LETTER

A Band Extension Technique for Narrow-Band Telephony Speech Based on Full Wave Rectification

Naofumi AOKI[†], ***Member***

SUMMARY This study investigates a band extension technique for narrow-band telephony speech. The proposed technique employs full wave rectification that nonlinearly generates high-band overtones from the low band. In order to improve the conventional technique, this study investigates a frame-by-frame gain control based on the estimation of gain parameter from narrow-band telephony speech. A subjective evaluation indicates that the proposed technique outperforms the conventional technique.

key words: *band extension, telephony speech, full wave rectification*

1. Introduction

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

In order to reconstruct artificially the high band, full wave rectification is employed in the conventional technique [1]. This technique nonlinearly generates high-band overtones from the low band.

As shown in Fig.1, the conventional technique applies full wave rectification to the band between 2 and 4 kHz in order to generate the high band over 4kHz [1]. Such high band is mixed with the low band after an appropriate gain control.

Although the conventional technique generally works well for voiced speech, it is pointed out that the conventional technique is not very appropriate for unvoiced speech [1]. Due to the constant gain parameter, the conventional technique does not appropriately take account of the fact that the high band of unvoiced speech tends to be prominent compared with that of voiced speech.

In order to improve the conventional technique, this study investigates a frame-by-frame gain control based on the estimation of gain parameter from narrow-band telephony speech.

2. Proposed Technique

In order to emphasize the gain for the high band of unvoiced speech, the proposed technique employs a parameter called gradient index for the detection of unvoiced speech [2].

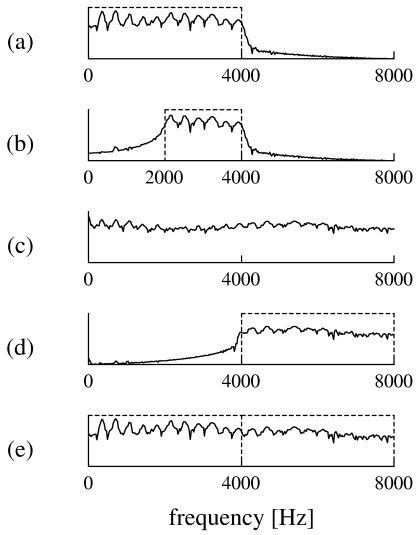

Fig. 1 Procedure of the band extension technique based on full wave rectification: (a) original speech, (b) band-pass filtering, (c) full wave rectification, (d) high-pass filtering, and (e) mixing the low and high bands.

The gradient index of speech data obtained from k-th frame is defined as follows.

$$d(k) = \frac{\sum_{n=2}^{N-1} \Delta\Psi(n)|s_n(n) - s_n(n-1)|}{\sqrt{\sum_{n=0}^{N-1} s_n^2(n)}} \tag{1}$$

with

$$\Delta\Psi(n) = \frac{|\Psi(n) - \Psi(n-1)|}{2} \quad (\in 0, 1) \tag{2}$$

$$\Psi(n) = \mathrm{sign}(s_n(n) - s_n(n-1)) \quad (\in -1, 1) \tag{3}$$

where $s_n(n)$ represents narrow-band speech data. In the proposed technique, N is chosen to be 160. It is equal to 20 ms at an 8 kHz sampling rate.

The gradient index measures the number and the magnitude of direction changes of speech data. It shows large values when the power of the high band is prominent. Therefore, the gradient index can be a useful

Manuscript received July 1, 2009.
Manuscript revised October 26, 2009.

[†]The author is with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060-0814 Japan.

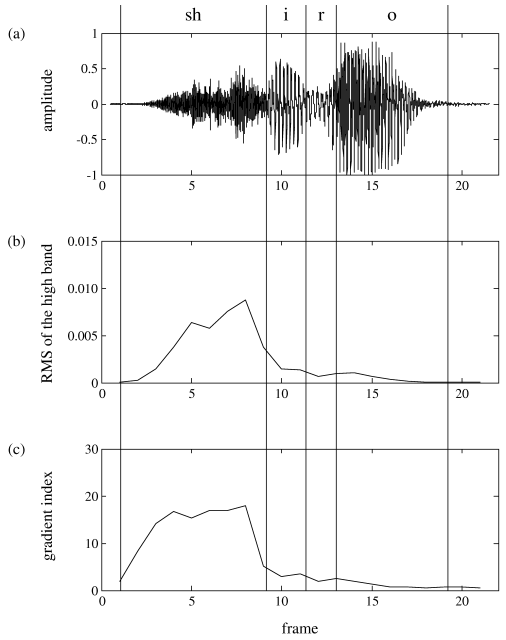

Fig. 2 A Japanese speech data "shiro": (a) waveform, (b) RMS of the high band, (c) gradient index.

indicator of unvoiced speech.

Figure 2 shows the RMS (Root Mean Square) of the high band and the gradient index of a Japanese speech data "shiro". As shown in this figure, there is positive correlation between these parameters. Similarly to the RMS of the high band, the gradient index shows large values in unvoiced speech.

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

$$g(k) = \alpha \cdot d(k) \tag{4}$$

where α defines the ratio of the power between the low and high bands. In the proposed technique, α is empirically chosen to be 1, so that the gain is equivalent to the gradient index itself.

The correlation coefficient of the RMS of the high band between the original speech data and band-extended speech data was calculated from 100 speech data obtained from a Japanese speech database [3]. The larger the correlation coefficient, the more the band extension technique works effectively.

The average of the correlation coefficient was 0.71 between the original speech data and band-extended speech data processed by the proposed technique where the gain was varied frame-by-frame.

On the other hand, the average of the correlation coefficient was 0.47 between the original speech data and band-extended speech data processed by the conventional technique where the gain was constant at 0.5 for all frames [1].

These results indicate that the proposed technique is more appropriate for generating the high band compared with the conventional technique.

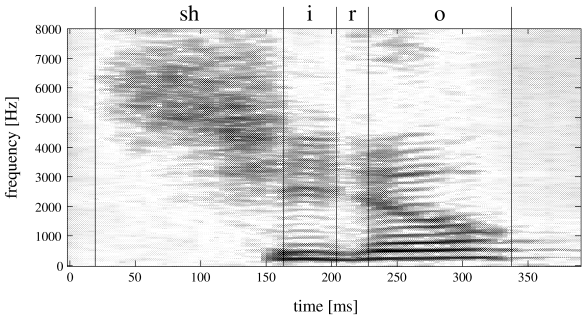

Fig. 3 Spectrogram of the original speech data "shiro".

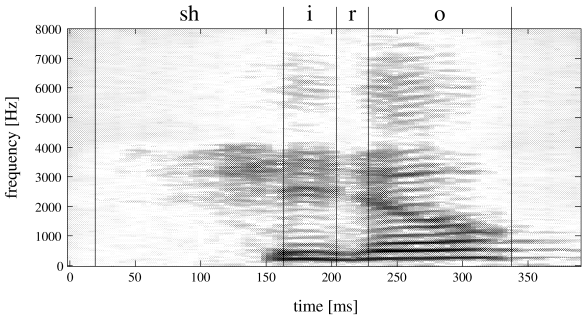

Fig. 4 Spectrogram of the band-extended speech data "shiro" with the conventional technique.

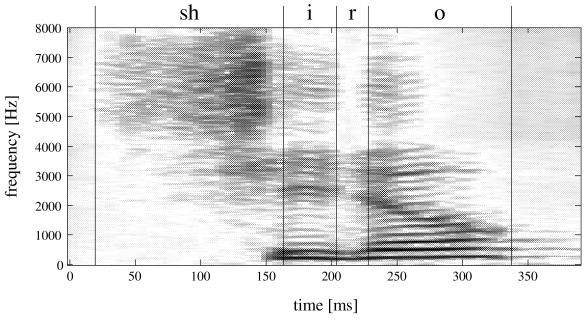

Fig. 5 Spectrogram of the band-extended speech data "shiro" with the proposed technique.

Instead of the gradient index, the zero crossing rate of speech data is often employed for the detection of unvoiced speech [4].

Similarly to the gradient index, the zero crossing rate may be employed for gain control [5]. However, the average of the correlation coefficient was 0.66 when the zero crossing rate is employed instead of the gradient index. This indicates that the gradient index may be a better candidate for the gain parameter compared with the zero crossing rate.

Figure 3 shows the spectrogram of the original speech data "shiro". The results of the band extension with the conventional technique and the proposed technique are shown in Figs.4 and 5, respectively. These fig-

Table 1 Seven-point score of CMOS.

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

ures indicate that the proposed technique outperforms the conventional technique especially in the case of unvoiced speech.

In order to mitigate the aliasing effect caused by the nonlinear processing, the full wave rectification is performed at a 64 kHz sampling rate by using an oversampling technique. In addition, in order to mitigate the tonal noise caused by artificial formant structure generated by the full wave rectification, a linear prediction technique is employed for flattening the spectral envelope of the high band.

These two techniques were applied to both the proposed technique and the conventional technique in a subjective evaluation described below.

3. Subjective Evaluation

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 [3].

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 data processed by means of either the proposed technique or the 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 6 shows the experimental result. This figure also shows the 95 % confidence intervals of the averages. This result indicates that the proposed technique outperforms the conventional technique.

4. Conclusions

Experiments indicated that the appropriate gain control may improve the conventional technique. Emphasizing the high band of unvoiced speech may enhance the intelligibility of the band-extended speech data. It seems that the gradient index is an appropriate candidate for the gain parameter.

In order to investigate the potential advantage of the proposed technique in actual environments, further verification is under consideration.

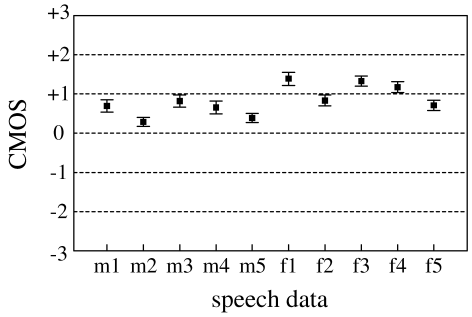

Fig. 6 CMOS score of the proposed technique compared with the conventional technique.

References

[1] R.M. Aarts, E. Larsen, and D. Schobben, "Improving perceived bass and reconstruction of high frequencies for band limited signals," IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002), pp.59–71, Nov.2002.
[2] J.W. Paulus, "Variable bitrate wideband speech coding using perceptually motivated thresholds," IEEE Workshop on Speech Coding for Telecommunications, pp.35–36, 1995.
[3] ATR Interpreting Telecommunications Research Laboratories, Speech dialogue database for spontaneous speech recognition, 1997.
[4] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.
[5] N. Aoki, "Improvement of band extension technique for G.711 telephony speech based on full wave rectification," 11th International Conference on Digital Audio Effects (DAFx-08), pp.161–164, 2008.
[6] ITU-T P.800, Methods for subjective determination of transmission quality, 1996.