Implementing and Evaluating an Audio Teleconferencing Terminal with Noise and Echo Reduction

Sumitaka Sakauchi, Akira Nakagawa, Yoichi Haneda, and Akitoshi Kataoka
NTT Cyber Space Laboratories, NTT Corporation
3–9–11, Midori–cho, Musashino–shi, Tokyo, 180–8585 Japan
Tel.: +81 422 59 4207; Fax.: +81 422 60 7811
Email: sakauchi.sumitaka@lab.ntt.co.jp

ABSTRACT

We have developed an audio teleconferencing terminal for use in noisy office environments. The terminal has an acoustic-echo cancellation filter in the time domain, noise and echo reduction in the frequency domain, and variable loss insertion control. The echo cancellation filter uses an exponentially weighted step-size projection algorithm and has a foreground/background structure. Noise and echo reduction, which is based on short-time spectral amplitude estimation, reduce the level of ambient noise and the amount of residual echo hidden in that noise. Evaluation of this terminal shows that the echo level is suppressed by 40 dB or more and the noise level is reduced by about 14 dB, levels that are good enough for an audio teleconference in noisy office.

1 Introduction

Acoustic echo cancellation is indispensable to hands-free telecommunications, for example in videoconferences. Conventional acoustic echo cancellers with adequate performance for quiet environments have previously been fabricated [1]. However, the demand for hands-free telecommunications in noisy environments, e.g. audio teleconferences between noisy offices, has increased. In a noisy office environment, the microphone of an audio teleconferencing terminal picks up ambient noise from noise sources such as air conditioners and the cooling fans of computers. This ambient noise not only reduces the clarity of the speech signals [2], but also negatively affects the acoustic-echo cancellation filter.

We thus developed an audio-teleconferencing terminal with noise and echo reduction to reduce the level of ambient noise and the amount of residual echo hidden in that noise. In this paper, we will describe the new terminal and show the results of an objective evaluation of its performance.

Figure 1: Photograph of proposed audio teleconferencing terminal.

2 Specifications

Fig. 1 shows the developed audio teleconferencing terminal. We used a fixed-point digital signal processor (DSP) to implement the terminal. The narrow frequency range, from 300 to 3400 Hz, is of course sufficient for telephony. The length of the adaptive filter is 160 ms. The terminal is for connection to the public switched telephone network (PSTN) via a network control unit (NCU). We thus equipped it with a line-echo canceller.

The responses of the four built-in-microphones at the corners of the terminal are made to form a null point at the built-in-loudspeaker at the center; this reduces acoustic coupling from the loudspeaker to the microphones [3]. External transducers increased the number of conference participants. The auxiliary input and output ports also let us place the terminal in the circuits of other telecommunications systems. We can thus use it as the echo-canceller unit of, for example, a video-teleconferencing system.

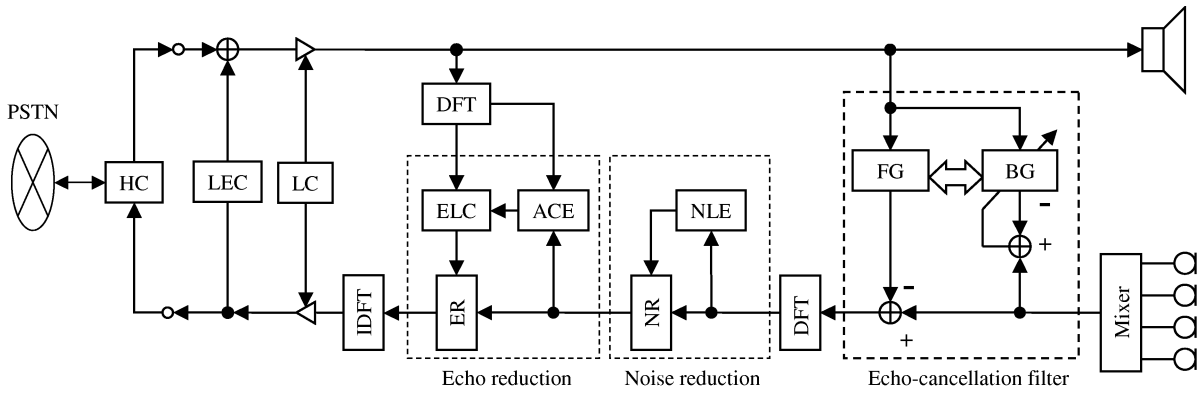


Figure 2: Block diagram of audio teleconferencing terminal.

3 System description

Fig. 2 shows a block diagram of the new terminal. Acoustic echo and ambient noise reducing are handled as follows. Firstly, we use an echo-cancellation filter (ECF) that brings acoustic echo levels down to almost the same level as the noise. Next, the level of ambient noise is lowered by noise reduction (NR) and residual echo is suppressed by echo reduction (ER).

3.1 Echo-cancellation filter (ECF)

To achieve fast convergence for speech input, ECF uses the second-order exponentially weighted step-size (ES) projection algorithm. This algorithm uses both on the expected variation in a room impulse response and the effect of speech whitening [4].

To distinguish between double talk and an echo path change, ECF has a foreground/background (FG/BG) structure, which has a fixed filter and an adaptive filter [5]. The fixed filter cancels the echo, and the adaptive filter estimates the room impulse response. The coefficients for the fixed filter are transferred from the adaptive filter when it converges with the adaptive filter. During double talk, the coefficients of the adaptive filter may become misadjusted, increasing the mean-squared error in the adaptive filter. If this occurs, the filter coefficients of the adaptive filter are not transferred to the fixed filter. The filter coefficients of the fixed filter are therefore not updated during double talk, and the echo canceling level before double talk begins is maintained.

3.2 Noise reduction (NR)

The noise reduction process is based on a short-time spectral amplitude (STSA) estimation [6]–[8]. Let $s(k)$ and $n(k)$ denote the near-end speech and the ambient noise. The microphone input signal $z_{\mathrm{n}}(k)$ is given by $z_{\mathrm{n}}(k) = s(k) + n(k)$. Let $S_\omega = |S_\omega|\, e^{j\phi}$, $N_\omega = |N_\omega|\, e^{j\psi}$, and $Z_{\mathrm{n}\omega} = |Z_{\mathrm{n}\omega}|\, e^{j\theta}$ denote the spectral components of the $s(k)$, $n(k)$, and $z_{\mathrm{n}}(k)$, where ω is frequency, ϕ, ψ, θ are phase, and $|\cdot|$ denotes amplitude.

In achieving noise reduction based on STSA estimation, the amplitude of the near-end speech is estimated as being obtained from $|Z_{\mathrm{n}\omega}|$ by a multiplicative nonlinear gain function $G_{\mathrm{n}\omega}$ in the frequency domain.

$$\hat{S}_\omega = |\hat{S}_\omega| e^{j\theta} = G_{\mathrm{n}\omega} \cdot |Z_{\mathrm{n}\omega}|\, e^{j\theta} \tag{1}$$

The gain function based on Wiener filtering $\hat{G}^{\mathrm{w}}_{\mathrm{n}\omega}$ is given by

$$\hat{G}^{\mathrm{w}}_{\mathrm{n}\omega} = \frac{\mathrm{E}[|S_\omega|^2]}{\mathrm{E}[|S_\omega|^2] + \mathrm{E}[|N_\omega|^2]} = \frac{|Z_{\mathrm{n}\omega}|^2 - \mathrm{E}[|N_\omega|^2]}{|Z_{\mathrm{n}\omega}|^2} \tag{2}$$

where $\mathrm{E}[|S_\omega|^2]$ and $\mathrm{E}[|N_\omega|^2]$ denote the ensemble average of $s(k)$ and $n(k)$.

A noise level estimator (NLE) estimates $\mathrm{E}[|N_\omega|^2]$ by updating the minimum value of the microphone input $|Z_{\mathrm{n}\omega}|^2$. However, ambient noise varies over time, so estimated $\mathrm{E}[|N_\omega|^2]$ is not completely correct. When $\mathrm{E}[|N_\omega|^2]$ is estimated to be larger than its real value, processed signal $\hat{S}_\omega$ is distorted and speech quality degrades. Therefore, speech distortion is reduced by adding a small amount of the untreated input signal to the output signal obtained by noise reduction [9].

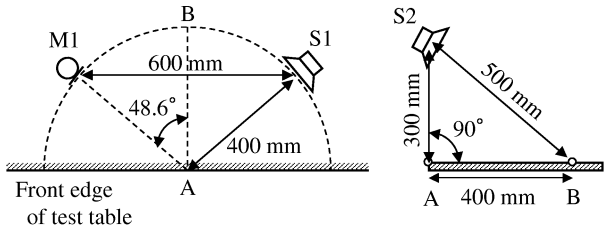
$$\hat{S}_\omega = (1-\alpha) Z_{\mathrm{n}\omega} + \alpha\, \hat{G}_{\mathrm{n}\omega}\, |Z_{\mathrm{n}\omega}|\, e^{j\theta} \tag{3}$$

The added untreated input signal masks distorted components of the speech so that speech quality improves in a poor SNR environment. When the untreated input signal adding ratio $(1-\alpha)$ is set to 0.2, ambient noise can theoretically be reduced a maximum amount of 14dB.

3.3 Echo reduction (ER)

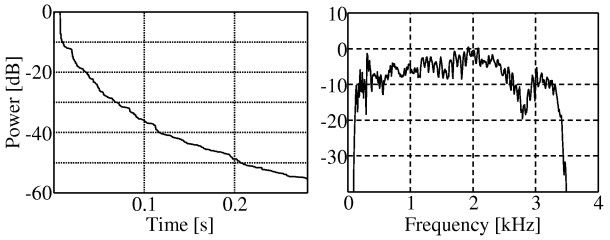
The echo reduction process is also based on STSA estimation [10]–[12]. Let $x(k)$ and $\boldsymbol{h}$ denote the received speech and the echo-path. The spectral components of the echo $y(k)$ is given by

$$Y_\omega = H_\omega \cdot X_\omega. \tag{4}$$



(a) Top view (b) Side view

Figure 3: Physical test arrangements for objective measurements.



(a) Reverberation curve (b) Frequency characteristics

Figure 4: (a) Reverberation curve and (b) frequency characteristics of the impulse response of the paths between a loudspeaker S1 to a microphone M1.

The gain function $G_{e\omega}$ for echo reduction can be calculated by transposing the ensemble average of noise $\mathrm{E}[|N_\omega|^2]$ to the value of residual echo $\mathrm{E}[|Y_\omega|^2]$ in Eq. (2). Hence, the echo level calculator (ELC) estimates the ensemble of the residual echo used by the spectral components of the received speech and the acoustic coupling as follows.

$$\mathrm{E}[|Y_\omega|^2] = \mathrm{E}[|\hat{H}_\omega|^2]\cdot|X_\omega|^2 + \beta\cdot\mathrm{E}[|Y_\omega^{f-1}|^2] \qquad (5)$$

By adding the value $\mathrm{E}[|Y_\omega^{f-1}|^2]$ in the previous processing frame recursively, when reverberation time is longer than the length of the processing frame, the ensemble of the residual echo $\mathrm{E}[|Y_\omega|^2]$ can be estimated correctly. The acoustic coupling estimator (ACE) in Fig. 2 estimates $\mathrm{E}[|\hat{H}_\omega|^2]$ by updating the minimum value of the ratio of the received speech $|X_\omega|^2$ and the microphone input $|Z_{e\omega}|^2$, where acoustic coupling $\mathrm{E}[|\hat{H}_\omega|^2]$ includes a linear canceling process by ECF in addition to a real acoustic echo-path.

Furthermore, to improve the subjective performance, the desired echo-return loss is set on the basis of subjective auditory characteristics [13].

4 Evaluation

4.1 Conditions

We connected D/A and A/D converters to the auxiliary input and output, and used a combination of directly input and electro acoustically reproduced digitized speech to test the performance of the terminal. Fig. 3 shows physical test arrangements for objective measurements. M1 and S1 represented the near-end microphone and loudspeaker, and loudspeaker S2 simulated the near-end talker. The loudspeaker and microphone levels were as prescribed by ITU Recommendation P. 34 [14]. Fig. 4 shows reverberation curve and frequency characteristics of the impulse response of the paths between loudspeaker S1 to microphone M1. The reverberation time in the test room was about 250 msec. All speech was in Japanese and the ambient noise was white and at about 55 dBA. Fig. 5 and Fig. 6 show the received speech and the near-end speech.

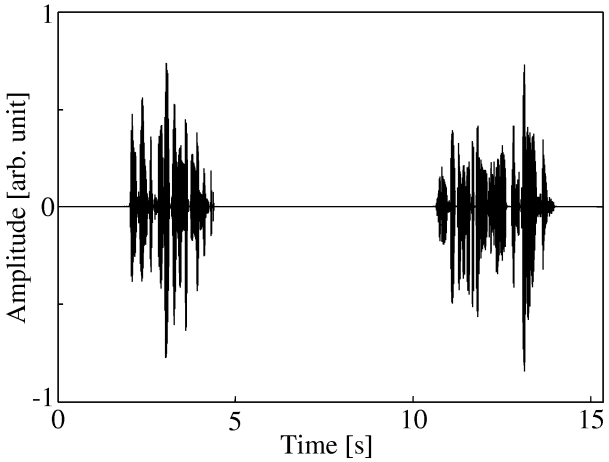


Figure 5: Received speech signal $x(k)$.

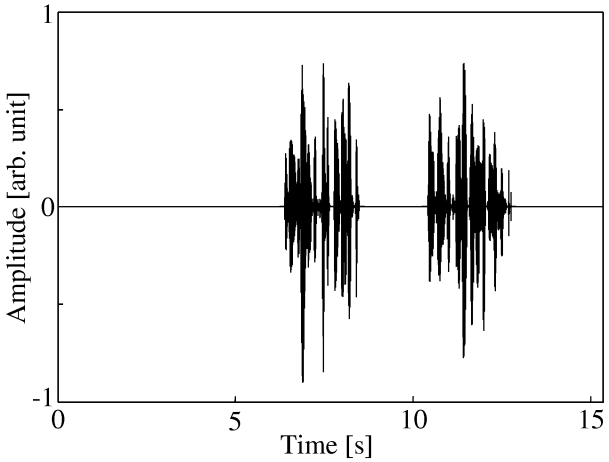


Figure 6: Near-end speech signal $nes(k)$.

4.2 Results

Fig. 7 shows the microphone input signal. During the whole period, the microphones are picking up the ambient noise. The waveform in period "A" shows echo-signal during a single-talk situation when the microphones are picking up the echo of the received signal from the far-end terminal. That in period "B" shows sent-signal during a single-talk situation where the microphones are picking up the near-end speech. In pe-

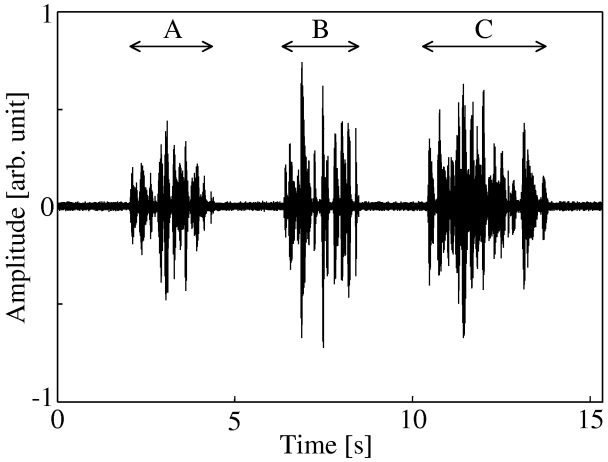


Figure 7: Microphone input signal $z(k)$. Period A: echo-signal during single-talk situation, Period B: sent-signal during single-talk situation, Period C: double-talk situation.

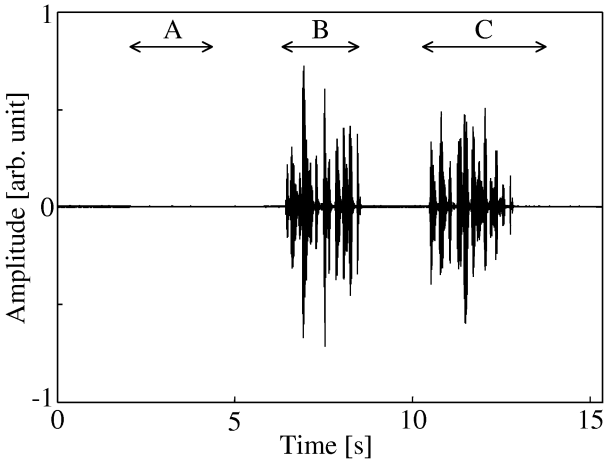


Figure 8: Line-output signal: signal of Fig. 7 after processing by proposed terminal.

riod "C" a double-talk situation is shown, where the microphones pick up both an echo of the far-end signal and near-end speech. Fig. 8 shows a line-output signal. The echo in the period "A" has been almost completely eliminated and the ambient noise has been greatly reduced over the whole period. The echo level is suppressed by 40 dB or more and the noise level is reduced by about 15 dB. The near-end speech was only slightly distorted and the subjective quality was good.

5 Conclusion

We have developed an audio teleconferencing terminal with noise and echo reduction capabilities. Evaluation of this terminal shows that it is effective in suppressing echo and reducing noise.

Acknowlegements

We are very grateful to Dr. Hisashi Ohara, Mr. Masashi Tanaka, Ms. Junko Sasaki, Mr. Suehiro Shimauchi for their support and constructive suggestions.

References

[1] Ch. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, Th. Schertler, G. Schmidt, and J. Tilp, "Acoustic Echo Control: An Application of Very-High-Order Adaptive Filters," IEEE Signal Processing Magazine, vol. 16, pp. 42–69, Jul. 1999.

[2] J. S. Collura, "Speech enhancement and coding in harsh acoustic environments," Proc. IEEE Workshop on Speech Coding, pp. 162–164, Finland, Jun. 1999.

[3] A. Nakagawa, S. Shimauchi, S. Aoki, S. Makino, and Y. Kaneda: "A study of multi-microphone system for hands-free teleconferencing units," in Proc. IWAENC99, pp. 104–107, 1999.

[4] S. Makino and Y. Kaneda: "Exponentially weighted step-size projection algorithm for acoustic echo cancellers," Trans. IEICE Japan, vol. E75-A, no. 11, pp. 1500–1508, Nov. 1992.

[5] Y. Haneda, S. Makino, J. Kojima, and S. Shimauchi, "Implementation and evaluation of an acoustic echo canceller using duo-filter control system," Proc. EUSIPCO'96, vol. 2, pp. 1115–1118, Sept. 1996.

[6] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. on ASSP, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[7] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. IEEE, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[8] R. J. McAulay, and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE Trans. on ASSP, vol. 28, no. 2, pp. 137–145, 1980.

[9] J. Sasaki, Y. Haneda, and S. Makino, "Noise reduction for subband acoustic echo canceller," Proc. Third Joint Meeting, Acoustical Society of America and Acoustical Society of Japan, pp. 1285–1290, Dec. 1996.

[10] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals," Signal Processing, vol. 64, no. 1, pp. 33–47, Jan. 1998.

[11] E. Hänsler, G. U. Schmidt, "Hands-free telephones - joint control of echo cancellation and postfiltering," Signal Processing, vol. 80, no. 11, pp. 2295–2305, 2000.

[12] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," IEEE Trans. on Speech and Audio processing, vol. 10, no. 5, pp. 245–256, 2002.

[13] S. Sakauchi, Y. Haneda, S. Makino, M. Tanaka, and Y. Kaneda: "Subjective assessment of the desired echo return loss for subband acoustic echo cancellers," IEICE Trans., Vol. E83-A, No. 12, pp. 2633–2639, 2000.

[14] International Telecommunication Union, "Transmission performance of hands-free telephones," Recommendation P. 34 CCITT Blue Book, 5, 1988.