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 prediction, 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.

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](image)

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 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 denotes pattern 00000, 1 denotes pattern 10000, 2 denotes 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:

$$d = \sum_{i=1}^{3} \sum_{j=1}^{5} w_j (b_{i,j} - \hat{b}_{i,j})^2$$  \hspace{1cm} (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, $\hat{b}_{i,j} = 0$ or $\hat{b}_{i,j} = 1$, 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 the super-pattern is 111111111111111 512-level codebook E the other super-pattern 512-level codebook F

TABLE II

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

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
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:

Step 1: Select the M-best candidates using the weighted squared Euclidean distance measure:
\[
d = \sum_{i=1}^{3} w_i |p_i - \hat{\lambda}_i| \quad (2)
\]
where the weighting coefficient is defined by:
\[
w_i = \begin{cases} 
1 & \text{for voiced frame} \\
0 & \text{for unvoiced frame} 
\end{cases} \quad (3)
\]

Step 2: Calculate the differentials of the unquantized log pitch values using
\[
\Delta p_i = \begin{cases} 
p_i - p_{i-1} & \text{if } i-th \text{ and } (i-1)-th \text{ frames are voiced} \\
0 & \text{otherwise} 
\end{cases}
\]
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 \(\Delta p_i\) and \(p_i\) by \(\hat{\Delta} p_i\) and \(\hat{\lambda}_i\) respectively in the equation above, where \(\hat{\lambda}_0\) is the quantized version of \(p_0\).

Step 3: Select the optimum index from the M-best candidates that minimizes:
\[
d' = \sum_{i=1}^{3} w_i |\hat{\lambda}_i - \hat{\lambda}_{i-1}|^2 + \delta \sum_{i=1}^{3} |\Delta p_i - \hat{\Delta} p_i|^2
\]
where \(\delta\) 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 \(lsf_{i-1}^{3}, lsf_{i-2}^{2}, lsf_{i-3}^{2}\) 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:
\[
[lfsf_{i-1}^{3}, lfsf_{i-2}^{2}, lfsf_{i-3}^{2}, lfsf_{i-1}^{2}, lfsf_{i-2}^{2}, lfsf_{i-3}^{2}, lfsf_{i-1}^{3}, lfsf_{i-2}^{3}, lfsf_{i-3}^{3}]
\]

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:
\[
G(X \mid M) = \sum_{i=1}^{m} c_i N_i(x, \mu_i; \Sigma_i) \quad (6)
\]
\[
N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (7)
\]
\[
M = [c_1, \ldots, c_m, \mu_1, \ldots, \mu_m, \Sigma_1, \ldots, \Sigma_m]
\]
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, \(\{\lambda_j^{(i)}\}_{j=1}^{m}\), where \(\lambda_i = \{\lambda_j^{(i)}\}_{j=1}^{m}\). And \(m\) sets of eigenvalue, \(\{v_j^{(i)}\}_{j=1}^{m}\), where \(v_i = \{v_j^{(i)}\}_{j=1}^{N}\). 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.

<table>
<thead>
<tr>
<th>Probability (%)</th>
<th>Patterns transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-X</td>
<td>21.679</td>
</tr>
<tr>
<td>1-X</td>
<td>5.006</td>
</tr>
<tr>
<td>2-X</td>
<td>0.986</td>
</tr>
<tr>
<td>3-X</td>
<td>0.410</td>
</tr>
<tr>
<td>4-X</td>
<td>0.259</td>
</tr>
<tr>
<td>5-X</td>
<td>0.708</td>
</tr>
<tr>
<td>X-0</td>
<td>4.189</td>
</tr>
<tr>
<td>X-1</td>
<td>8.217</td>
</tr>
<tr>
<td>X-2</td>
<td>2.255</td>
</tr>
<tr>
<td>X-3</td>
<td>1.152</td>
</tr>
<tr>
<td>X-4</td>
<td>0.760</td>
</tr>
<tr>
<td>X-5</td>
<td>2.282</td>
</tr>
</tbody>
</table>

### Table III

**PATTERN TRANSFER PROBABILITY DISTRIBUTION**
2. Bit allocation

In the encoding phase of the superframe GMM-based block quantizer, 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 \( b_{\text{tot}} \) 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:

\[
2^b_i = 2^{b_{\text{tot}}/m} \left( \frac{(c_i \lambda_i)}{N^2} \right) \quad i = 1, 2, \ldots, m \tag{8}
\]

Where

\[
\Lambda_i = \left( \prod_{j=1}^N \lambda_{ij} \right)^{1/N} \quad i = 1, 2, \ldots, m \tag{9}
\]

\( N \) is the dimension of the vectors, \( \lambda_{ij} \) is the \( j \)th eigenvalue of the \( i \)th 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:

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

where \( i = 1, 2, \ldots, m \), \( j = 1, 2, \ldots, N \).

3. Minimum distortion block quantization

To quantize a LSF vector, \( \mathbf{x} \), using a particular cluster \( i \), the cluster mean vector, \( \mathbf{\mu} \), is first subtracted and its components decorrelated using the orthogonal matrix, \( \mathbf{P}_i \), for that cluster. The variance of each component is then normalized to produce the super-frame GMM-based block quantizer is minimum distortion block quantization. If the target bitrate of the super-frame GMM-based block quantizer is \( b_{\text{tot}} \) 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:

\[
2^b_i = 2^{b_{\text{tot}}/m} \left( \frac{(c_i \lambda_i)}{N^2} \right) \quad i = 1, 2, \ldots, m \tag{8}
\]

Where

\[
\Lambda_i = \left( \prod_{j=1}^N \lambda_{ij} \right)^{1/N} \quad i = 1, 2, \ldots, m \tag{9}
\]

\( N \) is the dimension of the vectors, \( \lambda_{ij} \) is the \( j \)th eigenvalue of the \( i \)th 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:

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

where \( i = 1, 2, \ldots, m \), \( j = 1, 2, \ldots, N \).

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

\[
d(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{i=1}^{N} \sum_{j=1}^{N} w_j (x_{ij} - \hat{x}_{ij})^2 \tag{13}
\]

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

\( P(lsf_j) \) is the power spectrum of inverse linear prediction filter at the frequency \( lsf_j \) 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

<table>
<thead>
<tr>
<th>Cluster</th>
<th>bits</th>
<th>Avg.SD (dB)</th>
<th>2-4dB (%)</th>
<th>4dB (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>21</td>
<td>1.088</td>
<td>2.11</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>1.075</td>
<td>1.70</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>0.98</td>
<td>0.90</td>
<td>0.00</td>
</tr>
<tr>
<td>32</td>
<td>21</td>
<td>1.024</td>
<td>1.48</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>0.965</td>
<td>1.16</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>0.925</td>
<td>0.75</td>
<td>0.00</td>
</tr>
</tbody>
</table>

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
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 frame. The excitation signal \( e(n) \) is produced with the following equation:

\[
e(n) = \sum_{k=1}^{K} \left[ A(k) \cos(\omega_k n + \phi(n,k)) + B(i) \text{noise}(n) \right] * h(l)
\]

Where \( \omega_k \) is the fundamental 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.

\[
A(k,i) = \begin{cases} A(k) & \text{while sub-band } i \text{ is voiced} \\ 0 & \text{while sub-band } i \text{ is unvoiced} \end{cases}
\]

(15)

\[
B(i) = \begin{cases} B & \text{while sub-band } i \text{ is unvoiced} \\ 0 & \text{while sub-band } i \text{ is voiced} \end{cases}
\]

(16)

Where \( B \) is the noise amplitude. Phase \( \phi(k,n) \) will be reconstructed at the receiver under the construction of phase continuity. The \( h(l) \) 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.

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

<table>
<thead>
<tr>
<th>Coder</th>
<th>Quite</th>
<th>Office</th>
<th>Car</th>
<th>1% BRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC10e 2400bps</td>
<td>87.34</td>
<td>84.61</td>
<td>80.36</td>
<td>85.06</td>
</tr>
<tr>
<td>MELP 2400bps</td>
<td>93.27</td>
<td>90.71</td>
<td>86.57</td>
<td>91.76</td>
</tr>
<tr>
<td>Propose coder</td>
<td>91.15</td>
<td>87.68</td>
<td>84.22</td>
<td>86.38</td>
</tr>
<tr>
<td>600bps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

REFERENCES


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.