Acoustic Echo Canceller Software for VoIP Hands-free Application on Smartphone and Tablet Devices

Masahiro Fukui, *Member*, IEEE, Suehiro Shimauchi, *Member*, IEEE, Kazunori Kobayashi, Yusuke Hioka, *Senior Member*, IEEE, Hitoshi Ohmuro, *Member*, IEEE

Abstract — *An acoustic echo cancellation (AEC) method developed for voice over IP (VoIP) hands-free applications is proposed. This method can effectively reduce undesired acoustic echo arriving at a microphone from a loudspeaker and emphasize the target talker's voice when near-end and far-end talkers speak simultaneously (i.e. double-talk), irrespective of smartphone/tablet device models. This method mainly involves cancellation of non-linear acoustic echo caused by loudspeaker distortion, residual echo reduction robust against echo-path change, and estimation of pure delay resulting from both room echo and audio input/output buffers. The experimental results show that the proposed AEC method reduced more than 40 dB of undesired echo for every smartphone or tablet used for the evaluation. This indicates that the performance of the proposed AEC method does not depend on the difference in the acoustic characteristics of individual devices*[1].

Index Terms — VoIP hands-free application, smartphone, tablet, acoustic echo canceller, delay variation, non-linear echo, microphone-sensitivity change, doubletalk.

I. Introduction

Smartphones and tablets have rapidly become popular among internet users in recent years. Along with their popularization, voice over IP (VoIP) phone applications [1]-[3] that can run on such devices are also becoming popular, whose worldwide market size has been increasing. Hands-free conversation may be a more popular style of phone-call because it allows us to, for example, talk while looking at documents displayed on the screen.

Acoustic echo cancellation (AEC) [4]-[8] is an indispensable technology for hands-free VoIP applications because it prevents detrimental acoustic echo and howling caused by the acoustic coupling between loudspeakers and microphones. Various AEC techniques have been developed [9]-[14]. One of these AEC techniques consists of an adaptive filter (ADF) [10], [11] combined with echo reduction (ER) [12]-[14], has frequently been practically applied 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 (*error signal*) still remains in its output because in practice, there are limitations on the calculation and memory capacities required to accurately estimate the echo path, which occasionally changes in practical environments. Echo reduction is therefore used, which involves a non-linear post-filter that reduces the residual echo included in the error signal. Echo reduction estimates the power spectrum of the residual echo using the squared amplitude frequency response of the acoustic coupling (*acoustic coupling level*: ACL) then calculates the post-filter that reduces the residual echoes.

The combination of the ADF and ER are known to result in reasonable performance for conventional teleconferencing systems in which the linear characteristics between the loudspeaker and microphone and slow echo-path change are guaranteed. However, three model-specific problems arise when AEC is applied to various smartphones or tablets: i) loudspeaker distortion, ii) microphone sensitivity variation, and iii) audio input/output delay variation [15]. Problem i) causes non-linear distortions in the echo, and problems ii) and iii) cause frequent and abrupt echo-path changes. As a result, AEC performance will noticeably degrade when applied to VoIP hands-free applications on smartphones or tablets.

An AEC method that automatically tailors its performance to the acoustic characteristics of individual devices is proposed. The proposed AEC method uses three new techniques: 1) ADF with nonlinear echo path modeling and its identification algorithm to cancel distorted echo, 2) ER robust against echo-path change, and 3) delay estimation (DE) to track audio input/output delay. Technique 1) can cancel out not only linear, but also nonlinear echoes that result from loudspeaker distortion. Technique 2) can instantaneously track the residual echo level, which changes when the microphone sensitivity varies. Technique 3) can sequentially calculate the pure delay resulting from both room echo and the buffer process of audio input/output.

This paper, which is based on the conference paper [16], discusses further details of these three techniques and compares the conventional and proposed methods in terms of echo-cancellation performance with six different smartphone and tablet devices.

The remainder of this paper is organized as follows. Section II presents the principle of the conventional AEC methods,

[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).

K. Kobayashi is with NTT Media Intelligence Laboratories, Tokyo 180-8585 JAPAN (e-mail: kobayashi.kazunori@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).

H. Ohmuro is with NTT Media Intelligence Laboratories, Tokyo 180-8585 JAPAN (e-mail: ohmuro.hitoshi@lab.ntt.co.jp).

and section III provides details of the proposed AEC method. The overview of a VoIP-phone prototype equipped with the proposed AEC method is introduced in section IV. Experimental results using VoIP applications are described in section V, and this paper is concluded with remarks in section VI.

II. CONVENTIONAL AEC

A block diagram of the conventional AEC is shown in Fig. 1. 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 having an impulse response modeled as $\mathbf{h}(n) = [h_1(n), \ldots, h_R(n)]^T$, where R is the effective length of the impulse response and T is the transposition.

The ADF technique is a linear-processing technique; therefore, it can cancel out only the echo signal from a microphone signal by adaptively modeling an unknown acoustic echo path. Given the reference input vector $\mathbf{x}(n) = [x(n), \ldots, x(n-R+1)]^T$ and the adaptive filter vector $\mathbf{w}(n) = [w_1(n), \ldots, w_R(n)]^T$, the output signal of the ADF $y(n)$ can be written in terms of the reference input vector $\mathbf{x}(n)^T$ which is convoluted by the impulse response between the reference and ADF output signals (*residual echo path*) $\mathbf{h}'(n) = [h'_1(n), \ldots, h'_R(n)]^T$, including the near-end speech signal $s(n)$, as

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

The ER technique is a common frequency-domain post-filter technique based on short-time spectral amplitude (STSA) estimation [17]; it estimates residual echo levels and suppresses the residual echo using multiplicative gains in the frequency domain.

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

$$Y_i(\omega) = D_i(\omega) + S_i(\omega), \quad (2)$$

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

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

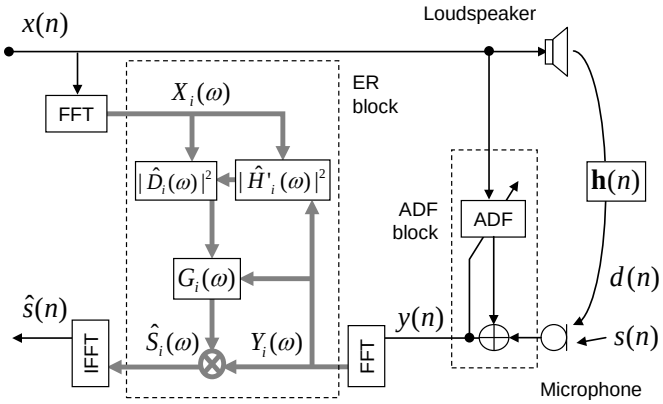

Fig. 1. Block diagram depicting conventional AEC method

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 [18] obtained by

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

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

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

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

$$|\hat{H}'_i(\omega)|^2 = \left(\frac{E[|X_i(\omega)||Y_i(\omega)|]}{E[|X_i(\omega)|^2]} \right)^2, \quad (6)$$

where $E[\cdot]$ is the ensemble average. Equation (6) is a simplified equation for the previously proposed technique [19]. In (6), all the complex number operations are approximately replaced with real number operations.

III. PROPOSED AEC METHOD

A block diagram of the proposed AEC method is illustrated in Fig. 2. Many smartphones and tablets available on the market are equipped with small inexpensive loudspeakers, built-in auto-gain control (AGC), and variable audio input/output buffers. Therefore, when a conventional AEC method is applied to a VoIP application on smartphone or

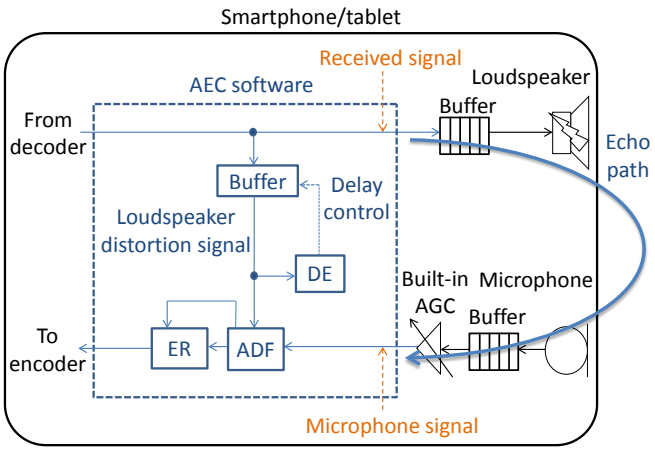

Fig. 2. Block diagram depicting proposed AEC method developed for smartphone and tablet devices.

tablet devices, the AEC performance will degrade because of loudspeaker distortion, microphone sensitivity variation, and audio input/output delay variation of the devices. The issues of distortion and variations in level and delay are adequately addressed with the ADF, ER, and DE techniques of the proposed AEC method. The ADF technique cancels out not only linear but also nonlinear echoes that result from loudspeaker distortion. The ER technique instantaneously tracks the residual echo level, which changes when the microphone sensitivity varies, and suppresses the residual echo. The DE technique sequentially calculates the pure delay resulting from both room echo and the buffer process of audio input/output. These techniques are described in the remainder of this section.

A. Nonlinear ADF

The cascade adaptive filtering scheme [20], [21] is used in the proposed AEC method, in which the adaptive hard clipping function is followed by the adaptive finite impulse response (FIR) filter. This scheme can compensate for the nonlinear echo caused by the saturation effects due to the small loudspeaker and/or poor amplifier.

The adaptive hard clipping function simulates the saturated loudspeaker output signal $u(n)$ as

$$u(n)=\begin{cases} a(n) & \left(x(n)>a(n)\right) \\ x(n) & \left(\left|x(n)\right|\le a(n)\right), \\ -a(n) & \left(x(n)<-a(n)\right) \end{cases} \tag{7}$$

where $a(n)$ denotes the nonnegative hard clipping threshold. The time domain signal $u(n)$ is transformed into the frequency domain signal $U(\omega)$ and the frequency domain estimate of the echo signal $V(\omega)$ is efficiently calculated as

$$V(\omega)=W(\omega)U(\omega), \tag{8}$$

where $W(\omega)$ is the frequency domain FIR filter coefficient of each frequency bin.

The adaptive parameters $a(n)$ and $W(\omega)$ are updated for every discrete frame i as follows:

$$a(n)\leftarrow a(n)+\mu\Delta a(n), \tag{9}$$

$$W(\omega)\leftarrow W(\omega)+\mu\frac{U(\omega)^{*}}{\left|U(\omega)\right|^{2}}\left[Y(\omega)-\Delta a(n)W(\omega)U'(\omega)\right], \tag{10}$$

where

$$\Delta a(n)=\text{real}\left[\frac{\sum_{\omega=0}^{N-1}\frac{\left(W(\omega)U'(\omega)\right)^{*}Y(\omega)}{\left|U(\omega)\right|^{2}}}{\beta+\sum_{\omega=0}^{N-1}\frac{\left|W(\omega)U'(\omega)\right|^{2}}{\left|U(\omega)\right|^{2}}}\right], \tag{11}$$

$U'(\omega)$ indicates the frequency domain signal of $\partial u(n)/\partial a(n)$, $Y(\omega)$ indicates the frequency domain output (or estimation error), N is the number of all frequency bins, μ is the adaptation step size, β is the regularization parameter, and $\text{real}[c]$ indicates the real value of c.

Unlike the original cascade scheme [20], [21], the adaptive FIR filter is implemented in the frequency domain [11] for computational efficiency. To do this, the parameter update equations shown in (9) and (10) are also differently formulated from those of the original scheme in order to jointly estimate the time domain threshold parameter $a(n)$ and the filter coefficient of each frequency bin $W(\omega)$ by commonly evaluating the frequency domain error $Y(\omega)$.

B. Instantaneous ER

The ER technique instantaneously estimates the residual echo level after separating the level and the spectral structure from the residual echo. With this technique, it is assumed that only the residual echo level is changed when the microphone sensitivity varies because the spectral structure is maintained even if the echo path is changed [22]. Under this assumption, the residual echo level can be estimated using not time but frequency spectral statistics, so the level variation can be tracked in a short observation time. The power spectrum of residual echo $|\hat{D}_i'(\omega)|^2$ can be derived by calculating the estimate of the residual echo level $\hat{g}_i$ as

$$|\hat{D}_i'(\omega)|^2=\hat{g}_i|\hat{H}_i'(\omega)|^2|X_i(\omega)|^2, \tag{12}$$

$$\hat{g}_i = \max\left[\frac{\sum_{m=0}^{Q-1}\sum_{\omega=0}^{N-1}|Y_{i-m}(\omega)|^2|\hat{H}'_{i-m}(\omega)|^2|X_{i-m}(\omega)|^2}{\sum_{m=0}^{Q-1}\sum_{\omega=0}^{N-1}\left(|\hat{H}'_{i-m}(\omega)|^2|X_{i-m}(\omega)|^2\right)^2},1\right], \quad (13)$$

where $|\hat{H}'_i(\omega)|^2$ is the estimated power frequency response of the residual echo path, Q is the number of frames, and $\max[\cdot]$ is the maximum value selection.

C. Delay Estimation

A block diagram of the DE technique is illustrated in Fig. 3. This technique sequentially calculates the pure delay resulting from both room echo and the buffer process. This technique first calculates the segment echo-path transfer functions $H''_0(\omega),\ldots,H''_{L-1}(\omega)$ by using a generalized cross correlation (GCC) method [23] consisting of a multi-delay filter (MDF) [24] as follows:

$$\begin{bmatrix} H''_0(\omega) \\ \vdots \\ H''_{L-1}(\omega) \end{bmatrix} \approx \frac{\begin{bmatrix} X_0(\omega) & 0 & 0 \\ \vdots & \ddots & 0 \\ X_{M-1}(\omega) & \ddots & X_0(\omega) \\ 0 & \ddots & \vdots \\ 0 & 0 & X_{M-1}(\omega) \end{bmatrix}^{\mathrm{H}} \begin{bmatrix} Z_0(\omega) \\ \vdots \\ Z_{M+L-2}(\omega) \end{bmatrix}}{\sum_{i=0}^{M-1} X_i^*(\omega)X_i(\omega)}, \quad (14)$$

where $Z_i(\omega)$ indicates the frequency-domain microphone signal, H is the complex conjugate transposition, and L and M are the numbers of the multi-delay filters and delay search frames, respectively. In the GCC method, the frame number i of $H''_i(\omega)$ corresponding to the initial increase in the echo-path response is considered as the pure delay. The frame number that shows the pure delay is sent to the buffers, and the received signal is adjusted for the ADF technique to maintain the delay size of the estimated echo path.

IV. Prototype Overview

Photographs of a VoIP-phone prototype implemented with the proposed AEC method are shown in Fig. 4. This prototype is a mobile VoIP softphone built using peer-to-peer (P2P) techniques and allows free VoIP calls only between prototypes. This software is implemented in the mobile operating system (OS) platform and some functions are optimized for the power-efficient central processing unit (CPU). This software can also be used with three speech/audio codecs: ITU-T Recommendations G.711 [25], G.711.1 [26], and G.711.1 Annex D [27]. The sampling frequencies of these

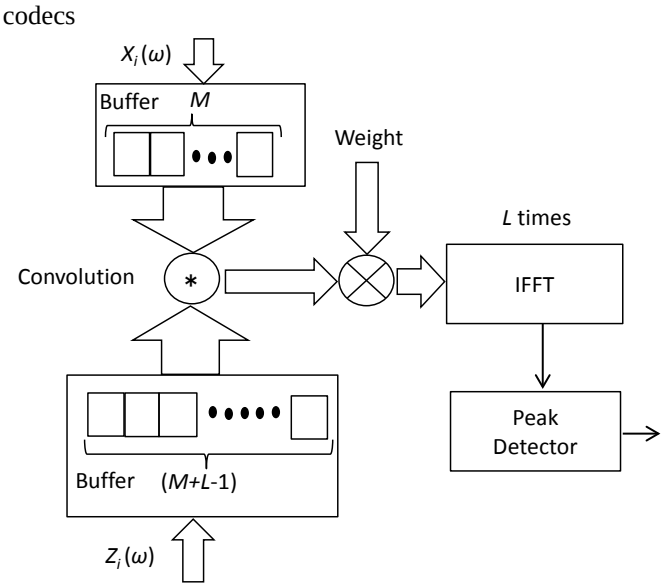

Fig. 3. Delay Estimation by MDF-GCC.

Fig. 4. Photograph of VoIP phone prototype equipped with proposed AEC method.

are 8, 16, and 32 kHz, respectively. The A/D and D/A converters are compatible with 8/16/32/44.1/48 kHz sampling. The frame-shift size is 20 ms, the frame size for a fast Fourier transform (FFT) is 40 ms, and the signal delay of the software is 30 ms. The maximal filter tap length in the echo path modeling is 200 ms. This software keeps memory consumption below 10 Mbytes, and the percentage of CPU usage is 10% or less.

A block diagram of the proposed AEC method is shown in Fig. 5. It consists of blocks for the following components: sampling frequency switch (SFS), analysis filter (AF), loss control (LC), synthesis filter (SF), sampling frequency converter (SFC), sound device control, delay estimator, buffer, and acoustic echo controller.

The received speech signal from the decoder enters the AEC software, and its sampling frequency is selected by the SFS according to the used codec. The signal after the SFS is split into two or three sub-band signals by the AFs if the sampling frequency is more than 16 kHz. The frequency ranges of the sub-band signals are 0–4, 4–8, and 8–16 kHz. These signals undergo gain controls by the LC and are re-synthesized by the SFs. The sampling frequency of the loudspeaker output is switched by the SFS. If the sampling

frequency is set to 44.1 or 48 kHz, the 32-kHz sampling is converted into 44.1 or 48 kHz by the SFC. The sound device control controls the sound buffers in the mobile device in order to play the far-end talker's voice through the loudspeaker and pick up the near-end talker's voice with the microphone.

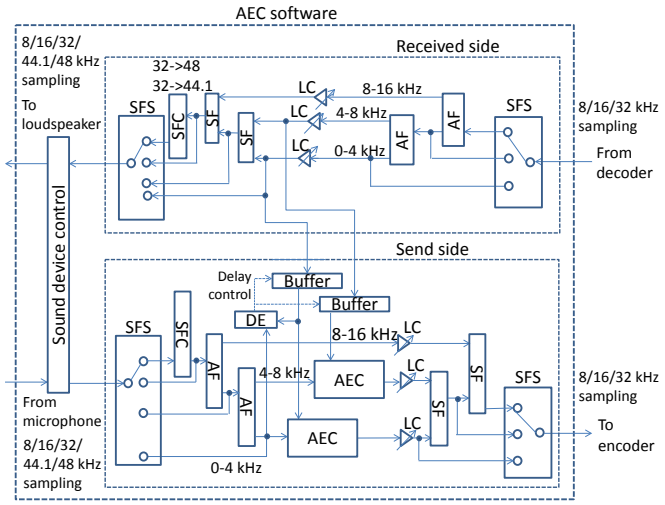

Fig. 5. Block diagram of AEC software implemented in VoIP application.

The sampling frequency of the microphone input signal is switched by the SFS, and the signal is split into sub-band signals by the AFs. The sampling frequency is converted into 32 kHz if the sampling frequency of the microphone signal is 44.1 or 48 kHz.

The delay estimator estimates the delays of both acoustic echo and sound device control to allow the acoustic echo canceller to work correctly. This controller is composed of an ADF and ER, and cancels out the undesired echo. The signal after AEC undergoes the gain control by the LC. The sub-band signals are re-synthesized by the SFs, the sampling frequency of synthesis signal is selected by the SFS, and the output signal is sent to the encoder.

V. Performance Evaluation

The echo cancellation performances of the proposed and conventional AEC methods were compared in a practical environment using smartphone and tablet devices. The conventional method was that in which the conventional ADF and ER techniques were used, as described in Section II, without a DE technique.

A. Test Conditions

The arrangement of the smartphone/tablet device and sound source is shown in Fig. 6. The loudspeaker shown in this figure simulated the near-end talker and background noise. The loudspeaker and microphone levels were as prescribed by ITU-T Recommendation P.340 [28]. The tests were based on specific test signals as prescribed in ITU-T Recommendation P.501 [29]. The following background noise types were used in the tests: pink noise at a 20-dB signal-to-noise ratio (SNR) and office noise at a 15-dB SNR. The reverberation time was set to 300 ms. Four smartphones (S-A, S-B, S-C, and S-D) and two tablets (T-A and T-B) were used in the tests.

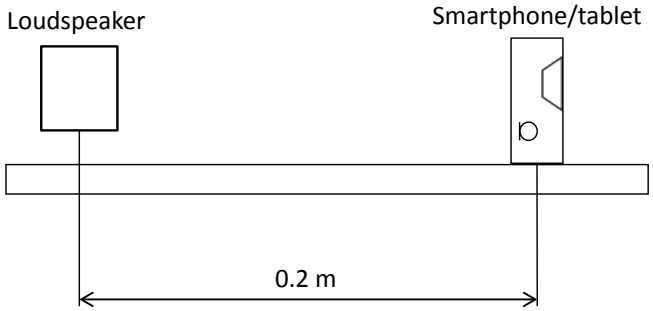

Fig. 6. Test arrangements for objective measurements.

B. Experimental Results

The signal-to-echo ratio (SER) was used as an evaluation metric of AEC performance. In the tests, the SERs in single-talk and double-talk periods were evaluated. Single talk is a situation where only the far-end speaker is talking. Double talk is a situation where both the near-end and far-end speakers are talking concurrently. The SERs in the single-talk and double-talk cases were computed using the following equations, respectively:

$$\mathrm{SER}_{\mathrm{ST}} = 10\log_{10}\frac{\sum_{n=0}^{K}|\hat{s}_S(n)|^2}{\sum_{n=0}^{K}|\hat{s}(n)|^2}, \tag{15}$$

$$\mathrm{SER}_{\mathrm{DT}} = 10\log_{10}\frac{\sum_{n=0}^{K}|\hat{s}_S(n)|^2}{\sum_{n=0}^{K}|\hat{s}(n)-s_S(n)|^2}, \tag{16}$$

K is the signal length and $\hat{s}_S(n)$ is the transmitted signal observed when only the near-end speaker is talking.

These results are shown in Figs. 7 to 12. Figs. 7 to 9 are the single-talk cases with and without background noise, and Figs. 10 to 12 are double-talk cases with and without background noise, respectively. The SNRs are 20 and 15 dB in the cases of the pink and office noises, respectively. As these results indicate, the proposed AEC method sufficiently suppressed the undesired acoustic echo compared with the conventional AEC method in both the single-talk and double-talk periods irrespective of the types of devices and background noise. Regarding the proposed AEC method, SERs of more than 40 and 20 dB were achieved in the single- talk and double-talk periods, respectively.

VI. Conclusion

An AEC method for VoIP hands-free application on

smartphones and tablets was proposed. The proposed method

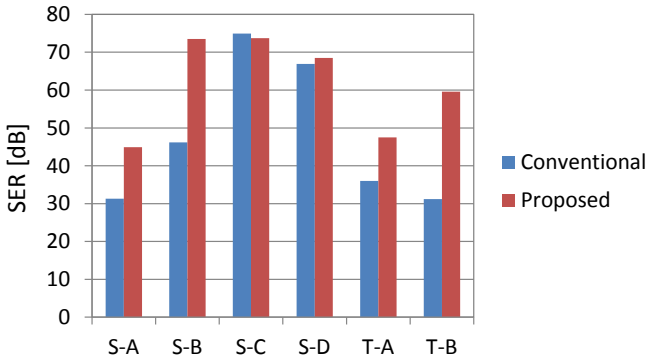

Fig. 7. Comparison of echo reduction performance by SER (single talk without background noise).

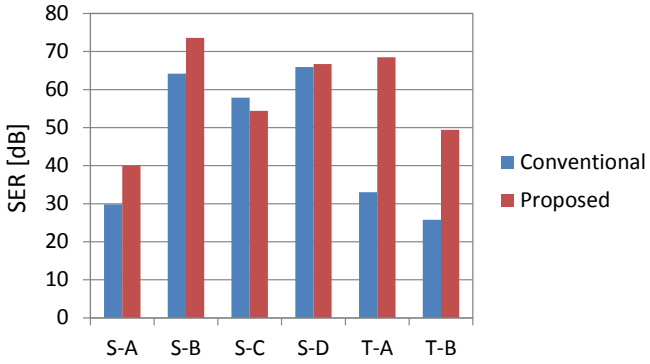

Fig. 8. Comparison of echo reduction performance by SER (single talk with pink noise at 20 dB SNR).

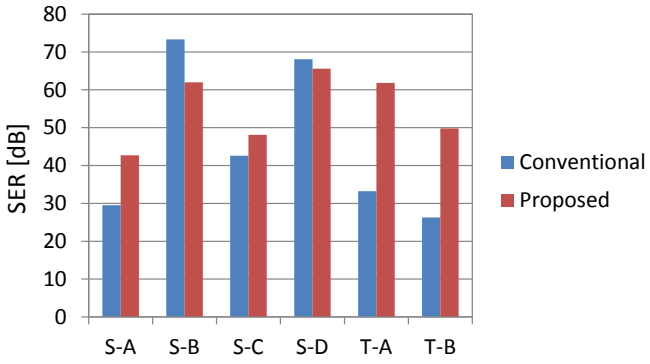

Fig. 9. Comparison of echo reduction performance by SER (single talk with office noise at 15 dB SNR).

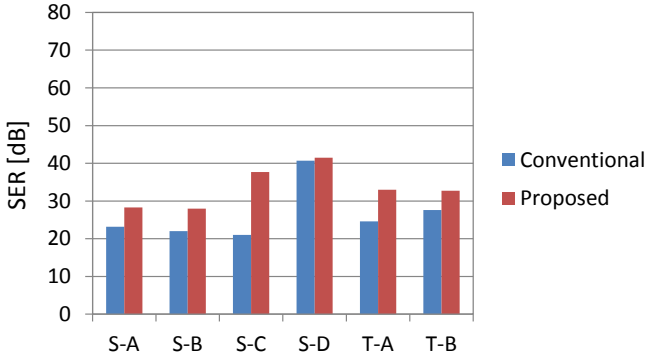

Fig. 10. Comparison of echo reduction performance by SER (double talk without background noise).

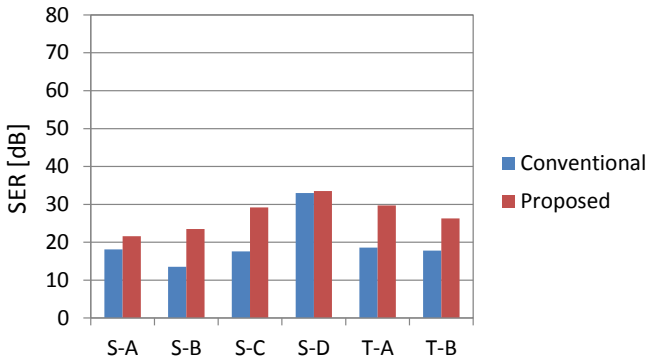

Fig. 11. Comparison of echo reduction performance by SER (double talk with pink noise at 20 dB SNR).

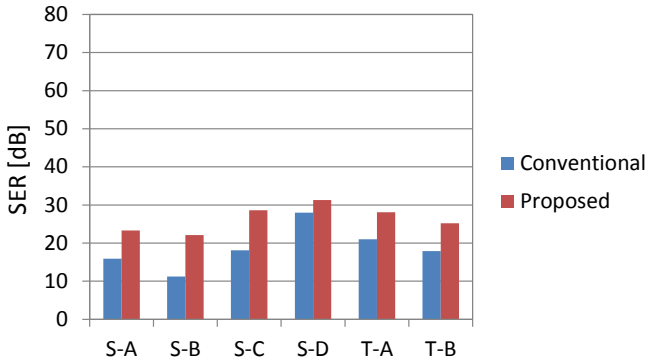

Fig. 12. Comparison of echo reduction performance by SER (double talk with office noise at 15 dB SNR).

can reduce the undesired acoustic echo and emphasize the target near-end speech during double-talk periods, irrespective of the smartphone and tablet models. This method can estimate the non-linear acoustic echo caused by loudspeaker distortion, instantaneous residual echo variation caused by echo-path change, and pure delay resulting from both room echo and audio input/output buffers. This method was implemented in a VoIP hands-free phone application used on smartphones and tablets. The experimental results demonstrated that the proposed method effectively reduces the undesired echo on various smartphone/tablet models and performed better than the compared conventional AEC method.

REFERENCES

[1] L. Caviglione, "A simple neural framework for bandwidth reservation of VoIP communications in cost-effective devices," *IEEE Trans. Consumer Electron.*, vol. 56, no. 3, pp.1252–1257, Aug. 2010.

[2] H.-H. Choi, J.-R. Lee and D.-H. Cho, "On the use of a power-saving mode for mobile VoIP devices and its performance evaluation," *IEEE Trans. Consumer Electron.*, vol. 55, no. 3, pp.1537–1545, Aug. 2009.

[3] H.-G. Kim and J.-H. Lee, "Enhancing VoIP speech quality using combined playout control and signal reconstruction," *IEEE Trans. Consumer Electron.*, vol. 58, no. 2, pp.562–569, May 2012.

[4] 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.

[5] 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.
[6] M. Borhani, 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.
[7] 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.
[8] 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. 20*08.
[9] 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.
[10] S. Haykin, *Adaptive filter theory*, 3rd ed., Prentice-Hall, Inc.: New Jersey, USA, pp. 365–444, 1996.
[11] J. J. Shynk, “Frequency-domain and multirate adaptive filtering,” IEEE Signal Processing Magazine, vol. 9, no. 1, pp. 14-37, 1992.
[12] 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.
[13] 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, Sep. 2005.
[14] 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.
[15] I. Papp, Z. Saric, S. Pal, and I. Velikic, “Hands-free VoIP solution for embedded platforms in consumer electronics,” IEEE International Conference on Consumer Electronics – Berlin 2011, Berlin, Germany, pp. 22–25, Sept. 2011.
[16] M. Fukui, S. Shimauchi, K. Kobayashi, Y. Hioka, and H. Ohmuro, “Acoustic echo canceller software for VoIP hands-free application on smartphone and tablet devices,” Proc. IEEE International Conference on Consumer Electronics, Las Vegas, USA, pp.10–13, Jan. 2014.
[17] 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.
[18] 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.
[19] 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.
[20] A. Stenger and W. Kellermann, “Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling,” *Signal Processing*, vol. 80, no.9, pp.1747-1760, 2000.
[21] S. Shimauchi and Y. Haneda, “Nonlinear acoustic echo cancellation based on piecewise linear approximation with amplitude threshold decomposition,” Proc. International Workshop on Acoustic Signal Enhancement, Aachen, German, pp. 1–4, Sept. 2012
[22] M. Fukui, S. Shimauchi, Y. Hioka, H. Ohmuro, and Y. Haneda, “Acoustic echo reduction robust against echo-path change with instant echo-power-level adjustment,” Proc. European Signal Processing Conference, Marrakech, Morocco, pp. 1–5, Sept, 2013.
[23] C. H. Knapp and C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 24, no. 4, pp. 320–327, 1976.
[24] J.-S. Soo and K. K. Pang, “Multidelay block frequency domain adaptive filter,” *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 38, no. 2, pp. 373–376,1990.
[25] ITU-T Recommendation G.711, “Pulse code modulation (PCM) of voice frequencies,” International Telecommunications Union, Geneva, Nov. 1988.
[26] ITU-T Recommendation G.711.1, “Wideband embedded extension for G.711 pulse code modulation,” International Telecommunications Union, Geneva, Mar. 2008.
[27] ITU-T Recommendation G.711.1 Annex D, “New Annex D with superwideband extension,” International Telecommunications Union, Geneva, Nov. 2010.
[28] ITU-T Recommendation P.340, “Transmission characteristics and speech quality parameters of hands-free terminals,” International Telecommunications Union, Geneva, May 2000.
[29] ITU-T Recommendation P.501, “Test signals for use in telephonometry,” International Telecommunications Union, Geneva, Aug. 1996.

BIOGRAPHIES

Masahiro Fukui (M’09) received his B.E. degrees 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 the 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 Nippon Telegraph and Telephone Corporation (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.

Kazunori Kobayashi received his B.E., M.E., and Ph.D. degrees in Electrical and Electronic System Engineering from Nagaoka University of Technology in 1997, 1999, and 2003. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 1999, he has been engaged in research on microphone arrays and hands-free systems. He is now affiliated with NTT Cyber Space 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.

Ohmuro Hitoshi (M’93) received his B.E. and M.E. degrees in electrical engineering from Nagoya University, Aichi, in 1988 and 1990. He joined NTT in 1990. He has been engaged in research on highly efficient speech coding and the development of VoIP applications. He is currently the manager of the Speech, Acoustics and Language Laboratory at NTT Media Intelligence Laboratories. He is a member of IEICE and ASJ.