ROBUST SPEECH FEATURE EXTRACTION USING SBCOR ANALYSIS

Shoji KAJITA and Fumitada ITAKURA
School of Engineering, Nagoya University
Furo-cho 1, Chikusa-ku, Nagoya, 464-01 JAPAN
Email: {kaji *or* ita}@itakura.nuee.nagoya-u.ac.jp

ABSTRACT

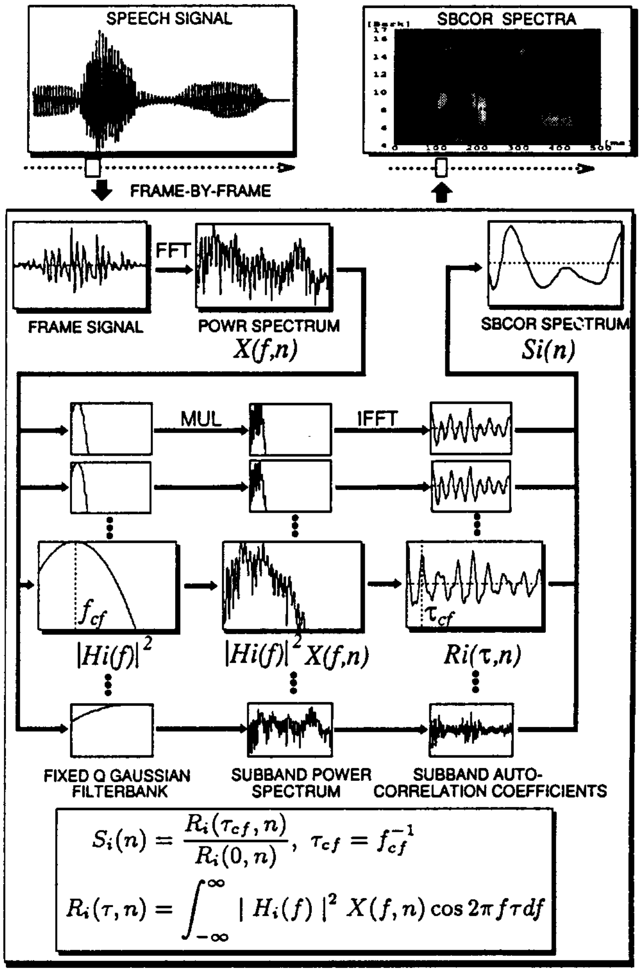
This paper describes to what extent the subband-autocorrelation(SBCOR) analysis is robust against waveform distortion and noises. The SBCOR analysis, which has been already proposed, is a signal processing technique based on subband processing and autocorrelation analysis so as to extracts periodicities present in speech signals. First, it is shown that SBCOR is robust against severe waveform distortions such as zero-crossing. Although the zero-crossing distortion deteriorates the performance of conventional recognition systems, such distorted signals are still intelligible for humans. The experimental results using a DTW word recognition show that the SBCOR(Q=1.0) performs about 19% higher than smoothed group delay spectrum(SGDS), when the test signals are distorted by zero-crossing. Second, it is shown that SBCOR is more robust against multiplicative signal-dependent white noise, Gaussian white noise, and a human speech noise than SGDS. The validity of the SBCOR is larger when the noise is white than when the noise is the human speech noise.

1. INTRODUCTION

In speech recognition systems, the speech analysis part is the front end for the acoustic environment. Since the acoustic features lost there cannot be recovered in later stages, what features and how to extract them from acoustic signals is one of the most important problems in speech recognition. It seems that insufficient investigations for the problem is the reason why good recognition performance has been accomplished only under "laboratory conditions". Nowadays, since some powerful recognition algorithms such as HMM are available, to overcome the problem above is the best way to improve robustness in speech recognizers[1].

In order to tackle the problem, we have proposed a new signal processing technique based on subband processing and autocorrelation analysis, namely, subband-autocorrelation(SBCOR) analysis[2, 3]. This SBCOR analysis has been developed so as to extract periodicities associated with the inverses of the center frequencies. The basic idea comes from the auditory models proposed by Seneff and Ghitza[4, 5]. 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 investigate to what extent the SBCOR analysis is robust against zero-crossing distortion and three types of noise, using another implementation of the SBCOR with a DTW word recognizer. Besides, it is also shown that SBCOR is robust under constant SNRs at any points using an HMM based phoneme recognizer.



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

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

Figure 1: The flow diagram of the SBCOR analysis.

2. SBCOR ANALYSIS

2.1. Method

The 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[4, 5].

Figure 1 shows an implementation of the SBCOR analysis used in this paper. The SBCOR analysis calculates an array $\{S_i(n), i = 1, \cdots, N\}$ of the autocorrelation coefficient at the lag τ_{cf}, which is associated with the f_{cf}^{-1}, of each

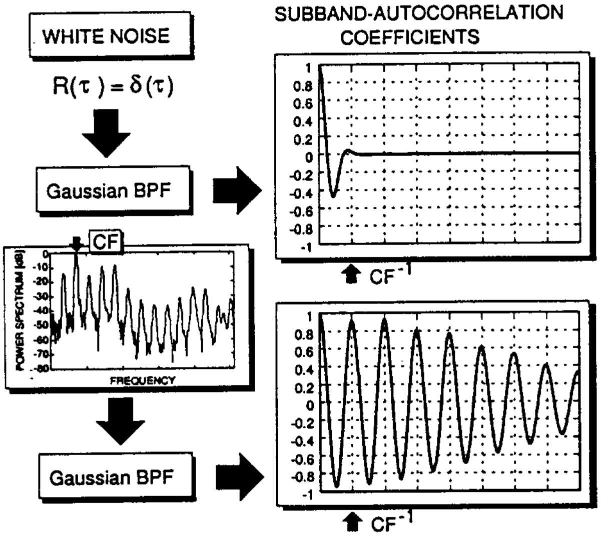


Figure 2: The explanation of the robustness. The CF means the center frequency of the Gaussian BPF. If the white noise is additive, the influence of the noise in detecting the formant's component is less at lag CF^{-1} than at lag 0.

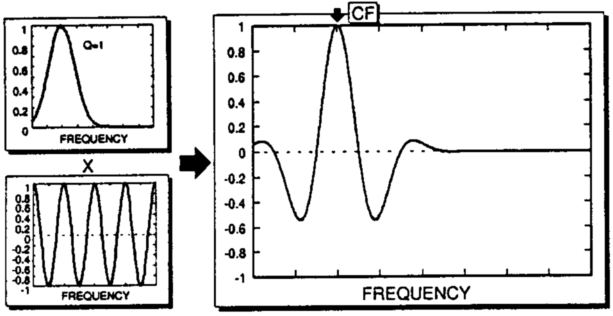


Figure 3: In frequency domain, the SBCOR analysis is interpreted as a weighted sum of power spectrum.

subband signal passed through the filter bank $\{H_i(f), i = 1, \cdots, N\}$. The array $\{S_i(n), i = 1, \cdots, N\}$ is interpreted as a "spectrum" and referred to as "SBCOR spectrum". As for the filter bank, a fixed Q one whose center frequencies are equally spaced on the Bark scale has been shown to be suitable for speech recognition under noisy conditions so far[2, 3]. In this paper, the filter bank consists of 128(in shown analysis examples) or 16(in recognition experiments) fixed Q gaussian bandpass filters(BPF) defined by

$$| H_i(f) |^2 = e^{-2C(f-f_{cf})^2} \quad | f | \geq 0, \; C = \frac{2Q^2 \ln 2}{f_{cf}^2}. \quad (1)$$

2.2. Why the SBCOR can be robust against noise

The robustness of the SBCOR against noise is explained in Figure 2. Here, suppose that the noise is white and the input signal has the spectrum whose first formant coincide with the center frequency of the BPF, depicted in the middle plot. Then, the subband-autocorrelation coefficients for each signal can be derived as the right-upper and the right-lower plots respectively. Assume that the noise is additive. Then, since the noisy coefficients affect additively the coefficients of the input signal, the influence of noise in detecting the formant is lower at lag CF^{-1} than at lag 0, which corresponds to a classic filter bank case. The extension of the SBCOR, considered the integral multiples of CF^{-1}, has been already described in [6].

Moreover, in the frequency domain, the SBCOR analysis is interpreted as a weighted sum of power spectrum from the equation in Figure 1, as shown in Figure 3, where the weighting function is

$$w(f) = e^{-2C(f-f_{cf})^2} \cos 2\pi f \tau_{cf}. \quad (2)$$

Further consideration in terms of this aspect is beyond the scope of this paper.

3. ROBUSTNESS AGAINST ZERO-CROSSING DISTORTION

In this section, we select zero-crossing distortion as an example of severe waveform distortion, and investigate the SBCOR robustness.

3.1. Zero-Crossing Distortion

We define a zero-crossing distorted signal as follows:

$$y(n) = \begin{cases} a \times \mathrm{sgn}(x(n)) & | x(n) | > 0 \\ 0 & | x(n) | = 0, \end{cases} \quad (3)$$

where $x(n)$ and $y(n)$ are the input signal and the zero-crossing distorted signal, respectively. The gain a is determined so that the power of the input signal is preserved.

Such zero-crossing signals are still intelligible for humans, but the performances of conventional speech recognizers deteriorate significantly. The reason seems to be that the speech features used in conventional recognizers do not always represent enough of the speech information contained in the speech signal. For example, Figure 4 shows speech features of the smoothed group delay spectrum. As can be seen, the zero-crossing distortion influences the formant structure significantly. The SBCOR analysis, however, is stable for such distortion, as shown in Figure 5. The following recognition experiments will quantitatively evaluate the robustness.

3.2. Experimental Conditions

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 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 test signals are distorted by Eq.(3). In extracting speech features, the length and shift of the analysis window are 20ms and 10ms respectively.

Moreover, the performance of the SBCOR spectrum is compared with that of the smoothed group delay spectrum(SGDS), already shown to be robust[7, 8]. The SGDS, as distinct from the SBCOR, is a speech representation based on the group delay characteristic of the speech signal, and is defined as the derivative of the phase for an all pole filter that has smoothed poles. In order to compare the performance of the SBCOR with that of the SGDS under exactly the same conditions, the analysis frequency points are chosen to be the same center frequencies of the SBCOR, which are equally spaced between 4 and 17Bark.

Table 1: Average recognition rate for zero-crossing signals

FEATURE	CLEAN	ZERO-CROSSING
SBCOR(Q1.0)	95.6%	87.8%
SBCOR(Q1.5)	96.8%	77.6%
SGDS	97.2%	68.5%

3.3. Results

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

4. OTHER TYPES OF NOISE

In this section, we show the robustness against three types of noise, namely, the multiplicative signal dependent white noise, the Gaussian white noise, and a human speech noise.

4.1. Three types of noise

The multiplicative signal dependent white noise is defined as follows:

$$s'(n) = s(n)(1 + a \cdot r(n)), \quad z[\text{dB}] = 10 \log_{10} \frac{3}{a^2} \qquad (4)$$

where $s(n)$ is the clean speech signal, $s'(n)$ is the noisy speech signal, a is the relative noise amplitude, z is the desired SNR, and $r(n)$ is an uniform distributed random number between -1 and 1. Since the SNR of the noisy speech signal is constant anywhere, we can demonstrate the quantitative characteristics of the robustness.

The Gaussian white noise is a white noise whose amplitude distribution is Gaussian. We generated it using a Gaussian random-number generator on computer.

Finally, in order to create a noise whose spectrum represents approximately the frequency characteristics of human speech, we added cyclically a long speech signal to a fixed-length buffer. The speech signal was created by concatenating the ATR phoneme balanced Japanese phrases(3,200 phrases, the longest phrase is about 15 seconds) spoken by 30 males and 34 females in the Continuous Speech Corpus for Research edited by the ASJ. The fixed-length buffer is 3 seconds long. Here, we refer to it as a "human speech noise". The power spectrum density is shown in Figure 6. This human speech noise is considered as a pink noise where there is no pitch sensation.

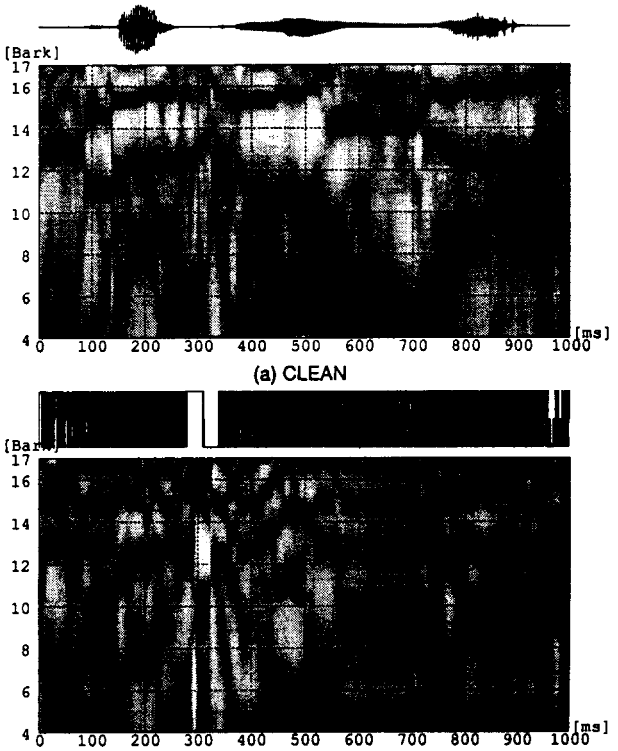
The robustness of the SBCOR against these types of noise is evaluated by following recognition experiments.

4.2. Experimental Conditions

The performed DTW word recognitions are the same as the one used in previous section. The Gaussian white noise and the human speech noise are added to the test signals, based on the global SNR.

4.3. Recognition Results

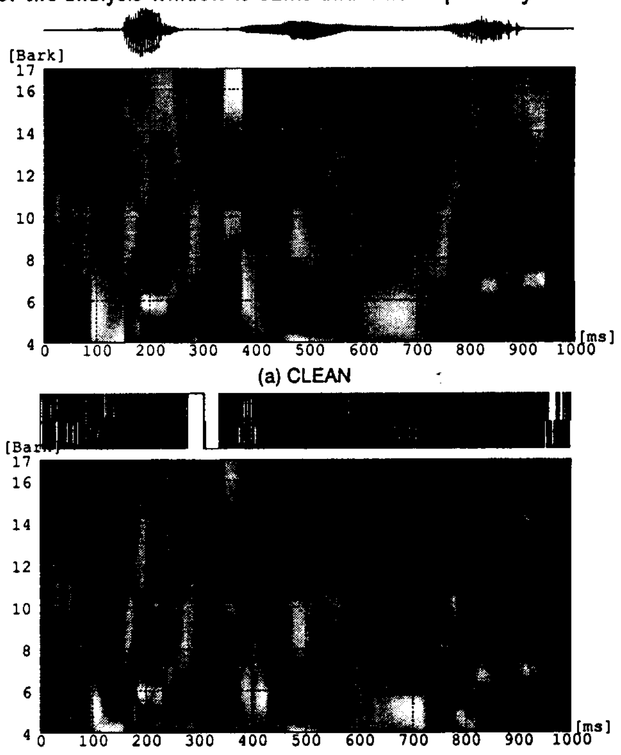
Figure 7 shows the recognition rates of the SBCOR, made as a function of Q(left side), and the comparison with that of the SGDS(right side). As shown in the figure, the SBCOR spectrum performs equally as well as the SGDS under clean conditions, and better than it under noisy conditions, for all noises. Besides, the best Q for the white noises is 1.5, while the best one for the human speech noise is 2.0. The reason seems to be that the noise components at the low frequencies can be attenuated by narrowing the band width.



(a) CLEAN

(b) Zero Crossing

Figure 4: Analysis Examples of the SGDS for clean(a) and zero-crossing(b) signals. The utterance is "bakuonga" in Japanese, spoken by a female speaker. The length and shift of the analysis window is 32ms and 4ms respectively.



(a) CLEAN

(b) Zero Crossing

Figure 5: Analysis Examples of the SBCOR(Q=1.0). The utterance and the analysis conditions is the same as Figure 4.

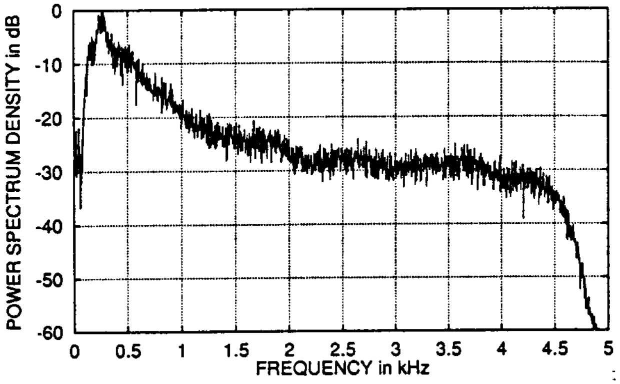


Figure 6: The power spectrum density of the human speech noise estimated by the Blackman-Tukey method. It has a peak at about 250Hz, and the slope is about 10 dB/Oct. The attenuation beyond 4.5kHz is due to the anti-aliasing filter.

Note that the following two facts suggest that the best performance may be related with the shape defined by Eq.(2); (i) the differences between the SBCOR and the SGDS are smaller when Q=1.5 than when Q=2.0, (ii) increasing the Q beyond 1.5 or 2.0 has no advantage.

5. HMM BASED PHONEME RECOGNITION

Finally, we evaluate the robustness at phonemic level.

5.1. Recognizer and Database

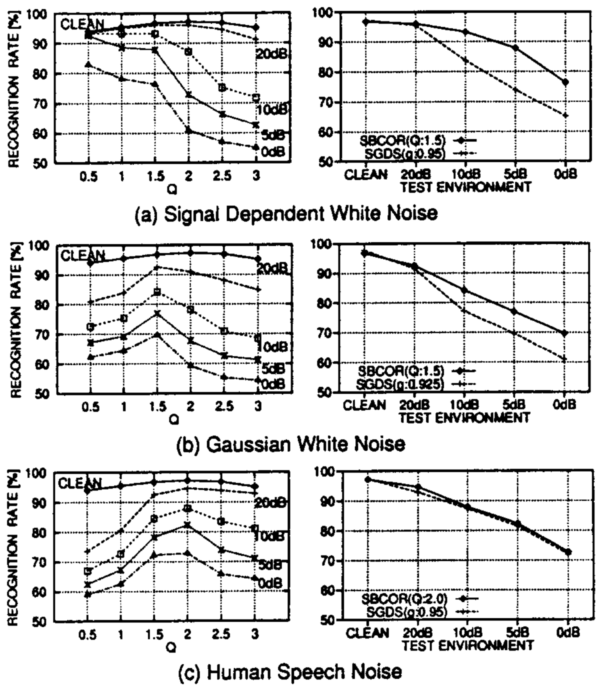
The task is 23 phoneme speaker-dependent recognition for the /a,i,u,e,o,b,d,g,m,n,N,p,t,k,s,h,r,y,w,z,ts,ch,sh/ using HMMs. Each HMM is left-to-right and seven mixture HMM. The parameter estimation was performed using the 2620 even-numbered words in the ATR Japanese 5240 speech database(two male and two female speakers). The speech data for tests were collected from the odd-numbered 2620. The sampling frequency is 10 kHz. To examine the robustness against noise, the multiplicative signal-dependent white noise is added to the database for tests.

5.2. Results

Figure 8 shows the average recognition rates of the SBCOR spectrum, plotted for the SNR of the test database. When the SNR falls, the best Q becomes low gradually. When it is taken into account that the best Q for low SNR is not the best for high SNR and vice versa, the best Q is 1.5. Moreover, although the performance of the SBCOR(Q=1.5) is slightly worse than that of SGDS under clean conditions, the SBCOR performs much better than the SGDS under SNR 20 and 10dB. Under SNR 0dB, since the absolute rate of both are about 10%, which is close to the chance level(about 6%), the evaluation makes no sense.

6. CONCLUSIONS

In this paper, we showed that the SBCOR is robust against severe waveform distortion such as zero-crossing and three types of noise using a DTW recognizer. This results indicate that the SBCOR extracts the speech features that are not captured sufficiently by conventional speech analyses. As for the robustness at phonemic level, we could verify it as long as the noise is the multiplicative signal-dependent white noise. For the other noises, we should investigate further. (The C source program of the SBCOR analysis used in this paper is available from ftp.itakura.nuee.nagoya-u.ac.jp or http://www.itakura.nuee.nagoya-u.ac.jp.)



(a) Signal Dependent White Noise

(b) Gaussian White Noise

(c) Human Speech Noise

Figure 7: Recognition Rates of SBCOR made as a function of BPF's Q(left side) and comparison with that of the SGDS(right side) for the signal-dependent white noise(a), the Gaussian white noise(b), and the human speech noise(c).

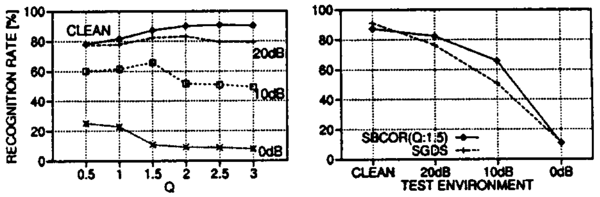


Figure 8: Recognition Rates of the SBCOR and the SGDS using the HMM recognizer.

REFERENCES

[1] J. H. L. Hansen, R. J. Mammone and S. Young: "Editorial for the special issue of the IEEE transactions on speech and audio processing on robust speech processing", IEEE Trans. on Speech and Audio Processing, **2**, pp. 549–550 (1994).

[2] 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).

[3] S. Kajita and F. Itakura: "Subband-autocorrelation analysis and its application for speech recognition", Proc. of **ICASSP**, Vol. II, pp. 193–196 (1994).

[4] S. Seneff: "A joint synchrony/mean-rate model of auditory speech processing", **JP**, **16**, pp. 55–76 (1988).

[5] O. Ghitza: "Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment", **JP**, **16**, pp. 109–123 (1988).

[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).