Research on low bit rate speech coding using STRAIGHT and Temporal Decomposition

Takao Ochi (010029)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 2002

Keywords: Temporal Decomposition, STRAIGHT, vector quantization, bit-rate.

1 Introduction

In recent years, the demand of speech coding which can perform efficient transmission or efficient record is increasing with development of cellular phones and multimedia communication. In order to build a low-bit-rate speech coding system, it is necessary to carry out skillful coding of phonetic information. Although researches on the low-bit-rate speech coding used various techniques now have been done, coding systems with sufficient quality in the bit rate below 2 kbit/s are not realized.

In order to raise high quality synthesized speech, Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram (STRAIGHT)[1] which can analyze and synthesize speech with high quality is used for a low bit rate speech coding system in this research. However, much information has to be transmitted for the coding system. It is disadvantageous in respect of information compression. Temporal Decomposition (TD) is used to archive low-bit-rate coding systems, after STRAIGHT extracts spectral information and fundamental frequency information from speech signals. Phonetic information which appears locally in time domain is decomposed. Then, a low-bit-rate speech coding system based on STRAIGHT and TD is constructed.

2 Coding System

The outline of a system is shown in Fig.1. In STRAIGHT, values sent to a encoder are smoothed spectra and a fundamental frequency (F_0). The spectra are changed into LSF. It is decomposed by using Modified Restricted Temporal Decomposition (MRTD)[3] into event functions which are time change patterns of spectrum parameters and event targets which are typical spectra. These values are compressed by vector quantization. Scalar quantization is applied to other parameters, such as F_0, Gain and Noise Ratio, as shows in Fig.1.

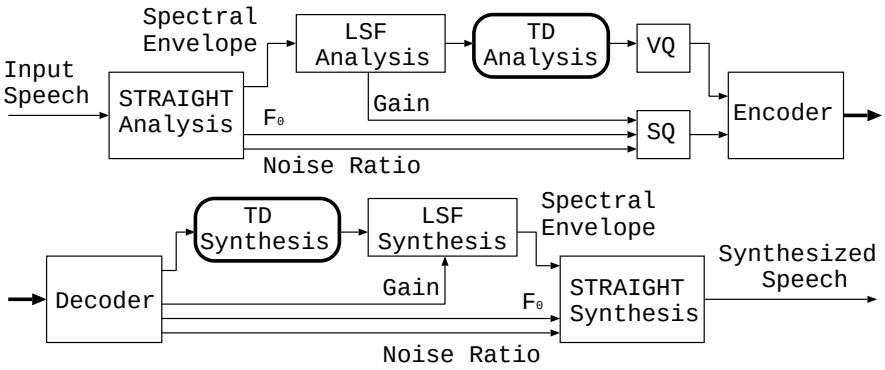

Figure 1: Low bit rate speech coding system using STRAIGHT and TD

3 STRAIGHT

Although analysis-by-synthesis coding systems of speech realize high compression performance based on a speech production model, it has a drawback that hight quality synthesized speech cannot be obtained as low-bit-rate coding systems. However, STRAIGHT proposed by Kawahara et al. is capturing the spotlight as a method obtaining a quality synthesis speech in analysis-by-synthesis coding system.

This research tries to improve quality of a synthesized speech by using STRAIGHT.

4 Coding of spectral information

4.1 LSF

A power spectrum is calculated using a amplitude spectrum obtained by STRAIGHT.

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

A correlation function is calculated from the power spectrum by inverse Fourier transform as follows,

$$R[n] = \frac{1}{N} \sum_{k=0}^{N-1} S[k] \exp\Big\{ j \frac{2\pi kn}{N} \Big\}$$

where, $S[k] = S[N-k]$. If a process $x(n)$ having this correlation function is an output from an all pole system (order L), the coefficient of the filter is set as $a_l^L, l = 1, 2, ..., L, 0 < L < N/2$L, and

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

where, P_L is estimation error. The coefficients a_l^L of a filter is determined as P_L becomes the minimum. Then, the filter coefficients are correspond to the LPC coefficients of the process $x(n)$. LSF is calculated using the prediction coefficients a_l^L.

4.2 MRTD[3]

The spectral information changed into LSF is further decomposed into an event function and an event target using MRTD. MRTD is a restricted and modified of original in order to fit TD[2] to low-bit-rate speech coding. TD involves the approximation of time varying spectral parameter vectors by a linear combination of event vectors as given in Equation (1).

$$\hat{\boldsymbol{y}}(n) = \sum_{k=1}^{K} \boldsymbol{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{1}$$

where, $\boldsymbol{a}_k$ and $\phi_k(n)$ are the k^{th} event vector and k^{th} event function, respectively. $\hat{\boldsymbol{y}}(n)$ is the approximation of $\hat{\boldsymbol{y}}(n)$, the n^{th} spectral parameter vector, produced by TD model.

In MRTD, two additional constraints on the event functions are enforced: (i) at any moment of time only two event functions, which are adjacent in time, are non-zero; and (ii) all event functions at any time sum up to one. With these constraints on the event functions, (1) becomes

$$\begin{aligned}\hat{\boldsymbol{y}}(n) &= \boldsymbol{a}_k\phi_k(n) + \boldsymbol{a}_{k+1}\phi_{k+1}(n) \\ &= \boldsymbol{a}_k\phi_k(n) + \boldsymbol{a}_{k+1}(1-\phi_k(n))\end{aligned}$$

for $C(k) \leq n \leq C(k+1)$, where $C(k)$ and $C(k+1)$ are the central positions of event k and event $k+1$, respectively. In addition,

$$\begin{gathered}\phi_k(C(k)) = 1,\ \phi_k(C(k+1)) = 0 \\ 0 \leq \phi_k(n) \leq 1 \ \text{ for } \ C(k) \leq n \leq C(k+1).\end{gathered}$$

Finally, determination of event functions can be written in mathematical form as

$$\phi_k(n) = \begin{cases} 1-\phi_{k-1}(n), & \text{if } C(k-1) < n < C(k) \\ 1, & \text{if } n = C(k) \\ \min(\phi_k(n-1), & \\ \max(0, \hat{\phi}_k(n))), & \text{if } C(k) < n < C(k+1) \\ 0, & \text{otherwise} \end{cases}$$

where,

$$\hat{\phi}_k = \frac{\langle (\boldsymbol{y}(n) - \boldsymbol{a}_{k+1}), (\boldsymbol{a}_k - \boldsymbol{a}_{k+1}) \rangle}{\|\boldsymbol{a}_k - \boldsymbol{a}_{k+1})\|^2}$$

4.3 Determination of order of LSF

4.3.1 Spectrum distortion vs. order of LSF

Logarithm spectrum distortion was adopted to determine order of LSF. In the coding system, 112 sentences were used as test data. This sentences are included in the phoneme balance 503 sentences uttered by speaker MMY in the ATR Japanese voice database. 8 kHz down sampling of the data was carried out. The result of two cases, applying LSF and LSF and MRTD, is shown in Fig.2. Quantization was not achieved. The horizontal axis expresses order of LSF. The vertical axis expresses logarithm spectrum distortion. Figure 2 shows that remarkable reduction of the spectrum distortion is not expectable after exceeding 22 of order of LSF, when LSF and MRTD are applied to a coding system.

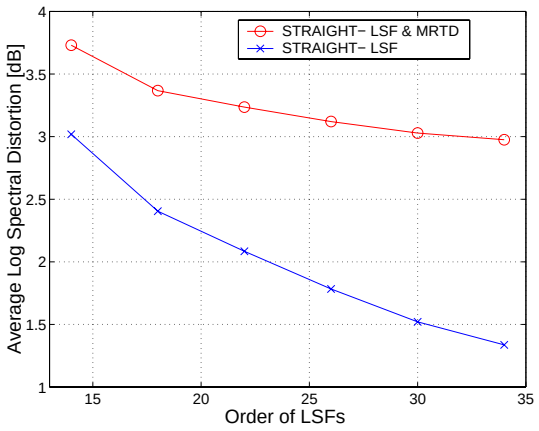

Figure 2: Change of the spectrum distortion to the order of LSF

4.3.2 Quality of synthesis speech vs. order of LSF

The listening experiment (Scheffe's method of paired comparison) subjectively estimated quality of a synthesis speech as a function of order of LSF.

Subjects judged in five grades whether of the distortion of speech (former and latter) is small. The subjects were six graduate students accepted to have normal hearing ability. 2 sentences were used as test data. These sentences were included in the phoneme balance 503 sentences uttered by the speaker MMY in the ATR Japanese voice database. 8 kHz down sampling of the data was carried out. These 2 sentences, analyzed and synthesized with 14, 18, 22, 26 and 30 of order of LSF. Quantization was not achieved. The experiment result is shown in Fig.3. The horizontal axis expresses the population parameter and the position expresses a relative distance of stimulus speech. The positive value indicates that distortion is small. The negative value is vice-versa. The number on the arrow expresses the order of LSF. The experiment result indicates that the improvement of distortion is not felt after exceeding 22 of order of LSF. Therefore, the order of LSF was determined as 22.

4.4 Vector Quantization

Vector quantization is used for quantizing the event function and the event target which were presumed by MRTD.

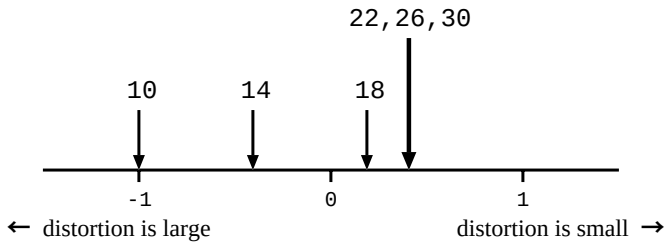

Figure 3: Change of the speech quality as the order of LSF

Quantization of event function
Since a time length of the event function differed for every event, the time length are normalized and vector-quantized by 5 bit.
Quantization of event target
Split vector quantization is adopted for the event target, since distribution of the low order LSF value is large. LSF was divided into three parts (6, 7, 8) and assigned 8 or 9 bit for every vector. The number of bit assigned to one event is 24 or 27 bit.

5 Coding of Fundamental Frequency

From the F_0 presumed by STRAIGHT, lengths of unvoiced and voiced intervals were quantized. An F_0 contour is sampled every 28 ms. Logarithmic scalar quantization was performed with 5 bit for each sampled value. In the decoder, F_0 of unvoiced interval is set to 0. F_0 of voiced interval is restored to the original size using linear interpolation. The RMS error was about 3.7 Hz.

6 Coding of gain

A gain contour is sampled every 20 ms. Logarithmic scalar quantization was performed with 6 bit for the data. The quantized value is restored using spline interpolation in the decoder. The RMS error was about 3.5 dB.

7 Coding of Noise Ratio

We reconstructed the noise ratio from the noise ratio targets and 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. The noise ratio targets were determined by considering the minimization of the squared error between the original noise ratio and interpolated noise ratio , as follows.

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

The noise ratio target was scalar quantized by 5 bit. The RMS error was 0.1.

8 Bit Allocation for Parameters

Bit allocation of each parameter is shown in Table 1. The number of events was set up so that it might become about 15 events/s. The value in the parenthesis of the bit allocation to an event target is the number of bits allocated to the divided event target.

Table 1: Bit allocation

Parameter	Proposed Coder 1	Proposed Coder 2
Event Target	24(8+8+8)	27(9+9+9)
Event Function	7	7
Distance between Events	8	8
Noise Ratio	5	5
Sub Total A (sum × the number of events)	660 [bit/s]	705 [bit/s]
F_0	215	215
Gain	300	300
Maximum of an Input Speech	9	9
Sub Total B	524 [bit/s]	524 [bit/s]
Total (A+B)	1194 [bit/s]	1229 [bit/s]

9 Quality Evaluation Experiment

In order to evaluate quality of decoded speech by the proposed method, the quality comparison experiment using synthesized speech waves by the proposed method and speech by other low-bit-rate speech coding systems was conducted.

The experiment was performed by Scheffe's method of paired comparison. Subjects judged in five grades whether distortion of speech (former and latter) is small. The subjects were eight graduate students accepted to have normal hearing ability. 108 sentences in the phoneme balance 503 sentences in the ATR Japanese speech database were used for learning data of the codebook. 8 kHz down sampling of the data is carried out. Speakers are three men and three women each. Two sentences (outside of learning) uttered by male and female used for speech data. Stimulus speech was synthesized by the six methods. The methods are LSF to STRAIGHT (with no quantization), LSF and MRTD (with no quantization), 4.8 kbit/s CELP, 2.4 kbit/s LPC10, Proposed 1 (1.19 kbit/s), and Proposed 2 (1.23 kbit/s).

Although the proposed method was lower than quality of 4.8 kbit/s CELP as the result of the evaluation experiment, it is better than 2.4 kbit/s LPC10.

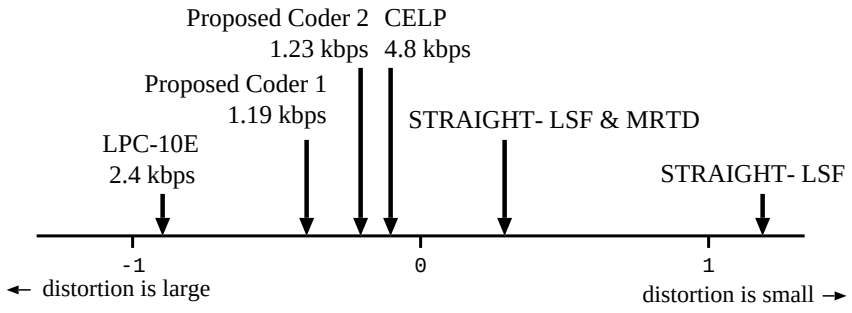

Figure 4: The result of a quality evaluation experiment

10 Conclusion

In this research, spectrum information obtained by STRAIGHT was compressed by using TD and vector quantization. Other parameters were scalar-quantized. Then, the low-bit-rate speech coding system with 1.2 kbit/s based on them was built.

Although it was lower than quality of 4.8 kbit/s CELP as the result of the evaluation experiment, it is better than 2.4 kbit/s LPC10. Therefore, although the proposed method is not hight quality in low bit rate speech coding, it can be said that the system can synthesize high quality synthesis speech with lower than 2 kbit/s created.

References

[1] Hideki KAWAHARA, "Scene analysis and STRAIGHT," Proc. ASJ, 1-2-1, pp.186-192, 1997-9.

[2] B.S.Atal, "Efficient coding of LPC parameters by temporal decomposition," Proc. ICASSP '83, pp.81-84, 1983.

[3] P.C.Nguyen and M.Akagi, "Improvement of the restricted temporal decomposition method for LSF parameters," Proc. 2001 Autumn Meeting of ASJ, pp.267-268, 2001.