Double-talk Robust Acoustic Echo Cancellation for CD-quality Hands-free Videoconferencing System

Masahiro Fukui, *Member*, IEEE, Suehiro Shimauchi, *Member*, IEEE, Yusuke Hioka, *Senior Member*, IEEE, Akira Nakagawa, and Yoichi Haneda, *Senior Member*, IEEE

Abstract — *A novel monaural acoustic echo cancellation method developed for a 20-kHz wideband hands-free video- or teleconferencing system is presented. The method can effectively reduce undesired acoustic echo included in a signal arriving at a microphone from a loudspeaker and can emphasize the target talker's voice when near-end and far-end talkers speak simultaneously (i.e. double-talk). The method estimates a power frequency response of an acoustic echo path between a loudspeaker and microphone (acoustic coupling level) whether or not double-talk has occurred, then calculates a post-filter that effectively reduces the undesired echo. In the echo cancellation processing, the calculation complexity is reduced to make the processing suitable for real-time implementation by using a low-complexity subband approach that employs a new subband filtering algorithm. Experiments were conducted to examine the performance of the proposed method. The results indicated that the proposed method delivered natural-sounding near-end speech even during double-talk periods while sufficiently suppressing the undesired echo*[1].

Index Terms — Acoustic echo canceller, acoustic coupling level, Wiener filtering, short-time spectral amplitude, double-talk, hands-free.

I. Introduction

An acoustic echo canceller (AEC) [1]-[3] is one of the key technologies for achieving hands-free telecommunication such as in video conferencing systems [4], [5]. It prevents detrimental acoustic echo and howling, which often occur in such hands-free systems, and emphasizes the desired talker's voice when near-end and far-end talkers speak simultaneously (double-talk).

In high-quality videoconferencing systems that have been introduced in recent years [4], the frequency band of the audio signal is supported up to compact disc (CD)-quality, i.e. 20-kHz wideband. Because of its high-fidelity quality, the AEC must also maintain the high sound quality that these systems provide to users.

Various conventional echo cancellation techniques have been developed [6]-[11]. The AEC consisting of an adaptive filter (ADF) [7], [8] combined with echo reduction (ER) [9]-[11], has frequently been applied to practical products because of its promising performance. The ADF estimates and cancels out the acoustic echo by adaptively identifying an unknown acoustic echo path. However, some residual echo still remains in its output (error signal) because in practice, there are limitations on the calculation and memory capacities required to accurately calculate the echo path, which frequently changes in practical environments. ER is therefore used as a non-linear post-filter that reduces the residual echo included in the error signal. The ER estimates the power spectrum of residual echo using the squared amplitude frequency response of the acoustic coupling (acoustic coupling level: ACL), and then calculates the post-filter that reduces the residual echoes.

The combination of the ADF and ER delivers reasonable performance for a narrowband teleconferencing system that limits audio frequencies to the range of 300 Hz to 3.4 kHz. However, two problems arise when the method is used in CD-quality wideband systems: i) the sound of a near-end talker's voice (near-end speech) is sometimes muffled in high frequency ranges during double-talk periods because of the incomplete ER process caused by low-accuracy ACL estimation, and ii) the amount of computation for the ADF increases exponentially as the sampling frequency becomes higher.

One of the objectives of this study was to improve the estimation accuracy of the ACL and to eliminate the muffled sound of near-end speech after applying ER. Another objective was to reduce the computational complexity of the whole process including the ADF in order to realize real-time implementation. To solve the first problem, this paper addresses the accuracy improvement of the ACL estimation by focusing on the assumption that echo and near-end speech are statistically independent not only in a time axial direction but also in a frequency axial direction. This is based on an algorithm similar to that of a previous technique [12], but different in that the target is 20-kHz wideband. Then, to improve the computational efficiency, this paper employs a low-complexity subband approach with a new subband filtering algorithm in the AEC. The fast Fourier transform (FFT) length of the new algorithm is shorter than the FFT length used in conventional subband filtering because up-and-down sampling is performed in the frequency domain. This paper evaluated the performance of the proposed method by

[1] M. Fukui is with NTT Media Intelligence Laboratories. Tokyo 180-8585 JAPAN (e-mail: fukuimas@ieee.org).

S. Shimauchi is with NTT Media Intelligence Laboratories. Tokyo 180-8585 JAPAN (e-mail: shimauchi.suehiro@lab.ntt.co.jp).

Y. Hioka is with the Department of Mechanical Engineering, University of Auckland, Auckland 1010 New Zealand (e-mail: yusuke.hioka@ieee.org).

A. Nakagawa is with NTT Media Intelligence Laboratories. Tokyo 180-8585 JAPAN (e-mail: nakagawa.akira@lab.ntt.co.jp).

Y. Haneda is with the Faculty of Informatics and Engineering, The University of Electro-Communications, Tokyo 182-8585 JAPAN (e-mail: haneda.yoichi@uec.ac.jp).

implementing it using an AEC-unit prototype that has a digital signal processor (DSP) board. Experimental results using the prototype showed that the proposed method delivered more natural sounding near-end speech compared to that of a conventional method, while sufficiently reducing the residual echo.

The rest of this paper is organized as follows. Sec. 2 presents the video conferencing system with an AEC unit. Sec. 3 describes the ordinary ER and the proposed ACL estimation method. The real-time implementation is introduced in Sec. 4. Experimental results using the AEC-unit prototype are described in Sec. 5, and this paper is concluded with remarks in Sec. 6.

II. Videoconferencing System with AEC Unit

Fig. 1 shows a videoconferencing system with an AEC unit; the flow of the audio signal processing of the AEC is illustrated in Fig. 2. The signal received from the far-end is played back through the loudspeaker, which is added as the undesired acoustic echo to the near-end speech observed by the microphone. The AEC generally removes such echo components from the microphone signal in order to transmit only the near-end speech to the far-end.

III. Improvement of Double-Talk Performance

The problem of near-end speech degradation is mainly caused by the low estimation accuracy of the ACL used in the ER. The conventional ACL estimation method [11] sometimes over-estimates the ACL when near- and far-end speeches exist simultaneously; also known as double-talk. The overestimation causes near-end speech in double-talk being muffled after ER is applied. To solve this problem, this paper proposes a new ACL estimation method that can accurately estimate the ACL even during double-talk periods. The ordinary ER process and the proposed double-talk-robust ACL estimation method are described in the remainder of this section.

The ER is a popular frequency-domain post-filter based on a short-time spectral amplitude (STSA) estimation [13]; it estimates residual echo levels and suppresses the residual echo using multiplicative gains. A block diagram of the AEC, which shows how the ER can be performed with the proposed method, is shown in Fig. 3.

The AEC receives the signal $x(n)$ from the far-end at a discrete time index n. This signal is picked up as an acoustic echo signal by the microphone after passing through the room echo path that has an impulse response modeled as $\mathbf{h}(n)=[h_1(n),\ldots,h_L(n)]^T$, where L is the effective length of the impulse response and T is the transposition, respectively. Given the reference input vector, $\mathbf{x}(n)=[x(n),\ldots,x(n-L+1)]^T$, and the adaptive filter vector, $\mathbf{w}(n)=[w_1(n),\ldots,w_L(n)]^T$, the output signal of ADF, $y(n)$, can be written in terms of the reference input

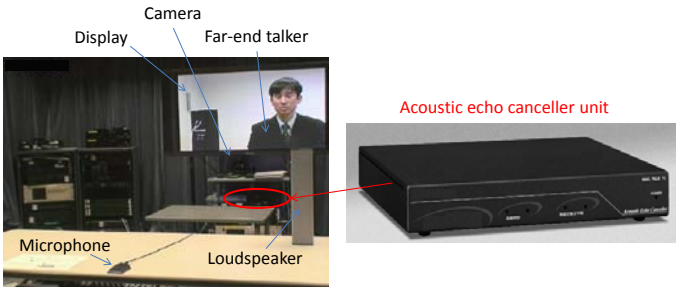

Fig. 1. Left: hands-free video conference system; Right: prototype of acoustic echo canceller unit.

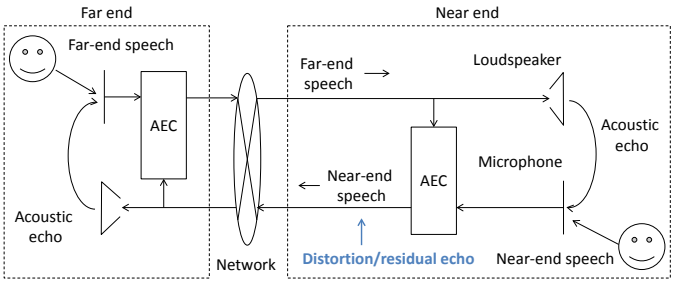

Fig. 2. Flow of acoustic echo cancellation.

vector, $\mathbf{x}(n)^T$, which is convoluted by the impulse response between reference and ADF output signals (*residual echo path*), $\mathbf{h}'(n)=[h'_1(n),\ldots,h'_L(n)]^T$, including the near-end speech signal $s(n)$ as

$$\begin{aligned} y(n) &= d(n)+s(n), \\ &= \mathbf{x}^T(n)\{\mathbf{h}(n)-\mathbf{w}(n-1)\}+s(n), \\ &= \mathbf{x}^T(n)\mathbf{h}'(n)+s(n). \end{aligned} \tag{1}$$

The short-time Fourier transform of $y(n)$ is represented as follows:

$$Y_i(\omega)=D_i(\omega)+S_i(\omega), \tag{2}$$

where ω is a discrete frequency index, i is a discrete frame index, and $D_i(\omega)$ and $S_i(\omega)$ are the short-time Fourier transforms of $d(n)$ and $s(n)$, respectively. The output of the ER is expressed as

$$\hat{S}_i(\omega)=G_i(\omega)Y_i(\omega), \tag{3}$$

where $\hat{S}_i(\omega)$ is the short-time Fourier transform of transmitted signal $\hat{s}(n)$, i.e. the estimate of $S_i(\omega)$. Here, $G_i(\omega)$ is the echo-reduction gain that is calculated according to the Wiener filtering method [14] obtained by

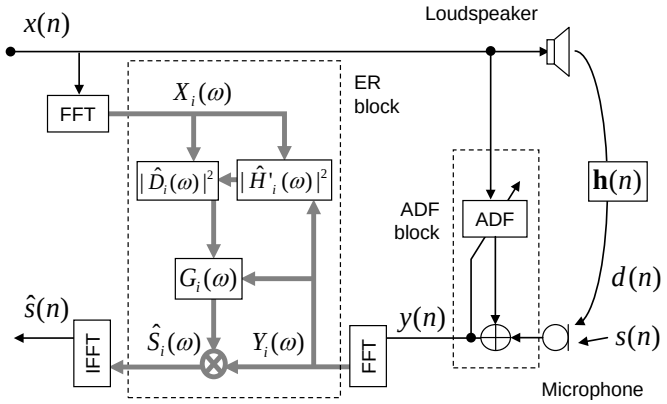

Fig. 3. Block diagram depicting AEC with proposed ER method.

$$G_i(\omega) = \frac{|Y_i(\omega)|^2 - |\hat{D}_i(\omega)|^2}{|Y_i(\omega)|^2}, \tag{4}$$

where $|\hat{D}_i(\omega)|^2$ is the estimate of the residual echo level $|D_i(\omega)|^2$. Here, $|D_i(\omega)|^2$ is calculated as

$$|\hat{D}_i(\omega)|^2 = |\hat{H}_i'(\omega)|^2 |X_i(\omega)|^2, \tag{5}$$

where $|\hat{H}_i'(\omega)|^2$ is the estimate of ACL $|H_i'(\omega)|^2$, which is the power spectrum of $h'(n)$, and $|X_i(\omega)|^2$ is the power spectrum of $x(n)$.

The proposed method assumes that the ideal estimation of ACL can be defined by the following equation:

$$|\hat{H}_i'(\omega)|^2 = \frac{E[|D_i(\omega)|^2]}{E[|X_i(\omega)|^2]}, \tag{6}$$

where $E[\cdot]$ is the ensemble average. Equation (6) cannot be directly calculated during double-talk periods, so the proposed method obtains an estimate of ACL using the following equation:

$$|\hat{H}_i'(\omega)|^2 = \left(\frac{\sum_{r=-R_\omega}^{R_\omega} E[|X_i(\omega+r)||Y_i(\omega+r)|]}{\sum_{r=-R_\omega}^{R_\omega} E[|X_i(\omega+r)|^2]} \right)^2, \tag{7}$$

where R_ω indicates the number of frequency bins required for calculating the correlation between $X_i(\omega)$ and $Y_i(\omega)$ in a frequency axial direction. Equation (7) is a new simplified equation for the previously proposed technique [12]. In Eq. (7), all the complex number operations are approximately replaced with real number operations.

If the reference and near-end speech signals are uncorrelated, the near-end speech component can be removed by calculating the amplitude correlation between both signals, and Eq. (7) is then approximated to the following equation:

$$\begin{aligned} |\hat{H}_i'(\omega)|^2 &= \left(\frac{\sum_{r=-R_\omega}^{R_\omega} E[|X_i(\omega+r)||D_i(\omega+r)+S_i(\omega+r)|]}{\sum_{r=-R_\omega}^{R_\omega} E[|X_i(\omega+r)|^2]} \right)^2 \\ &\approx \left(\frac{\sum_{r=-R_\omega}^{R_\omega} E[|X_i(\omega+r)||D_i(\omega+r)|]}{\sum_{r=-R_\omega}^{R_\omega} E[|X_i(\omega+r)|^2]} \right)^2 \\ &\approx \frac{\sum_{r=-R_\omega}^{R_\omega} E[|D_i(\omega+r)|^2]}{\sum_{r=-R_\omega}^{R_\omega} E[|X_i(\omega+r)|^2]}. \end{aligned} \tag{8}$$

This means the ER is able to directly estimate the ACL even during double-talk periods.

IV. Real-time Implementation

Because a large delay seriously obstructs bidirectional communication, the AEC should run on a real-time basis. For real-time implementation in 20-kHz wideband system, this paper proposes a cost-effective AEC algorithm achieved using a subband approach with low-complexity subband filtering. The specifications, subband approach, and the proposed subband filtering algorithm are described in this section.

A. Specifications

The proposed AEC algorithm was utilized for a hands-free voice communication unit. A circuit board of the unit is shown in Fig. 4, and the specifications are listed in Table I.

The AEC was implemented on a single fixed-point DSP chip. The sampling frequency of the A/D or D/A converters is 48 kHz. Its frequency range is from 100 Hz to 20 kHz, which corresponds to the frequency range of widely used wideband transmission systems (CD quality). The total signal delay is 30 ms. The maximum filter tap length in the ADF is 140 ms. The total calculation performance of the AEC is 193 MIPS, which means the system can achieve the real-time acoustic echo cancellation on a low-cost DSP chip while maintaining the wide signal bandwidth.

B. Subband Approach

In the AEC, the ADF has a significant impact on the overall

TABLE I
SPECIFICATIONS OF ACOUSTIC ECHO CANCELLER UNIT

Item	Description
Size	200 mm (W) × 150 mm (D) × 35 mm (H)
Weight	700 g
Sampling frequency	48 kHz
Frequency range	100 HZ – 20 KHZ
Interface	RCA jacks:
	Line input × 1
	Line output × 1
	Audio input × 1
	Speaker output × 1

Fig. 4. Circuit board of acoustic echo canceller prototype.

complexity because the degree of computational complexity required in the ADF increases in proportion to the square of the sampling frequency because of the convolution operation of adaptive filtering. For example, the amount of computation required in the ADF increases 16-fold when the sampling frequency increases 3-fold from 8 kHz to 48 kHz. This means that it is essential to reduce the computational complexity of the ADF to achieve a real-time CD-quality wideband AEC.

This paper has introduced a subband approach that first divides a signal into several narrow bandwidths and then applies the AEC to the signal in each divided frequency band. The advantage of this approach is that the computational complexity of the convolution operation can be reduced. Figure 5 shows the flow chart of the AEC realized in the high frequency range from 100 Hz to 20 kHz at a 48-kHz sampling frequency. The subband AEC consists of the ADF, ER, decimation (DEC), interpolation (INT), low-pass filter (LPF), band-pass filter (BPF), and high-pass filter (HPF) processes. The impulse response and frequency responses of LPF, BPF, and HPF are shown in Fig. 6.

The subband approach performed in analysis and synthesis filter blocks can reduce the computational complexity of the convolution operation by $1/M$, where M is the DEC ratio.

In this unit, the computational complexity required in the ADF was reduced by 75% by quartering the band. Moreover, if the ADF is not processed for frequencies above

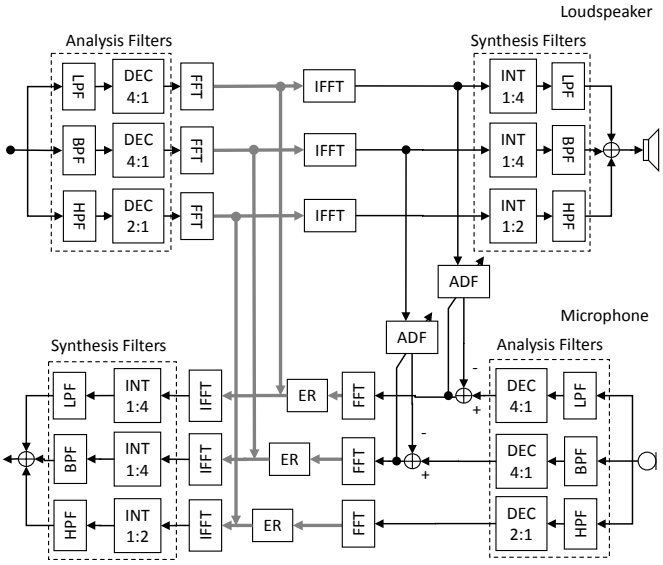

Fig. 5. Signal flow of implemented AEC.

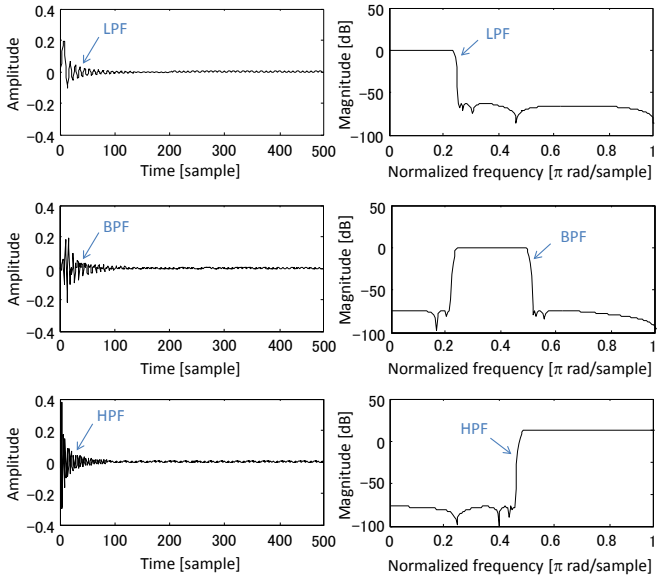

Fig. 6. Impulse responses and frequency responses of low-pass, band-pass, and high-pass filters.

12 kHz, the computational cost can be further reduced by about 50%. Because the power of speech components above 12 kHz is extremely low, the acoustic echo component above 12 kHz can be sufficiently suppressed without the ADF but only by the ER. In total, the computational complexity required in the ADF was reduced by about 90% using this configuration.

C. Subband Filtering

To further reduce the computational complexity of the AEC, the analysis and synthesis filters used in the subband filtering were revised. These filters incur a high computational cost because the subband filtering uses a convolution operation between an input signal and a band-division filter whose length is a quarter of the adaptive filter length. To avoid the convolution operation in the time domain, a frequency-domain

filtering method is usually used. The problem with this approach is that the computational complexity required in the FFT is high because long FFT lengths are needed to avoid aliasing.

This paper proposes frequency-domain DEC and INT implemented in short FFT lengths. First, the analysis filter consists of the LPF, BPF, HPF, and DEC processes, as shown in Fig. 7. The frequency ranges of the LPF, BPF, and HPF are 0 Hz – 6 kHz, 6 kHz – 12 kHz, and 12 kHz – 24 kHz, respectively. An input signal is transformed into the frequency domain and then is band-limited by frequency-domain DEC filters (LPF, BPF, and HPF) as

$$Z_i(e^{j\omega\tau}) = U_i(e^{j\omega\tau}) I_i(e^{j\omega\tau}), \tag{9}$$

where τ is the sampling period, $Z_i(e^{j\omega\tau})$ is the frequency-domain band-limited input signal, $U_i(e^{j\omega\tau})$ is the frequency-domain DEC filter, LPF, BPF or HPF, and $I_i(e^{j\omega\tau})$ is the frequency-domain input signal. The proposed method realizes a precise DEC of $Z_i(e^{j\omega\tau})$ in the frequency domain by adding the component reflected into the lower frequencies, which is called the aliasing component below, to the band-limited input signal as follows:

$$Z_i^{\mathrm{D}}(e^{j\omega\tau'}) = \frac{1}{M}\sum_{k=0}^{M-1} Z\left(e^{j\frac{\omega\tau'-2\pi k}{M}}\right), \tag{10}$$

where $\tau' = M\tau$. $Z_i^{\mathrm{D}}(e^{j\omega T'})$ is the frequency-domain band-limited input signal decimated by M. This means that when $Z_i^{\mathrm{D}}(e^{j\omega\tau'})$ is transformed into the time domain by inverse-FFT (IFFT), the required IFFT points are reduced by $1/M$. The example of DEC is given in Fig. 7. For example, if the DEC ratio M is two, Eq. (10) is convertible as follows:

$$\begin{aligned} Z_i^{\mathrm{D}}(e^{j\omega\tau'}) &= \frac{1}{2}\sum_{k=0}^{2-1} Z\left(e^{j\frac{\omega\tau'-2\pi k}{2}}\right), \\ &= Z\left(e^{j\omega\tau}\right) + Z\left(e^{j\omega\tau-\pi}\right). \end{aligned} \tag{11}$$

Here, the second term denotes the aliasing component shifted only by π and with the addition of the first term.

Next, the synthesis filter consists of LPF, BPF, HPF, and INT processing, as shown in Fig. 5. This filter performs filtering and INT in the frequency domain. The aim is to reduce the FFT points required in a frequency-domain transform as effectively as that achieved with the DEC filter. An input signal is transformed into the frequency domain and interpolated in that domain by adding the aliasing component as

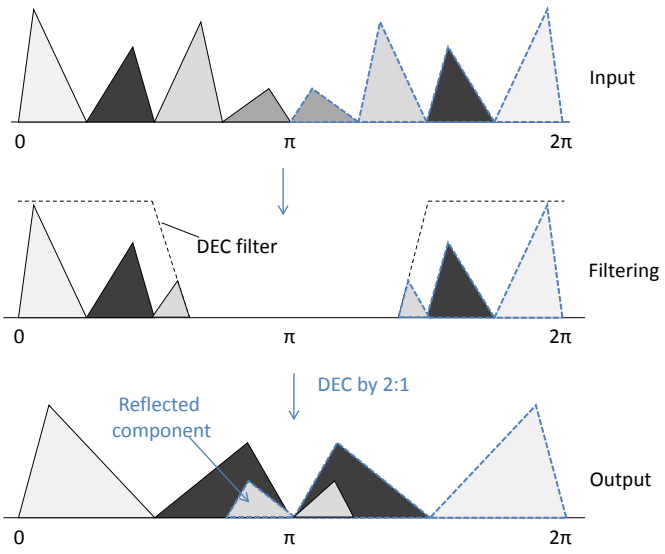

Fig. 7. Illustration showing how decimation can be performed in the frequency domain.

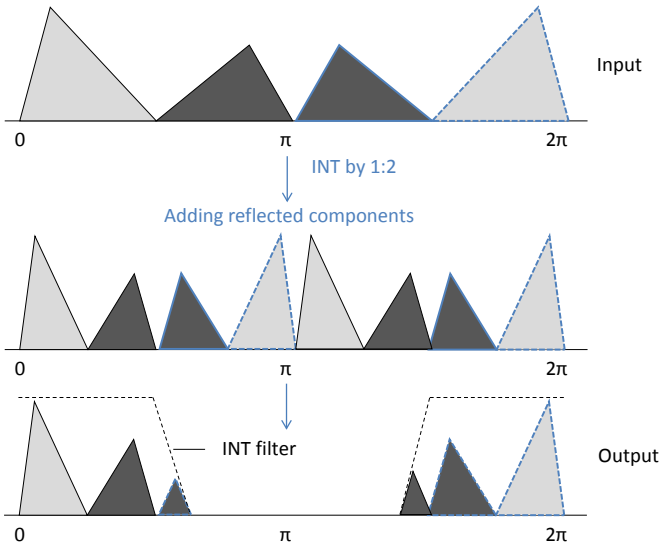

Fig. 8. Illustration showing how interpolation can be performed in the frequency domain.

$$I_i^U(e^{j\omega\tau}) = I_i(e^{j\frac{\omega\tau'}{M}}). \tag{12}$$

The required FFT points are reduced by $1/M$ by performing the INT after FFT processing. The interpolated signals are band-limited by INT filters, LPF, BPF, or HPF, in the frequency domain, resynthesized, and transformed into the time domain. An example of INT with $M = 2$ is given in Fig. 8.

V. Performance Evaluation

The performance of the proposed method implemented in the AEC-unit prototype was evaluated in a practical environment. This paper examined how the near-end ambient sound is processed and transmitted under various conversation states, particularly during double-talk periods. If the

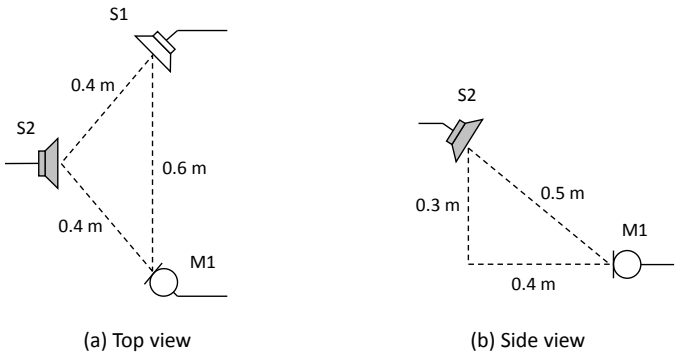

Fig. 9. Test arrangements for objective measurements.

difference between the near-end speech and transmitted signals after processing is reduced, it means that the system can provide more natural-sounding teleconferencing. In the evaluation, AECs using the conventional and proposed methods were compared. The conventional method was an AEC employing the conventional ACL estimation method [11].

A. Test Conditions

The arrangement of the microphones and sound sources used in the test is shown in Fig. 9. Here, M1 and S1 represent the near-end microphone and loudspeaker, respectively, and loudspeaker S2 simulated the near-end talker. The loudspeaker and microphone levels were as prescribed by ITU Recommendation P. 34 [15]. The room reverberation time was about 300 ms. Japanese speech was used in all conditions. The reference and near-end speech signals are shown in Fig. 10.

B. Experimental Results

The microphone input signal and the transmitted signals received after processing by the conventional and proposed AECs are shown in Fig. 11, respectively. The waveform during the single-talk period shows the acoustic echo signal when the microphone picks up the acoustic echo of the signal from the far end. In the double-talk period, the microphone picks up both an acoustic echo and near-end speech. The conventional and proposed AECs sufficiently suppress echo components over the entire period, as seen in Fig. 11.

For the evaluation of the echo-suppression level, the echo return loss enhancement (ERLE) measure, which is defined by ITU-T Recommendation G.168 [16], was used. The ERLE gives large values when the echo component is well suppressed by the AEC. The average ERLEs of the conventional and proposed methods are 36.1 dB and 37.6 dB in the single-talk period, respectively.

The levels of transmitted signals processed with the conventional and proposed methods during a double-talk period are shown in Figs. 12 and 13, respectively. The level of the transmitted signal processed by the proposed method is closer to that of the near-end speech component than that of the conventional method. In the transmitted signal processed with the proposed method, the acoustic echo was sufficiently suppressed, and the near-end speech passed through the ER with low degradation.

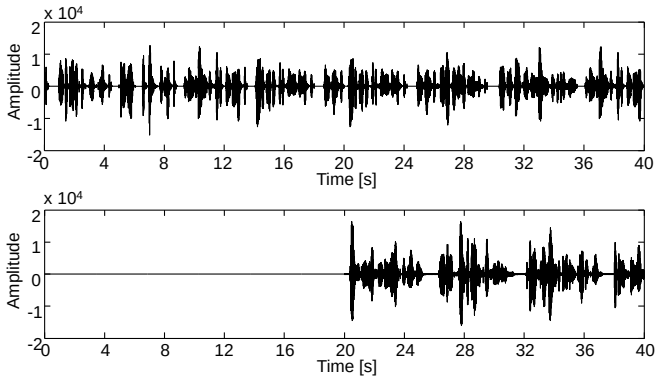

Fig. 10. Top: reference signal; bottom: near-end speech signal.

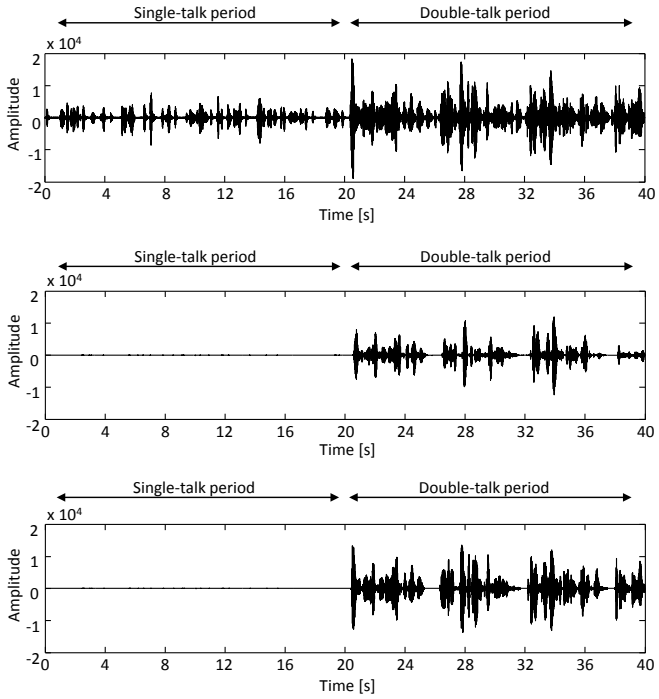

Fig. 11. Top: microphone input signal; middle: transmitted signal processed using conventional method; bottom: transmitted signal processed using proposed method.

This paper also evaluated the amount of speech distortion during a double-talk period using a linear predictive coding (LPC) cepstral distance [17] between the near-end speech and transmitted signals. The LPC cepstral distance is computed as

$$\mathrm{CD}(n) = \frac{10}{\log 10}\sqrt{2\sum_{k=1}^{16}[c(k,n)-\hat{c}(k,n)]^2}, \qquad (13)$$

where $\mathrm{CD}(n)$ is the LPC cepstral distance and n denotes the discrete time. Terms $c(k,n)$ and $\hat{c}(k,n)$ are the k-th cepstral coefficients of near-end speech and transmitted signals, respectively. The LPC cepstral distance gives a small

value when the transmitted signal does not suffer a loss of the

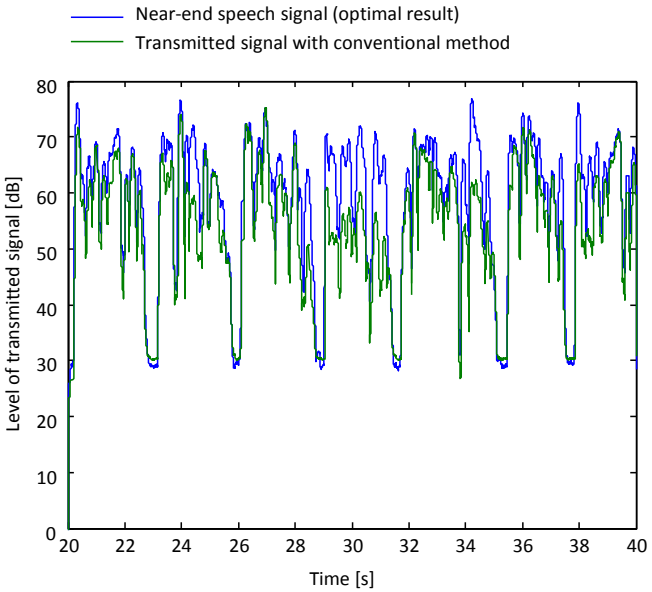

Fig. 12. Level of transmitted signal processed with conventional method during double-talk period.

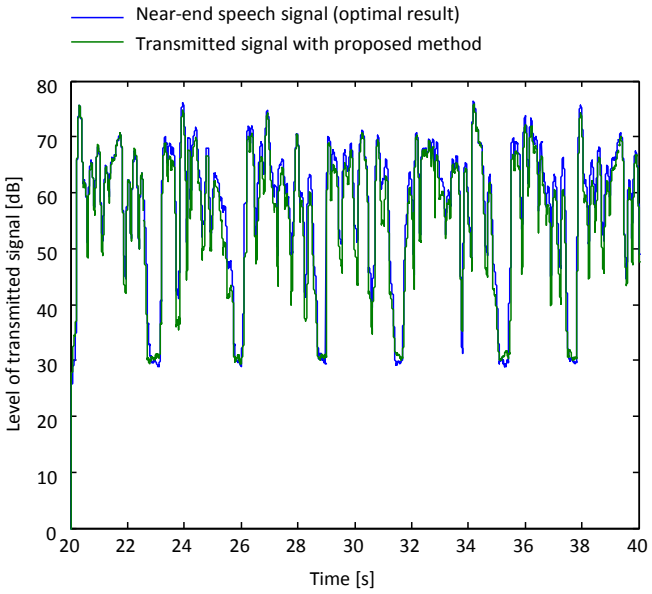

Fig. 13. Level of transmitted signal processed with proposed method during double-talk period.

near-end speech and when the echo is well suppressed. The time transitions of the LPC cepstral distance of the conventional and proposed methods are shown in Fig. 14. These results show that in the proposed method, the speech distortion was suppressed over the entire period, while the echo component was reduced sufficiently, in contrast to the conventional method.

VI. Conclusion

This paper proposed an acoustic echo cancellation method that can emphasize the target near-end speech with low degradation during double-talk periods while reducing undesired acoustic echo. The method estimates the ACL

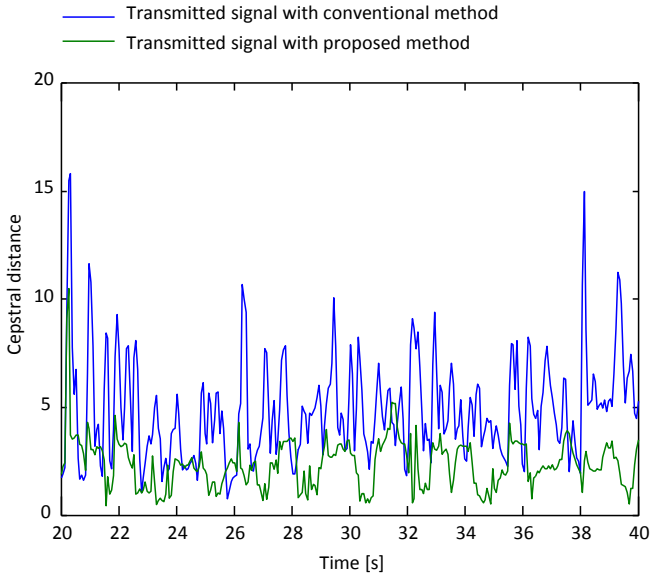

Fig. 14. Cepstral distances of transmitted signals processed with conventional and proposed methods during double-talk period.

between reference and ADF-output signals whether or not double-talk has occurred and calculates the post-filter based on the estimated ACL for the ER process. The AEC with proposed method was then modified for real-time implementation in a 20-kHz wideband videoconferencing system. To reduce the calculation complexity, the ACL estimation algorithm of the proposed method was simplified by approximately replacing all the complex number operations with real number operations. In addition, a subband approach of AEC process and a low-cost subband filtering algorithm were introduced. Experiments conducted with an AEC-unit prototype showed that the proposed method delivers natural sounding near-end speech even during double-talk periods.

References

[1] J. Casar-Corredera and J. Alcazar-Fernandez, "An acoustic echo canceller for teleconference systems," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Tokyo, Japan, vol. 11, pp. 1317-1320, Apr. 1986.

[2] C. C. Kao, "Design of echo cancellation and noise elimination for speech enhancement," *IEEE Trans. Consumer Electron.*, vol. 49, no. 4, pp. 1468-1473, Nov. 2003.

[3] M. Borhani and V. Sedghi, "An acoustic echo canceller chip," Proc. IEEE International Workshop on System-on-Chip for Real-Time Applications, Alberta, Canada, pp. 193-198, July 2005.

[4] Y. Hioka, M. Okamoto, K. Kobayashi, Y. Haneda, and A. Kataoka, "A display-mounted high-quality stereo microphone array for high-definition videophone system," *IEEE Trans. Consumer Electron.*, vol. 54, no. 2, pp. 778-786, May 2008.

[5] K. Kobayashi, Y. Haneda, K. Furuya, and A. Kataoka, "A hands-free unit with noise reduction by using adaptive beamformer," *IEEE Trans. Consumer Electron.*, vol. 54, no. 1, pp. 116-122, Feb. 2008.

[6] R. L. B. Jeannes, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: a survey," *IEEE Trans. Speech and Audio*, vol. 9, no. 8, pp. 808-820, Nov. 2001.

[7] S. Haykin, *Adaptive filter theory*, 3rd ed., Prentice-Hall, Inc.: New Jersey, USA, pp. 365-444, 1996.

[8] S. Shimauchi, Y. Haneda, and A. Kataoka, “A robust NLMS algorithm for acoustic echo cancellation,” *IEICE Trans. Fundamentals*, vol. J89-A, no. 8, pp. 926-934, Aug. 2005.
[9] C. Avendano, “Acoustic echo suppression in the STFT domain,” IEEE Workshop Sig. Proc. to Audio and Acoust., New York, USA, vol. 21, no. 24, pp. 175-178, Oct. 2001.
[10] C. Faller and J. Chen, “Suppressing acoustic echo in a spectral envelope space,” *IEEE Trans. Speech and Audio*, vol. 13, no. 5, pp. 1048-1062, Sept. 2005.
[11] S. Sakauchi, A. Nakagawa, Y. Haneda, and A. Kataoka, “Implementing and evaluating an audio teleconferencing terminal with noise and echo reduction,” Proc. International Workshop on Acoustic Echo and Noise Control, Kyoto, Japan, pp. 191-194, Sept. 2003.
[12] M. Fukui, S. Shimauchi, A. Nakagawa, Y. Haneda, and A. Kataoka, “Acoustic-coupling level estimation for performance improvement of echo reduction,” Proc. International Workshop on Acoustic Echo and Noise Control, Seattle, USA, pp. 1-4, Sept. 2008.
[13] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 27, no. 2, pp. 113-120, Apr. 1979.
[14] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proc. of the IEEE*, vol. 67, no. 12, pp. 1586-1604, Dec. 1979.
[15] ITU-T Recommendation P.34, “Transmission performance of hands-free telephones,” International Telecommunication Union, Geneva, Mar. 1993.
[16] ITU-T Recommendation G.168, “Digital network echo cancellers,” International Telecommunications Union, Geneva, Apr. 1997.
[17] A. H. Gray Jr. and J. D. Markel, “Distance measures for speech processing,” *IEEE Trans. Acoust. Speech Signal Process.* vol. 24, no. 5, pp. 380-391, Oct. 1976.

BIOGRAPHIES

Masahiro Fukui (M’09) received his B.E. degree in information science from Ritsumeikan University, Shiga, Japan, in 2002. He received his M.E. degree in information science from Nara Institute of Science and Technology, Nara, Japan, in 2004. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 2004, he has been engaged in research on acoustic echo cancellers and speech coding. He is now a research engineer at NTT Media Intelligence Laboratories. He received the best paper award of ICCE conference and the technical development award from the Acoustic Society of Japan (ASJ) in 2014. He is a member of the Institute of Electronics, Information, and Communication Engineers of Japan (IEICE), and ASJ.

Suehiro Shimauchi (M’95) received his B.E., M.E., and Ph.D. degrees from Tokyo Institute of Technology in 1991, 1993, and 2007. Since joining NTT in 1993, he has been engaged in research on acoustic signal processing for acoustic echo cancellers. He is now a senior research engineer at NTT Media Intelligence Laboratories. He is a member of IEICE and ASJ.

Yusuke Hioka (S’04–M’05–SM’12) received his B.E., M.E., and Ph.D. degrees in engineering in 2000, 2002, and 2005 from Keio University, Yokohama, Japan. From 2005 to 2012, he was with the NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories), Nippon Telegraph and Telephone Corporation (NTT). From 2010 to 2011, he was also a visiting researcher at Victoria University of Wellington, New Zealand. In 2013 he moved to New Zealand and was appointed as a Lecturer at the University of Canterbury, Christchurch. Then in 2014, he joined the Department of Mechanical Engineering, the University of Auckland, Auckland, where he is currently a Senior Lecturer. His research interests include microphone array signal processing and room acoustics. He is also a member of IEICE and ASJ.

Akira Nakagawa received his B.E. and M.E. degrees from Kyushu Institute of Technology in 1992 and 1994. Since joining NTT in 1994, he has been investigating acoustic signal processing and acoustic echo cancellers. He is now a senior research engineer at NTT Media Intelligence Laboratories. He received a paper award from the ASJ in 2001. He is a member of ASJ.

Yoichi Haneda (A’92–M’97–SM’06) received his B.S., M.S., and Ph.D. degrees from Tohoku University, Sendai, in 1987, 1989, and 1999. From 1989 to 2012, he was with the NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories), NTT. In 2012, he joined the University of Electro-Communications, where he is a professor. His research interests include modeling of acoustic transfer functions, microphone arrays, loudspeaker arrays, and acoustic echo cancellers. He received paper awards from the ASJ and from the IEICE of Japan in 2002. He is a member of ASJ.