NOISY-SPEECH ENHANCEMENT FOR AUDIO-TELE-CONFERENCE SYSTEMS

Michiko Kazama, Waseda University, Tokyo, Japan
Mikio Tohyama, Waseda University, Tokyo, Japan
University of York, York, UK

1. INTRODUCTION

Speech signal enhancement is a fundamental issue in audio telecommunication network, since both direct and spatial sounds are required for immersive audio communication [1]. This paper describes a method for noise reduction of a received signal in a noisy environment such as an audio teleconference room.

In general noisy speech is not helpful for audio communication. Therefore noise reduction using a single microphone has already a long research-history. In particular, estimation of STSA [2] which stands for short time spectrum amplitude, sub-space approach [3], and spectrum subtraction [4] have been intensively investigated. However frame-dependent speech or noise level estimation is still under study in order to reduce the processing noise so-called musical noise [4]. Frame-by-frame estimation of signal to noise ratio using an updated equation is proposed in STSA, but the noise level is assumed to be stationary and thus noise estimation is performed only in the initial frames without speech contents. Estimation of frame dependent noise levels instead of the signal to noise ratio is also studied based on the minimum-statistics of frame-dependent signal energy where the minimum-level of the frame energy is assumed to be the noise level [5]. This estimation, however, might be frame-length dependent and not be flexible enough to pursue abrupt changes in the noise levels (such as noisy machines start and stop). The noise level is also assumed to be stationary in spectrum subtraction. However since the no-correlation hypothesis between noise and speech signals is not always accepted on the short-frame basis, over-subtraction is proposed taking account of the frame variances of the noise characteristics.

The authors introduce frame-dependent noise estimation into the conventional spectrum subtraction, since spectrum subtraction does not require fixed statistical models in the framework. Noise and speech signals are assumed to be un-correlated in almost all of conventional noise reduction methods. However we assume those can be in-phase in a short frame rather than un-correlated. Kazama et al [6] already demonstrated that an intelligible speech can be reconstructed with the magnitude spectrum and random phase instead of speech phase when the analysis and synthesis frame is short (say within 8-256 ms). This can be just the case when in-phase-subtraction of noise is performed in the spectrum subtraction.

We will propose an updated equation for frame-by-frame noise-spectrum estimation based on the dissimilarity of magnitude-spectrum envelopes between the noise and speech instead of conventional energy statistics. Correlation coefficient is used for a measure of dissimilarity. This dissimilarity measure can be also used for an indicator of noise (or speech) dominancy frame-by-frame basis.

The proposed noise reduction method will be estimated using energy analysis and narrow-band temporal envelopes of signals after noise reduction for noisy signals recorded in a teleconference room. This is because speech intelligibility is highly sensitive to the narrow-band (ex. 1/4 oct. band) temporal envelopes [6][7]. A schematic of our proposed method is presented in Section 2, and

numerical studies will be described in Section 3 in order to reconfirm that our procedure is effective, and finally noise reduction effects expected in a practical situation will be discussed in Section 4.

2. PROPOSED PROCEDURE FOR NOISE REDUCTION

The proposed procedure is summarized in Fig.1. The noise reduction goes on frame-by-frame by magnitude-spectrum subtraction of estimated noise from the noisy (composed of speech and noise) spectrum where speech and noise are assumed to be in-phase each other. Spectrum analysis is performed by conventional STFT. A key issue is noise spectrum estimation.

Suppose that we have a noise spectrum estimate at (l-1)th frame as $N(k,l-1)$, and an observation of l-th frame magnitude spectrum as $X(k,l)$. We take an updated rule of noise spectrum estimation $N(k,l)$ for the l-th frame such as [2]

$$N(k,l) \equiv aN(k,l-1)+bX(k,l) \quad \text{(1a)}$$

when speech is absent (noise is dominant) in the l-th frame, otherwise we set

$$N(k,l) \equiv N(k,l-1) \quad \text{(1b)}$$

where $a+b \equiv 1$, and these are updated parameters(or functions) defined later. We need a classifier of noise and speech in a frame for managing the updated process above.

We introduced a probability which represents the noise dominancy instead of hard decision of (1a) or (1b). Let us express the probability as $P(N)$. The pair of equations (1a) and (1b) can be combined as

$$N(k,l) \equiv P(N)\left[aN(k,l-1)+bX(k,l)\right]+(1-P(N))N(k,l-1). \quad \text{(2)}$$

We will express the probability using the magnitude-envelope correlation between the noise estimate $N(k,l-1)$ and the l-th frame spectrum $X(k,l)$. Suppose that the l-th frame spectrum envelope $X_E(k,l)$ is composed of

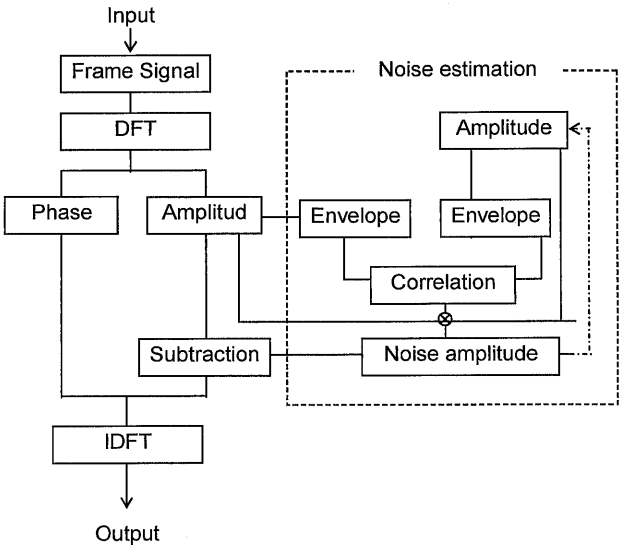

Figure 1 A schematic of the proposed procedure for noise reduction

$$X_E(k,l) \cong \hat{N}_E(k,l) + S_E(k,l) \quad (3)$$

where $\hat{N}_E(k,l)$ and $S_E(k,l)$ denote noise and speech envelopes in the frame, respectively. Here we assume $N_E(k,l-1) \cong \hat{N}_E(k,l)$ where $N_E(k,l-1)$ is the envelope of the noise estimate in the (l-1)-th frame.

Taking the cross-correlation coefficients $\rho(l)$ between $X_E(k,l)$ and $N_E(k,l-1)$, we can get

$$\rho^2(l) \cong \frac{U^2(l)}{U^2(l)+V^2(l)} \quad (4)$$

where $U^2(l) \equiv \left\langle N_E{}^2(k,l) \right\rangle$, $V^2(l) \equiv \left\langle S_E{}^2(k,l) \right\rangle$, $\langle * \rangle$ denotes the frequency average, and the noise and speech envelopes are assumed to be uncorrelated each other. Thus we can interpret Eq. (4) gives the probability of the noise dominancy (noise dominant frame or not), that is, we set

$$\rho^2(l) \cong \frac{U^2(l)}{U^2(l)+V^2(l)} \equiv P(N). \quad (5)$$

Consequently if we substitute Eq.(5) for Eq.(2), we obtain

$$N(k,l) = \left[1 - b\rho^2(l)\right] N(k,l-1) + b\rho^2(l) X(k,l) \equiv \alpha N(k,l-1) + \beta X(k,l). \quad (6)$$

Here we simply set the pair of updated functions such as

$$\alpha \to \beta \to 1/2 \ \text{ when } \rho(l) \to 1. \quad (7)$$

Following this requirement we rewrite the updated equation (6) as

$$N(k,l) = \left[1 - \rho^q(l)\right] N(k,l-1) + \rho^q(l) X(k,l) \quad (8a)$$

where q can be defined so that

$$\rho^q(l) \to 1/2 \text{ when } \rho(l) \to 1. \quad (8b)$$

These updated functions depend on the temporal-weighting factor for the noise estimates in previous frames. Thus we can control the functions according to the temporal characteristics of noise.

According to the noise estimation process described above, noise subtraction is performed in the l-th frame as

$$\hat{S}(k,l) \equiv X(k,l) - N(k,l) \quad (9a)$$

when the estimated speech magnitude spectrum $\hat{S}(k,l)$ is non-negative, otherwise we newly set

$$\hat{S}(k,l) \equiv 0. \quad (9b)$$

The noise suppressed speech signal can be synthesized by inverse STFT of the estimated speech magnitude $\hat{S}(k,l)$ with the observation phase of the noisy signal in every frame, since we assume that speech and noise are in-phase each other in a short frame.

3. NUMERICAL ESTIMATION OF NOISE REDUCTION EFFECT

We will estimate noise reduction effect expected by our proposed method using numerical samples. We used projector's fan noise recorded in a teleconference room even in this numerical study; however noisy speech was synthesized by in-phase superposition of the noise and speech in every short frame so that our basic proposition holds well.

3.1 Frame Processing Conditions

We will describe processing parameters. The signals are sampled at sampling rate of 16 kHz. We take a frame of 256 ms every 128 ms using a rectangular window. A triangular window is used for synthesizing the signal after subtraction. The envelope of spectrum is obtained by smoothing the STFT magnitude spectrum every frame. If we describe this smoothing process of the sequence in the frequency domain in terms of time-sequence processing, then it corresponds to low-pass filtering with the cutoff frequency of 1,000 Hz.

3.2 Noise Estimation Errors

The noise spectrum to be subtracted from every frame spectrum is estimated and updated following Eq.(8). An initial estimate of the noise spectrum can be obtained using initial frames where we can assume no speech dominant frames are contained. Figure 2 is an example of distribution of frame-noise estimates. We can see the estimates nicely follow the really contained noise samples when the signal to noise ratio(S/N) is low. As the signal noise ratio is high, the estimates give us over-estimates. Figure 3 similarly shows the percentile of the averaged noise-estimate in the observed frame-noise distribution. Over 50% of the frame-noise samples are distributed lower than the average of noise-estimates, when the S/N is greater than 6 dB. Only for S/N conditions below 6 dB, more than 50% of frames are distributed upper than the estimates.

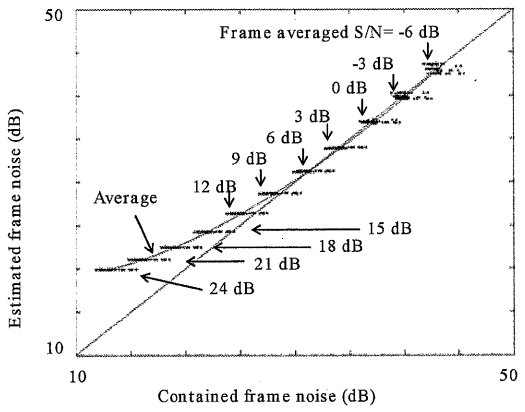

Figure 2 Distribution between estimated and really contained noise levels in every frame
S/N: frame average of the signal to noise ratio in the noisy signal
Frame length: 256 ms

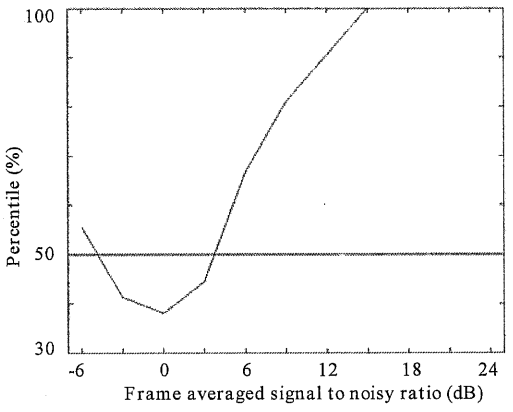

Figure 3 Percentile of the averaged noise-estimate in the really contained noise samples

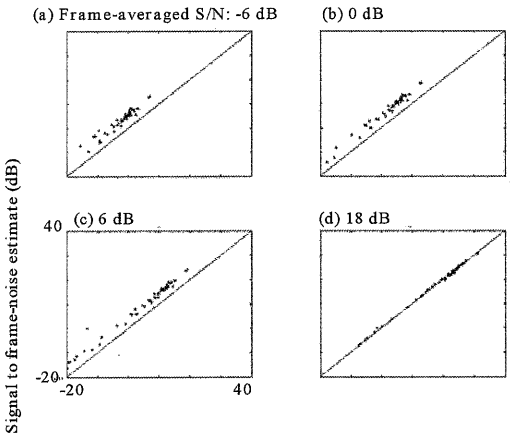

Fig.4 Distribution of frame-noise estimates using the signal to noise ratio

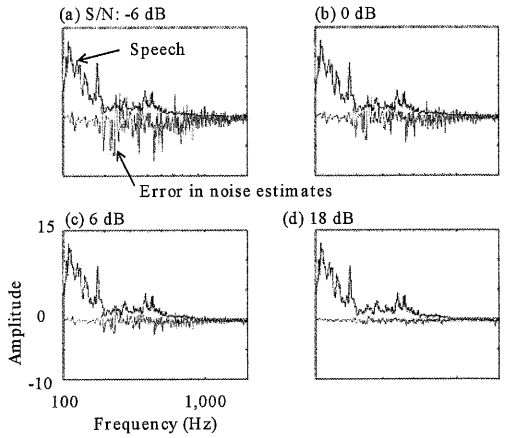

Figure 5 Frequency characteristics of frame-averaged speech and noise□estimation error (observed - estimate)

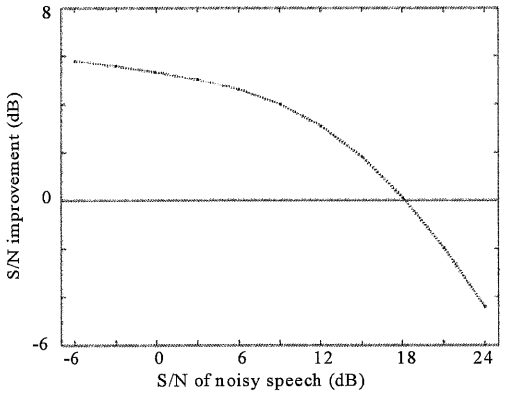

Figure 6 Improvement in frame-averaged S/N in dB

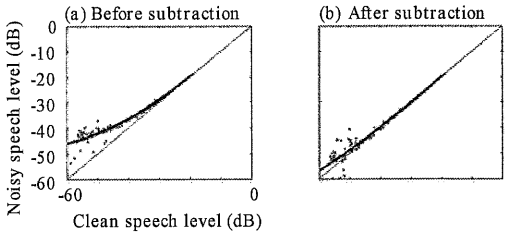

Figure 7 Distributions between clean speech and noisy speech levels
S/N for the noisy speech: 12 dB
Speech and noisy signal levels are averaged every 20 ms

Figures 4(a)-(d) draw the noise-estimates distributions using the S/N parameter. We can see scattergrams between the frame-noise and its estimate by comparison with the signal level. The estimates suitably correspond to frame noise variances without frame-wise S/N dependency. When the averaged S/N is below 0 dB, the noise estimates show slightly under-estimates. Figures 5(a)-(d) are displays of the frequency characteristics of speech levels and noise estimation errors by taking a frame average. We can see random nature of the estimation errors in the frequency domain. Consequently we decided over-subtraction factor used in conventional spectrum subtraction [4] is not necessary in our case.

3.3 Improvement of the Signal to Noise Ratio

Figure 6 shows improvement of S/N by our noise subtraction. Subtraction is performed following Eq. (9). Here we defined the frame averaged S/N in dB as the frame average of frame-based S/N in

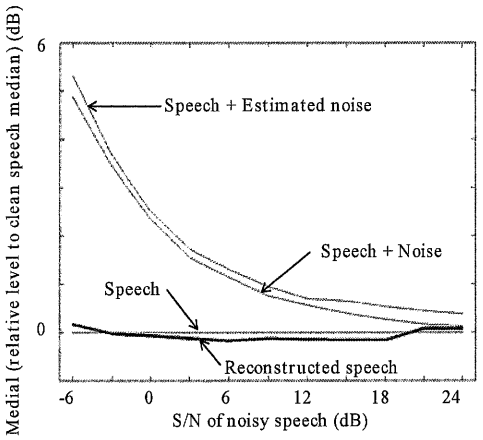

Figure 8 Median of frame signal levels
0 dB: Median level of the clean speech

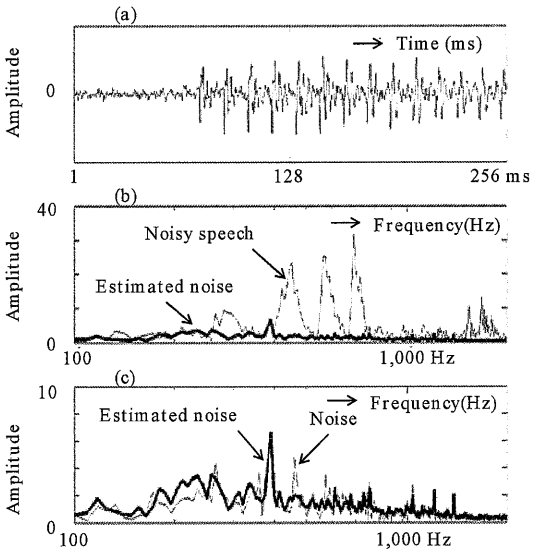

Figure 9 A sample of frame-noise estimation
(a) A frame of noisy speech waveform
(b) Magnitude spectrum of (a) and noise spectrum estimate
(c) Close-up of noise spectrum and its estimate

dB. We can expect improvement when the S/N of noisy speech is lower than 18 dB. Figure 7 is scattergram between the clean speech and noisy signal levels when S/N is 12 dB. Here the signal levels are averaged every 20 ms. We can confirm the subtraction effect by decreasing in the signal levels after subtraction.

This change of signal level distributions can be also characterized using the median of frame levels. Figure 8 demonstrates the median of the distributions of frame levels. The median of the noisy speech is quite similar to that for levels of sum of the clean speech and estimated noise. Both are higher than that for clean speech normalized to be 0 dB. The median of the noisy speech can be reduced to around 0 dB after subtraction. This reduction is similar to S/N improvement as shown in Fig.6. Consequently we can surmise that noise suppression can be expected without severe deformation in signal dynamics, since both the median and average of noise levels are reduced and the distribution of noise-subtracted signal levels comes close to clean speech dynamics.

4. EXPERIMENTAL RESULTS

Speech and noise signals are not always in-phase each other in a practical situation. We will estimate the noise reduction effect expected under the condition close to the practical situation. That is, the noisy signal is composed of noise and speech without phase-manipulation in this study.

4.1 Samples of frame-wise and averaged noise estimation

Figure 9 is an example of frame-wise noise spectrum estimation. Figure 9(a) is a frame-waveform of noisy signal, Figure 9(b) shows its magnitude spectrum with noise spectrum estimate, and close-up of the noise-estimate is displayed in Fig.9(c). We can confirm a nice trace of the frame-noise spectrum by its estimate even in the speech-contained frame.

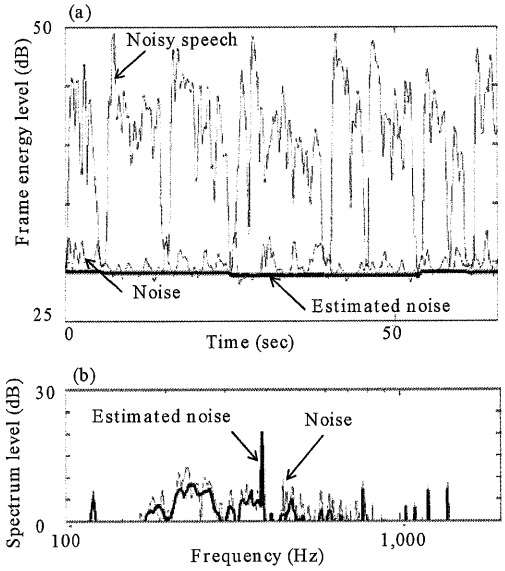

Figure 10 Temporal dynamics of signal energy and averaged spectrum
(a) Temporal characteristics of noisy signal energy and noise estimate
(b) Frame averaged noise-spectrum and its estimate

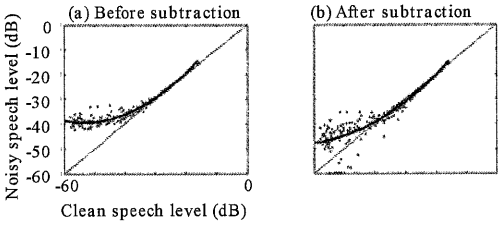

Figure 11 Distributions similar to Fig.7 between clean and noisy speech levels
S/N for the noisy speech: 12 dB
Noisy speech signal is composed of the noise and speech without phase manipulation.

Figure 10 presents long-term characteristics of the noisy signal and noise estimate. Figure 10(a) is the noisy signal dynamics of frame energy with the noise estimate, and Fig. 10(b) shows the frame-averaged noise spectrum with its estimate. The averaged noise spectrum can be reasonably estimated as well as the frame-wise estimate which we could see in Fig. 9(b). Therefore we can expect noise reduction effect by noise subtraction.

4. 2 Noise reduction effect

Similar to Fig.7, the scattergrams between the clear and noisy speech are shown in Fig.11. We can see noise reduction effect of around 10 dB when the signal level is not so high. If we compare Fig. 11 (without phase manipulation) to Fig. 7 (with phase manipulation), we can see a little larger variance of the noisy signal levels in Fig.11. In particular some of the noisy signal levels becomes lower than the clean speech levels in Fig.11 (b). This might be due to frame-wise phase relationship between speech and noise. Figure 12 shows the entire waveforms of before (Fig.12 (a)) and after subtraction(Fig.12(b)), and subtracted noise signal is shown in Fig.12(c).

4.3 Temporal Envelope Recovery

Speech intelligibility is highly sensitive to narrow-band envelope rather than signal to noise ratio [6] [7]. Therefore we have to be careful with the temporal characteristics in the waveform after subtraction, even if the noise level can be reduced. We divide the signal into 1/4 oct. band sub-signals through a filter bank of which center frequencies are located between 250 and 3363 Hz. Each sub-signal is rectified and put into a low-pass filter with the cutoff frequency of 40 Hz, so that the narrow-band envelope might be obtained every 1/4 oct. band.

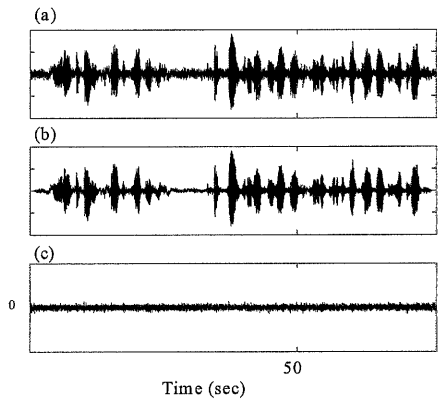

Figure 12 Waveforms before (a) and after (b), and subtracted noise (c)

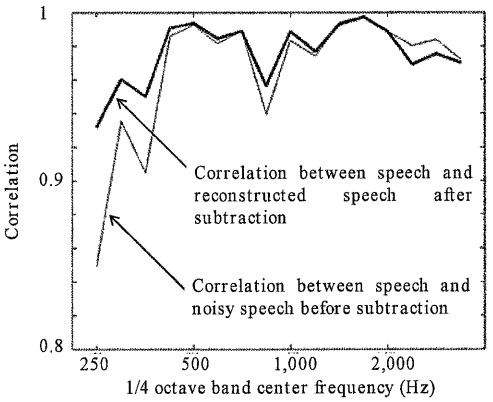

Figure 13 Correlation coefficients for 1/4 oct.-band temporal envelopes between the clean and noisy speech before and after subtraction

We assume that if the narrow-band envelope is quite similar to that for the clean speech every 1/4 oct. band even after subtraction, speech intelligibility can be preserved. We will use the correlation coefficient as a measure of the similarity. We calculate the envelope correlations between the clean and noisy speech signals every 1/4 oct.-band. Figure 12 shows the calculated results before and after subtraction. The noisy signal used for this experiment has the S/N of 12 dB as we already described. Therefore the signal is noisy but still intelligible, and the correlation is reasonably high without subtraction. But the correlation becomes a little higher by subtraction at the frequency bands lower than 500 Hz. We can expect noise reduction is performed without loss of intelligibility. and noisy speech signals every 1/4 oct.-band. Figure 12 shows the calculated results before and after subtraction. The noisy signal used for this experiment has the S/N of 12 dB as we already described. Therefore the signal is noisy but still intelligible, and the correlation is reasonably high without subtraction. But the correlation becomes a little higher by subtraction at the frequency bands lower than 500 Hz. We can expect noise reduction is performed without loss of intelligibility.

5. SUMMARY

A method for spectrum subtraction including frame-wise noise estimation process has been proposed in this article. A frame-by-frame updated equation is derived using the spectrum envelope correlation between the previous estimate of noise and present frame-signal. The correlation shows dissimilarity between speech and noise spectrum, and thus it could be a good indicator of noise dominancy in every frame. The frame-wise noise spectrum could be reasonably estimated following the updated equation. Consequently around 10 dB of S/N improvement was obtained for a noisy signal with 12 dB of S/N. This noise reduction can be performed without loss of intelligibility, since the narrow-band temporal envelope characteristics of speech is preserved. This proposed method can be extended into 2-channel teleconferencing systems, keeping sound source localization information. This study is based on signal analysis of magnitude spectrum. However, noise or concurrent speech effect on target speech in teleconference systems must be estimated from a point of view of information masking. A dissimilarity measure between a speech and other environmental sounds including concurrent speech is a future problem. The authors would thank Prof. T. Houtgast for his suggestions, and they are most grateful to Prof. Y. Yamasaki for his constant encouragement.

6. REFERENCES

1. C.□Kyriakakis, et al., "Surrounded by Sound", IEEE Signal Processing Magazine, vol. 1, pp.55-66 1999
2. Y. Ephraim and D. Mahlar, "Signal Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator," IEEE ASSP 32 (6) pp.1109-1121, 1984
3. Y. Ephraim and H. L. Van, "A Signal Subspace Approach for Signal Enhancement," IEEE SAP 3 (4), pp. 251-266, 1995
4. M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," ICASSP 79, pp,208-211, 1979
5. I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," IEEE SAP 11(5), pp.466-475 (2003)
6. M. Kazama, M. Tohyama and T. Houtgast, "Speech Reconstruction by Using Only Its Magnitude Spectrum or Only Its phase", 17th ICA, 2001,□7P.51
7. R. Drullman, "Temporal Envelope and Fine Structure Cues for Speech Intelligibility," J. Acoust. Soc. Am. 97(1), pp. 585-592, 1995