EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

NOISE REDUCTION BY PAIRED MICROPHONES

Masato Akagi and Mitsunori Mizumachi

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-12, Japan
Tel. +81-761-51-1236, Fax. +81-761-51-1149
email: {akagi, mizumati}@jaist.ac.jp

ABSTRACT

This paper proposes a front-end method for enhancing the target signal by subtracting estimated noise from a noisy signal using paired microphones, assuming that the noise is unevenly distributed with regard to time, frequency, and the direction. Although the Griffiths-Jim type adaptive beamformer has been proposed using the same concept, this method has some drawbacks. For example, sudden noises cannot be reduced because the convergence speed of the adaptive filter is slow; also the signal is distorted in a reverberated environment. The proposed method, however, can overcome the above drawbacks by formulating noises using arrival time differences between paired microphones and by estimating noises analytically using the directions of the noises. The simulated results show that the method with one paired microphone can increase signal-to-noise ratios (SNR) by 10 ~ 20 dB in simulations and can reduce log-spectrum distances by about 5 dB in real noisy environments.

1. INTRODUCTION

Although recent speech recognition systems for clean speech give high recognition rates, adverse environments reduce speech recognition accuracy. Therefore, robust recognition systems for noisy environments are required. One solution to the problem is to equip these systems with a noise reduction system as a front-end. Some methods such as adaptive filters for reducing noise [1] and large-scale microphone arrays for sharpening beam widths [2] have been proposed. However, these methods are not useful for controlling a handy phone or a navigation system by speech in a car because these methods need high-power DSPs to process them in real time and many microphones to construct large-scale microphone arrays. For these applications, small-scale noise reduction systems are needed.

This paper proposes a front-end method for enhancing the target signal by subtracting estimated noise from a noisy signal using paired microphones, assuming that the noise is unevenly distributed with regard to time, frequency, and the direction. Although the Griffiths-Jim type adaptive beamformer [3] has been proposed using the same concept, it has some drawbacks: sudden noises or moving noises cannot be reduced because the convergence speed of the adaptive filter is slow, and the signal is distorted in a reverberated environment because of the assumption that signal and noise are uncorrelated. The proposed method, while possessing the advantages of the Griffiths-Jim type adaptive beamformer, can overcome the above drawbacks using the following techniques: noises are formulated by using arrival time differences between paired microphones, and the largest noise not having the signal component is estimated analytically by using the direction of the largest noise in each short time period and narrow frequency band window.

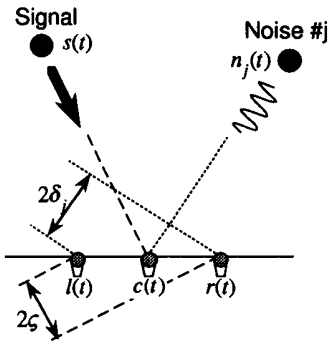

Fig. 1. Schematic illustration of a noisy environment.

The experiments described here show that (1) the proposed method using one paired microphone (i.e., two microphones and one sub-microphone) can estimate more than one noise existing in different frequency bands and coming from different directions, and (2) it can increase signal-to-noise ratios (SNR) by 10 ~ 20 dB in simulations and reduce log-spectrum distances by about 5 dB in real noisy environments. Thus this method can improve noise reduction performances with a small number of microphones.

2. FORMULATION

Let us assume that there are two main microphones and a centered sub-microphone in a noisy environment as shown in Fig. 1. $s(t)$ is a signal coming from a direction such that the difference in arrival time between the two main microphones is 2ζ, and $n_j(t)$ is a noise coming from another direction such that its time difference is $2\delta_j$ $(j = 1, 2, \ldots, N)$. The arrival sounds at the two main microphones are assumed as follows:

$$\text{left}: \quad l(t) = s(t-\zeta) + \sum_{j=1}^{N} n_j(t-\delta_j) \tag{1}$$

$$\text{right}: \quad r(t) = s(t+\zeta) + \sum_{j=1}^{N} n_j(t+\delta_j). \tag{2}$$

Additionally, the arrival sound at the center sub-microphone is assumed to be

$$\text{center}: r(t) = s(t) + \sum_{j=1}^{N} n_j(t). \tag{3}$$

Let us consider the microphone pair with two main microphones. $l(t)$ in Eq. (1) is shifted ζ, $r(t)$ in Eq. (2) is shifted $-\zeta$ in time, and both are transformed by short-term Fourier transformation (STFT). Then,

$$\mathcal{F}[l(t+\zeta)] = L(\omega)e^{j\omega\zeta} = S(\omega) + \sum_{j=1}^{N} N_j(\omega)e^{-j\omega\delta}e^{j\omega\zeta} \tag{4}$$

$$\mathcal{F}[r(t-\zeta)] = R(\omega)e^{-j\omega\zeta} = S(\omega) + \sum_{j=1}^{N} N_j(\omega)e^{j\omega\delta}e^{-j\omega\zeta}. \tag{5}$$

Next, $l(t+\zeta)$ and $r(t-\zeta)$ are shifted $\pm\tau$ in time, τ is a cer-

tain constant ($\tau \neq 0$), and their difference $g_{lr}(t)$ is transformed by STFT. The result $G_{lr}(\omega)$ is then

$$G_{lr}(\omega) = \mathcal{F}[g_{lr}(t)] = \sin\omega\tau \sum_{j=1}^{N} N_j(\omega)\sin\omega(\delta_j - \zeta), \quad (6)$$

where

$$g_{lr}(t) = \frac{\{l(t+\zeta+\tau) - l(t+\zeta-\tau)\} - \{r(t-\zeta+\tau) - r(t-\zeta-\tau)\}}{4} \quad (7)$$

and $N_j(\omega)$ is the STFT of the noise $n_j(t)$. Dividing $G_{lr}(\omega)$ by $\sin\omega\tau$ ($\omega\tau \neq n\pi$) gives

$$F_{lr}(\omega) = \sum_{j=1}^{N} N_j(\omega)\sin\omega(\delta_j - \zeta). \quad (8)$$

$F_{lr}(\omega)$ is the weighted-sum of the noises and is zero to the signal direction.

Since $F_{lr}(\omega)$ becomes infinite when $\omega\tau = n\pi$ in Eq. (8), $F_{lr}(\omega)$ is actually calculated as

$$F_{lr}(\omega) = \begin{cases} G_{lr}(\omega)/\sin\omega\tau, & |\sin\omega\tau| > \varepsilon \\ G_{lr}(\omega), & |\sin\omega\tau| \leq \varepsilon \end{cases} \quad (9)$$

where ε is a certain small value.

On the other hand, the STFT of the sum of the arrival sounds at the two main microphones is

$$\begin{aligned} U_{lr}(\omega) &= \mathcal{F}\left[\frac{l(t+\zeta) + r(t-\zeta)}{2}\right] = \frac{L(\omega)e^{j\omega\zeta} + R(\omega)e^{-j\omega\zeta}}{2} \\ &= S(\omega) + \sum_{j=1}^{N} \frac{N_j(\omega)\left(e^{-j\omega(\delta_j-\zeta)} + e^{j\omega(\delta_j-\zeta)}\right)}{2} \\ &= S(\omega) + \sum_{j=1}^{N} N_j(\omega)\cos\omega(\delta_j - \zeta) \end{aligned} \quad (10)$$

Thus in this method noise reduction is achieved by subtracting the second term in Eq. (10) by using $F_{lr}(\omega)$ in Eq. (8):

$$s(t) = \mathcal{F}^{-1}\left[U_{lr}(\omega) - \sum_{j=1}^{N} N_j(\omega)\cos\omega(\delta_j - \zeta)\right]. \quad (11)$$

Note that $\zeta = 0$ indicates that the signal direction is the front.

3. ESTIMATION OF NOISE DIRECTION

For simplicity let us assume $\zeta = 0$ and the noise coming from the direction such that the difference in arrival time between the two main microphones is $2t_r$, which is the largest in the frequency range $\omega_0 < |\omega| \leq \omega_1$.

In the period in which there is no signal and only noise exists, the noise direction can be estimated easily as follows.

Assuming that

$$\hat{L}(\omega) = \begin{cases} L(\omega) & \omega_0 < |\omega| \leq \omega_1 \\ 0 & other \end{cases}$$

$$\hat{R}(\omega) = \begin{cases} R(\omega) & \omega_0 < |\omega| \leq \omega_1 \\ 0 & other \end{cases} \quad (12)$$

and calculating

$$\mathcal{F}^{-1}\left[\frac{\hat{L}(\omega)\hat{R}^*(\omega)}{|\hat{L}(\omega)||\hat{R}(\omega)|}\right] \equiv d_{\omega_0,\omega_1}(t), \quad (13)$$

then

$$2t_r = \arg\max_t \left[d_{\omega_0,\omega_1}(t)\right]. \quad (14)$$

However, in the period in which both signal and noise exist, interaction between them yields estimation errors. We thus propose a noise direction estimation method that eliminates the influence of the signal and estimates the direction using the noise only.

Applying Eqs. (6)-(9) to the microphone pair with the left and center microphones,

$$g_{lc}(t) = \frac{\{l(t+\tau) - l(t-\tau)\} - \{c(t+\tau) - c(t-\tau)\}}{4}$$

$$G_{lc}(\omega) = \mathcal{F}[g_{lc}(t)] = \sin\omega\tau\left\{N_{t_r}(\omega)\sin\omega\frac{t_r}{2}e^{-\omega\frac{t_r}{2}} + E_{lc}(\omega)\right\}$$

$$F_{lc}(\omega) = N_{t_r}(\omega)\sin\omega\frac{t_r}{2}e^{-\omega\frac{t_r}{2}} + E_{lc}(\omega). \quad (15)$$

Similarly, to the microphone pair with the center and right microphones,

$$g_{cr}(t) = \frac{\{c(t+\tau) - c(t-\tau)\} - \{r(t+\tau) - r(t-\tau)\}}{4}$$

$$G_{cr}(\omega) = \mathcal{F}[g_{cr}(t)] = \sin\omega\tau\left\{N_{t_r}(\omega)\sin\frac{t_r}{2}\omega e^{-\frac{t_r}{2}\omega} + E_{cr}(\omega)\right\}$$

$$F_{cr}(\omega) = N_{t_r}(\omega)\sin\frac{t_r}{2}\omega e^{\frac{t_r}{2}\omega} + E_{cr}(\omega), \quad (16)$$

where $E_{\bullet\bullet}(\omega)$ is the sum of the noises from other directions.

The values given by Eqs. (15) and (16) have noise components and noise direction terms, $e^{\pm t_r\omega/2}$. Thus, calculating the following equation instead of Eq. (12),

$$\hat{L}(\omega) = \begin{cases} F_{lc}(\omega) & \omega_0 < |\omega| \leq \omega_1 \\ 0 & other \end{cases}$$

$$\hat{R}(\omega) = \begin{cases} F_{cr}(\omega) & \omega_0 < |\omega| \leq \omega_1 \\ 0 & other \end{cases}. \quad (17)$$

Then, using Eq. (13), the arrival time difference between the two main microphones $2t_r$ is given by

$$2t_r = 2 \cdot \arg\max_t \left[d_{\omega_0,\omega_1}(t)\right]. \quad (18)$$

4. ESTIMATION AND REDUCTION OF NOISE

Assuming that the arrival time difference of the largest noise is $2t_r$ and transforming Eq. (8) into

$$F_{lr}(\omega) = N_{t_r}(\omega)\sin\omega t_r + E_{lr}(\omega), \quad (8)'$$

the estimated noise $\tilde{N}_j(\omega)$ in the frequency band $\omega_0 < |\omega| \leq \omega_1$ can be represented as

$$\tilde{N}_{t_r}(\omega) = \begin{cases} \dfrac{F_{lr}(\omega)}{\sin\omega t_r} = N_{t_r}(\omega) + \dfrac{E_{lr}(\omega)}{\sin\omega t_r}, & |\sin\omega t_r| > \varepsilon \\ F_{lr}(\omega), & |\sin\omega t_r| \leq \varepsilon. \\ \omega_0 < |\omega| \leq \omega_1, \end{cases} \quad (19)$$

Thus the largest noise can be estimated by calculating Eq. (19) over all the frequency range. Substituting Eq. (19) into Eq. (11),

$$\tilde{s}(t) = \mathcal{F}^{-1}\left[R_{lr}(\omega) - \tilde{N}_{t_r}(\omega)\cos\omega t_r\right], \quad \omega_0 < |\omega| \leq \omega_1. \quad (20)$$

Figure 2 shows a block diagram of the proposed method.

5. EVALUATION

5-1. Sound Data

Two types of noise-added speech waves are prepared for evaluation. Both sound data are sampled at 48 kHz with 16 bit accuracy.

(1) *Sound data A*

The sound data A-(I) shown in Fig. 3(b), which is used for the estimation and reduction of time-limited noises, is a synthesized sound waveform on a computer using a sentence in the ATR speech database in which one band noise is included. The noise is an intermittent band noise, with a center frequency of 1.0 kHz and a bandwidth of 1.6 kHz, and it moves from 90 deg right to 90 deg left.

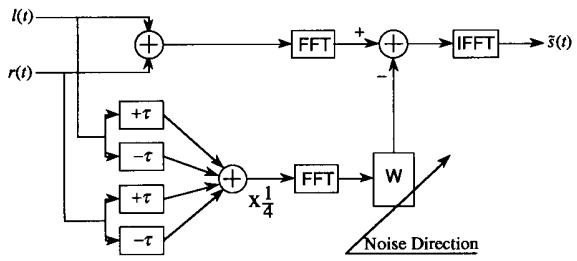

Fig. 2. Block diagram of the proposed method.

The sound data A-(II) shown in Fig. 4(b), which is used for the estimation and reduction of frequency-limited noises, is also a synthesized sound waveform on a computer using a word in the ATR speech database in which two band noises coming from two different directions are included. One is a continuous band noise (30 deg to the right) with a center frequency of 1 kHz and a bandwidth of 200 Hz, and the other is an intermittent band noise (45 deg to the left) with a center frequency of 0.5 kHz and a bandwidth of 100 Hz.

(2) *Sound data B*

Sound data B is real sound waveforms presented by two speakers in a soundproof room (reverberation time: about 50 ms); speech and noises come from 0 deg and 30 degs to the right, respectively, and both are 3 meters from the microphones. The noise is wide-band white noise from 125 Hz to 6 kHz, and three SNRs, -10, 0, and 10 dB, are prepared. Speech waves without noise (a) and the speech wave with SNR = 0 (b), are illustrated in Fig. 6.

5-2. Simulation conditions

The frame length and frame period for Sound data A are respectively 1024 and 512 points and $\varepsilon = 0.05$. Additionally, for Sound data A, the frequency bandwidths of partitions for noise direction estimation and noise reduction are all of the frequency range for A-(I) and 46.9 Hz each for A-(II).

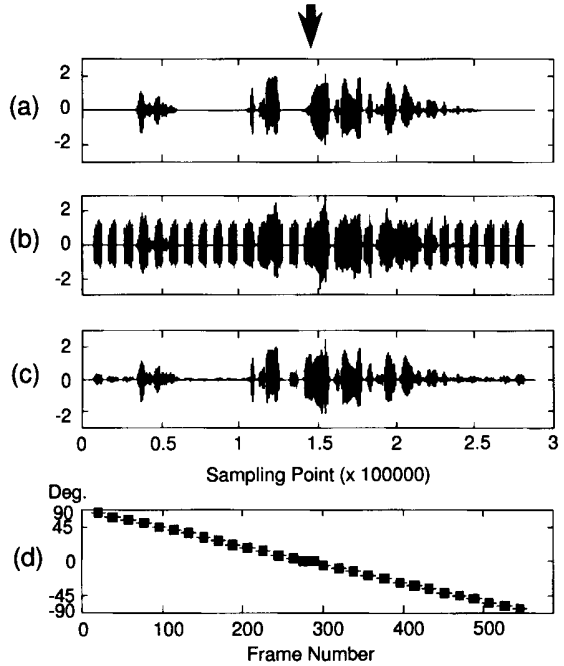

Fig. 3. Simulated results for time-limited noise using Sound data A-(I). (a) original noise-free speech wave (ATR, mhtsc101), (b) noise-added speech wave, (c) noise-reduced speech wave, and (d) estimated noise direction.

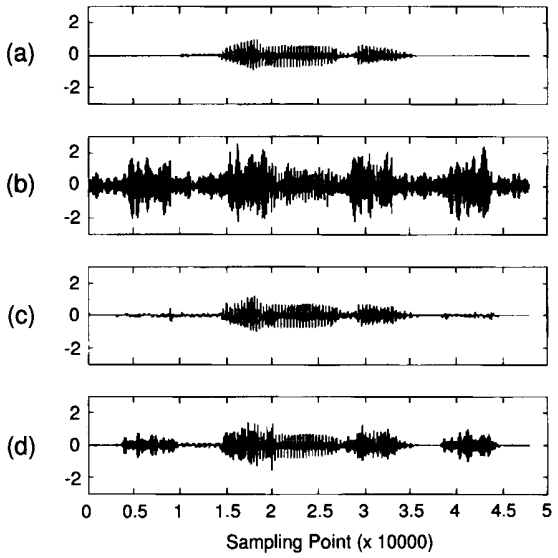

Fig. 4. Simulated results for frequency-limited noise using Sound data A-(II), (a) original noise-free speech wave (ATR, mht14348 /bunri/), (b) noise-added speech wave, (c) noise-reduced speech wave (noise direction estimation at 46.9 Hz each), and (d) noise-reduced speech wave (noise direction estimation using all the frequency range).

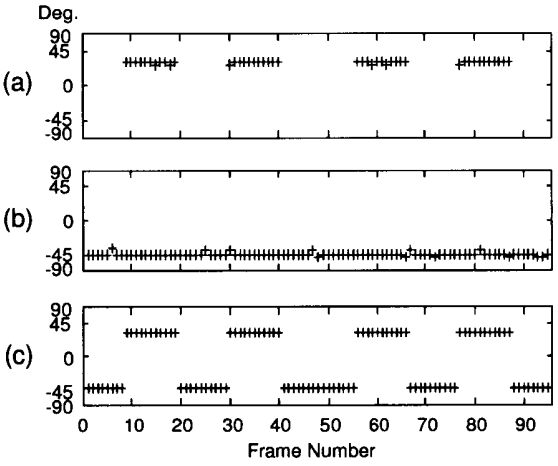

Fig. 5. Estimated results of noise direction for Sound data A-(II), (a) estimated noise direction at about 500 Hz frequency partition, (b) estimated noise direction at about 1 kHz frequency partition, and estimated noise direction using all the frequency range.

The frame length and frame period for Sound data B are respectively 4096 and 2048 points, $\varepsilon = 0.1$, and the frequency bandwidth of partitions for noise direction estimation and noise reduction is all of the frequency range.

The window types are Hamming for the direction estimation and a triangle for the noise reduction.

For the evaluation, the following signal-to-noise ratio (SNR) is adopted for Sound data A,

$$SNR = 10\log_{10}\frac{\sum_n s^2(t_n)}{\sum_n \{s(t_n) - \tilde{s}(t_n)\}^2} \quad (dB), \tag{21}$$

where $s(t_n)$ is an original sound wave and $\tilde{s}(t_n)$ is a noise-reduced sound wave. For Sound data B, the following log-spectrum distance (LSD) is adopted:

$$LSD = \sqrt{\frac{1}{W}\sum_{\omega}^{W}\left(20\log_{10}|S(\omega)| - 20\log_{10}|\tilde{S}(\omega)|\right)^2} \quad (dB), \quad (22)$$

where $20\log_{10}|S(\omega)|$ and $20\log_{10}|\tilde{S}(\omega)|$ are noise-free and noise-reduced log-spectrums normalized by their root mean squares, the frame length and frame period are respectively 1024 and 512 points, and $W = 6$ kHz.

5-3. Results

For Sound data A-(I), all the frequency range was used for the noise direction estimation by Eqs. (15)-(18) and the directions were well estimated, as shown in Fig. 3(d). Comparing Figs. 3(a), (b) and (c), it is seen that the noise components were reduced by the method. The SNRs calculated by Eq. (21) are -2.9 dB for the noise-added speech wave and 9.6 dB for the noise-reduced speech wave, which is an increase of about 12 dB. At the position indicated by the arrow in Fig. 3, however, noise cannot be reduced because the arrival time difference between the two microphones $2t_r$ approaches zero. Additionally, $2t_r$ is large as the noise direction comes close to ± 90 deg. The angular frequency satisfying $\omega t_r = n\pi$ then goes to low and noise still remains at the low frequency region. Noises at both ends of the noise-reduced speech wave shown in Fig. 3(c) are the noises caused by the situation mentioned above.

For Sound data A-(II), Figs. 4(c) and 4(b) show that two noise components in the noise-added sound were reduced when noise direction was estimated at every 46.9 Hz frequency partition. The continuous noise was reduced in all portions, and the intermittent noises placed roughly at points 5000, 15000, 30000, and 45000 were also reduced. Under these conditions, noise direction in each frequency portion was well estimated, as shown in Figs. 5(a) and (b), and the SNR increased from -10.3 to 10.9 dB, an increase of about 20 dB. On the other hand, when the noise direction was estimated [see Fig. 5(c)] using all of the frequency range shown in Fig. 4(d), only the direction of the largest noise was estimated. Thus, two noises coming from different directions cannot be reduced and the non-reduced noise adversely affects the results. Under these conditions, the SNR was -1.2 dB, which is about a 10 dB increase over the noise-added speech sound.

The noise direction for Sound data B was estimated using all of the frequency range by Eqs. (15)-(18) and the result was $t_r = 7$ points, i.e. 30 deg to the right, in all frames. Figure 6(c) shows the noise-reduced sound from the 0-dB SNR. The amplitude of the noise was reduced and was almost the same as that of the 10-dB SNR speech wave [Fig. 6(d)]. These figures indicate that the method can reduce noises in all portions. Figure 7 illustrates the mean LSD for each SNR in the speech portion calculated by Eq. (22), where No Process, Delayed-sum, and NORPAM mean noise-added speech, addition of two microphones, and noise-reduced speech by the proposed method. The figure indicates that the method can reduce the LSD by about 5 dB and can increase SNR by about 10 dB in real noisy environments.

4. CONCLUSION

This paper proposed a method that subtracts estimated noises from a noisy signal using paired microphones, assuming that the noises exist in different frequency bands and come from different directions. The simulated results indicate that the method with one paired microphone can increase the SNR by 10 to 20 dB in simulations and can reduce the LSD by about 5 dB in real noisy environments. This method can improve noise reduction performances with a small number of microphones. Additionally, this method can follow the Auditory Scene Analysis based noise reduction system [4].

References

[1] Kaneda, Y. and Ohga, J. (1986). "Adaptive microphone-array system for noise reduction," IEEE trans. ASSP, 34, 6, 1391-1400.

[2] Flanagan, J. L. et.al. (1991). "Autodirective microphone systems," Acoustica, 73, 2, 58-71.

[3] Griffiths, L. and Jim, C. (1982). "An Alternative Approach to Linearly Constrained Adaptive Beamforming," IEEE AP-30, 1, 27-34.

[4] Unoki, M. and Akagi, M. (1997). "A method of signal extraction from noisy signal," Proc. EUROSPEECH97.

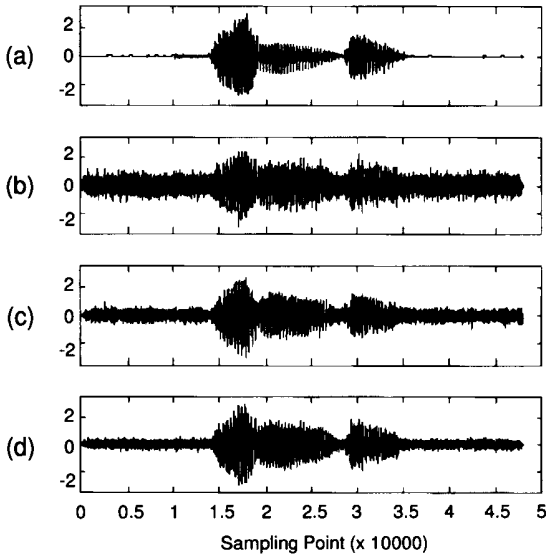

Fig. 6. Simulated results for Sound data B, (a) noise-free speech wave presented by a speaker (ATR, mht14348 /bunri/), (b) noise-added speech wave: SNR = 0 dB, (c) noise-reduced speech wave, and (d) noise-added speech wave: SNR = 10 dB.

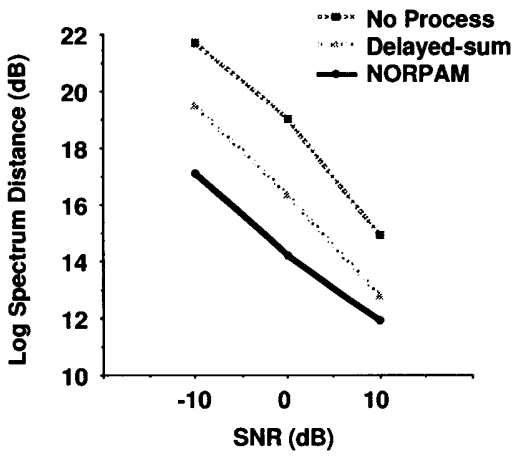

Fig. 7. Mean log-spectrum distances.