IMPROVEMENT OF BAND EXTENSION TECHNIQUE FOR G.711 TELEPHONY SPEECH BASED ON FULL WAVE RECTIFICATION

Naofumi Aoki

Graduate School of Information Science and Technology,
Hokkaido University
Sapporo, Japan
aoki@nis-ei.eng.hokudai.ac.jp

ABSTRACT

This study investigates a band extension technique for the narrow-band speech encoded with G.711, the most common codec for digital speech communications such as VoIP. The proposed technique is based on the full wave rectification that generates high-band harmonics by nonlinear processing. In order to improve the conventional technique, this study focuses on the parameter control according to the characteristics of speech data. From the subjective evaluation, it is indicated that the proposed technique may potentially outperform the conventional technique.

1. INTRODUCTION

Band extension of the narrow-band speech may improve the intelligibility of speech communications.

This study focuses on band extension of the narrow-band speech encoded with G.711 [1], the most common codec for digital speech communications such as VoIP (Voice over IP). G.711 is an ITU (International Telecommunication Union) standard of the narrow-band speech codec that encodes speech data into a stream of 8 bit speech samples at an 8 kHz sampling rate.

The proposed technique is based on the full wave rectification that generates high-band harmonics by nonlinear processing [2], [3]. This technique is based on the fact that there is a strong correlation between the low and high frequency bands of most speech material.

However, it is pointed out that the band extension of unvoiced speech is not very appropriate with conventional technique [2]. Due to the constant parameter, the conventional technique cannot take account of the fact that the high frequency band of unvoiced speech is remarkable compared with that of voiced speech.

In order to improve the conventional technique, this study investigates frame-by-frame parameter control according to the characteristics of speech data.

2. BAND EXTENSION BASED ON FULL WAVE RECTIFICATION

The proposed technique employs a band extension technique based on the full wave rectification [2]. As shown in Fig.1, full wave rectification is applied to the band between 2 and 4 kHz in order to generate high-band harmonics over 4kHz. Such high frequency band is then mixed with the original low frequency band after the gain control. The block diagram of the band extension technique is shown in Fig.2.

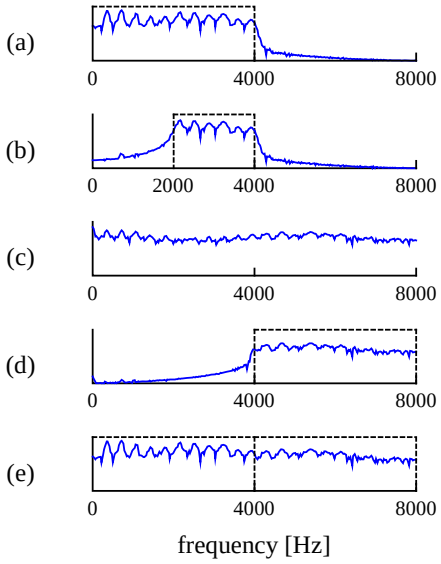

Figure 1: *Procedure of the band extension technique: (a) original speech, (b) band-pass filtering, (c) full wave rectification, (d) high-pass filtering, and (e) mixing the low and high frequency bands.*

3. SIDE INFORMATION FOR THE GAIN CONTROL

In the conventional technique, the gain of the high frequency band is kept constant at 0.5 [2]. Although this works well for voiced speech in which the high frequency band is not dominant, it is not very appropriate for unvoiced speech. It should be considered that the high frequency band is remarkable in unvoiced speech such as fricatives and sibilants.

It seems that the appropriate gain control may potentially improve the quality of band-extended speech. A solution is the transmission of the actual gain information from the sender in which the original wide-band speech is available.

The proposed technique calculates the mean amplitude of the high frequency band in each frame. The frame length is chosen to be 20 ms, 160 speech samples at an 8 kHz sampling rate. This gain information is quantized with 7 bits according to the logarithmic characteristic of G.711. This is the side information in the proposed technique.

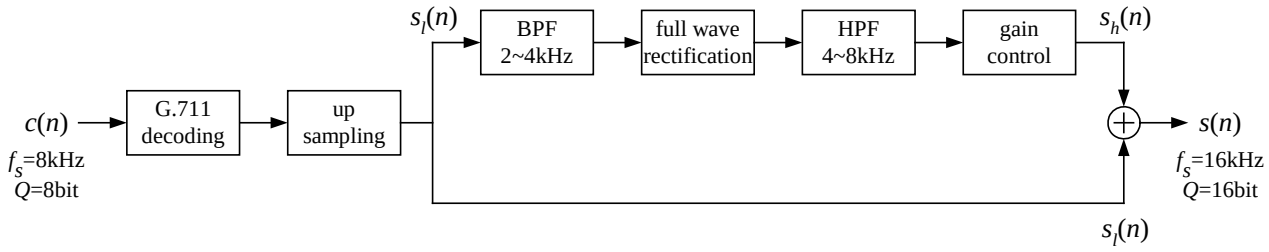

Figure 2: *Procedure of the band extension technique.*

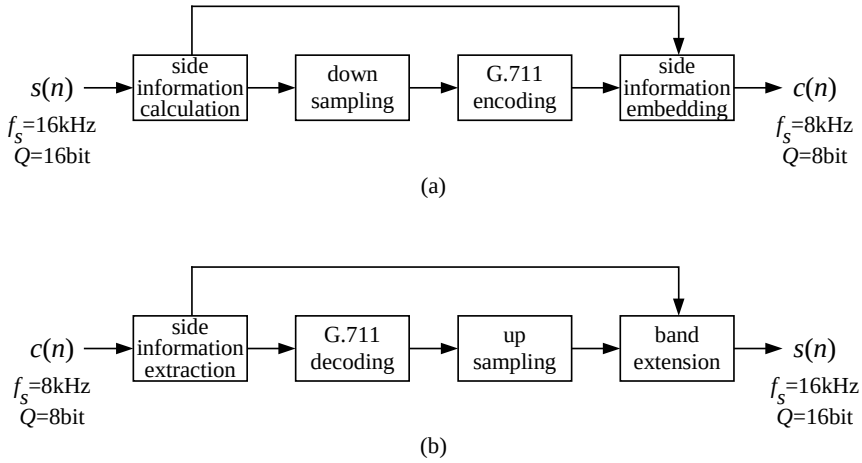

Figure 3: *Procedure of the proposed technique at (a) sender and (b) receiver.*

4. PROPOSED TECHNIQUE BASED ON THE TRANSMISSION OF THE SIDE INFORMATION

The procedure of the proposed technique is shown in Fig.3. The proposed technique employs steganography to transmit the side information [4].

In each frame, the side information is directly embedded into speech samples by replacing their LSB (Least Significant Bit). Although embedding the side information causes some degradation, it is almost negligible in the proposed technique since the side information is embedded into the LSB of just 7 speech samples in each frame consisting of 160 speech samples.

Furthermore, in order to decrease the degradation as much as possible, the proposed technique embeds the side information into the speech samples whose absolute amplitude is relatively small. This is based on the fact that the weight of the LSB in G.711 is proportional to the absolute amplitude of speech sample due to the logarithmic quantization [4].

The SNR (Signal-to-Noise Ratio) between the original and decoded speech data was investigated from 10 speech data obtained from a speech database [5].

The average of the SNR calculated by the speech data with embedding the side information was 37.903 dB, while the speech data without embedding the side information was 37.908 dB. This indicates that the degradation by embedding the side information is almost negligible, so that the degradation is hardly perceived by the human ear.

Figure 4 shows the spectrogram of the original speech of a Japanese word “shiro”. The results of the band extension with the conventional and proposed technique are shown in Fig.5 and 6, respectively.

Compared with the conventional technique, the proposed technique may appropriately emphasize the high frequency band of unvoiced speech. This indicates that the proposed technique may outperform the conventional technique especially in the case of unvoiced speech.

Subjective evaluation was performed in order to examine how effectively the proposed technique makes the speech quality intelligible. 10 speech data consisting of 5 male voice (m1 - m5) and 5 female voice (f1 - f5) were obtained from the speech database [5].

The evaluation employed CMOS (Comparison Mean Opinion Score) [6]. In each trial of comparison test, stimulus A and stimulus B were presented to the listeners in this order, where stimulus A and B were the band-extended speech processed by means of either the proposed or conventional technique.

10 listeners rated the quality of stimulus B compared with stimulus A according to Table 1. Each combination of stimulus A and B was presented twice by reversing the order, so that each condition was evaluated 20 times by 10 listeners.

Figure 7 shows the experimental result. This figure also shows the 95% confidence intervals of the averages. This shows that the proposed technique may outperform the conventional technique.

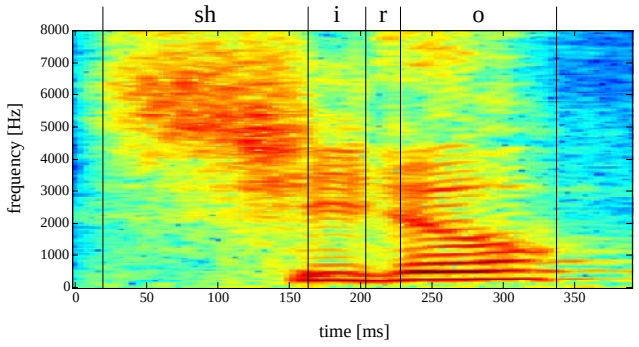

Figure 4: *Spectrogram of the original speech "shiro".*

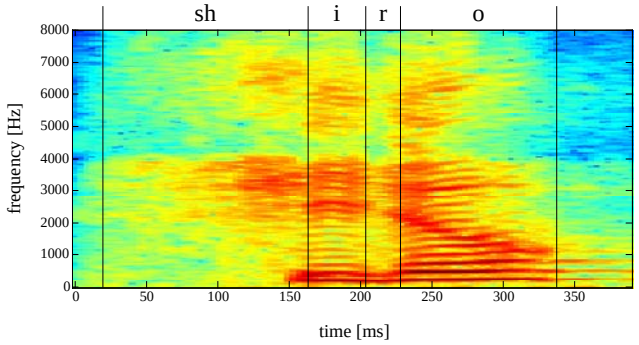

Figure 5: *Spectrogram of the band-extended speech "shiro" with the conventional technique.*

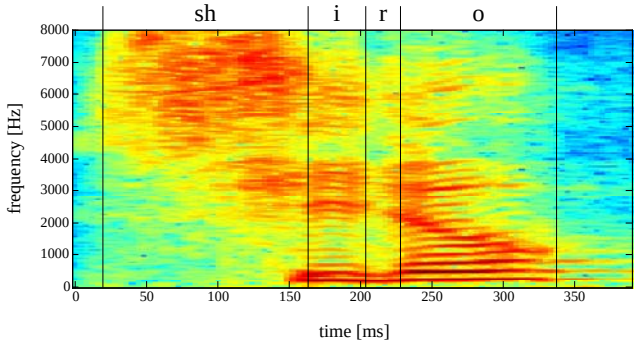

Figure 6: *Spectrogram of the band-extended speech "shiro" with the proposed technique.*

Table 1: *Seven-point scale in CMOS.*

point	quality
+3	much better
+2	better
+1	slightly better
+0	about the same
-1	slightly worse
-2	worse
-3	much worse

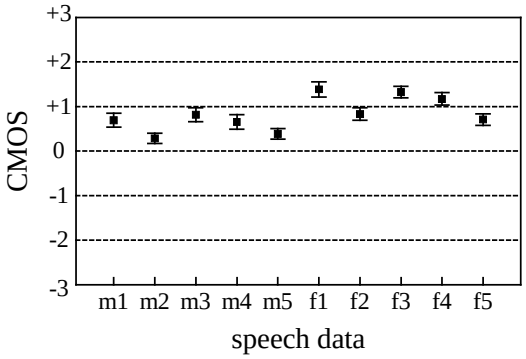

Figure 7: *Experimental result of the subjective evaluation (proposed technique vs. conventional technique).*

5. PROPOSED TECHNIQUE BASED ON THE ESTIMATION OF THE SIDE INFORMATION

Although the speech quality may be enhanced with the appropriate side information, it is required to modify both the sender and receiver for the transmission of the side information. It is much desirable to modify just only the receiver if it is possible to estimate appropriately the side information at the receiver.

As shown in Fig.8, there is a correlation between the mean amplitude of the high frequency band and the zero-crossing rate of the speech data. Similarly to the mean amplitude of the high frequency band, the zero-crossing rate tends to be large in unvoiced speech [7].

Based on this characteristic, the proposed technique defines the gain of k-th frame as follows.

$$g(k) = \alpha z(k) \tag{1}$$

where $z(k)$ is the zero-crossing rate of k-th frame, and α is a constant. In the proposed technique, α is empirically chosen to be 5.

The correlation coefficient of the mean amplitude in the high frequency band was investigated from 100 speech data obtained from the speech database [5].

The average of the correlation coefficient was 0.71 between the original speech and band-extended speech processed by the proposed technique. On the other hand, the average of the correlation coefficient was 0.52 between the original speech and band-extended speech processed by the conventional technique. This indicates that the zero-crossing rate may be a candidate for the parameter estimation.

Figure 9 shows the spectrogram of the band-extended speech of a Japanese word "shiro" processed by the proposed technique. Compared with the conventional technique shown in Fig.5, the proposed technique may emphasize the high frequency band of unvoiced speech. This indicates that the proposed technique may outperform the conventional technique especially in the case of unvoiced speech.

Subjective evaluation was performed in order to examine how effectively the proposed technique makes the speech quality intelligible. 10 speech data consisting of 5 male voice (m1 - m5) and 5 female voice (f1 - f5) were obtained from the speech database [5].

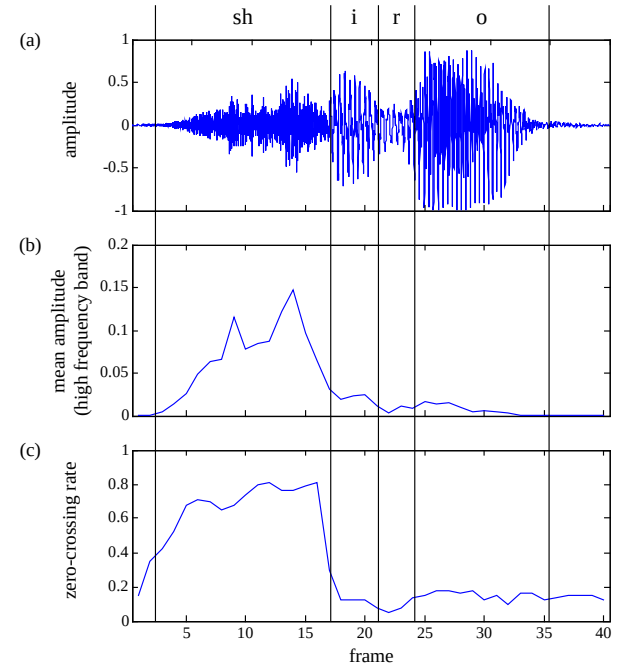

Figure 8: *Characteristics of speech data: (a) speech data "shiro", (b) mean amplitude in the high frequency band, and (c) zero-crossing rate.*

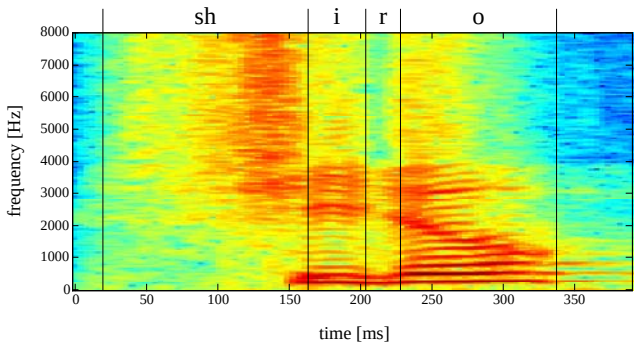

Figure 9: *Spectrogram of the band-extended speech "shiro" with the proposed technique.*

The evaluation employed CMOS [6]. In each trial of comparison test, stimulus A and stimulus B were presented to the listeners in this order, where stimulus A and B were the band-extended speech processed by means of either the proposed or conventional technique.

10 listeners rated the quality of stimulus B compared with stimulus A according to Table 1. Each combination of stimulus A and B was presented twice by reversing the order, so that each condition was evaluated 20 times by 10 listeners.

Figure 10 shows the experimental result. This figure also shows the 95% confidence intervals of the averages. This shows that the proposed technique may outperform the conventional technique.

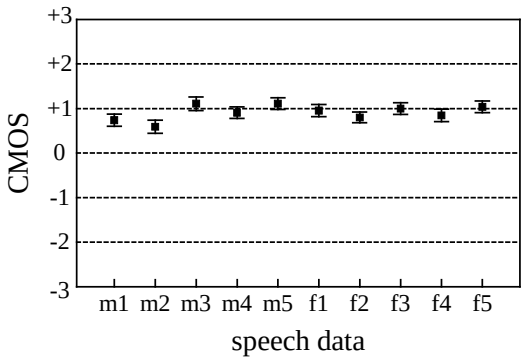

Figure 10: *Experimental result of the subjective evaluation (proposed technique vs. conventional technique).*

6. CONCLUSIONS

The experimental results indicate that the appropriate gain control may improve the conventional technique. Emphasizing the high frequency band of unvoiced speech may potentially enhance the intelligibility of the band-expanded speech.

Even though the side information is not transmitted, speech quality may be enhanced if the parameter is appropriately estimated. As indicated from the experimental results, the zero-crossing rate may be a candidate for the parameter estimation.

In order to investigate the potential advantage of the proposed technique for the actual digital speech communications system, further verification is under consideration.

7. ACKNOWLEDGEMENT

The author would like to express the gratitude to the Ministry of Education, Culture, Sports, Science and Technology of Japan for providing a grant (no.18760263) toward this study.

8. REFERENCES

[1] ITU-T G.711, Pulse code modulation (PCM) of voice frequencies, 1988.

[2] R.M. Aarts, E. Larsen, and D. Schobben, "Improving perceived bass and reconstruction of high frequencies for band limited signals," Proc. IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002), pp.59-71, Nov.2002.

[3] U. Zölzer (ed.), DAFX - Digital Audio Effects, John Wiley & Sons, 2002.

[4] N. Aoki, "A band extension technique for G.711 speech using steganography," IEICE Transactions on Communications, vol.E89-B, no.6, pp.1896-1898, 2006.

[5] ATR Interpreting Telecommunications Research Laboratories, Speech dialogue database for spontaneous speech recognition, 1997.

[6] ITU-T P.800, Methods for subjective determination of transmission quality, 1996.

[7] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.