SPECTRAL WEIGHTING OF SBCOR FOR NOISE ROBUST SPEECH RECOGNITION

Shoji KAJITA, Kazuya TAKEDA and Fumitada ITAKURA

Graduate School of Engineering, Nagoya University
Furo-cho 1, Chikusa-ku, Nagoya 464-8603 JAPAN
kajita@nuee.nagoya-u.ac.jp

ABSTRACT

Subband-autocorrelation (SBCOR) analysis is a noise robust acoustic analysis based on filter bank and autocorrelation analysis, and aims to extract periodicities associated with the inverse of the center frequency in a subband. In this paper, it is derived that SBCOR results in the lateral inhibitive weighting (LIW) processing of power spectrum, and shown that the LIW is significantly effective for noise robust acoustic analysis using a DTW word recognizer. An interpretation of LIW is also described. In the second half of this paper, a flattening technique of noise spectral envelope using LPC inverse filter is applied to speech degraded with noise, and DTW word recognition is performed. The idea of this inverse filtering technique comes from weakening the strong periodic components included in noise. The experimental results using 32*th* order LPC inverse filter show that the recognition performance of SBCOR (or LIW) is improved for computer room noise.

1. INTRODUCTION

A major difficulty encountered in current ASR is that the recognition performance degrades rapidly in the presence of noise and distortion, due primarily to the acoustic mismatches in training and recognizing conditions. Considerable effort has been made to overcome this problem. The research has been focused primarily on three areas; (1) noise robust feature extraction, (2) speech enhancement and (3) speech model compensation for noise[1]. Among these areas, we have been focusing on (1) to address the noise robust speech recognition problem.

From this point of view, we have proposed subband-autocorrelation (SBCOR) analysis[2]. The SBCOR is a type 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 conventional methods like smoothed group delay spectrum (SGDS) and mel-filterbank cepstral coefficients (MFCC) under noisy conditions[3].

In this paper, firstly, we derive that SBCOR results in the lateral inhibitive weighting (LIW) processing of power spectrum, and investigate the effectiveness of LIW using a DTW word recognizer. Second, a flattening technique of noise spectral envelope using LPC inverse filter is applied to speech degraded with noise.

This paper is constructed as follows. The following section reviews SBCOR analysis and derives the weighting function of power spectrum. Section 3 investigates the effectiveness of the weighting under noisy conditions, and provides an interpretation of LIW. Section 4 describes an inverse filtering of noise spectral envelope to improve noise robustness of SBCOR. Section 5 summarizes this research.

2. SBCOR AND LATERAL INHIBITIVE WEIGHTING PROCESSING

2.1. Method

SBCOR analysis is based on filter bank and autocorrelation analysis, and is defined as follows:

$$S_n(i) = \frac{R_n^i(\tau_{cf_i})}{R_n^i(0)},\ \tau_{cf_i} = f_{cf_i}^{-1} \quad (1)$$

$$R_n^i(\tau) = \int_{-\infty}^{\infty} \mid H_i(f) \mid^2 X_n(f) \cos 2\pi f \tau df, \quad (2)$$

where, $S_n(i)$ is the SBCOR coefficient of ith channel, $R_n^i(\tau)$ is ith subband autocorrelation function, $H_i(f)$ is the transfer function of BPF for ith channel, f_{cf_i} the center frequency of $H_i(f)$, $X_n(f)$ is the power spectrum of speech signal at nth analysis frame.

A subband filter bank of fixed Q Gaussian filter whose center frequencies are equally spaced on the Bark scale has been shown to be suitable for speech recognition under noisy conditions[4]. Each BPF is defined by

$$\mid H_i(f) \mid^2 = \begin{cases} \mathrm{e}^{-2C_i(f-f_{cf_i})^2},\ f \geq 0 \\ \mid H_i(-f) \mid^2,\ f < 0, \end{cases} \quad (3)$$

where, $C_i = (2Q^2 \ln 2)/f_{cf_i}^2$.

As a straightforward extension, SBCOR with multi-delay weighting (MDW) processing has been also proposed[5]. SBCOR with MDW uses the autocorrelation coefficient at not only $f_{cf_i}^{-1}$ but also the integral multiples of $f_{cf_i}^{-1}$ with exponential weighting as follows:

$$\hat{S}_n(i) = \frac{1}{A} \sum_{k=0}^{\infty} \alpha^k R_n^i((k+1)\tau_{cf_i})/R_n^i(0) \quad (4)$$

where $0 \leq \alpha < 1$ and $A = \sum_{k=0}^{\infty} \alpha^k$. Note that $\hat{S}_n(i)$ is equal to the basic SBCOR $S_n(i)$ when $\alpha = 0$.

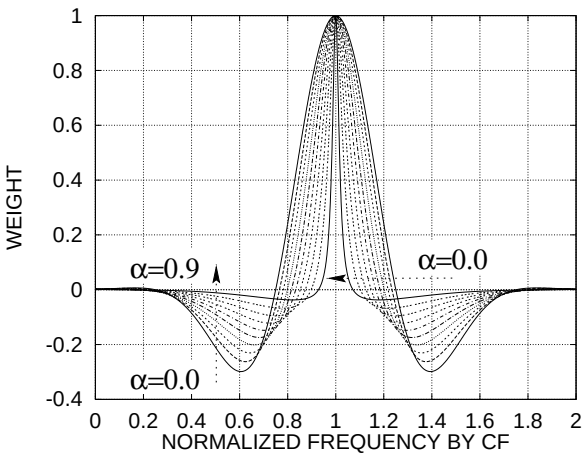

Figure 1: Weighting function $W_i(f)$ of the power spectrum of analysis frame signal. Q=1.5. The horizontal axis is normalized by the center frequency (CF).

2.2. Interpretation of SBCOR with MDW processing in Frequency Domain

When $v(\tau)$ is defined as

$$v(\tau) = \sum_{k=0}^{\infty} \alpha^k R_n^i(\tau + (k+1)\tau_{cf_i}), \quad (5)$$

Equation (4) can be calculated by $\hat{S}_n(i) = v(\tau)|_{\tau=0} / \{A R_n^i(0)\}$. Thus, MDW processing can be seen as a linear filter whose input and output are $R_n^i(\tau)$ and $v(\tau)$ respectively. Changing the range of the summation,

$$v(\tau) = \frac{1}{\alpha} \left\{ \sum_{k=0}^{\infty} \alpha^k R_n^i(\tau + k\tau_{cf_i}) - R_n^i(\tau) \right\}.$$

Let the Fourier transform of $R_n^i(\tau)$ and $v(\tau)$ be $X_n^i(f)$ and $V(f)$, then

$$\begin{aligned} V(f) &= \frac{1}{\alpha} \left\{ \sum_{k=0}^{\infty} \alpha^k X_n^i(f) \mathrm{e}^{j2\pi\tau_{cf_i} kf} - X_n^i(f) \right\} \\ &= \frac{\mathrm{e}^{j2\pi\tau_{cf_i} f}}{1 - \alpha \mathrm{e}^{j2\pi\tau_{cf_i} f}} X_n^i(f). \end{aligned}$$

From the even property of $R_n^i(\tau)$, $X_n^i(f)$ only has real components. Hence,

$$V(f) = \frac{\cos 2\pi\tau_{cf_i} f - \alpha}{1 - 2\alpha \cos 2\pi\tau_{cf_i} f + \alpha^2} X_n^i(f). \quad (6)$$

Using the inverse Fourier transform and setting $\tau = 0$, we have

$$v(\tau)|_{\tau=0} = \int_{-\infty}^{\infty} \frac{\cos 2\pi\tau_{cf_i} f - \alpha}{1 - 2\alpha \cos 2\pi\tau_{cf_i} f + \alpha^2} X_n^i(f) df.$$

Putting $X_n^i(f) = \mid H_i(f) \mid^2 X_n(f)$ and $A = \sum_{k=0}^{\infty} \alpha^k = 1/(1-\alpha)$, Equation (4) can be expressed in the frequency domain as follows:

$$\hat{S}_n(i) = \int_{-\infty}^{\infty} W_i(f) X_n(f) df / R_n^i(0) \quad (7)$$

$$\hat{W}_i(f) = \frac{(1-\alpha)(\cos 2\pi\tau_{cf_i} f - \alpha)}{1 - 2\alpha \cos 2\pi\tau_{cf_i} f + \alpha^2} \mid H_i(f) \mid^2 . \quad (8)$$

Thus, SBCOR analysis with MDW processing results in the weighting processing of power spectrum $X_n(f)$ by the weighting function $\hat{W}_i(f)$.

Figure 1 shows $\hat{W}_i(f)$ normalized by center frequency. As shown in the figure, both (1) frequency resolution and (2) spectral contrast enhancement by the lateral inhibitive weighting are controllable by α. As α tends closer to 1, the frequency resolution becomes high, and the spectral contrast enhancement becomes weak.

The contribution of this effect on recognition performance will be experimentally shown in the following recognition experiments.

3. EXPERIMENT 1

To investigate to what extent the lateral inhibitive weighting for power spectrum shown in Figure 1 is effective for noise robust speech recognition, DTW word recognition performs for the case of (1) the lateral inhibitive weights are removed and (2) the lateral inhibitive weights are used.

3.1. Experimental Conditions

3.1.1. Two Additive Noises

Gaussian white noise and a computer room noise were added to clean speech. The white noise was generated using a Gaussian random-number generator on computer. The computer room noise was recorded in a computer room using single microphone, as an example of realistic environmental noises. The power spectrum has several sharp peaks (Figure 5).

3.1.2. DTW word recognition

The standard DTW speaker-dependent isolated word recognizer is used. The recognition task is a 68 word-pair discrimination. Each pair is a phonetically similar city name pair, selected from a 550 Japanese city name database recorded twice by five Japanese male speakers[4]. 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 five speakers.

3.1.3. LIW (or SBCOR) and Comparative Feature Parameters

Under the above experimental conditions, the effectiveness of LIW is investigated by removing the inhibitive (or negative) weight, as shown in Figure 2. The LIW is applied directly to power spectrum of each analysis frame signal calculated by FFT. Moreover, in order to show the effectiveness under noisy conditions, the performance of LIW is compared with those of SGDS[7] and MFCC[8].

3.2. Results

The experimental results are shown in Figure 3 for each noise. The results of LIW are the best case. The best pair (α_r, α_t) in figures stands for αs in extracting reference patterns and test

Table 1: Analysis conditions.

SBCOR	The Q value is investigated for 1.0,1.5,2.0, and the center frequencies are equally spaced on the Bark scale between 4Bark and 17Bark. The α value in MDW processing is from 0.0 to 0.9 every 0.1.
SGDS	The analysis frequency points of the SGDS are chosen to be the same center frequencies of SBCOR.
MFCC	The filter bank consists of 28 triangle BPF whose center frequencies are equally spaced on the Mel scale.
COMMON	The analysis frame length and shift is 20ms and 10ms respectively. The dimension of all features is 16.

patterns respectively. The results of "no LIW" were obtained by the same αs of the best LIW except for using positive only weights.

As shown in figures, the LIW performs better than the non LIW under all conditions. These results clarify that the robustness of SBCOR is the effect of the LIW for power spectrum.

3.3. Discussions

When $W_i(f)$ is considered as an impulse response applied to power spectrum, LIW can be seen as a weighting procedure in the autocorrelation domain, like liftering procedure in the cepstrum domain. As shown in Figure 4, LIW suppresses low order autocorrelation unlike the case of positive only weighting. This implies that LIW is qualitatively equivalent to the spectral tilt elimination. In addition, the higher order autocorrelation is also suppressed for smaller α. The recognition results that smaller α (0.3 for white noise, 0.1 for computer room noise) is preferred under noisy conditions coincide with the fact that higher order autocorrelation is more influenced by noise. These effects, i.e., (1) spectral tilt elimination and (2) noise variability elimination, would be the essence of lateral inhibitive weighting, and lead to a more robust recognition under noisy conditions.

4. EXPERIMENT 2

In this section, a flattening technique of noise spectral envelope using LPC inverse filter is applied to speech degraded with noise, specially the computer room noise. As shown in Figure 5, the power spectrum of the computer room noise has several sharp peaks, in other words, several strong periodic components. The idea of this inverse filtering technique comes from weakening the strong periodic components included in noise.

4.1. Recognition Experiment

The same recognition experiment in the previous section was performed. The second, 8th, 16th and 32th order LPC filters were calculated from the computer room noise whose length is 3 seconds. In extracting both reference and test patterns, the inverse filtering using the same order LPC filter was performed.

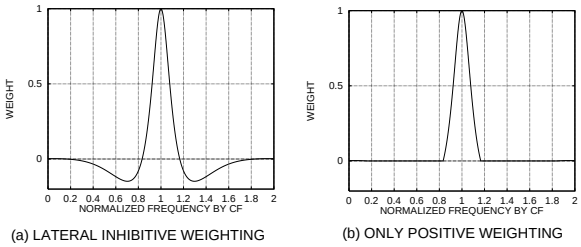

Figure 2: (a) Lateral inhibitive weighting, and (b) positive only weighting of (a). (Q=1.5,$\alpha = 0.5$)

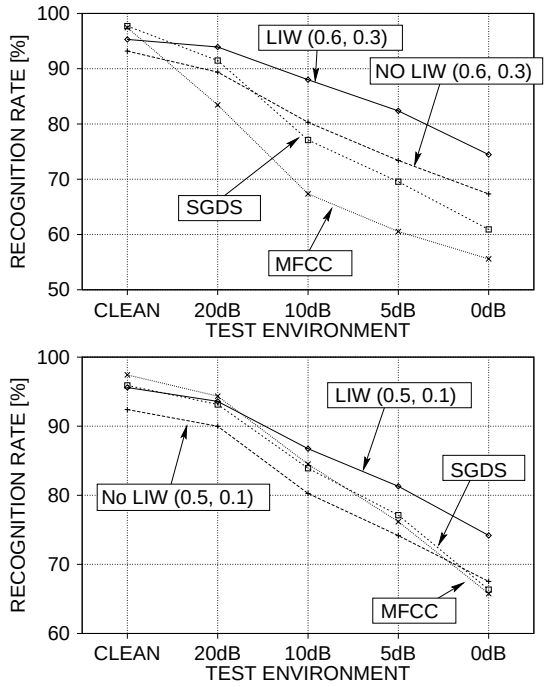

Figure 3: Averaged recognition rates for each features (upper: white noise, lower: computer room noise).

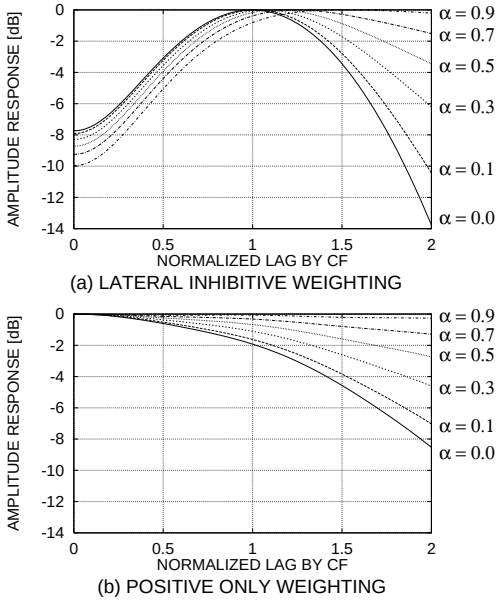

Figure 4: Amplitude response of LIW normalized by τ_{cf_i}.

The recognition results of LIW and MFCC are shown in Figure 6. The LIW performs well under noisy conditions as the order of the inverse LPC filter, while the performance of MFCC does not change. These results indicate that the de-correlation of strong periodic components is crucial for SBCOR that extracts periodicity included in speech.

5. CONCLUSIONS

In this paper, it is derived that SBCOR with MDW results in the lateral inhibitive weighting (LIW) processing of power spectrum, and shown that the LIW is significantly effective for noise robust acoustic analysis using a DTW word recognizer. These results clarify that the robustness of SBCOR is based on the LIW of power spectrum. In addition, by removing strong periodic components of noise from noisy speech using a LPC inverse filter, SBCOR (or LIW) performs well under noisy conditions, while MFCC does not.

6. REFERENCES

[1] Y. Gong: "Speech recognition in noisy environments: A survey", Speech Communication, **16**, pp. 261–291 (1995).

[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: "Robust speech feature extraction using SBCOR analysis", Proc. of ICASSP, Vol. 1, pp. 421–424 (1995).

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

[6] D. Kobayashi, S. Kajita, K. Takeda and F. Itakura: "Extracting speech features from human speech like noise", Proc. of ICSLP, Vol. 1, pp. 422–425 (1996).

[7] T. Umezaki, S. Harald and F. Itakura: "Evaluation of the smoothed group delay spectrum distance measure in speaker-independent speech recognition", Institute of Electronics, Information and Communication Engineers, **J74-A**, 4, pp. 610–618 (1991). in Japanese.

[8] 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).

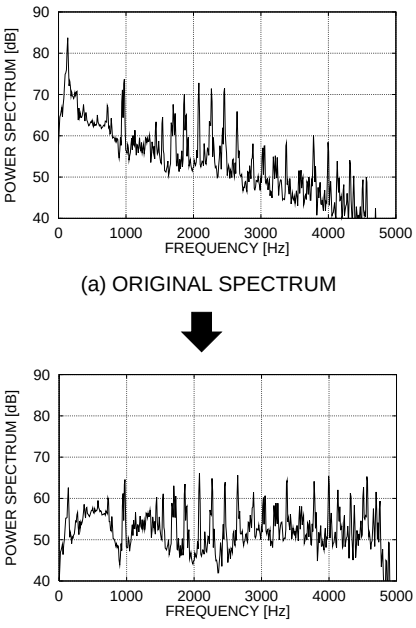

Figure 5: The power spectrum density of computer room noise estimated by Welch's averaged periodogram method. (a) original spectrum, (b) inversed spectrum by 16 *th* order LPC filter.

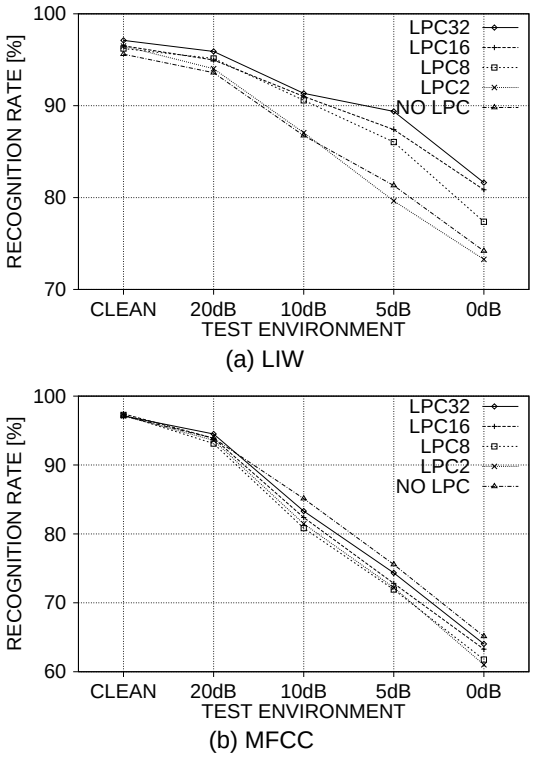

Figure 6: Averaged recognition rates of (a) LIW ($Q = 1.5$, $(\alpha_r, \alpha_t) = (0.5, 0.1)$) and (b) MFCC (CH=28) for inverse filtered noisy speech.