Acoustical Society of America and Acoustical Society of Japan

Third Joint Meeting

Honolulu, Hawaii, 2-6 December 1996

NOISE REDUCTION FOR SUBBAND ACOUSTIC ECHO CANCELLER

Junko SASAKI, Yoichi HANEDA, and Shoji MAKINO
NTT Human Interface Laboratories,
3-9-11 Midori-cho, Musashino-shi, Tokyo 180, Japan

1. INTRODUCTION

In teleconferences using loudspeakers and microphones, such as audio-conference or video-conference systems, echoes and background noise degrade the communication quality. Our aim is to achieve a good teleconference system combining a subband echo canceller and noise reduction, where processing delay and required computational power are small. Ways of combining an acoustic echo canceller with noise reduction have been studied recently [1] [2].

We used the noise-reduction techniques based on the short time spectral amplitude (STSA) estimation such as MMSE [3], Wiener filtering [4], maximum likelihood envelope estimation [5], and spectral subtraction [6], for combination with a subband echo canceller. These noise-reduction techniques have a processing delay because they use a speech spectrum obtained by discrete Fourier transform (DFT) with finite duration sequence. A subband echo canceller also has a processing delay. There will be a long delay if noise reduction and a subband echo canceller are linearly combined. One reason for this is that they each transform the input signal into the frequency domain for processing and transform the output signal back into the time domain.

In this paper, we propose a combination system that divides the input signals only once, using a poly-phase filter bank and performs both echo cancellation and noise reduction on the subband signals. We set the number of subbands to 32 from the viewpoint of reducing the processing delay. This number is adequate for an echo canceller, but much smaller than usual for noise reduction. To evaluate the performance using only 32 subbands, we performed subjective tests, and identified the most effective combination among the above mentioned noise-reduction techniques. Furthermore, we propose a better noise-reduction technique, which uses masking by original noisy speech. Adding a small amount of original noisy speech improves its performance.

We also present an improvement in echo cancellation in a noisy environment. When using an echo canceller in a noisy environment, there is another problem that the adaptive

algorithm in the echo canceller cannot work stably for small far-end signals. We propose to use the noise level estimated in the noise-reduction part in the adaptive algorithm to solve this problem.

2. NOISE-REDUCTION TECHNIQUES

The noise-reduction techniques based on STSA estimation multiply each spectral component by a noise-reduction gain factor. This gain factor is adaptively determined using the signal-to-noise ratio (SNR) calculated at each spectral component [3] [7].

Let $S_k(n)$ and $W_k(n)$ denote the k-th spectral component of speech and background noise signal, respectively. The k-th spectral component of noisy observed signal $V_k(n)$ is written as

$$V_k(n) = S_k(n) + W_k(n) . \tag{1}$$

In the noise-reduction technique based on STSA estimation, the k-th spectral component of noise-reduced signal $\hat{S}_k(n)$ is obtained by

$$\hat{S}_k(n) = G(SNR_k(n)) \times V_k(n) , \tag{2}$$

where $G(SNR_k(n))$ is the gain factor calculated using the SNR of $V_k(n)$.

Here, the $SNR_k(n)$ are defined in two ways: a posteriori SNR and a priori SNR. These two SNRs can be estimated by the following equations.

$$\text{a posteriori SNR:} \quad SNR_k(n)' = P_{V,k}(n)/P_{N,k}(n) \tag{3}$$

$$\text{a priori SNR:} \quad SNR_k(n) = (1- \beta)\, P[SNR_k(n)' - 1] + \beta\, [SNR_k(n-1)'] , \tag{4}$$

where $P_{N,k}(n)$ and $P_{V,k}(n)$ are the noise signal power and noisy observed signal power at each spectral component and P[*] denotes half-wave rectification. Ephraim showed an estimation method of a priori SNR in the MMSE method [3]. Scalart showed that the noise-reduction effect is increased by using the a priori SNR instead of the a posteriori SNR in Wiener filtering, maximum likelihood envelope estimation, and spectral subtraction [7]. We confirmed

Noise reduction method	$G(SNR_k(n))$
MMSE[3]	$\frac{\mu k \exp(\upsilon)}{(1+SNR_k(n)')+\mu_k \exp(\nu)} \cdot \frac{\sqrt{\pi}}{2}\sqrt{\frac{SNR_k(n)}{SNR_k(n)[1+SNR_k(n)]}} \cdot F[SNR_k(n)'[\frac{SNR_k(n)}{1+SNR_k(n)}]]$
Wiener filtering[7]	$\frac{SNR_k(n)}{1+SNR_k(n)}$
Maximum likelihood envelope estimation[7]	$\frac{1}{2}[1+\sqrt{\frac{SNR_k(n)}{1+SNR_k(n)}}]$
Spectral subtraction[7]	$\sqrt{\frac{SNR_k(n)}{1+SNR_k(n)}}$

Table 1

Gain defined in each noise-reduction method based on short time spectral amplitude (STSA) estimation. $SNR_k(n)'$: a posteriori SNR, $SNR_k(n)$: a priori SNR(n), $\mu_k = (1-q_k)/q_k$, q_k is probability of signal absence in the k-th spectral component, $\nu = (SNR_k(n)'\ SNR_k(n))/(1+SNR_k(n))$ F: confluent hypergeometric function.

this finding, so we used the a priori SNR calculated by Eq. (4) in our system. Table 1 shows the gain factor using a priori SNR in four techniques, which we evaluated in our system.

3. COMBINATION OF SUBBAND ECHO CANCELLER AND NOISE REDUCTION

We combined a subband echo canceller with each of the four noise-reduction techniques, where noise reduction uses the subband signals divided in the subband echo canceller in order to avoid the long delay. Our combination system is shown in Fig. 1. The far-end signal $x(n)$ and the microphone output signal $y_{all}(n)$ are divided into 32 subband signals by using a poly-phase filter banks that are generally used in subband echo cancellers. The microphone output signal includes echo signal $y(n)$, near-end speech signal $s(n)$, and background noise signal $w(n)$. The echo canceller reduces the echo signal in the microphone output signals in each subband. When the echo canceller works well, we can assume

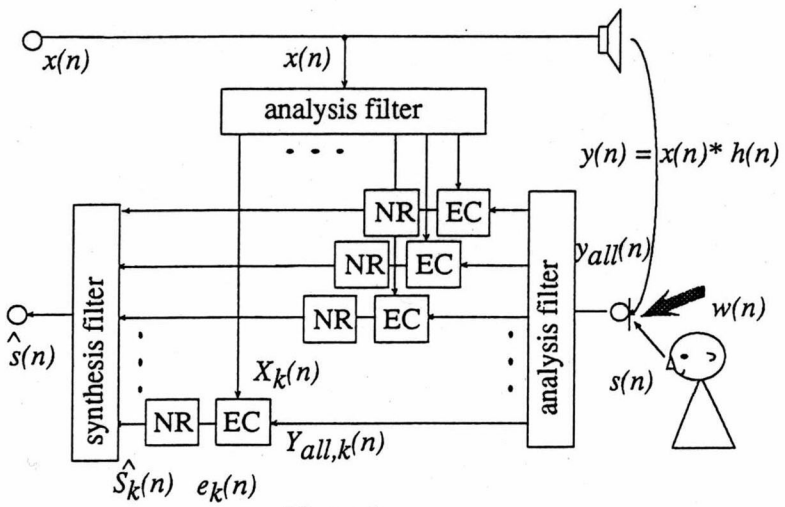

Figure 1
System combining subband echo cancellation with noise reduction. $x(n)$: far-end signal, $y_{all}(n)$: microphone output signal , $h(n)$: room transfer function, $y(n)$: echo signal, $s(n)$: near-end speech signal, $w(n)$: background noise, $X_k(n)$ and $Y_{all,k}(n)$: k-th subband signals of $x(n)$ and $y_{all}(n)$, $e_k(n)$: k-th subband echo canceller output signal, $\hat{S_k}(n)$: noise-reduced signal , $\hat{s}(n)$: near-end speech signal, NR: noise reduction part, EC: echo canceller part.

$$e_k(n) \doteqdot S_k(n) + W_k(n) = V_k(n), \tag{5}$$

where $e_k(n)$, $S_k(n)$, $W_k(n)$, and $V_k(n)$ are the k-th subband echo canceller output signal, the k-th subband near-end speech signal, the k-th subband background noise signal, and the k-th subband noisy speech signal, respectively. $e_k(n)$ is sent to the noise-reduction part denoted NR in Fig. 1. In the noise-reduction part, the gain factor is calculated according to each of the four definitions in Table 1 and noise is then reduced by Eq. (2). Finally, noise-reduced signal $\hat{S}_k(n)$ is synthesized using a synthesis filter bank, resulting in near-end speech signal $\hat{s}(n)$.

Noise power $P_{N,k}(n)$ used to calculate the a priori SNR is estimated in each subband by making histograms that show the level distribution of $V_k(n)$ [5]. The voice signal levels vary over a wide range, while the static noise has a much narrower distribution. Consequently, a large peak appears as the average noise level in the histogram. We estimated noise power level $P_{N,k}(n)$ by using this characteristic.

4. SUBJECTIVE EVALUATION

In this system, we used an excessively small number of subbands, 32, in order to make the delay time small. We performed subjective evaluations of this combination system for each of the four noise-reduction techniques listed in Table 1, and determined the most effective one under this condition.

We evaluated the noise-reduced signal by using the mean opinion score (MOS). The stimulated signals included the original speech signal, three noise-added signals corresponding to different SNRs, and noise-reduced signals processed using the four noise-reduction techniques. These signals were stimulated through a loudspeaker and evaluated as a received signal in a teleconferencing situation on a five-point opinion scale: 4 (excellent), 3 (good), 2 (fair), 1 (poor), and 0 (bad). We used both a female's speech and a male's speech for the original speech and used air-conditioner noise for added noise. This subjective evaluation was performed in a reverberation room with a volume of 87 m^3 and a reverberation time of 300 ms. The background noise level in the room was set to 46 dB(A). Thirty listeners (16 speech and audio researchers and 14 ordinary people) evaluated them. Actual testing was done after a practice session using eight samples.

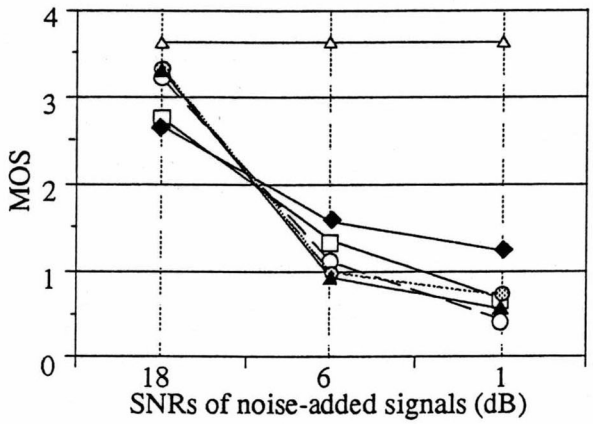

Figure 2
Comparison of noise-reduction methods. —△— original signal, —□— noise-added signal, —○— signal with noise reduced by MMSE, —▲— by Wiener filtering, —◆— by maximum likelihood envelope estimation, and ···◍··· by spectral subtraction.

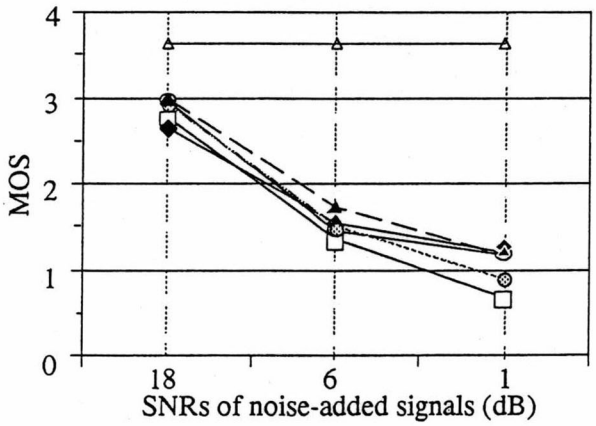

Figure 3
Masking by input signal. —△— original signal, —□— noise-added signal, —○— signal with noise reduced by MMSE, —▲— by Wiener filtering, —◆— by maximum likelihood envelope estimation, and ···◍··· by spectral subtraction.

The relationship between the MOS and the SNR is shown in Fig. 2. Comparing the MOS before and after noise reduction, noise reduction generally increased the MOS when the SNR was high. When SNR was low, noise reduction decreased the MOS, except for maximum likelihood envelope estimation.

In maximum likelihood envelope estimation, the gain factor is given by

$$G(SNR_k(n)) = \frac{1}{2}\left[1+\sqrt{\frac{SNR_k(n)}{1+SNR_k(n)}}\right] . \tag{6}$$

The first term, which adds an input signal, apparently causes masking of the distortion caused by the second term. When the number of subbands is small, the decrease in quality due to speech distortion and residual noise is considered to outweigh the increase in quality due to noise reduction. Therefore, masking of the distorted signal by the input signals increases the

quality.

To confirm the effectiveness of masking by the input signal, we added an input signal term to the gain factor for each noise-reduction technique. Figure 3 shows the MOS when using gain factor $G(SNR_k(n))'$,

$$G(SNR_k(n))' = \alpha + (1-\alpha)\, G(SNR_k(n)), \tag{7}$$

instead of $G(SNR_k(n))$ in Eq. (2). For each technique with $\alpha = 0.3$, the MOS decreased when SNR was high; however, the MOS was improved between 0.5 and 1 when the SNR was low, comparing the values before and after adding the input signal.

Our results show that the masking by adding the input signal is effective and every technique has almost the same performance when the SNR is low and the number of subbands is small. Using Wiener filtering with an added input signal is especially effective for an echo canceller in real-time communication, because it produces a bigger effect with relatively less calculation according to the definitions in Table 1. When the SNR is high, this approach is adequate without masking; therefore, it is better to add an input signal based on the SNR.

5. IMPROVEMENT IN ECHO CANCELLATION

We have already proposed an exponentially weighted stepsize (affine) projection algorithm for echo cancellers [8] . We used this algorithm in our subband echo canceller, because it achieves four times faster convergence than the conventional normalized LMS algorithm without additional computational power. The updating equation in the second-order ES projection algorithm using intermediate variable z is given as

$$z_k(n+1) = z_k(n) + \gamma A_k[\beta_{1,k}(n-1)+\beta_{2,k}(n-1)]X_k(n-1) \tag{8}$$

$$\beta_{1,k}(n) = [e_k(n)r_{11,k} - (1-\gamma_k)e_k(n-1)r_{10,k}]/[r_{00,k}r_{11,k} - r_{10,k}r_{10,k} + \delta_k] \tag{9}$$

$$\beta_{2,k}(n) = [(1-\gamma_k)e_k(n-1)r_{00,k} - e_k(n)r_{10,k}]/[r_{00,k}r_{11,k} - r_{10,k}r_{10,k} + \delta_k], \tag{10}$$

where

$$e_k(n) = Y_k(n) - \hat{Y}_k(n) + W_k(n), \tag{11}$$

$$\hat{Y}_k(n) = z_k(n)^T X_k(n) + \gamma\, \beta_{1,k}(n-1)r_{10,k}, \tag{12}$$

$$r_{00,k} = X_k(n)^T A_k X_k(n), \tag{13}$$

$$r_{10,k} = X_k(n-1)^T A_k X_k(n), \tag{14}$$

$$r_{11,k} = X_k(n-1)^T A_k X_k(n-1), \text{ and} \tag{15}$$

$$A_k = \begin{pmatrix} \alpha 1 & & & 0 \\ & \alpha 2 & & \\ & & \ddots & \\ 0 & & & \alpha n \end{pmatrix}, \tag{16}$$

with $X_k(n)$, $Y_k(n)$, $\hat{Y}_k(n)$, $W_k(n)$, and $e_k(n)$ defined as the far-end signal, the echo signal, the echo replica, the background noise, and the subband echo canceller output signals in the k-th subband, respectively. δ in $\beta_{1,k}(n)$ and $\beta_{2,k}(n)$ is a regularized parameter to avoid dividing a small denominator. Previously, we used a small positive constant value as δ [9]. In this paper, we use an estimated $\delta_k(n)$, which depends on time. $\delta_k(n)$ in the second-order projection algorithm is written as

$$\delta_k(n) = (L_k \times \overline{[W_k(n)]^2})^2 \tag{17}$$

based on the Kalman filter theory. Here, L_k is the number of filter taps in each subband. We applied the noise level estimated in the noise-reduction part to Eq. (17).

We confirmed the effectiveness of the proposed echo cancellation technique using computer simulation with a speech signal. This signal was sampled at a rate of 16 kHz. The echo signal to noise ratio was 15 dB, the number of filter taps in each subband was 40, and the impulse response length in the room was 80 ms.

Figure 4 shows the effect of applying the estimated noise level. The echo signal was more stably reduced when the time-dependent variable was used rather than a constant value, even when noise level was high.

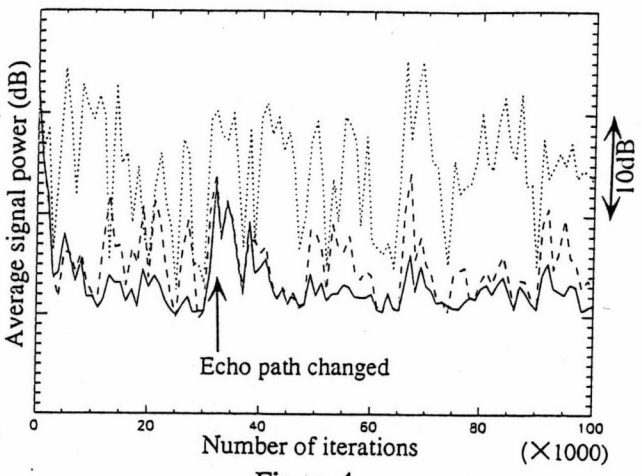

Figure 4
Effect of applying estimated noise level. Solid line: power of residual signal after cancelling echo by using estimated noise level for $\delta\, k(n)$, Dashed line: power of residual signal after cancelling echo by using a small constant value for $\delta\, k$. Dotted line: power of microphone output signal.

6. SUMMARY

We investigated the combination of subband echo canceller and noise reduction, where they use the same subband signals divided by a poly-phase filter bank. Considering the subjective results and amount of calculation of four different noise-reduction techniques, we propose to use the Wiener filtering based on the a priori SNR with an input signal added when the SNR is low. We also improved the echo cancellation techniques by using a time-dependent variable based on the noise level to avoid the divergence of the filter coefficient in the second-order affine projection algorithm; this stably reduced the echoes for high noise level. Our system can stably reduce noise and echoes with only small resulting distortion and delay in a noisy environment.

Acknowledgments

We are grateful to Dr. Nobuhiko Kitawaki, Mr. Junji Kojima, Dr. Yutaka Kaneda and Dr. Kenzo Ito for their discussions and technical advice. In addition, we thank those who helped with the listening tests.

References

1) Y. Guelou, A. Benamar & P. Scalart, in Proc. IEEE ICASSP'96, pp. 637-640, Atlanta, USA, (1996).
2) B. Ayad & G. Faucon, in Proc. International Workshop on Acoustic Echo and Noise Control, pp. 91-94, Roros, Norway, (1995).
3) Y. Ephraim & D. Malah, IEEE Trans. on ASSP, vol.32, no. 6, pp. 1109-1121, Dec (1984).
4) J. S. Lim & A. V. Oppenheim, in Proc. IEEE, vol.67, no. 12, pp. 1586-1604, Dec (1979).
5) R. J. McAulay & M. L. Malpass, IEEE Trans. on ASSP, vol.28, no. 2, pp. 137-145, Apr (1980).
6) S. F. Boll, IEEE Trans. on ASSP, vol.27, no. 2, pp. 113-120, Apr (1979).
7) P. Scalart & J. V. Filho, in Proc. IEEE ICASSP'96, pp. 629-632, Atlanta, USA, (1996).
8) S. Makino & Y. Kaneda, Trans. IEICE Jalan, vol, E75-A. no. 11, pp.1500-1508, 1992.
9) Y. Haneda, S. Makino, J. Kojima, & S. Shimauchi, in Proc. International Workshop on Acoustic Echo and Noise Control, pp. 79-82, Roros, Norway, (1995).