Acoustic Analysis of Dialog Speech: Subband-Autocorrelation Analysis for Robust Speech Recognition

Fumitada ITAKURA and Shoji KAJITA
Graduate School of Engineering, Nagoya University Furo-cho 1, Chikusa-ku, Nagoya 464-01, JAPAN e-mail: ita@nuee.nagoya-u.ac.jp

1. Introduction

In dialog speech recognition systems, the speech analysis part is an essential front-end using the acoustic signal processing. Since the acoustic features lost there cannot be easily recovered in later stages, extraction of effective features from acoustic signals is one of the most important problems in dialog speech recognition. Therefore, intensive research on acoustic analysis is required to achieve a robust speech recognition in a realistic acoustic environment.

We have proposed subband-autocorrelation(SBCOR) analysis as a signal processing technique. This SBCOR analysis has been developed so as to extract periodicities associated with the delay time equal to inverses of the center frequencies. The SBCOR has been shown to be robust under the multiplicative signal-dependent white noise that has constant SNRs at any points.

In this paper, we describe (1) an investigation to what extent the SBCOR analysis is robust against more realistic adverse environments; existence of additive noise and waveform distortion, (2) an elaborated SBCOR analysis, which is calculated from the weighted sum of autocorrelation coefficients at multiples of 1/center-frequency, (3) subband-crosscorrelation (SBXCOR) analysis in order to improve the robustness of SBCOR.

2. SBCOR Analysis

The SBCOR analysis is based on subband 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 field of auditory modeling[1][2]. In those models, subband spectral analysis and detection of periodicity or synchronicity are combined, based on the auditory periphery.

2.1 SBCOR implementation

The SBCOR analysis used in this paper is implemented by the fast Fourier transform(FFT) for the subband filtering and the calculation of autocorrelation coefficients(Figure 1). For an analysis frame, the i-th subband filtering is conducted by multiplying the power spectrum by the characteristics of the subband filter. The subband autocorrelation coefficients are calculated by the inverse FFT of the subband power spectrum. Then, the autocorrelation coefficient at the lag, which is associated with the inverse of the center frequency(CF), is picked up. The array is interpreted as "spectrum" and referred to as "SBCOR spectrum" for the frame.

It has been shown to be suitable for speech recognition under noisy conditions that the filter bank is composed of fixed Q filters whose center frequencies are equally spaced on the Bark scale[3]. In this paper, the filter bank consists of 128 or 16 band pass filters(BPFs) in analysis examples of Figures 5, and the following recognition respectively.

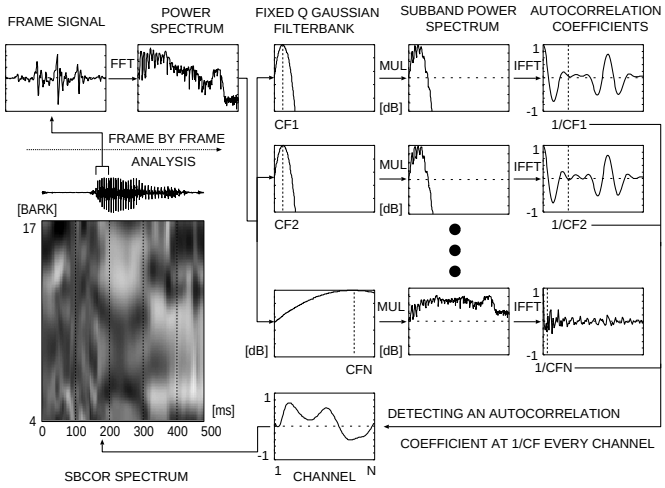

Figure 1: Flow diagram of SBCOR analysis

2.2 Robustness of SBCOR against the additive noise

The robustness of the SBCOR against the additive noise can be depicted in Figure 2. Suppose that the noise is white and the speech signal has a spectrum whose first formant coincides with the center frequency of the BPF, as shown in the middle plot. The subband-autocorrelation coefficients for each signal can be derived as shown in the right-upper and right-lower plots. If the noise is additive, the influence of noise in detecting the formant is weak at lag 1/CF than at lag 0. Hence, SBCOR analysis is expected to be robust against the additive noise when the periodicity of the noise is low.

3. Robustness against Noises

In this section, we investigate the robustness against additive noises such as the Gaussian white noise, the human speech like noise or the bubble noise in multiple telephone cross talk, and a computer room noise. In order to discuss the effect of back-end pattern recognition methods, the investigation is conducted using DTW word recognizer and HMM phoneme recognizer.

3.1 Experimental conditions

3.1.1 Noises

As for the adding noise, we used three different types of noise; Gaussian white noise, human speech-like noise and computer room noise. The three noise signals are generated by the following ways.

[Gaussian white noise]

It was generated using a Gaussian random-number generator on computer.

[Human speech-like noise(HSN)]

The HSN is a kind of bubble noise simulating that a lot of people are talking simultaneously, as in the cocktail party[7]. The HSN was generated by overlapping cyclically a long speech signal to a fixed-length buffer. The long speech signal was prepared by concatenating 3,200 phrases of the ASJ continuous speech corpus. The signal is about 4.6 hours long. The fixed-length buffer is three seconds long.

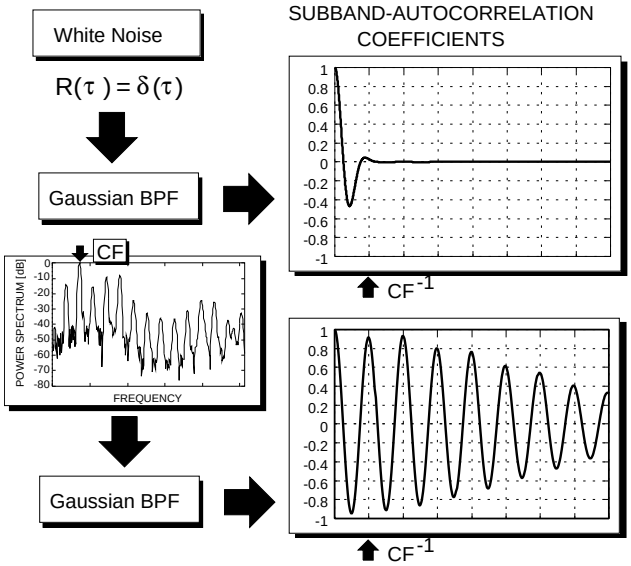

Figure 2: An explanation of the robustness. The CF means the center frequency of the Gaussian BPF.

[Computer room noise]

We recorded the noise in a computer room(8 workstations and their peripherals such as HDD and CD-ROM drives).

3.1.2 Recognition systems

The robustness of SBCOR against these types of noise is evaluated by the following speaker-dependent recognition experiments using DTW and HMM. The global signal-to-noise ratios(SNRs) used in the test phase are 20, 10, 5 and 0dB. Moreover, in order to show the effectiveness under noisy conditions, the performance of the SBCOR spectrum is compared with that of the smoothed group delay spectrum(SGDS)[4][5].

[DTW word recognition]

A standard DTW speaker-dependent isolated word recognizer is used. The recognition task is a 68 word-pair discrimination[5]. 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 sampling rate is 10 kHz. 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. The recognition rate is given by the average for the 5 speakers.

[HMM phoneme recognition]

The task is the speaker-dependent 23 phoneme(a,i,u,e,o,b,d,g,m,n,N,p,t,k,s,h,r,y,w,z,ts,ch,sh) recognition. Each phone model has 3 states of 7 component mixture Gaussians. The parameter estimation was performed using the 2620 even-numbered words from the ATR Japanese 5240 word speech database(two male and two female speakers). The speech data for tests were collected from the odd-numbered 2620 words. The sampling rate is 10 kHz. The recognition rate is given by the average for the 4 speakers.

3.2 Recognition Results

The performances of SBCOR and SGDS are shown in Figure 3 for the DTW word recognition and Figure 4 for the HMM phoneme recognition, respectively. In comparing the performances between SBCOR and SGDS, the parameters Q and gamma that attained the best results under all conditions are selected.

As shown in Figure 3 for the Gaussian white noise, the SBCOR spectrum performs as well as the SGDS under clean conditions, and better than it under noisy conditions. At the SNR 10dB or lower, the recognition rates are improved about 7 to 9%. Therefore, it indicates that the robustness of SBCOR is considerably better than that of the SGDS. However, for the case of human speech like noise and computer room noise, the differences of performance between SBCOR and SGDS are small. Therefore, the robustness of SBCOR against human speech like-noise and computer room noise is comparable to that of SGDS.

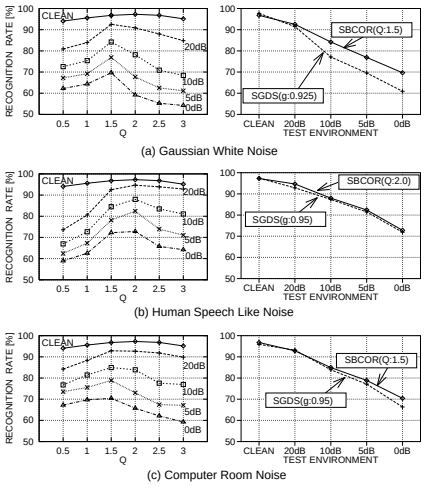

Figure 3: Recognition Results (DTW)

In the experiments of HMM phoneme recognition, almost the same as those of the DTW word recognition were obtained. As shown in Figure 4, for the Gaussian white noise, the performances of SBCOR are about 4 to 8% better than those of SGDS. For the case of the human speech like noise and computer room noise, the robustness of SBCOR is comparable to that of SGDS.

From these results, robustness of SBCOR analysis against additive noise is clarified.

4. Robustness against Waveform Distortion

Next, we investigate the robustness of SBCOR against waveform distortion using zero-crossing signal of speech.

4.1 Zero-crossing Distortion

Zero-crossing distortion is a hard limiting distortion that is caused, for example, when a signal is clipped due to the excessive input power. It is often called infinite peak clipping distortion. In the zero-crossing distorted signal, the amplitude information of the original signal is lost, and only the zero-crossing points are preserved, as shown in Figure 5. Such zero-crossing signals are still intelligible for human, but the performance with conventional speech recognizers deteriorates significantly as shown later. Figure 5 shows SGDS and

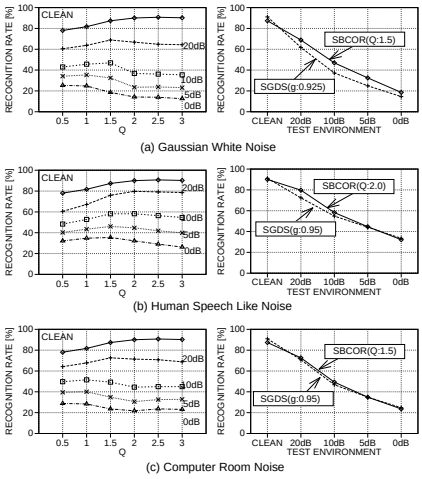

Figure 4: Recognition Results (HMM)

SBCOR spectrum of the original signal and the zero-crossing signals. As can be seen, the zero-crossing distortion influences the formant structure significantly in SGDS spectrum. The SBCOR analysis, however, is stable for such a distortion.

4.2 Recognition Experiments and Results

The same standard DTW speaker-dependent isolated word recognizer as in the previous section is used here.

Although the performance of SGDS(with pole smoothing parameter gamma=0.925) for zero-crossing signals deteriorates significantly, the performance of SBCOR is about 19%(Q=1.0) and 9%(Q=1.5) higher than that of SGDS(See Table 1). These results indicate that the SBCOR spectrum is much more robust against the zero-crossing distortion than the SGDS.

5. Multi-Delay Weighting in SBCOR Analysis

An elaborated SBCOR spectrum is calculated from the weighted sum of autocorrelation coefficients at multiples of 1/center-frequency[8]. This is based on the ringing phenomenon in firing patterns of an auditory nerve fiber. This processing is referred as Multi-Delay Weighting(MDW). The evaluation is performed by the same DTW word recognition system in section 3 under the signal dependent white noise. The results indicate that the SBCOR spectrum with MDW is more robust against noise than the conventional SBCOR spectrum, as shown in Figure 6.

6. Subband-Corsscorrealtion Analysis

Finally, Subband-Crosscorrelation(SBXCOR) analysis is proposed in order to improve the robustness of SBCOR. In the SBXCOR analysis, crosscorrelation coefficients of two input signals recorded by two microphones are used instead of autocorrelation coefficients in SBCOR. Assume that speech is uttered right in front of the two microphones. Since the speech signals recored by the two microphones have almost same amplitude and phase, SBXCOR analysis extracts the same feature as SBCOR. On the other hand, the background

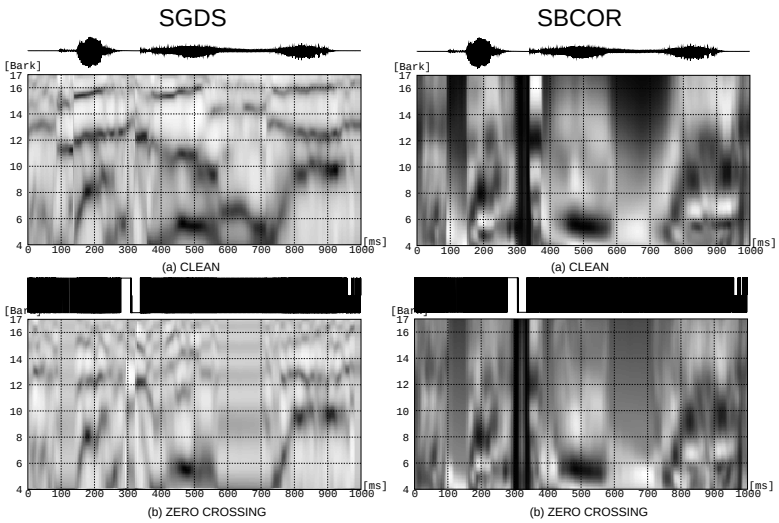

Figure 5: Analysis examples of SGDS and SBCOR for clean(a) and zerocrossing(b) signals. The utterance is "bakuonga" in Japanese, spoken by a female speaker. The length and shift of the analysis window are 32ms and 4ms respectively.

Table 1: Recognition rates(%).

	CLEAN	Zero-Crossing
SBCOR(Q1.0)	95.6	87.8
SBCOR(Q1.5)	96.8	77.6
SGDS	97.2	68.5

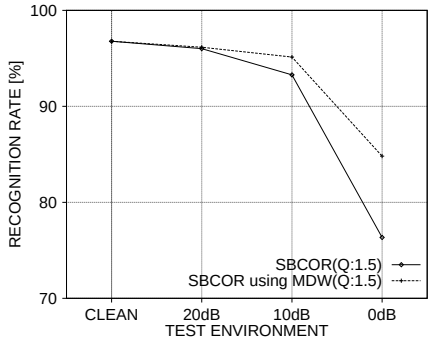

Figure 6: Improvement by MDW.

noise is suppressed due to the low correlation between the two input signals(Figure 7). Hence, SBXCOR analysis is expected to be more robust against noise than SBCOR.

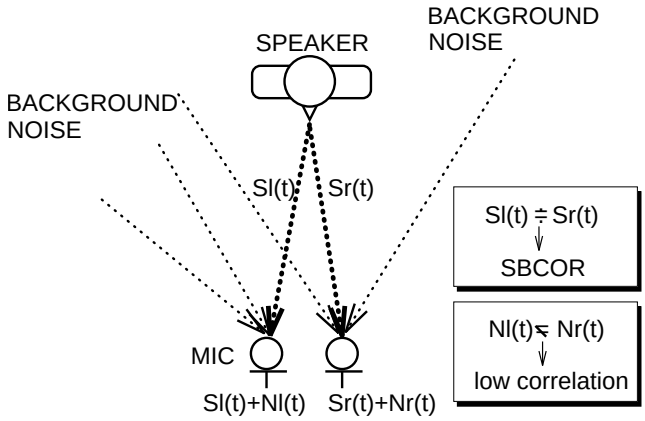

Figure 7: Concept of SBXCOR.

The preliminary evaluation of SBXCOR is performed by the same DTW word recognition system in section 3. The database was recorded by a dummy head in a computer room. The SBCOR spectrum was extracted from the superimposed signal of the left and right ear signals, while the SBXCOR spectrum 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, as shown in Figure 8. Thus, we can conclude that SBXCOR can be more robust than SBCOR.

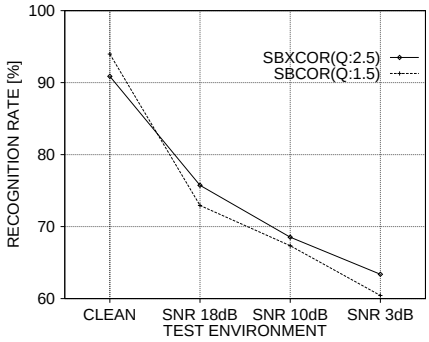

Figure 8: Comparison of SBCOR and SBXCO

7. Conclusions

In this paper, we investigated the robustness of SBCOR analysis, together with developing the elaborated processing, such as MDW and SBXCOR. In our future works, we intend to investigate the robustness against the other noises. Our plans also include the further investigations of SBXCOR analysis based on the cross correlation of binaural signals.

References

[1] S.Seneff: "A joint synchrony/mean-rate model of auditory speech processing", J. Phonetics, vol.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, vol.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), vol.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] 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).

[6] 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).

[7] D.Kobayashi, S.Kajita, K.Takeda and F.Itakura: "Extracting speech features from human speech-like noise", Tecnical Report of IEICE, SP95-105, pp.85-92 (1995), in Japanese.

[8] 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).

Keywords: SBCOR analysis, waveform distortion, noise, DTW, HMM