SUBBAND-CROSSCORRELATION ANALYSIS FOR ROBUST SPEECH RECOGNITION

Shoji Kajita, Kazuya Takeda and Fumitada Itakura

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

ABSTRACT

This paper describes subband-crosscorrelation (SBXCOR) analysis using two channel signals. The SBXCOR analysis is an extended signal processing technique of subband-autocorrelation (SBCOR) analysis that extracts periodicities present in speech signals. In this paper, the performance of SBXCOR is investigated using a DTW word recognizer, under simulated acoustic conditions on computer and a real environmental condition. Under the simulated condition, it is assumed that speech signals in each channel are perfectly synchronized while noises are not correlated. Consequently, the effective signal-to-noise ratio of the signal generated by simply summing the two signals is raised about 3dB. In such a case, it is shown that SBXCOR is less robust than SBCOR extracted from the two-channel-summed signal, but more robust than the conventional one-channel SBCOR. The resultant performance was much better than that of smoothed group delay spectrum and mel-frequency cepstral coefficient. In a real computer room, it is shown that SBXCOR is more robust than the two-channel-summed SBCOR.

1. INTRODUCTION

For practical speech recognition systems, room noise and room reverberation significantly influence recognition performance. On the other hand, in real human speech recognition, such influences can be compensated by the auditory system. Our aim is to develop a practical front-end system for speech recognition using available knowledge of human auditory processing.

Based on the importance of periodicities in the auditory nerve firing, shown in the auditory modeling proposed by Seneff[1] and Ghitza[2], we have proposed subband-autocorrelation (SBCOR) analysis, and applied it to speech recognition. The experimental results show that SBCOR spectrum performs equally as well as the smoothed group delay spectrum under clean conditions, and much better than it under heavy noise conditions[3, 4, 5].

In this paper, subband-crosscorrelation(SBXCOR) analysis is proposed in order to improve the robustness of SBCOR. In the SBXCOR analysis, the crosscorrelation coefficients of two input signals recorded by two microphones are used instead of the autocorrelation coefficients used in SBCOR.

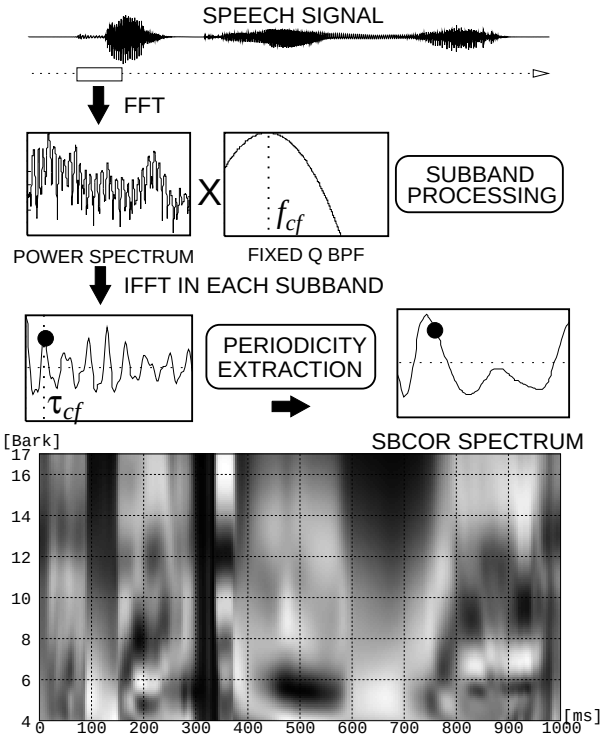


Figure 1: SBCOR analysis.

This paper is constructed as follows. The following section reviews SBCOR analysis and describes the proposed SBXCOR analysis in detail. Sections 3 and 4 investigate the robustness under simulated acoustic conditions on computer and a real environmental condition, respectively. Section 5 concludes the whole paper.

2. SUBBAND-CROSSCORRELATION ANALYSIS

2.1. Subband-Autocorrelation Analysis

SBCOR analysis is based on filter bank and autocorrelation analysis, and aims to extract periodicities included in speech signals. The importance of such information for speech recognition has been shown by Seneff and Ghitza in the research of auditory modeling[1, 2].

Figure 1 shows an implementation of SBCOR analysis used in this paper. It calculates $\{S_i(n), i = 1, \cdots, N\}$ of the autocorrelation coefficient at the lag τ_{cf_i}, which is associated with the $f_{cf_i}^{-1}$, of each subband signal passed through the filter bank $\{H_i(f), i = 1, \cdots, N\}$.

$$S_i(n) = \frac{R_i(\tau_{cf_i}, n)}{R_i(0, n)},\ \tau_{cf_i} = f_{cf_i}^{-1}$$

$$R_i(\tau, n) = \int_{-f_n}^{f_n} \mid H_i(f) \mid^2 X(f, n) \cos 2\pi f \tau df,$$

where $R_i(\tau, n)$ and $X(f, n)$ are *i*th subband autocorrelation function and the power spectrum of *n*th analysis frame, respectively, and f_n is Nyquist frequency. The $\{S_i(n), i = 1, \cdots, N\}$ is interpreted as a "spectrum" and referred to as "SBCOR spectrum".

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[3, 4]. In the following experiments, the filter bank consists of 16 fixed Q gaussian bandpass filters(BPF) defined by

$$\mid H_i(f) \mid^2 = \begin{cases} \mathrm{e}^{-2C_i(f - f_{cf_i})^2},\ f \geq 0 \\ \mid H_i(-f) \mid^2,\ f < 0, \end{cases}$$

where

$$C_i = \frac{2Q^2 \ln 2}{f_{cf_i}^2}.$$

2.2. SBXCOR Analysis

SBCOR analysis is signal processing based on "autocorrelation" of a speech signal so as to extract periodicity in terms of the inverse of the center frequency. As seen in the auditory system, however, binaural signal processing seems to be more important in the real environment. Therefore, in order to improve the performance of speech recognition, we extend SBCOR analysis so that the autocorrelation analysis is replaced by crosscorrelation analysis, and refer to it as "subband-crosscorrelation" analysis, or SBXCOR analysis in the abbreviated form.

The robustness of SBXCOR against noise can be explained as shown in Figure 2. Since the speech signals recorded by two microphones, which is 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.

2.3. Implementation of SBXCOR Analysis

SBXCOR analysis is implemented using FFT in this research. The *i*th SBXCOR coefficient for *n*th analysis frame is calculated as follows:

$$Sc_i(n) = \frac{\mathrm{Re}\, R_{ixy}(\tau_{cf}, n)}{\sqrt{R_{ixx}(0, n) R_{iyy}(0, n)}},\ \tau_{cf} = f_{cf}^{-1}$$

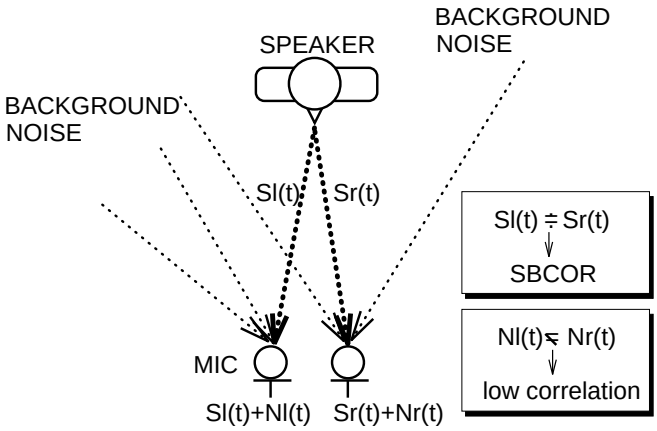


Figure 2: Concept of SBXCOR analysis.

$$R_{ixy}(\tau_{cf}, n) = \int_{-f_n}^{f_n} \mid H_i(f) \mid^2 F_x(f, n) F_y^*(f, n) \times \mathrm{e}^{j2\pi f \tau_{cf}} df,$$

$$R_{ixx}(0, n) = \int_{-f_n}^{f_n} \mid H_i(f) \mid^2 \mid F_x(f, n) \mid^2 df,$$

$$R_{iyy}(0, n) = \int_{-f_n}^{f_n} \mid H_i(f) \mid^2 \mid F_y(f, n) \mid^2 df,$$

where $R_{ixx}(\tau, n)$, $R_{iyy}(\tau, n)$ and $R_{ixy}(\tau, n)$ are the autocorrelations and crosscorrelation function of *i*th subband signal respectively, and $F_x(f, n)$ and $F_y(f, n)$ are FFT spectrum of $x(t, n)$, $y(t, n)$ respectively.

3. EVALUATIONS UNDER SIMULATED ACOUSTIC CONDITIONS

In this section, in order to investigate the upper-bound performance of SBXCOR, we conduct recognition experiments under simulated acoustic conditions. We assume that speakers utter just in front of the two microphones. This assumption means:

1. the speech signals recorded by two microphones are perfectly synchronized,
2. the reverberation in the real environment is ignored,
3. noises are not correlated between the signals recorded by two microphones.

Of course, this assumption is not realistic in the real environment where speech recognizer is used. However, since there are a lot of uncontrollable factors in the real environment, we start the investigation under this assumption.

3.1. Experimental Conditions

The above condition is implemented on computer by adding gaussian white noise to speech signals and processing such signals. In the experiment, using DTW word recognition, we compare the robustness of SBXCOR analysis with that of SBCOR. As a further

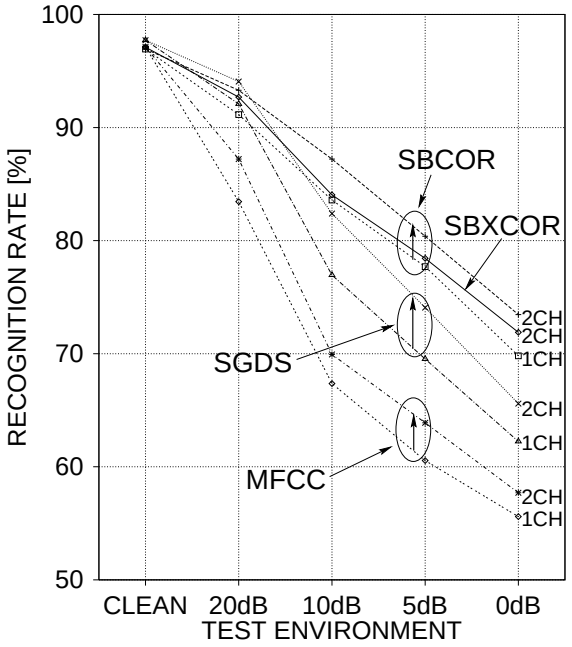


Figure 3: Recognition results under simulated acoustic conditions of SBXCOR, SBCOR, SGDS and MFCC. The arrows show the improvement by using two-channel-summed signal.

reference, we also compare SBXCOR with the smoothed group delay spectrum[6, 7] and the mel-frequency cepstral coefficient[8] extracted from one-channel signal by simply summing the two signals.

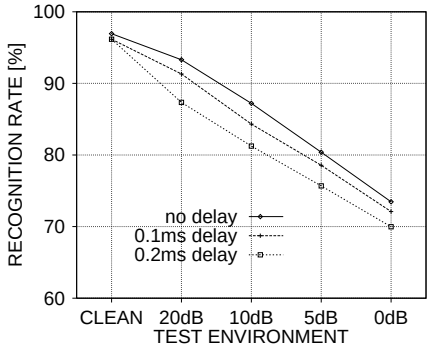
DTW word recognizer. A standard DTW speaker-dependent isolated word recognizer is used. The recognition task is a 68 pair discrimination[6]. Each pair is a phonetically similar city name pair, selected from a 550 Japanese city name database recorded twice by 5 Japanese male speakers. The first set is used as the reference pattern and the second set, which was spoken a week later, is used as the test pattern.

Generation of two-channel signals. In order to simulate the environment described above, two gaussian white noises generated by changing seed are added to the speech database. The global signal-to-noise ratios(SNRs) used in the test phase are 20, 10, 5 and 0dB.

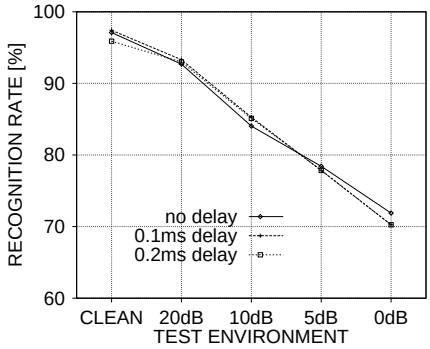
Generation of one-channel signal by simply summing the two signals. The two-channel-summed signals are generated by simple-summation of the above two signals. By doing this processing, the effective SNR improvement of the two-channel-summed signals is about 3dB.

Smoothed group delay spectrum(SGDS). SGDS has been shown to be robust against noise, and it is calculated as the derivative of phase of a *p*th order all pole filter that has smoothed poles[6, 7]. In order to compare the performance of SBCOR with that of SGDS under exactly the same conditions, the analysis frequency points of SGDS were chosen to be the same as the center frequencies of SBCOR.

Mel-filterbank cepstral coefficient(MFCC). MFCC is commonly used speech feature in speech recognizer[8]. In recent research, the noise robustness of MFCC is almost the same as the auditory models proposed by Seneff and Ghitza[9]. In this experiment, MFCC is calculated using a 28 triangular shape mel-filterbank.



(a) SBCOR extracted from two-channel-summed signal



(b) SBXCOR

Figure 4: Recognition results when the speech signal in each channel are not synchronized.

SBCOR and 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 simulated under conditions precisely, two-times oversampling and polynomial interpolation were used. The center frequencies of the BPFs are equally spaced on the Bark scale between 4 and 17 Bark.

Common analysis conditions. The analysis frame length and shift are 20ms and 10ms, respectively. The dimension of each feature is 16. The sampling rate is 10 kHz.

3.2. Experimental Results

The recognition rates of SBXCOR, SBCOR, SGDS and MFCC are shown in Figure 3. In SBXCOR and SBCOR analysis, the best Qs are 2.0 and 1.5 respectively.

These results are summarized to three points. First, SBXCOR is more robust than the conventional one-channel SBCOR under any test conditions. However, the improvement was less than about 2%. Second, the performance of SBXCOR is less than SBCOR extracted from the two-channel-summed signal. Third, SBXCOR performs better than SGDS and MFCC even if two-channel-summed signals are used.

3.3. Discussions

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 is worse than that of SBCOR 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, that is, the direction of speaker cannot be estimated precisely. Here, we show the same experimental results as the above when a 0.1ms or 0.2ms sample delay exists. When the distance between the two microphones is 20cm, the sound velocity is 340m/s and the sound propagates as a plain wave, the delay results in the incorrect direction estimation of about 10 degrees for a 0.1ms delay and about 20 degrees for a 0.2ms delay.

As shown in Figure 4, SBXCOR is more robust against such delays while SBCOR extracted from the two-channel-summed signal degrades significantly. As a result, we can expect that SBXCOR can be robust under more realistic acoustic conditions.

4. EVALUATIONS IN A REAL ENVIRONMENT

Finally, we investigate the performance of SBXCOR in a real environment. This preliminary evaluation of SBXCOR is performed by the same DTW word recognition system in section 3.

4.1. Database Recording

The 68 city name pairs were output from a speaker, and recorded using a dummy head(B&K type 4128 head-torso simulator) in a sound proof room and a computer room. In this study, the dummy head was placed exactly in front of the speaker. The SNR was changed so that the sound level of the 1kHz sinusoidal wave is about 90 dBA in the sound proof room, and about 90dBA, 79dBA and 70dBA in the computer room. The noise level in each room was about 28dBA and 60dBA respectively. The resultant SNRs were about 15dB, 8dB and 0dB.

4.2. Experimental Results

Figure 5 shows the recognition rates of SBXCOR and SBCOR. SBCOR was extracted from the summed signal of the left and right ear signals, while SBXCOR was extracted from the left and right ear signals. The results showed that the performance of SBXCOR(Q:2.5) is about 3% higher than that of SBCOR(Q:1.5) under noisy conditions. Thus, we can conclude that SBXCOR is more robust than the conventional SBCOR in the real environment.

5. SUMMARY

In this paper, we proposed subband-crosscorrelation analysis and investigated using a DTW word recognizer, under the simulated acoustic condition on computer and a real environmental condition. Under the simulated condition, we clarified that SBXCOR is less robust than SBCOR extracted from the two-channel-summed signal, but more robust than the conventional one-channel SBCOR.

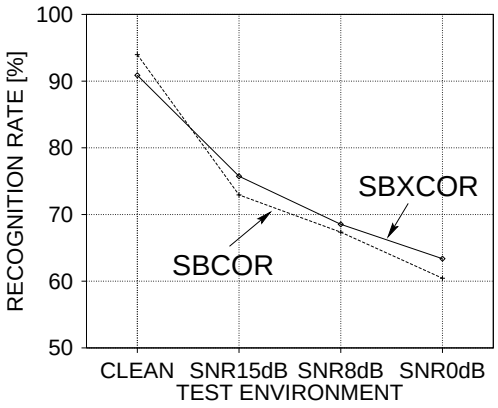


Figure 5: Recognition results in a real computer room

The resultant performance was much better than that of smoothed group delay spectrum and mel-frequency cepstral coefficient. In a real computer room, we showed that SBXCOR is more robust than the two-channel-summed SBCOR.

6. REFERENCES

1. S. Seneff: "A joint synchrony/mean-rate model of auditory speech processing", J. Phonetics, **16**, pp. 55–76 (1988).
2. O. Ghitza: "Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment", J. Phonetics, **16**, pp. 109–123 (1988).
3. 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).
4. S. Kajita and F. Itakura: "Subband-autocorrelation analysis and its application for speech recognition", Proc. of ICASSP, Vol. II, pp. 193–196 (1994).
5. S. Kajita and F. Itakura: "Robust speech feature extraction using SBCOR analysis", Proc. of ICASSP, Vol. 1, pp. 421–424 (1995).
6. 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).
7. 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).
8. S. B. Davis and P. Mermelstein: "Comparison of parametric representations for monosyllabic word recognition in countinuously spoken sentences", IEEE Trans. on Acoustics, Speech and Signal Processing, pp. 357–366 (1980).
9. C. R. Jankowski Jr., H.-D. H. Vo and R. P. Lippmann: "A comparison of signal processing front ends for automatic word recognition", IEEE Trans. on Speech and Audio Processing, **3**, pp. 286–294 (1995).