3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SBCOR SPECTRUM TAKING AUTOCORRELATION COEFFICIENTS AT INTEGRAL MULTIPLES OF 1/CF INTO ACCOUNT

Shoji KAJITA and Fumitada ITAKURA

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

ABSTRACT

This paper describes an extension of subband-autocorrelation (SBCOR) analysis that has been already proposed and shown to be robust against noise. The extended SBCOR analysis is defined by a weighted sum of the autocorrelation coefficients at the integral multiples of 1/CF(center frequency of band pass filter), in order to capture more information about the periodicity included in the speech signal. The experimental results performed by a DTW word recognition indicate that the extended SBCOR analysis is more robust against noise than the conventional SBCOR. The performance of the extended SBCOR is about 8% higher than that of the conventional SBCOR(when Q=1.5), and about 20% higher than that of the smoothed group delay spectrum under SNR 0dB. The characteristics of the SBCOR and its extension are also described.

I. INTRODUCTION

What information included in the speech signal is important for speech recognition? The conventional speech analysis techniques such as filter bank, LPC and cepstrum analyses are, broadly speaking, "formant oriented". Of course, formants and their dynamic changes are an important feature to characterize a speech signal, and such analyses have been shown to be successful under clean conditions. However, in noisy conditions, their robustness against noise is not satisfactory by any means. Thus, an alternative analysis technique with the robustness is required.

One of the possible approaches is to learn from the human auditory process and its model. In the peripheral auditory system, the speech signals are converted into auditory nerve firing patterns and transferred to the higher auditory system. Although it is not known well how the higher auditory system utilizes such patterns in the recognition process, a generalized synchrony detector(GSD) and an ensemble interval histogram(EIH) in the auditory models proposed by Seneff and Ghitza. respectively, give a hint for the above question[1, 2].

A key to their success seems to be that the GSD and EIH capture the extent of the dominance of periodicities in the auditory nerve firing; because noises that have no correlation with speech do not influence the periodicity so much. For example, suppose that the periodicity is expressed by the autocorrelation function. If the noise is white, it does not influence the autocorrelation function except for the zero lag; the periodicities is a key to the robustness against noise.

We have proposed a new signal processing technique based on subband processing and autocorrelation analysis, i.e., subband-autocorrelation(SBCOR) analysis. This SBCOR analysis was shown to be robust against noise[3, 4].

In this paper, we introduce an extension of the SBCOR analysis in order to capture more information about the periodicity included in the speech signal. And we also describe the characteristics of the SBCOR and its extension to explain the reason of the robustness.

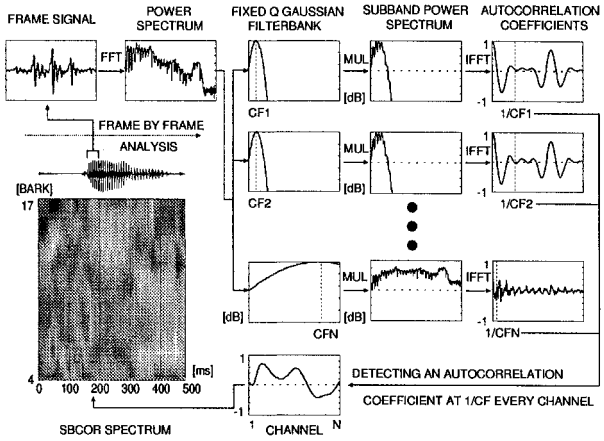

Figure 1. An implementation of the SBCOR analysis.

II. SBCOR ANALYSIS

2.1. Method

The SBCOR analysis is based on filter bank and autocorrelation analysis, and is defined by

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

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

where

$s_i(n)$: a SBCOR coefficient of ith channel at nth analysis frame
$R_i(\tau, n)$: the autocorrelation function of ith channel
$H_i(f)$: the transfer function of ith channel's BPF
f_{cf} : the center frequency of $H_i(f)$
$X(f, n)$: the power spectrum of speech signal.

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 inverse of the center frequency f_{cf}^{-1}, of each 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". It is easy to understand from equation (1) that the SBCOR spectrum is independent of the signal power, and its dynamic range is between -1 and 1. 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 so far[3, 4].

Figure 1 shows an implementation of the SBCOR analysis used in this paper. The filter bank consists of 128 or 16 fixed Q gaussian band pass filters(BPF) defined by

$$| H_i(f) |^2 = e^{-2C(f-f_{cf})^2} \quad | f | \geq 0 \tag{3}$$

$$C = \frac{2Q^2 \ln 2}{f_{cf}^2}. \tag{4}$$

2.2. Channel Frequency Response of SBCOR

In order to investigate the characteristics of the SBCOR in analyzing the signal whose spectrum is spread in a wide band like speech, we calculate the channel frequency response(CFR) by means of a tone signal to which white noise is added as follows:

$$S(f) = \frac{N_0}{2} + \frac{S_p}{2}\{\delta(f - f_s) + \delta(f + f_s)\} \tag{5}$$

where N_0 is the power spectrum density of the noise, S_p and f_s are the power and the frequency(in Hz) of the tone signal, respectively. To calculate the CFR analytically, the BPF is re-defined by

$$| H_i(f) |^2 = e^{-2C(f-f_{cf})^2} + e^{-2C(f+f_{cf})^2}. \tag{6}$$

This modification only affects the CFR at the low frequencies when Q is less than about 1.5.

From equations (1)-(6), the CFR of the ith channel, made as a function of the tone frequency f_s, is derived as follows:

$$CFR_i(f_s) = \frac{R_i(f_s, \tau_{cf})}{R_i(f_s, 0)}, \quad \tau_{cf} = {f_{cf}}^{-1}, \tag{7}$$

where

$$R_i(f, \tau) = \int_{-\infty}^{\infty} | H_i(f) |^2 S(f) \cos 2\pi f \tau df \tag{8}$$

$$= N_0 \sqrt{\frac{\pi}{2C}} e^{-(\pi\tau)^2/(2C)} \cos 2\pi f_{cf} \tau$$

$$+ S_p\{e^{-2C(f-f_{cf})^2} + e^{-2C(f+f_{cf})^2}\} \cos 2\pi f \tau \tag{9}$$

Figure 2 shows the CFR for the normalized frequency by f_{cf} as a function of "signal-to-critical-noise-ratio(SNR)" in dB. The power of the critical noise is calculated by (i) the critical noise level N_c of the white noise is determined as the one which causes the response to be the mean value between 1 and bias b when $f_s = f_{cf}$, (ii) the noise components passed through the BPF of the analysis channel are integrated. The bias defined by equation (10) is produced by the band pass characteristic of the BPF centered at f_{cf}, and depends on the Q.

$$b = \lim_{f_s \to \infty} CFR_i(f_s) = e^{-(\pi\tau_{cf})^2/(2C)} \tag{10}$$

As shown in Figure 2, the CFR shows that the SBCOR analysis has the lateral inhibition centered at f_{cf}, depending on its SNR. Since the emphasis of spectral contrast by such lateral inhibition improves the robustness against noise[5], it indicates that the robustness of the SBCOR is due to this lateral inhibition.

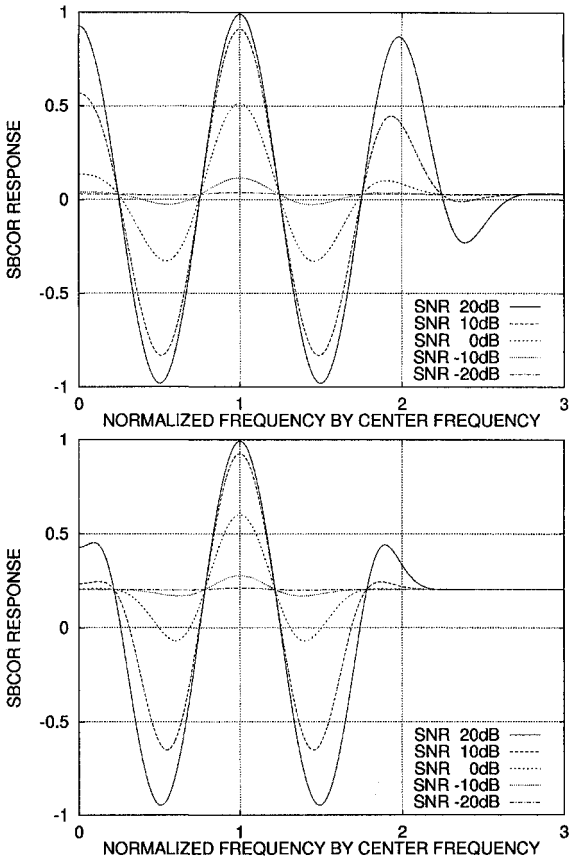

Figure 2. The channel frequency responses of the SBCOR, depending on its "signal-to-critical-noise-ratio(SNR)". The upper and lower plots are for Q=1.0 and Q=1.5 respectively.

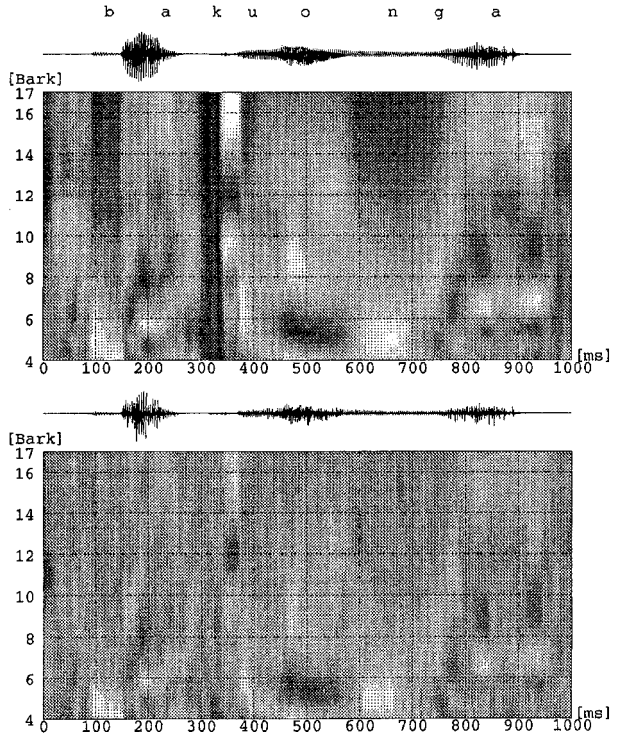

Figure 3. Analysis examples of the SBCOR analysis in analyzing the same speech signal under different SNRs. The upper and lower sides are under clean and SNR 0dB respectively. The number of BPF is 128(Q=1.0).

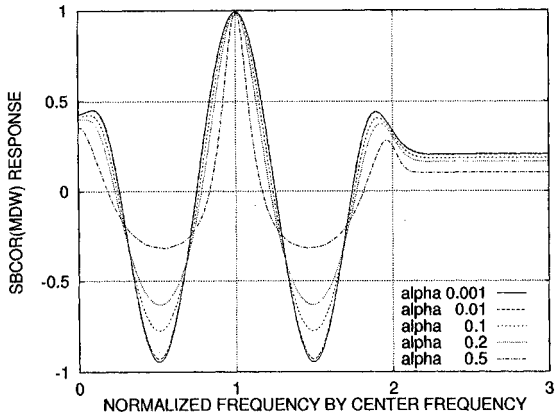

Figure 4. The channel frequency responses of the SBCOR by introduced the MDW(Q=1.5). The SNR is 20dB. When $\alpha \rightarrow 0$, the CFR agrees with that of the conventional SBCOR.

2.3. Analysis Examples

Figure 3 shows two examples of the SBCOR spectrum in analyzing the same speech signal under different SNRs. The utterance is spoken by a female speaker, and its sampling rate is 8kHz. The added noise is the multiplicative signal-dependent white noise(See section IV). The analysis frame length and shift are 32ms and 4ms respectively.

It can be seen that the SBCOR analysis extracts important speech properties, like spectral lines related to the first formant, and shows sharpness of both onset and offset for different speech segments under clean conditions(the upper side of Figure 3). However, the SBCOR spectrum does not necessarily extract higher formants. At SNR 0dB, the SBCOR spectrum preserves the information up to about 10 Bark(the lower side of Figure 3).

III. TAKING AUTOCORRELATION COEFFICIENTS AT INTEGRAL MULTIPLES OF 1/CF INTO ACCOUNT

In order to capture more information about the periodicity included in the speech signal, we introduce an extension of the SBCOR.

3.1. Multi-Delay Weighting

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

$$\hat{s}_i(n) = \frac{\sum_{k=0}^{M-1} \alpha^k \frac{R_i(\tau_{cf}(k+1), n)}{R_i(0,n)}}{\sum_{k=0}^{M-1} \alpha^k}, \quad 0 < \alpha < 1. \qquad (11)$$

We refer to it as the Multi-Delay Weighting(MDW) processing.

3.2. The Effect of MDW

Figure 4 shows the CFR of the SBCOR introduced the MDW. It was calculated in the same way of section 2.2.

The most important effect by the MDW is that the lateral inhibition centered at f_{cf} is controllable by α. As far as a speech recognition system performing under several conditions is concerned, it seems that the performance can be better by controlling the lateral inhibition, because the lateral inhibition of the SBCOR depends on the SNR of the signal(See Figure 2).

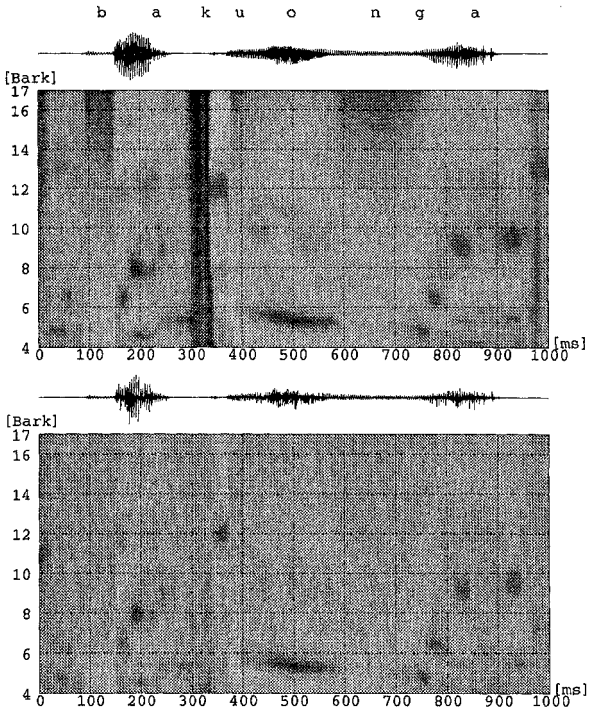

Figure 5. Analysis examples of the SBCOR(MDW,α=0.5) analysis. The upper and lower sides are under clean and SNR 0dB respectively.

Figure 5 shows analysis examples of the SBCOR processed by the MDW using the same speech signals of section 2.3. As shown in Figure 5, by controlling the lateral inhibition with the MDW, the pattern under clean conditions(the upper side of Figure 5) is closer to the pattern of the conventional SBCOR under SNR 0dB(the lower side of Figure 3) than that of the conventional one under clean conditions(the upper side of Figure 3).

IV. EXPERIMENTAL RESULTS

In this section, we quantitatively demonstrate the effect of introducing the MDW processing, using the SBCOR analysis as a front-end of a speech recognizer.

4.1. Recognizer and Database

A standard DTW speaker-dependent isolated word recognizer is used; it has a symmetric path and performs DP matching with fixed starting and ending points.

The basic database consists of two sets of 550 Japanese city names recorded twice by 5 Japanese male speakers. The sampling frequency is 10kHz. 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. Since the following experiments are speaker-dependent word recognition, we will get a very high recognition rate. Therefore, so as to clarify the differences of the performance, we selected 68 pairs of city names with phonetically similar names and performed DP matching between each pair [3, 4, 6]. Each pair is assumed to be easily mistaken in recognition. The recognition rate is given by the average of the 5 speakers.

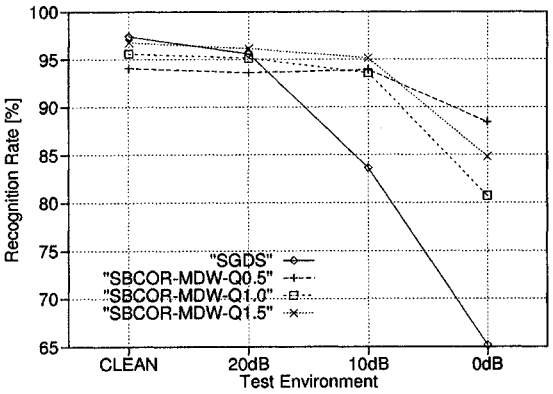

Figure 6. The best recognition rates of the SBCOR attained by introducing the MDW and the SGDS.

To examine the robustness against noise, white noise is added to the test patterns. The white noise used here is the multiplicative signal-dependent white noise which is defined as follows:

$$x'(n) = x(n)(1 + a \cdot r(n)), \quad z[\mathrm{dB}] = 10\log_{10}\frac{3}{a^2} \quad (12)$$

where $x(n)$ is the clean speech signal, $x'(n)$ is the noisy speech signal, a is the relative noise amplitude, z is the desired signal to noise ratio(SNR) and $r(n)$ is a 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 SNR is examined for four cases, namely $z = \infty$, 20, 10 and 0dB.

4.2. SBCOR Spectrum and Smoothed Group Delay Spectrum

The fixed Q filter bank consists of 16 BPFs defined by equation (3) and their center frequencies are equally spaced on the Bark scale between 4Bark and 17Bark. The Q values of 0.5, 1.0 and 1.5 are investigated. The α in processing the MDW is investigated for values between 0.001 and 0.9(M=8 in equation (11)).

To evaluate the performance under noisy conditions, the performance of the SBCOR spectrum is compared with that of the smoothed group delay spectrum(SGDS), already shown to be robust[6, 7]. The SGDS as distinct from the SBCOR, is the speech representation based on the group delay characteristic of the speech signal, and is defined as the derivative of the phase for a pth order all pole filter that has smoothed poles:

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} \hat{\alpha}_k z^{-k}} \quad (13)$$

$$\hat{\alpha_k} = \gamma^k \alpha_k \quad 0 < \gamma < 1 \quad for \quad k = 1, \cdots, p \quad (14)$$

where $H(z)$ is the transfer function of the smoothed all pole filter, α_k is the kth LPC coefficient and γ is the smoothing parameter. In order to compare the performance of the SBCOR with that of the SGDS under exactly the same conditions, the analysis frequency points of the SGDS are chosen to be the same center frequencies of the SBCOR.

4.3. Results

Table 1 shows the best recognition rates of the SBCOR processed by the MDW(SBCOR-MDW) and the recognition rates of the conventional SBCOR under each SNR.

Table 1. Average recognition rates among 5 speakers(%).

Q		CLEAN	20dB	10dB	0dB
0.5	conv.	94.07	93.46	93.31	82.83
	MDW	94.08	93.62	93.92	88.42
1.0	conv.	95.59	94.97	93.29	78.16
	MDW	95.59	95.13	93.59	80.70
1.5	conv.	96.78	96.01	93.28	76.34
	MDW	96.78	96.16	95.14	84.81

Table 2. the value of α in MDW processing.

Q		CLEAN	20dB	10dB	0dB
0.5	ref	0.01	0.1	0.1	0.6
	tes	0.01	$\leq$0.01	$\leq$0.01	0.01
1.0	ref	$\leq$0.01	$\leq$0.01	0.1	0.6
	tes	$\leq$0.01	$\leq$0.01	$\leq$0.01	0.01
1.5	ref	$\leq$0.01	0.01	0.6	0.6
	tes	$\leq$0.01	0.01	0.01	0.01

Table 2 shows the α of the reference and test patterns in the case of attaining the performance of Table 1.

These results are summarized to two points. First, by the MDW, the performance of the SBCOR is equal to or more than that of the conventional SBCOR for all Q (See Table 1). When the SNR becomes worse, the rate of the improvement becomes higher(specially, +8.5% for Q=1.5 under SNR 0dB). Second, it can be seen that the best performance is attained when the appropriate α according to the SNR of the test patterns is used. Therefore, it can be concluded that the robustness of the SBCOR improves by the appropriate MDW processing for each SNR.

Moreover, as shown in Figure 6, the performance of the SBCOR-MDW(Q=1.5) is equally as well as that of the SGDS(best γ:0.95) under clean conditions and much better under noisy conditions(Specially, +19.6% under SNR 0dB).

V. CONCLUSIONS

We described an extension of the SBCOR analysis that is defined by a weighted sum of autocorrelation coefficients at the integral multiples of 1/CF, in order to capture the periodicity included in the speech signal. It was shown that the robustness of the SBCOR by means of the appropriate MDW processing for each SNR becomes better than the conventional SBCOR.

In this paper, we investigated only the exponential weight as the weight of the MDW processing. In order to know how to control the lateral inhibition for the improvement of the robustness, we should investigate other weights.

REFERENCES

[1] S. Seneff: "A joint synchrony/mean-rate model of auditory speech processing", **JP**, **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", **JP**, **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) (1994). (in printing).

[4] S. Kajita and F. Itakura: "Subband-autocorrelation analysis and its application for speech recognition", **ICASSP** (1994).

[5] K. Obara and T. Hirahara: "Evaluation of auditory front-ends in DTW word recognition system", J. Acoust. Soc. Jpn., **50**, 6, pp. 452–464 (1994). (in Japanese).

[6] F. Itakura and T. Umezaki: "Distance measure for speech recognition based on the smoothed group delay spectrum", **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",**ICASSP**, Vol. 2, pp. 761–764 (1990).