CODING SPEECH AT VERY LOW RATES USING STRAIGHT AND TEMPORAL DECOMPOSITION

Phu Chien Nguyen, Takao Ochi and Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
E-mail: {chien, t-ochi, akagi}@jaist.ac.jp

ABSTRACT

This paper presents a new method for speech coding at rates around 1.2 kbps based on STRAIGHT, a high quality speech analysis-synthesis method. For encoding spectral information, Modified Restricted Temporal Decomposition (MRTD) based vector quantization is used, where MRTD is a method of temporal decomposition for line spectral frequency parameters. Meanwhile, pitch and gain parameters are coded using linear and spline interpolation, respectively. Subjective test results indicate that the performance of the proposed speech coding method is close to that of the 4.8 kbps US Federal Standard (FS-1016) CELP coder.

1. INTRODUCTION

Temporal decomposition (TD) [1, 2, 4, 6, 7] is an efficient technique of modeling speech parameter trajectories in terms of a sequence of event vectors and an associated sequence of event functions. TD can also be considered as an effective method of decorrelating the inherent inter-frame correlation presented in any frame based parametric representation of speech, and therefore is versatile for low bit rate speech coding systems. In previous work with TD [7], we have proposed a method of temporal decomposition that is capable of decomposing the line spectral frequency (LSF) parameters. The proposed method, termed Modified Restricted Temporal Decomposition (MRTD), can be used for efficient coding of spectral information of speech.

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of veiGHTed spectrum) has been proposed by Kawahara et al., which is a high quality vocoder type algorithm [3]. STRAIGHT decomposes a speech waveform into a spectral envelope, i.e. amplitude spectrum, pitch information and noise ratios. Those parameters and the maximum value of amplitude are required for re-synthesizing speech.

To make STRAIGHT applicable to very low rate speech coding, the bit rate required to represent the spectral envelope must be minimized. Since the spectral envelope can be further analyzed into spectral parameters and gain information, TD can be incorporated with STRAIGHT to create a high quality speech coder working at very low rates.

In this paper, we propose a low bit rate speech coding method based on STRAIGHT and TD. The proposed speech encoder and decoder block diagrams are shown in Fig. 1 and Fig. 2, respectively. The paper is organized as follows. The next section presents STRAIGHT and the derivation of LSF parameters from the amplitude spectrum. Section 3 briefly describes the MRTD of LSF parameters and Section 4 shows the determination of LSFs’ order. The way of coding spectral information of speech based on MRTD is presented in Section 5, while pitch, gain and noise ratio quantization techniques are described in Section 6. Section 7 summarizes the bit allocation for all parameters and an evaluation of the proposed speech coding method is presented in Section 8.

2. STRAIGHT

2.1. STRAIGHT

STRAIGHT [3] is a high quality analysis-synthesis method, which uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region, and an excitation source design based on phase manipulation. This method extracts F0 (fundamental frequency) by using TEMPO (Time domain Excitation extractor using Minimum Perturbation Operator), and the extracted F0 information is used to control the spectral envelope extraction procedure.
2.2. Derivation of LSF parameters

The amplitude spectrum \( X[k] \), where \( 0 \leq k \leq N/2 \) with \( N \) is the number of samples in the frequency domain, obtained from STRAIGHT analysis is transformed to the power spectrum:

\[
S[k] = |X[k]|^2, \quad 0 \leq k \leq N/2
\]  

(1)

The \( n^{th} \) autocorrelation coefficient, \( R[n] \), is then calculated using the inverse Fourier transform of the power spectrum.

\[
R[n] = \frac{1}{N} \sum_{k=0}^{N-1} S[k] \exp\left\{\frac{2\pi j kn}{N}\right\}, \quad 0 \leq n \leq N - 1
\]  

(2)

where \( S[k] = S[N-k] \). Assume that the speech sample \( x(n) \) can be estimated by a L-th order all-pole model, where \( 0 < L < N \), the reconstruction error is calculated as follows.

\[
P_L = R[0] - \sum_{l=1}^{L} a_l^2 R[l]
\]  

(3)

where \( \{a_l^2\} \), \( l = 1, 2 \cdots L \), are the corresponding linear predictive coding (LPC) coefficients, \( P_L \) refers to gain in this work. By minimizing \( P_L \) with respect to \( \{a_l^2\} \), where \( l = 1, 2 \cdots L \), \( \{a_l^2\} \) could be evaluated. They are then transformed to LSF parameters since LSF is known to be the best parametric representation of speech in terms of both interpolation and quantization properties.

3. TEMPORAL DECOMPOSITION

Temporal decomposition initiated by Atal [1] involves the decomposition of a sequence of spectral parameter vectors, i.e. LPC parameters, into a series of overlapping event functions and an associated series of event vectors as given in Equation (4).

\[
\hat{y}(n) = \sum_{k=1}^{K} a_k \phi_k(n), \quad 1 \leq n \leq N
\]  

(4)

where, \( a_k \) and \( \phi_k(n) \) are the \( k^{th} \) event vector and \( k^{th} \) event function, respectively. \( \hat{y}(n) \) is the approximation of \( y(n) \), the \( n^{th} \) spectral parameter vector, produced by the TD model. \( N \) and \( K \) are the number of frames and number of events in the block of spectral parameters under consideration, respectively.

The restricted second order TD model was utilized in [2, 4, 7], where only two adjacent event functions overlap and all event functions at any time sum up to one. The argument for imposing this constraint on the event functions can be found in [2, 7]. Equation (4) is rewritten as

\[
\hat{y}(n) = a_k \phi_k(n) + a_{k+1} (1 - \phi_k(n)), \quad n_k \leq n < n_{k+1}
\]  

(5)

where, \( n_k \) and \( n_{k+1} \) are the central positions of event \( k \) and event \( k + 1 \), respectively.

To order to apply TD to decomposing LSF parameters, the stability of the corresponding LPC synthesis filter after spectral transformation performed by TD must be ensured. The restricted temporal decomposition (RTD) method [4] intends to make LSF parameters possible for TD by enforcing the LSF ordering property on the event vectors. However, RTD has not completely solved this problem as indicated in [7]. Moreover, some event functions derived from the RTD method are ill-shaped, i.e. they have more than one peak, which is undesirable from speech coding point of view. Thus, the modified RTD (MRTD) method [7] has been proposed to overcome the drawbacks imposed on the RTD method. The reader is referred to [4, 7] for a detailed mathematical treatment.

4. DETERMINATION OF LSFs’ ORDER

4.1. Spectral distortion vs. LSFs’ order

Log spectral distortion [8] was used as the objective measure of performance to determine the suitable order of LSFs. A set of 112 phoneme balanced sentences uttered by speaker MMY of the ATR Japanese speech database were used as the speech data. This speech data set were re-sampled at 8 kHz sampling frequency and then STRAIGHT analyzed. In the following, the spectral envelopes obtained from STRAIGHT analysis were transformed to LSF parameters using the procedure described in Section 2 and those LSF parameters were TD analyzed by the MRTD method.

Spectral distortion results obtained from STRAIGHT analysis were transformed to LSF parameters using the procedure described in Section 2 and those LSF parameters were TD analyzed by the MRTD method. The restricted second order TD model was utilized in [2, 4, 7], where only two adjacent event functions overlap and all event functions at any time sum up to one. The argument for imposing this constraint on the event functions can be found in [2, 7]. Equation (4) is rewritten as

\[
\hat{y}(n) = a_k \phi_k(n) + a_{k+1} (1 - \phi_k(n)), \quad n_k \leq n < n_{k+1}
\]  

(5)

where, \( n_k \) and \( n_{k+1} \) are the central positions of event \( k \) and event \( k + 1 \), respectively.

In order to apply TD to decomposing LSF parameters, the stability of the corresponding LPC synthesis filter after spectral transformation performed by TD must be ensured. The restricted temporal decomposition (RTD) method [4] intends to make LSF parameters possible for TD by enforcing the LSF ordering property on the event vectors. However, RTD has not completely solved this problem as indicated in [7]. Moreover, some event functions derived from the RTD method are ill-shaped, i.e. they have more than one peak, which is undesirable from speech coding point of view. Thus, the modified RTD (MRTD) method [7] has been proposed to overcome the drawbacks imposed on the RTD method. The reader is referred to [4, 7] for a detailed mathematical treatment.

![Fig. 3. Spectral distortion vs. the order of LSFs.](image-url)
distortion is small  distortion is large

Fig. 4. Speech quality vs. the order of LSFs.

listener was asked to grade from −2 to 2 the degradation perceived in speech quality when comparing the second stimulus to the first, in each pair. Two phoneme balanced sentences uttered by speaker MMY of the ATR Japanese speech database were used as the testing data. These utterances were re-sampled at 8 kHz sampling frequency. Next, they were STRAIGHT analyzed and the resulting spectral envelopes were then transformed to LSF parameters of orders 10, 14, 18, 22, 26 and 30. In the following, the LSF parameters were analyzed into event vectors and event functions by the MRTD method. Those event vectors and event functions were used to re-synthesize speech.

Fig. 4 shows the results of the listening experiment. In this figure, the horizontal axis indicates the distribution and the positions show the relative distances of stimuli. Here, the positive values mean that the distortion is small while the negative values indicate the high distortion. The number on each arrow is the order of LSFs. Results also show that an increase of distortion is not easily realized when the order of LSFs exceeds 22.

For the above reasons, the order of LSFs was finally determined as 22.

5. MRTD BASED VECTOR QUANTIZATION OF LSF PARAMETERS

The reason for interpolating the vector trajectory of LSF parameters by using MRTD is that the updating rate of events is much lower than that of LSF parameters, and both event vectors and event functions can be quantized efficiently. In other words, LSF parameters can be quantized efficiently by transforming them into the event sequences first, and then quantizing event vectors and event functions.

5.1. Vector quantization of event vectors

Since the event vectors obtained from the MRTD method are valid LSF parameter vectors [7], they can be quantized by usual quantization methods for LSF parameters. Here, the split vector quantization method [8] was adopted. Due to the distribution of LSFs, the event vectors were divided into three subvectors of dimensions 7, 7, 8 and each subvector was quantized independently. We assigned 8 or 9 bits to each subvector, resulting in the number of bits allocated to one event vector was 24 or 27, respectively.

5.2. Vector quantization of event functions

In the case of event functions, normalization of the event functions is necessary to fix the dimension of the event function vector space. Notice that only quantizing φk(n) in the interval [n, n + 1] is enough to reconstruct the whole event function ˆφk(n). Moreover, φk(n) always starts from one and goes down to zero in that interval, and the type of decrease (after normalizing the length of φk(n)) can be vector quantized. In this work, we took 10 equidistant samples from an event function for length-normalization. Considering that all intervals between two consecutive central positions are less than 256 frames long (note that the frame period used in STRAIGHT analysis is 1 ms long), we used 8 bits for quantizing the length of each event function. Shortly speaking, each φk(n) was quantized by its length L(k) = nk+1 − nk and the type of decrease.

6. CODING EXCITATION PARAMETERS

6.1. Coding pitch parameters

For encoding pitch information estimated by STRAIGHT, the lengths of unvoiced and voiced segments were scalar quantized first. Next, linear interpolation was used within the unvoiced segments to form a continuous pitch contour. The continuous pitch contour was then re-sampled at 28 ms intervals. Logarithmic quantization was performed with 5 bits for each sampled value. In the decoder, pitch values of unvoiced intervals were set to zero. Meanwhile, pitch values of voiced intervals were reconstructed from the quantized samples using the linear interpolation. The root mean square (RMS) pitch error was found to be about 3.7 Hz for the speech data set used in Section 4.1.

6.2. Coding gain parameters

The gain contour was re-sampled at 20 ms intervals. Logarithmic quantization was performed with 6 bits for each sampled value. The quantized samples and the spline interpolation were used in the decoder to form the reconstructed gain contour. The RMS gain error was found to be about 3.5 dB for the speech data set used in Section 4.1.

6.3. Coding noise ratio parameters

The noise ratio parameters were estimated from the noise ratio targets and the event functions as follows.

\[ \hat{i}(n) = \sum_{k=1}^{K} i_k \phi_k(n), \quad 1 \leq n \leq N \]

where, \( \hat{i}(n) \) and \( i_k \) are the reconstructed noise ratio parameter for the \( n^{th} \) frame and the \( k^{th} \) noise ratio target, respectively. The noise ratio targets were determined by minimizing the sum squared error, \( E_i \), between the original and the interpolated noise ratio parameters.

\[ E_i = \sum_{n=1}^{N} \left( i(n) - i_k \right)^2 = \sum_{n=1}^{N} \left( i(n) - \sum_{k=1}^{K} i_k \phi_k(n) \right)^2 \]

where, \( i(n) \) is the original noise ratio parameter for the \( n^{th} \) frame. The noise ratio targets were scalar quantized by using 5 bits each. The RMS noise ratio error was found to be about 0.1 for the speech data set used in Section 4.1.
Table 1. Bit allocation for the proposed speech coders.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Proposed Coder 1</th>
<th>Proposed Coder 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event vector</td>
<td>24 bits (8+8+8)</td>
<td>27 bits (9+9+9)</td>
</tr>
<tr>
<td>Event function</td>
<td>7 bits</td>
<td>7 bits</td>
</tr>
<tr>
<td>Event location</td>
<td>8 bits</td>
<td>8 bits</td>
</tr>
<tr>
<td>Noise ratio</td>
<td>5 bits</td>
<td>5 bits</td>
</tr>
<tr>
<td>(sum x event rate)</td>
<td>660 bps</td>
<td>705 bps</td>
</tr>
<tr>
<td>Pitch</td>
<td>215 bps</td>
<td>215 bps</td>
</tr>
<tr>
<td>Gain</td>
<td>300 bps</td>
<td>300 bps</td>
</tr>
<tr>
<td>Maximum amplitude</td>
<td>5 bps</td>
<td>5 bps</td>
</tr>
<tr>
<td>of input speech</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal A</td>
<td>520 bps</td>
<td>520 bps</td>
</tr>
<tr>
<td>Subtotal B</td>
<td>1180 bps</td>
<td>1225 bps</td>
</tr>
<tr>
<td>Total (A+B)</td>
<td>1700 bps</td>
<td>1745 bps</td>
</tr>
</tbody>
</table>

7. BIT ALLOCATION

The bit allocation for the proposed speech coding method is shown in Table 1. The average number of events per second, i.e. the event rate, was set as 15 events/sec. We allocated 8 bits and 9 bits to each subvector of the event vectors, which resulted in 1.18 kbps and 1.23 kbps speech coders, respectively.

8. PERFORMANCE EVALUATION

In order to evaluate the performance of the proposed speech coding method, the quality of reconstructed speech was compared to that of other low bit rate speech coders such as the 4.8 kbps FS-1016 CELP and 2.4 kbps LPC-10E coders.

A listening experiment was carried out by using the Scheffe’s method of paired comparison [9] similarly to that in Section 4.2. A set of 108 phoneme balanced sentences of the ATR Japanese speech database were used for training the codebooks. Speakers were three male and three female reading each of sentences. These speech utterances were re-sampled at 8 kHz sampling frequency and then STRAIGHT analyzed. LSF transformation was performed and the resulting LSF parameters were TD analyzed by the MRTD method. Two phoneme balanced sentences, which are out of training set, uttered by a male and a female were used as the testing data. Stimuli were synthesized by using the following methods: STRAIGHT-LSF (no quantization), STRAIGHT-LSF & MRTD (no quantization), 4.8 kbps FS-1016 CELP, 2.4 kbps LPC-10E, proposed 1.18 kbps speech coder 1, and proposed 1.23 kbps speech coder 2. The original and the reconstructed speech files are located at the following URL: http://www.jaist.ac.jp/~chien/OF/

Results of the listening experiment are shown in Fig. 5. It can be seen from this figure that the quality of reconstructed speech obtained from the proposed speech coder 2 is close to that of the 4.8 kbps FS-1016 CELP and is much better than that of the 2.4 kbps LPC-10E.

9. CONCLUSION

In this paper, a low bit rate speech coding method based on STRAIGHT and Temporal Decomposition was described.

For encoding spectral information of speech, MRTD based vector quantization was used. Other speech parameters were scalar quantized. As a result, low bit rate speech coders operating at rates around 1.2 kbps were produced. Although the quality of reconstructed speech is little bit lower than that of the 4.8 kbps FS-1016 CELP coder according to the listening experiment, it is much better than that of the 2.4 kbps LPC-10E coder. Moreover, it can be stated that the proposed speech coding method can produce high quality speech with less than 2 kbps.

10. REFERENCES