A BINAURAL SPEECH PROCESSING METHOD USING SUBBAND-CROSSCORRELATION ANALYSIS FOR NOISE ROBUST RECOGNITION

Shoji KAJITA, Kazuya TAKEDA and Fumitada ITAKURA

Graduate School of Engineering, Nagoya University
Furo-cho 1, Chikusa-ku, Nagoya, 464-01 JAPAN
Email: kaji@nuee.nagoya-u.ac.jp

ABSTRACT

This paper describes an extended subband-crosscorrelation(SBXCOR) analysis to improve the robustness against noise. The SBXCOR analysis, which has been already proposed, is a binaural speech processing technique using two input signals and extracts the periodicities associated with the inverse of the center frequency(CF) in each subband. In this paper, by taking an exponentially weighted sum of crosscorrelation at the integral multiples of the inverse of CF, SBXCOR is extended so as to capture more periodicities included in two input signals. The experimental results using a DTW word recognizer showed that the processing improves the performance of SBXCOR for both that of the white noise and a computer room noise. For white noise, the extended SBXCOR performed significantly better than the smoothed group delay spectrum and the mel-frequency cepstral coefficient(MFCC) extracted from both monaural and binaural signals. However, for the computer room noise, it outperformed only at SNR 0dB.

1. INTRODUCTION

In any speech recognition system, the speech analysis part is the front-end for the acoustic environment, and the acoustic features lost in this part cannot be easily recovered in any later stage. Especially, for constructing a robust recognition system used in real acoustic environments where noise, reverberation and the other interferences significantly affect the acoustic signal, the acoustic front-end has to be robust against such influences and extract effective speech features.

In order to address the problem, we have proposed subband-autocorrelation(SBCOR) analysis[1]. The SBCOR is a kind of filter bank analysis, and aims to extract periodicities associated with the inverse of the center frequency included in speech signals. The experimental results for speech recognition showed that SBCOR performs better than the conventional method under noisy conditions.

Furthermore, as an extension of SBCOR, we have proposed subband-crosscorrelation(SBXCOR) analysis using binaural signal to improve the performance of speech recognition under noisy conditions[2]. Speech feature extraction by such binaural processing is also a current topic in recent studies related to the auditory modelling, and several binaural processing models have been proposed to improve the recognition performance under the adverse conditions and shown the effectiveness [3][4][5]. Unlike those binaural auditory models, SBXCOR is a signal processing method based on filter bank and crosscorrelation analysis.

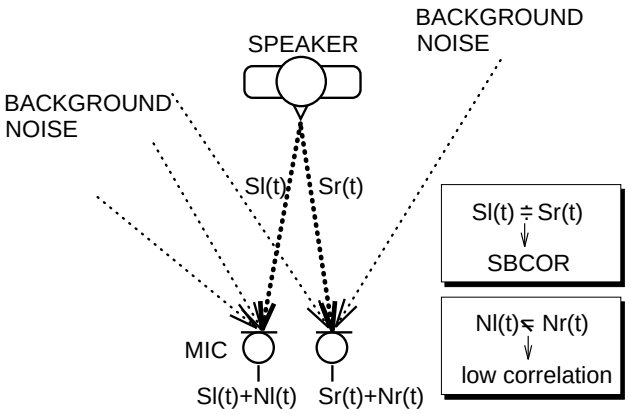

Figure 1. Concept of SBXCOR analysis.

In this paper, in order to improve the robustness of the SBXCOR, multi-delay weighting(MDW) processing[6] that utilizes more periodicities included in binaural signal is incorporated in SBXCOR analysis.

This paper is constructed as follows. The following section reviews SBXCOR analysis and introduces the multi-delay weighting processing. Section 3 investigates the robustness against white noise and a computer room noise under simulated acoustic conditions on computer. Section 4 summarizes this research.

2. SBXCOR ANALYSIS AND MDW PROCESSING

2.1. SBXCOR Analysis

SBXCOR analysis is a binaural speech processing technique using two input signals, and extracts the periodicities associated with the inverse of the center frequency $f_{cf_i}^{-1}$in each subband. The ith SBXCOR coefficient $Sc(\tau_{cf_i}, n)$ for nth analysis frame is calculated as follows:

$$Sc(\tau_{cf_i}, n) = \frac{\mathrm{Re}\ R^i_{x_1x_2}(\tau_{cf_i}, n)}{\sqrt{R^i_{x_1x_1}(0,n) R^i_{x_2x_2}(0,n)}} \quad (1)$$

$$\tau_{cf_i} = f_{cf_i}^{-1} \quad (2)$$

$$R^i_{x_1x_2}(\tau_{cf_i}, n) = \int_{-f_n}^{f_n} \mid H_i(f) \mid^2 F_{x_1}(f,n) F^*_{x_2}(f,n)$$

$$\times\ e^{j2\pi f\tau_{cf_i}}\,df, \tag{3}$$

$$R^i_{x_1x_1}(0,n) = \int_{-f_n}^{f_n} |\,H_i(f)\,|^2 |\,F_{x_1}(f,n)\,|^2\,df, \tag{4}$$

$$R^i_{x_2x_2}(0,n) = \int_{-f_n}^{f_n} |\,H_i(f)\,|^2 |\,F_{x_2}(f,n)\,|^2\,df, \tag{5}$$

where $R^i_{x_1x_1}(\tau,n)$, $R^i_{x_2x_2}(\tau,n)$ and $R^i_{x_1x_2}(\tau,n)$ are the autocorrelation and crosscorrelation functions of ith subband signal respectively, and $F_{x_1}(f,n)$ and $F_{x_2}(f,n)$ are the FFT spectrum of two input signals $x_1(t,n)$ and $x_2(t,n)$ respectively.

As for the filter bank, a fixed Q filter bank whose center frequencies are equally spaced on the Bark scale has been shown to be suitable for speech recognition under noisy conditions so far[1]. In the following experiments, the filter bank consists of 16 fixed Q gaussian bandpass filters(BPFs) defined by

$$|\,H_i(f)\,|^2 = \begin{cases} e^{-2C_i(f-f_{cf_i})^2}, & f \geq 0 \\ |\,H_i(-f)\,|^2, & f < 0, \end{cases} \tag{6}$$

where

$$C_i = \frac{2Q^2 \ln 2}{f^2_{cf_i}}. \tag{7}$$

The robustness of SBXCOR against noise can be explained as shown in Figure 1. Since the speech signals recorded by two microphones, which are uttered just in front of two microphones, have the same amplitude and phase, SBXCOR extracts the same spectrum as SBCOR. On the other hand, since noises are low correlation, their influences are canceled in the processing. In the following experiments, we investigate the performance of SBXCOR under the assumption that speakers utter just in front of two microphones.

2.2. Multi-Delay Weighting

If both of binaural signal are periodic with a period T, the crosscorrelation coefficients show several peaks at integral multiples of T. In conventional SBXCOR analysis defined by Equation (1), however, only one crosscorrelation coefficient at T is used to extract the periodicity included in the subband signal. Therefore, we extend the SBXCOR to capture the other peaks of the crosscorrelation coefficients by taking a weighted sum of them with the power of α, i.e. the exponential weighting(Fig.2) as follows:

$$\hat{Sc}(\tau_{cf_i},n) = \frac{\sum_{k=0}^{M-1} \alpha^k Sc((k+1)\tau_{cf_i},n)}{\sum_{k=0}^{M-1} \alpha^k} \tag{8}$$

$$(0 < \alpha < 1). \tag{9}$$

We have referred to it as multi-delay weighting(MDW) processing[6]. The MDW processing has been shown to be effective in SBCOR analysis[6].

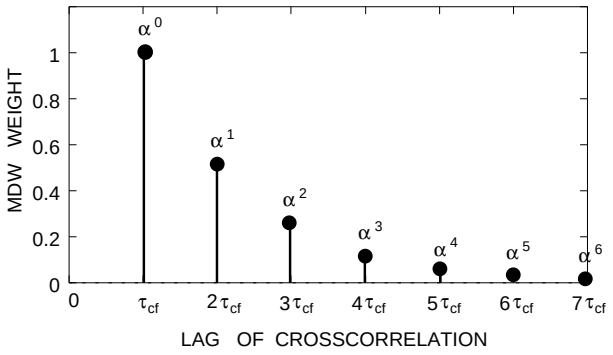

Figure 2. MDW weighting.

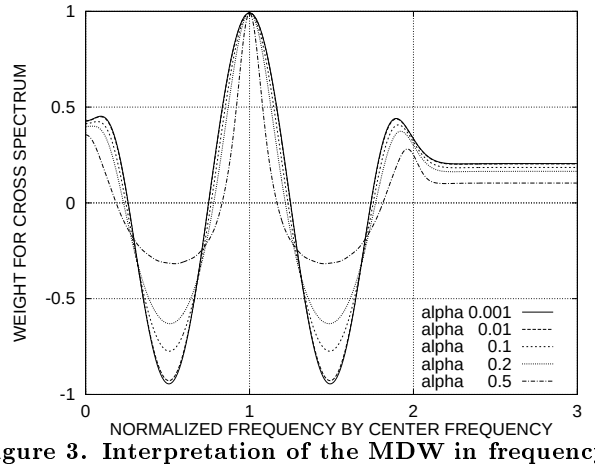

Figure 3. Interpretation of the MDW in frequency domain.

2.3. The Effect of MDW

Figure 3 shows the interpretation of MDW in the frequency domain as a weighting for the cross-spectrum of the binaural signal[6]. As shown in the figure, by MDW processing, the frequency resolution and the Mexican hat weighting of SBXCOR are controllable. The frequency resolution is higher as α is closer to one. On the other hand, the emphasis of spectral contrast by the Mexican hat weighting is lower as α is closer to one. The contribution of these effects to recognition performance will be experimentally shown in the following recognition experiments.

3. EXPERIMENTS

In this section, the robustness of the proposed method against white noise and a computer room noise is evaluated using a DTW word recognizer. From the experiment, the following points will be clarified:

1. how much the proposed method improves better than the conventional SBXCOR,
2. as a processing method using 2 channel signals, whether the proposed method is more effective or not than the delay-and-sum processing usually used in the multi-microphone system,
3. to what extent the performance of the proposed method is better than those of the smoothed group de-

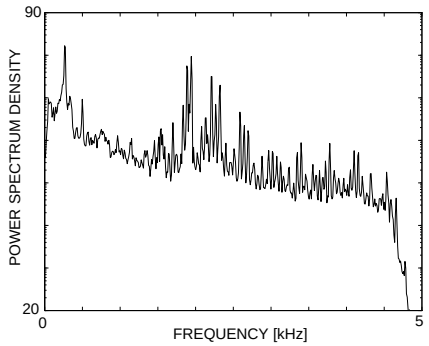

Figure 4. power spectrum of a computer room noise(3 seconds).

lay spectrum(SGDS)[7, 8] and the mel-frequency cepstral coefficient(MFCC)[9].

3.1. Experimental Conditions

3.1.1. DTW word recognizer

A standard DTW speaker-dependent isolated word recognizer is used. The recognition task is a 68 pair discrimination[7]. Each pair is a phonetically similar city name pair, selected from a 550 Japanese city name database recorded twice by 5 Japanese male speakers. The first set is used as the reference pattern and the second set, which was spoken a week later, is used as the test pattern.

3.1.2. Noisy Signals Used in Test

As for white noise, two gaussian white noises generated by different seeds are added to the test signals. Furthermore, as a realistic environmental noise, the noise in a computer room was recorded using three microphones and added to the test signals on computer. The distance between microphones is 10cm. The computer room noise is a non-stationary noise and its power spectrum is shown in Figure 4. The left and right channel signals are used as two channel signals, and the middle channel is used as one channel signal. In all cases, the global signal-to-noise ratios(SNRs) were set to be 20, 10, 5 and 0dB.

3.1.3. Two-Channel-Summed Signal

The two-channel-summed signals are generated by simple-summation of the above two channel signals because of no delay between two channel. By doing this, the effective SNR improvement is about 3dB.

3.1.4. SBXCOR

The Q values of 1.0, 1.5, 2.0, 2.5 and 3.0 are investigated. FFT-point is 1024. In order to calculate coefficients of the correlation function precisely, polynomial interpolation was used. The center frequencies of the BPFs are equally spaced on the Bark scale between 4 and 17 Bark. In MDW processing, the αs of 0.0-0.9 are investigated.

3.1.5. SGDS and MFCC

SGDS has been shown to be robust against noise, and it is calculated as the derivative of phase of a pth order all pole filter that has smoothed poles. In order to compare the performance of SBCOR with that of SGDS under exactly the same conditions, the analysis frequency points of SGDS

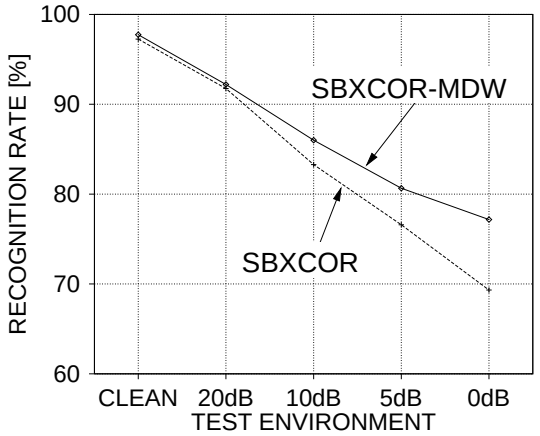

Figure 5. Comparison with SBXCOR with and without MDW(White Noise).

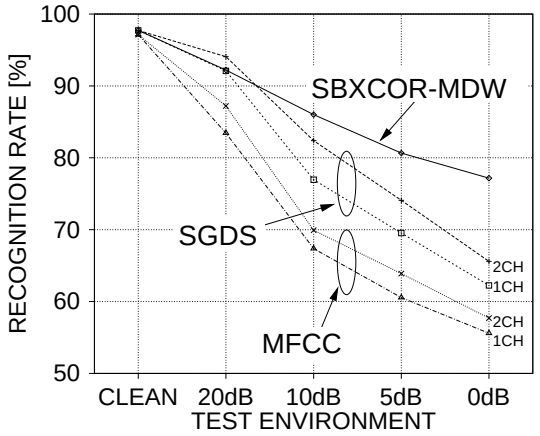

Figure 6. Comparison with SBXCOR with MDW, SGDS and MFCC(White Noise).

were chosen to be the same as the center frequencies of SBXCOR.

MFCC is a commonly used speech feature in speech recognizers. In this experiment, MFCC is calculated using a 28 triangular shape mel-filterbank.

As for the other common analysis conditions, the analysis frame length and shift are 20ms and 10ms respectively, and the dimension of each feature is 16. The sampling rate is 10 kHz.

3.2. Experimental Results

3.2.1. The Robustness Against White Noise

Figure 5 shows the best recognition rates of SBXCOR(Q=2.0) processed by MDW and SBXCOR without MDW. These results indicate that MDW improves the performance of SBXCOR, and the equivalent SNR improvement is about 6dB at SNR 0dB. The αs for reference and test patterns were 0.5 and 0.0 respectively. It indicates that the Mexican hat weighting centered at CF is important to improve the robustness against the white noise. Furthermore, as shown in Figure 6, SBXCOR with MDW outperforms SGDS and MFCC below SNR 10dB.

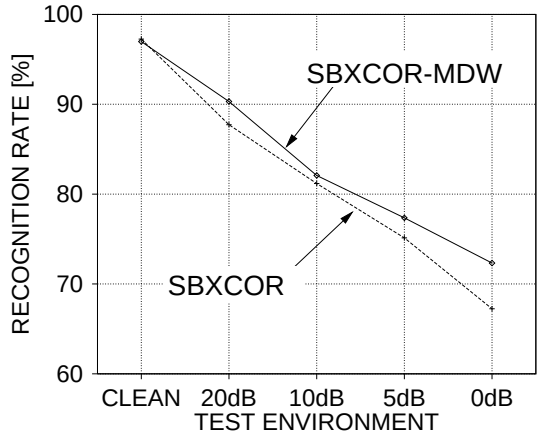

Figure 7. Comparison with SBXCOR with and without MDW(Computer Room Noise).

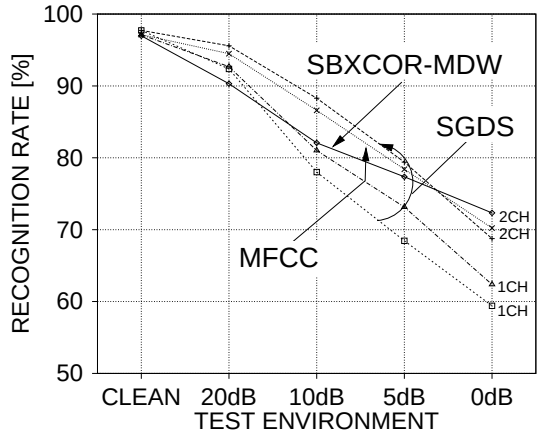

Figure 8. Comparison with SBXCOR with MDW, SGDS and MFCC(Computer Room Noise).

3.2.2. The Robustness Against Computer Room Noise

Figure 7 shows the best recognition rates of SBXCOR(Q=2.0) processed by MDW and SBXCOR without MDW. As shown in the figure, MDW also improves the performance of SBXCOR for the computer room noise. However, the equivalent SNR improvement was less than the case of the white noise, and it was about 3dB at SNR 0dB. This result indicates that the robustness of SBXCOR against noise depends on the degree of correlation between noises in each channel. The αs for reference and test patterns were 0.7 and 0.4 respectively. It indicates that narrower frequency resolution is important under the computer room noise than under white noise.

Furthermore, as shown in Figure 8, although SBXCOR with MDW outperforms SGDS and MFCC extracted from one channel signal below SNR 10dB, it only performs better those extracted two-channel-summed signal at SNR 0dB. As shown in these results, in the case of the simulated acoustic conditions that the speech signals in each channel are perfectly synchronized, the performance of SBXCOR with MDW is not necessary better than those of SGDS and MFCC extracted from the two-channel-summed signal. Such a situation, however, is not realistic. Therefore, it is necessary to investigate the performance when the speech signal in each channel are not synchronized.

4. SUMMARY

In this paper, we introduced the multi-delay weighting(MDW) processing in SBXCOR analysis so as to utilize more periodicities included in binaural signal. The experimental results using a DTW word recognizer are summarized as follows:(1) the MDW processing improves the performance of SBXCOR for both of white noise and a computer room noise, (2) for white noise, SBXCOR with MDW performs significantly better than SGDS and MFCC extracted from monaural and binaural signals, (3) for the computer room noise, SBXCOR with MDW outperforms SGDS and MFCC extracted from monaural signal below SNR 10dB, but it only outperforms SGDS and MFCC extracted from binaural signal at SNR 0dB. From these results, we conclude that MDW processing is effective, but the degree of improvement depends on the kind of noise.

REFERENCES

[1] S. Kajita and F. Itakura: "Speech analysis and speech recognition using subband-autocorrelation analysis", J. Acoust. Soc. Jpn.(English), **15**, 5, pp. 329–338 (1994).

[2] S. Kajita, K. Takeda and F. Itakura: "Subband-crosscorrelation analysis for robust speech recognition", Proc. of ICSLP, Vol. 1, pp. 422–425 (1996).

[3] M. P. DeSimio, T. R. Anderson and J. J. Westerkamp: "Phoneme recognition with a model of binaural hearing ", IEEE Trans. on Speech and Audio Processing, **4**, pp. 157–166 (1996).

[4] M. Bodden and K. Rateitschek: "Noise-robust speech recognition based on a binaural auditory model", Proc. of Workshop on the Auditory Basis of Speech Perception, pp. 291–296 (1996).

[5] R. D. Patterson, T. R. Anderson and K. Francis: "Binaural auditory images and a noise-resistant, binaural auditory spectrogram for speech recognition", Proc. of Workshop on the Auditory Basis of Speech Perception, pp. 245–252 (1996).

[6] S. Kajita and F. Itakura: "SBCOR spectrum taking autocorrelation coefficients at integral multiples of 1/CF into account", Proc. of ICSLP, Vol. 3, pp. 1051–1054 (1994).

[7] F. Itakura and T. Umezaki: "Distance measure for speech recognition based on the smoothed group delay spectrum", Proc. of ICASSP, Vol. 3, pp. 1257–1260 (1987).

[8] H. Singer, T. Umezaki and F. Itakura: "Low bit quantization of smoothed group delay spectrum for speech recognition", Proc. of ICASSP, Vol. 2, pp. 761–764 (1990).

[9] S. B. Davis and P. Mermelstein: "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on Acoustics, Speech and Signal Processing, **ASSP-28**, pp. 357–366 (1980).