where 1 denotes voiced decision, and 0 for unvoiced decision. The U/V states in these five bands have orderliness in sequences of low sub-bands to higher sub-bands from left to right.

Sub-band U/V state transitions with larger probability distributions are described in Table 3, where Patterns transfer 0 denote pattern 00000, 1 denotes pattern 10000, 2 denotes

TABLE II
SUB-BAND U/V PATTERNS AND PROBALITY DISTRIBUTION

U/V Pattern	Probability (%)	U/V Pattern	Probability (%)
11111	29.6963	10111	0.8358
00000	29.0470	11010	0.6898
10000	15.6004	11001	0.6078
11110	6.5293	10110	0.4066
11000	6.0491	10010	0.3817
11100	5.8639	10001	0.3380
11101	1.7212	10011	0.2353
11011	0.9424	10101	0.1475
10100	0.9080		

pattern 11000, 3 denotes pattern 11100, 4 denotes pattern 11110, 5 denotes pattern 11111.

Table 2 and 3 demonstrate that the probability distributions of sub-band U/V pattern are not uniform, such that fewer patterns could be reserved to reduce the bitrate. In the coder, we denote a super-pattern as following that the band-pass U/V decisions parameters of three consecutive frames are grouped together into a vector, and the larger 16 probabilities distributions of super-pattern are reserved and use 4-bit codebook to quantize per superframe. The VQ algorithm uses the weighted Euclidean distance as the distortion measure:

$$d=\sum_{i=1}^{3}\sum_{j=1}^{5}w_j(b_{i,j}-\hat{b}_{i,j})^2 \tag{1}$$

Where i is the i-th frame of the current superframe, j is the j-th band-pass of the current frame, $b_{i,j}$ is the band-pass voice/unvoiced decision, $b_{i,j}=0$ or $b_{i,j}=1$, and $\hat{b}_{i,j}$ is the quantized value, and w_j is the weighted factor:

$$w=\{1.0,0.7,0.4,0.2,0.1\}$$

B. Multi-frame pitch quantization

The pitch information is transmitted only for voiced frames. Different pitch quantization schemes are used for different U/V combinations in the superframe. If the first sub-band U/V of a frame is unvoiced, then the frame is regarded as U pattern, else it is regarded as V pattern. For a superframe consists of 3 consecutive frames, there will be 8 U/V pattern in a superframe as shown in Table 4. The joint quantization scheme is summarized in Table 4:

From Table 4, it can be shown that within those superframes

TABLE IV
JOINT QUANTIZATION OF PITCH

U/V Pattern	9-bit codebook
UUU	No pitch information
UUV	The pitch value is quantized with 9-bit uniform quantizer.
UVU	
VUU	
VVU	512-level codebook A
VUV	512-level codebook B
UVV	512-level codebook C
VVV	512-level codebook D the super-pattern is 111111111111111
	512-level codebook E the super-pattern is 111101111011100
	512-level codebook F the other super-pattern

where the voicing pattern contains either two or three voiced frames, the pitch parameters are vector quantized. For voicing patterns containing only one voiced frame, the scalar quantizer used in the MELP standard [6] is applied for the pitch of the voiced frame. For UUU voicing pattern, no pitch information is transmitted, and the unused bits are used to the error protection.

The pitch values, P_i (i=1,2,3), calculated in 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. For voicing patterns with two or three voiced frames, the pitch vector is quantized using a VQ algorithm with a new distortion measure. This distortion measure incorporates pitch

TABLE III
PATTERN TRANSFER PROBABILITY DISTRIBUTION

Transfer Probability (%)		Patterns transfer					
		X-0	X-1	X-2	X-3	X-4	X-5
Patterns Transfer	0-X	21.679	4.189	1.043	0.695	0.385	1.057
	1-X	5.006	8.217	1.894	1.087	0.686	1.963
	2-X	0.986	2.255	1.893	1.035	0.602	1.518
	3-X	0.410	1.152	1.105	1.979	0.998	1.941
	4-X	0.259	0.760	0.664	0.960	1.592	2.295
	5-X	0.708	2.282	1.690	1.829	2.266	20.922

differentials into the codebook search, which makes it possible to consider the time evolution of the pitch. This feature is motivated by the perceptual importance of adequately tracking the pitch trajectory.

The pitch VQ algorithm has three steps for obtaining the best index:

Step1: Select the M-best candidates using the weighted squared Euclidean distance measure:

$$d = \sum_{i=1}^{3} w_i \left| p_i - \hat{p}_i \right|^2 \tag{2}$$

where the weighting coefficient is defined by:

$$w_i = \begin{cases} 1 & for \quad voiced \quad frame \\ 0 & for \quad unvoiced \quad frame \end{cases} \tag{3}$$

where p_i and $\hat{p}_i$ are the unquantized and quantized log pitch values, respectively. The above equation indicates that only voiced frames are taken into account in the codebook.

Step 2: Calculate the differentials of the unquantized log pitch values using

$$\Delta p_i = \begin{cases} p_i - p_{i-1} & if \quad i-th \quad and \quad (i-1)-th \quad frames \quad are \quad voiced \\ 0 & otherwise \end{cases} \tag{4}$$

for i = 1,2,3. where p_0 is the last log pitch value of the previous superframe. For the pitch candidates selected in step 1, calculate the quantized differentials by replacing Δp_i and p_i by $\hat{\Delta} p_i$ and $\hat{p}_i$ respectively in the equation above, where $\hat{p}_0$ is the quantized version of p_0.

Step 3: Select the optimum index from the M-best candidates that minimizes:

$$\begin{aligned} d' &= \sum_{i=1}^{3} w_i \left| p_i - \hat{p}_i \right|^2 + \delta \sum_{i=1}^{3} \left| \Delta p_i - \Delta \hat{p}_i \right|^2 \\ &= d + \delta \sum_{i=1}^{3} \left| \Delta p_i - \Delta \hat{p}_i \right|^2 \end{aligned} \tag{5}$$

where δ is a parameter to control the contribution of pitch differentials which is set to be 1 in the proposed coder.

C. Multi-frame LSF joint quantization

In this paper, we proposed a modified version of the fixed-rate GMM-based block quantizer that operates on superframe and uses the weighted minimum mean squared error as distortion criterion. This modified scheme exploits inter-frame correlation by concatenating 3 successive frames into a larger vector. Let the 10 dimension vector $lsf_n^1, lsf_n^2, lsf_n^3$ be LSF parameter vector of 1st frame, 2nd frame, 3rd frame in current superframe respectively. Then the LSF parameters vector of a superframe could be reordered:

$$\left[lsf_1^1, lsf_2^1, \cdots, lsf_{10}^1 \right]^T + \left[lsf_1^2, lsf_2^2, \cdots, lsf_{10}^2 \right]^T + \left[lsf_1^3, lsf_2^3, \cdots, lsf_{10}^3 \right]^T$$

$$\Rightarrow \left[lsf_1^1, lsf_2^1, lsf_3^1, lsf_1^2, lsf_2^2, lsf_3^2, \cdots, lsf_1^{10}, lsf_2^{10}, lsf_3^{10} \right]$$

The reordered vector with the dimension of 30 is then processed by the GMM-based block quantizer. In the following sections, we would provide a description of the training and encoding phase of the superframe GMM-based block quantizer.

1. Training Phase

The PDF model, as a mixture of multivariate Gaussians $N(x, \mu, \Sigma)$ can be given by:

$$G(X \mid M) = \sum_{i=1}^{m} c_i N_i \left(x; \mu_i; \Sigma_i \right) \tag{6}$$

$$N(x; \mu; \Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}} \left| \Sigma \right|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \tag{7}$$

$$M = \left[c_1, \cdots, c_m, \mu_1, \cdots, \mu_m, \Sigma_1, \cdots, \Sigma_m \right]$$

where m is the number of mixture components, and p is the dimension of the vectors. M is the set of model parameters consisting of $\{c_i, \mu_i, \Sigma_i\}$, which are the weight, mean, and covariance matrix of the ith mixture component respectively. Using the Expectation-Maximization (EM) algorithm, the maximum-likelihood estimation of the parametric model is computed iteratively until the log likelihood converges, where a final set of means, covariance matrices, and weights are produced.

An eigenvalue decomposition is calculated for each of the covariance matrices, producing m sets of the eigenvalue, $\{\lambda_i\}_{i=1}^{m}$, where $\lambda_i = \{\lambda_{i,j}\}_{j=1}^{N}$, and m sets of eigenvalue, $\{v_i\}_{i=1}^{m}$, where $v_i = \{v_{i,j}\}_{j=1}^{N}$. The ith set of eigenvalue form the rows of the orthogonal transformation matrix, P_i, which will be used for the K-L transform in the encoding phase.

2. Bit allocation

In the encoding phase of the superframe GMM-based block quantiser, the bit allocation is initially determined, given the fixed target bitrate, and vectors are then encoded using minimum distortion block quantization. If the target bitrate of the super-frame GMM-based block quantizer is b_{tot} bits, these bits need to be divided among the m cluster block quantizers. The number of bits, b_i, allocated to the block quantizer of cluster i, is given by:

$$2^{b_i} = 2^{b_{tot}} \frac{\left(c_i \Lambda_i\right)^{\frac{N}{N+2}}}{\sum_{k=1}^{m} \left(c_k \Lambda_k\right)^{\frac{N}{N+2}}} \quad i = 1, 2, \cdots, m \tag{8}$$

Where

$$\Lambda_i = \left(\prod_{j=1}^{N} \lambda_{i,j} \right)^{\frac{1}{N}} \qquad i = 1, 2, \cdots, m \tag{9}$$

N is the dimension of the vectors, $\lambda_{i,j}$ is the jth eigenvalue of the ith cluster. Then for each block quantizer, the high resolution formula form is used to distribute the b_i bits to each of the vector components:

$$b_{i,j} = \frac{b_i}{N} + \frac{1}{2} \log_2 \frac{\lambda_{i,j}}{\left(\prod_{j=1}^{N} \lambda_{i,j} \right)^{\frac{1}{N}}} \tag{10}$$

where $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, N$

3. Minimum distortion block quantization

To quantize a LSF vector, $\boldsymbol{x}$, using a particular cluster i, the cluster mean vector, $\boldsymbol{\mu}_i$ is first subtracted and its components decorrelated using the orthogonal matrix, $\boldsymbol{P}_i$, for that cluster. The variance of each component is then normalized to produce a decorrelated, mean-subtracted, and variance-normalized vector, z_i:

$$z_i = \frac{\boldsymbol{P}_i(\boldsymbol{x} - \boldsymbol{\mu}_i)}{\boldsymbol{\sigma}_i} \tag{11}$$

where $\boldsymbol{\sigma}_i = \boldsymbol{\lambda}_i^{1/2}$ is the standard deviation vector of ith cluster. They are then quantized [8] using a set of n Gaussian quantizer with its respective bit allocation $\left\{ b_{i,j} \right\}_{j=1}^{N}$. Then the indices, q_i, from the quantizer are decoded, multiplied by the standard deviation and correlated again by multiplying with the inverse of the orthogonal matrix $\boldsymbol{P}_i^{-1}$. The cluster mean is then added back to give the reconstructed vector $\hat{\boldsymbol{x}}$:

$$\hat{\boldsymbol{x}}_i = \boldsymbol{P}_i^{-1} \sigma_i \hat{z}_i + \boldsymbol{\mu}_i \tag{12}$$

The distortion between the reconstructed vector and the original is then calculated $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$. The kth cluster which gives the least distortion is chosen. The weighted minimum mean squared distortion, $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$, between the original vector $\boldsymbol{x}$ and the quantization vector $\hat{\boldsymbol{x}}$ is given by:

$$d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \sum_{i=1}^{3} \sum_{j=1}^{10} w_j^i (x_j^i - \hat{x}_j^i)^2 \tag{13}$$

$$w_j^i = \begin{cases} P(lsf_j^i)^{0.3} & i \in 1 \sim 8 \quad j \in 1 \sim 3 \\ 0.64 P(lsf_j^i)^{0.3} & i \in 9 \quad j \in 1 \sim 3 \\ 0.16 P(lsf_j^i)^{0.3} & i \in 10 \quad j \in 1 \sim 3 \end{cases} \tag{14}$$

$P(lsf_j^i)$ is the power spectrum of inverse linear prediction filter at the frequency lsf_i of frame j.

Table 5 shows the performance of the proposed GMM-based block quantizer of 16 and 32 clusters. It can achieve transparent quality approximately with 21 bits.

D. Energy quantization

Considering the inter-frame redundancy of energy parameter, vector quantization is suitable for their quantization. In order to prevent sensitivity to speech input level, we use a gain-shape method for the three gain values in each superframe. Firstly, the mean value of the three gain values, denoted as G_m is computed and then G_m is transformed into logarithmic value $g_m^{'} = \log G_m$. The logarithmic value $g_m^{'}$ is quantized to 6 bits using a 64-level uniform quantizer. Secondly, three gains are normalized by quantized $\hat{G}_m$ and formed into a vector. The normalized gain vector is quantized to 3 bits with full VQ using unweighted Euclidean distance.

TABLE V
SPECTRAL DISTORTION OF GMM-BASED BLOCK QUANTIZER

Cluster	bits	Avg.SD (dB)	2-4dB (%)	4dB (%)
16	21	1.088	2.11	0.01
	22	1.075	1.70	0.00
	23	0.98	0.90	0.00
32	21	1.024	1.48	0.00
	22	0.965	1.16	0.00
	23	0.925	0.75	0.00

III. Decoder Description

The decoder follows a reverse process of the encoder and is shown in fig.2. The 4 bits allocated to quantize voicing patterns are first decoded, and the corresponding U/V pattern is then determined. Similarly to the encoding phase, the LSF

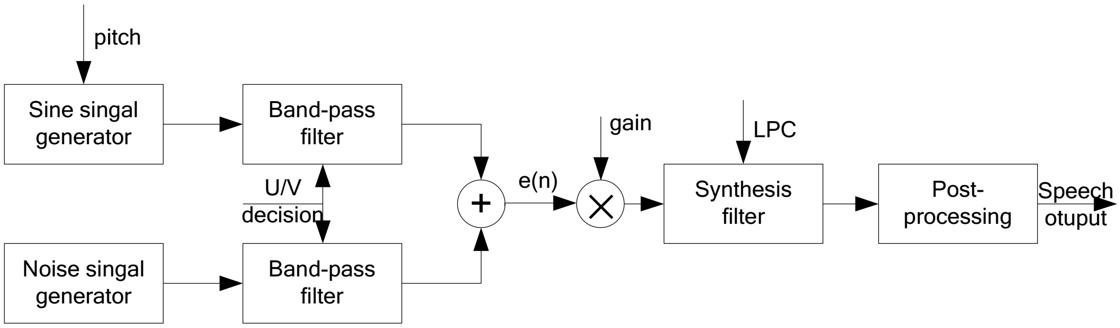

Fig. 2 The overall decoding structure

coefficients, gain coefficients, pitch are decoded.

The multi-sub-band mixed excitation is applied to improve the quality of synthetic speech. In the decoding end, the pitch and the sub-band U/V pattern will be used to generate a mixed excitation signal. The excitation components are generated by filtered harmonics of fundamental frequency located in voiced sub-bands and normalized random noise spectra in unvoiced sub-bands according to the U/V pattern a frame by a frame. The excitation signal e(n) is produced with the following equation:

$$e(n)=\sum_{i=1}^{5}\{\sum_{k=1}^{K}A(k,i)\cos[\omega_0 kn+\varphi(n,k)]+B(i)noise(n)\}*h_i(l) \quad (15)$$

$$A(k,i)=\begin{cases}A(k) & while \quad sub-band \quad i \quad is \quad voiced\\ 0 & while \quad sub-band \quad i \quad is \quad unvoiced\end{cases} \quad (16)$$

Where ω_0 is the fundmental frequency, and A(k) is the harmonic amplitude, k=1,…,K, K is the number of harmonics within 0~4000Hz. B(i) is the mark of unvoiced sub-band.

$$B(i)=\begin{cases}B & while \quad sub-band \quad i \quad is \quad unvoiced\\ 0 & while \quad sub-band \quad i \quad is \quad voiced\end{cases} \quad (17)$$

Where B is the noise amplitude. Phase $\varphi(k,n)$ will be reconstructed at the receiver under the construction of phase continuity. The $h_i(l)$ is the impulse response of the sub-band filter i. Then the e(n) will pass the LP synthesis filter to produce reconstructed speech signals.

IV. TEST RESULT

The Diagnostic Rhyme Test (DRT) is used to measure speech intelligibility. For comparison purpose, the 2.4kbps MELP standard coder [9] and LPC10e [10] were used. The coders were tested on speech containing quite, office and car background, and 1% random bit error channel. All of the coders scored higher for male talkers than female talkers, and the average results of male and female are shown in Table 6. From the result, It can be seen that the quality of the proposed coder is found better than that of LPC10e and achieves approximately the same as that of 2.4kbps MELP standard in informal tests with high robustness.

V. CONCLUSION

This paper described a very low bit rate vocoder at 600 bit/s and the important aspects of the algorithm were described. The algorithm was based on a sinusoidally excited linear prediction model, and the parameters of three frames were quantized together. Efficient vector quantization scheme was employed depending on the different U/V decision for the superframe, taking into account of statistical properties of voiced and unvoiced speech. A GMM-based LSF parameters block quantizer that operated on superframe was proposed, which could achieve transparent quality approximately with 21 bits. Experimental results demonstrated that this vocoder could obtain synthetic speech of high intelligibility with high naturalness as well as robustness.

TABLE VI
SUBJECTIVE INTELLIGIBILITY TEST

Coder	Quite	office	car	1% BRE
LPC10e 2400bps	87.34	84.61	80.36	85.06
MELP 2400bps	93.27	90.71	86.57	91.76
Propose coder 600bps	91.15	87.68	84.22	86.38

REFERENCES

[1] Ovens, M.J..Ponting, Turner.M.E, "Ultra low bit rate voice coding," *IEE Seminar*, Vol.4, pp 911 – 920, 2000

[2] Gwenael guilmin, Francois Capman, and et.al, "New NATO STANAG narrow band voice coder at 600 bit/s", *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol.3, pp.689-692, 2006

[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,* Vol.1, pp. 122-126, 2002

[4] O.Gottesman, A.Gersho, "Enhanced Waveform Interpolative Coding at Low Bit-rate", *IEEE Trans.Speech Audio Processing*, vol.9, No.8, pp.242-250, 2001

[5] Minoru Kohata, "A New 1.2kbit/s speech coding method based on a sinusoidal harmonic vocoder," *Systems and Computers in Japan*, vol.31, No.14, pp.64-73, 2000

[6] Jian Cong, Suo Cong, "New speech encoding algorithm for ultra low bit rate at 600/300," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol.2, pp.709-712, 2006

[7] Ehsan Jahangiri, Shahrokh Ghaemmaghami, “Scalable speech coding at rates below 900 bps”, *IEEE International Conference on Multimedia & Expo,* Vol.1, pp.85-88, 2008
[8] A.D.Subramaniam, B.D.Rao, “PDF Optimized Parametric Vector Quantization of Speech Line Spectral Frequencies,” *IEEE Trans. Speech Audio Processing,* Vol. 11, No. 2, pp. 130–142, Mar. 2003.
[9] L.M. Supplee, R.P.Cohn, J.S.Collura, A.V.McCree, “MELP: The new federal standard at 2400 bits/s,” *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol.4, pp.1591-1954, 1997
[10] Thomas E.Tremain, “The Government Standard Linear Predictive Coding Algorithm: LPC-10,” *Speech Technology*, No.2, pp.40-49, 1982

Yong Zhang was born in 1980, and received the B.Eng. degree from Wuhan University (first class honors), Wuhan, China, and the M.eng. degree from Wuhan University in 2003 and 2005, respectively. Now He is a Ph.D candidate of National Engineering Research Center for Multimedia software, Wuhan University, China. His main research interests include speech and audio signal processing, data compression and digital signal processing.

Ruimin Hu is a Full Professor, the dean of the National Engineering Research Center for Multimedia software , Wuhan University. He received the Ph.D. degree in Communication and Information System from Huazhong University of Science and Technology in 1994, and the Master Degree and Bachelor Degree in Communication and Information System from Nanjing University of Posts & Telecommunications in 1984 and 1990. His research interests include multimedia signal processing, multimedia communication system theory and application, pattern recognition, QoS over heterogeneous network, etc.