Acoustic echo and noise canceller for personal hands-free video IP phone

doi: 10.1109/TCE.2016.7838099

Acoustic Echo and Noise Canceller for Personal Hands-Free Video IP Phone

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

Abstract — *This paper presents implementation and evaluation of a proposed acoustic echo and noise canceller (AENC) for videotelephony-enabled personal hands-free internet protocol (IP) phones. This canceller has the following features: noise-robust performance, low processing delay, and low computational complexity. The AENC employs an adaptive digital filter (ADF) and noise reduction (NR) methods that can effectively eliminate undesired acoustic echo and background noise included in a microphone signal even in a noisy environment. The ADF method uses the step-size control approach according to the level of disturbance such as background noise; it can minimize the effect of disturbance in a noisy environment. The NR method estimates the noise level under an assumption that the noise amplitude spectrum is constant in a short period, which cannot be applied to the amplitude spectrum of speech. In addition, this paper presents the method for decreasing the computational complexity of the ADF process without increasing the processing delay to make the processing suitable for real-time implementation. The experimental results demonstrate that the proposed AENC suppresses echo and noise sufficiently in a noisy environment; thus, resulting in natural-sounding speech[1].*

Index Terms — Videophone, hands-free telecommunication, acoustic echo and noise canceller, adaptive digital filter, noise reduction.

I. INTRODUCTION

The popularity of videotelephony-enabled personal internet protocol (IP) phones has been growing in recent years, and the hands-free communication function implemented in these videophones is often used because of its convenience. Such terminals therefore require acoustic echo cancellers because acoustic echoes, which result from acoustic coupling between loudspeakers and microphones, must be removed for better speech quality and for avoidance of acoustic howling in hands-free telecommunication. Noise cancellers are also required to eliminate undesired background noise included in microphone signals because personal videophones are usually used in noisy environments such as offices.

[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 1142 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).

A previous study [1] discussed an implementation of an acoustic echo and noise canceller (AENC) [2] for voice over IP phones on smartphone and tablet devices. On the other hand, this paper addresses the AENC for personal hands-free videophone; it has to be integrated into a single fixed-point digital signal processor (DSP). In general the DSP for the videophone has a constraint on a computational complexity; its processing performance is lower than that of the smartphone and tablet devices. Therefore, it is important to satisfy required specifications of both performance and computational complexity of the AENC in order to make real-time implementation possible when using a low-performance DSP chip.

The AENC is mainly composed of an adaptive digital filter (ADF) [3]-[7] process, an echo reduction (ER) [8]-[11] process, and a noise-reduction (NR) [12]-[15] process. The ADF process identifies an unknown acoustic echo path, produces the echo replica, and cancels out the acoustic echo. However, in general the ADF accumulates prediction errors of the acoustic echo path when it is computed in noisy environments, for example, in an open-plan office. In addition, its computational complexity may be particularly high because the ADF of the AENC has to handle a digital filter with large number of taps, which is required to model the acoustic echo path accurately. Hence it is important to decrease not only the prediction error but also the computational complexity when the target chip performance is limited. The ER process is used to reduce the residual echo, which remains in the output of the ADF (often called *error signal*), by applying a non-linear filter to the error signal. The NR process lowers the level of background noise by multiplicative gains in the frequency domain. This process is based on a short-time spectral amplitude (STSA) estimation [16]. In recent years, high-performance NRs with a microphone array that consists of multiple microphones placed at different spatial locations have been proposed [14], [15]. However, it is often difficult to adopt the high-performance NRs into personal videophones because the videophones have constraints to its casing size and DSP-chip performance. Due to budget constraint, monaural NRs [12], [13], the number of components of which is small and the computational complexity is relatively low, are still widely used in the videophones.

This paper addresses two issues in using the AENC for personal videophones; i) to maintain the AENC performance in a noisy environment such as the open-plan office, and ii) to attain the performance while reducing the computational complexity to make real-time implementation possible. For the issue i), this paper introduces two methods. One of them is

controlling the step size of the ADF [5]-[7]. With this method, the step size is adaptively adjusted according to the level of disturbance such as background noise. It can minimize the effect of the disturbance in a noisy environment. Another method is estimating the noise level under the assumption that the noise amplitude spectrum is constant in a short period [17], which cannot be applied to the amplitude spectrum of speech. This method estimates the noise level even during speech periods without suspending the calculation. However, its estimation accuracy might decrease when the disturbance such as the residual echo exists. In the proposed system, the combination of the proposed ADF and NR solves the issue; the disturbance can be removed before the noise level is estimated because the noise robust ADF with the step-size control sufficiently eliminates the acoustic echo as the previous processing of the NR. This combination results in natural near-end speech even when the near-end and the far-end talkers speak simultaneously (i.e. *double-talk*).

For the issue ii), this study attempts to reduce the computational complexity of the AENC largely not by optimizing the code for the DSP platform but by distributing the operations of the ADF. The ADF employed in this study distributes the filter update and convolution processes across different frame times, whereas the ADF used in the previous study [1] updates and convolves the filter coefficients within each single frame. As a result of such modification, the computational complexity is drastically decreased, more details of which will be discussed in section IV.A.

To validate the performance of the proposed algorithm, the speech quality of the AENC was evaluated by implementing the algorithm to a personal hands-free videophone prototype that has a DSP. Experimental results using the videophone prototype showed that the method could deliver natural sounding speech while sufficiently reducing the acoustic echo and background noise.

The rest of this paper is organized as follows. Section II presents the specifications of the developed hands-free videophone prototype. Section III describes the proposed AENC that employs the ADF and NR methods robust against the noisy environment in detail. The experimental results are explained using the AENC implemented in the videophone prototype in section IV, and this paper is concluded with remarks in section V.

II. Specifications

An external view of a personal hands-free videophone prototype equipped with the AENC is shown in Fig. 1. This prototype is a wideband IP phone and can be used as a hands-free phone. A block diagram of the audio processing part of the prototype is shown in Fig. 2. The circuit uses a fixed-point DSP. Specifications of the prototype are listed in Table I. The prototype includes an omni-directional microphone, a loudspeaker, and a DSP board. All processing for the videophone are integrated into the single DSP; it exhibits a maximum speed of 600 MHz and 148 kbytes of on-chip random access memory (RAM); Off-chip memory has 128 Mbyte of synchronous dynamic RAM. These regions allocated to the AENC are less than 30%.

The implemented AENC is a fixed-point DSP software; a part of the software uses optimized assembly codes and dual multiple access channel (MAC) instructions. The dual MAC may be recognized as the duplication of two single MACs; it reduces the computational complexity of operation instructions such as the convolution and the complex arithmetic, and efficiently uses load and store instructions. In addition, processor-integrated FFT and division accelerators are also used. Most of these codes have been written from scratch when replacing the floating-point operations with the fixed-point operations. The required specifications of the AENC are as follows. The sampling frequency is 16 kHz and the frequency range is 100-7000 Hz, which realize superior speech quality and voice naturalness compared to narrowband speech of 300-34000 Hz used in conventional telephones. The processing frame size is 10 ms and the processing delay is 40 ms. This low latency guarantees that the delay will not cause degradation in an interactive conversation. The filter length of the ADF is 90 ms, the echo-reduction level is more than 35 dB, and the noise-reduction level is about 20 dB.

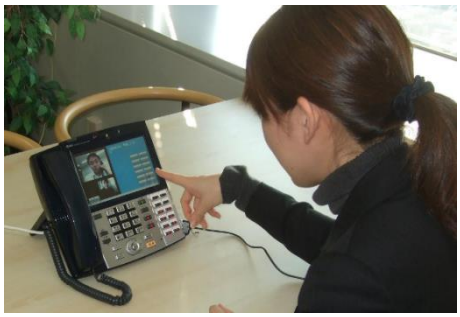


Fig. 1. External view of videophone prototype.

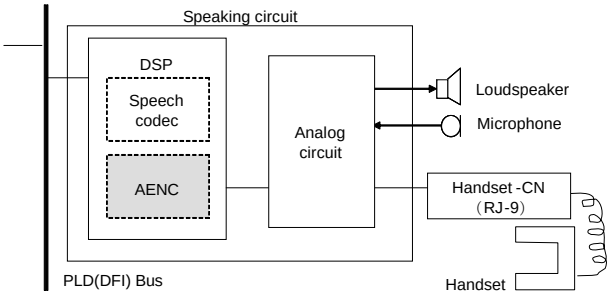


Fig. 2. Block diagram of speaking circuit.

TABLE I
Specifications of Videophone Prototype

Item	Description
Dimensions	242 mm (W) × 239 mm (D) × 102.4 mm (H)
Weight	1360 g
Display	7.0" TFT screen
Microphone	Omni-directional condenser microphone
Loudspeaker	Single-cone electrodynamic loudspeaker
Connectors	- Two USB interfaces (Type A) - Two Ethernet ports (RJ-45)

III. System Description of AENC

A block diagram of the proposed AENC that employs the

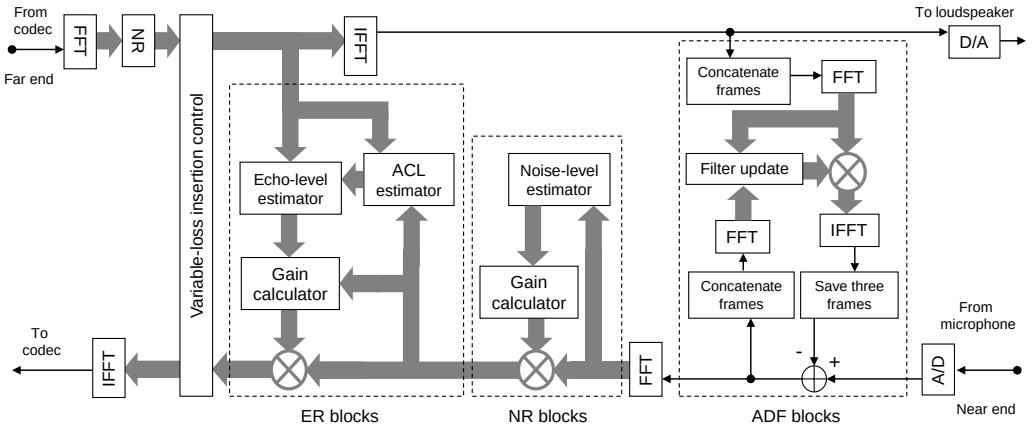


Fig. 3. Block diagram of new AENC. FFT: fast Fourier transform, IFFT: inverse fast Fourier transform.

ADF and NR methods robust against the noisy environment is shown in Fig. 3. It consists of four blocks with the following functions: ADF process, NR process, ER process, and variable-loss insertion control (VLIC) [4] process. These processes are carried out in the frequency domain. In this section, the detailed features of the ADF and NR are described first then the outlines of the ER and VLIC processes are summarized.

A. ADF process

A block diagram of the proposed ADF is shown in Fig. 4. The received speech signal $x(k)$ from the far-end at a discrete time index *k* is picked up as an echo signal $d(k)$ by the microphone after passing through the room echo path, which has an impulse response denoted as $\mathbf{h}_L = [h_1, \ldots, h_L]^T$, where L is the filter length and T is the transposition. The microphone input signal $y(k)$ is expressed as

$$y(k) = d(k) + \upsilon(k), \tag{1}$$

where $\upsilon(k)$ is the signal, called the outlier, which includes the near-end speech, background noise, and so on. In the digital-to-analogue (D/A) converter, the received speech signal $x(k)$ is stored as the received speech vector $\mathbf{x}_M(i) = [x(iM - M + 1), \ldots, x(iM)]^T$, where i is the frame index and M denotes the buffer size of the signal. The microphone input signal is also stored as vector $\mathbf{y}_M(i)$ with length M. The ADF calculates the echo-replica vector $\hat{\mathbf{d}}_M(i)$ corresponding to vector $\mathbf{y}_M(i)$ and achieves block echo cancellation as

$$\mathbf{e}_M(i) = \mathbf{y}_M(i) - \hat{\mathbf{d}}_M(i), \tag{2}$$

where

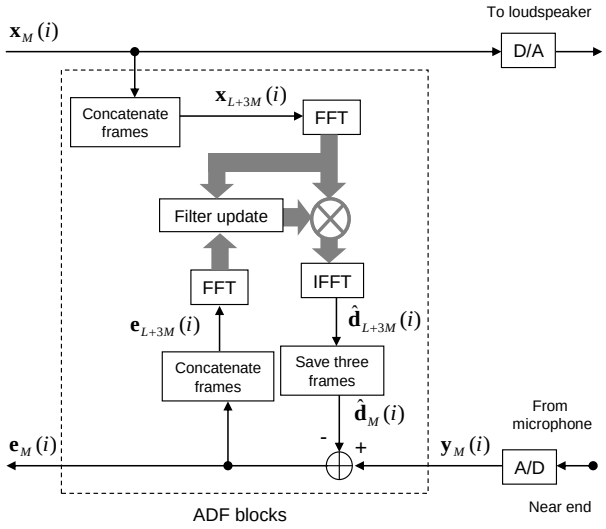


Fig. 4. Block diagram of ADF.

$$\hat{\mathbf{d}}_M(i) = \begin{bmatrix} x(iM - M + 1) & \ldots & x(iM - M - L + 2) \\ \vdots & \ddots & \vdots \\ x(iM) & \ldots & x(iM - L + 1) \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_L \end{bmatrix} \tag{3}$$

and $\mathbf{e}_M(i)$ is a residual signal vector and is fed back to the ADF to update the filter coefficients $w_1, \ldots, w_l, \ldots, w_L$. The filter-coefficient update is based on a frequency-domain adaptive filter algorithm [3] given by

$$W_{i+1}(\omega) = W_i(\omega) + \mu \frac{\varepsilon_i(\omega) X_i^*(\omega)}{P\left[\left|X_i(\omega)\right|^2\right]}, \tag{4}$$

where $W_i(\omega)$, $\varepsilon_i(\omega)$ and $X_i(\omega)$ are the short-time Fourier transforms of the w_l, $e(k)$ and $x(k)$, respectively.

Note that ω is the discrete frequency bin, μ is the step size, * denotes the conjugate and $P[\cdot]$ is the smoothing function [3].

The filter-coefficient accuracy becomes important when using the ADF. Therefore, an optimal value for the step size must be specified; its value affects the convergence speed, steady state error, and stability. However, the optimal step size is time variant according to outliers such as background noise. The filtering scheme of the robust ADF adaptively calculates the step size based on the Gaussian-Laplacian mixture assumption for the signals normalized by the reference input signal amplitude [5] because the speech spectra have Laplacian distributions but the normalized echo spectra tend to become Gaussian distributions; the other hand the normalized unwanted outlier spectra have Laplacian distributions regardless of types of outliers. Therefore the Gaussian-Laplacian mixture assumption is suitable for modeling many actual situations in the telecommunications. The step size is calculated as follows:

$$\mu = \alpha \cdot \Psi\left(\frac{\left|\varepsilon_i(\omega)\right|}{\left|X_i(\omega)\right|}, \frac{\sigma_\omega^2}{\lambda_\omega}\right)\frac{\left|X_i(\omega)\right|}{\left|\varepsilon_i(\omega)\right|+\delta}, \tag{5}$$

where

$$\Psi(b, b_{\text{th}}) = \begin{cases} |b| & \text{if } |b| \le b_{\text{th}} \\ b_{\text{th}} & \text{otherwise} \end{cases}, \tag{6}$$

α is a fixed step size, and δ is a stability parameter. σ_ω denotes a standard deviation of the following Gaussian distribution model,

$$\text{Gaussian}\left(R_i(\omega)\right) = \frac{1}{\sqrt{2\pi\sigma_\omega^2}} e^{-\frac{|R_i(\omega)|^2}{\sigma_\omega^2}}, \tag{7}$$

where $R_i(\omega)$ is the residual echo spectra normalized by $\left|X_i(\omega)\right|$. λ_ω denotes a hyperparameter in the following Laplacian distribution model,

$$\text{Laplacian}\left(V_i(\omega)\right) = \frac{1}{2\lambda_\omega} e^{-\frac{|V_i(\omega)|}{\lambda_\omega}}, \tag{8}$$

where $V_i(\omega)$ is an unwanted outlier spectra normalized by $\left|X_i(\omega)\right|$. Note that $R_i(\omega)$ and $V_i(\omega)$ have the following relationship,

$$\frac{\varepsilon_i(\omega)}{\left|X_i(\omega)\right|} = R_i(\omega) + V_i(\omega). \tag{9}$$

The normalized residual echo spectra are estimated by the maximum a posteriori (MAP) estimate maximizing the posteriori probability $p\left(R_i(\omega)\middle|\varepsilon_i(\omega)\middle|X_i(\omega)\right|^{-1}\right)$, and used to derive the optimal step size in Eq. (5). The optimal step size is regarded as optimal on the assumption that the normalized residual echo and outlier components in the error signal can be represented by Gaussian and Laplacian distributions, respectively. This approach steadily decreases the prediction error of the filter coefficients and improves adaption in non-stationary and noisy environments.

The ADF normally requires the calculation of filter coefficients in a long vector length, $L+M$ samples per frame, to calculate echo-replica vectors. On the other hand, to achieve cost-effective processing, the proposed ADF makes effective use of two buffer delays caused by A/D and D/A converters and reduces the vector length required to calculate the filter. This method focuses on these two buffers $\mathbf{x}_M(i-2)$, $\mathbf{x}_M(i-1)$ from the far-end receive side arriving late at the near-end send side; thereby, the two-frame-future echo can be reasonably foreseeable using these two buffer delays. The proposed ADF calculates not only the present echo replica vector $\hat{\mathbf{d}}_M(i)$ but also future echo-replica vectors $\hat{\mathbf{d}}_M(i+1)$ and $\hat{\mathbf{d}}_M(i+2)$ simultaneously.

A concatenated received speech vector $\mathbf{x}_{L+3M}(i)$ is obtained by connecting it with the past frames as follows: $\mathbf{x}_{L+3M}(i) = [x(iM-L-3M+1),\ldots,x(iM)]^T$, where the length of $\mathbf{x}_{L+3M}(i)$ is $L+3M$. The corresponding echo-replica vector $\hat{\mathbf{d}}_{L+3M}(i)$, including the last $3M$ elements, is composed of three echo-replica vectors as follows:

$$\hat{\mathbf{d}}_{L+3M}(i) = \begin{bmatrix} \mathbf{c}_L(i) \\ \hat{\mathbf{d}}_M(i) \\ \hat{\mathbf{d}}_M(i+1) \\ \hat{\mathbf{d}}_M(i+2) \end{bmatrix}, \tag{10}$$

where $\hat{\mathbf{d}}_M(i)$, $\hat{\mathbf{d}}_M(i+1)$, and $\hat{\mathbf{d}}_M(i+2)$ correspond to the microphone input signal frames $\mathbf{y}_M(i)$, $\mathbf{y}_M(i+1)$, and $\mathbf{y}_M(i+2)$, respectively, and $L\times 1$ vector $\mathbf{c}_L(i)$ is unimportant because it is not used here. Thus, by waiting for $\mathbf{y}_M(i+2)$ until the $(i+2)$-th frame, the concatenated error vector

$$\mathbf{e}_{L+3M}(i)=\begin{bmatrix}\mathbf{0}_L\\ \mathbf{e}_M(i)\\ \mathbf{e}_M(i+1)\\ \mathbf{e}_M(i+2)\end{bmatrix} \tag{11}$$

is obtained.

The conventional ADF updates and convolves the filter coefficients for each one frame time for obtaining the echo-replica vector. On the other hand, the proposed ADF can decentralize the filter update and convolution processes at different frame times because the echo-replica vectors are calculated only once every three frames. As a result, the calculation complexity of the ADF process was reduced without increasing the processing delay.

B. NR process

A block diagram of the proposed NR process is shown in Fig. 5. This method assumes that noisy speech $z(k)$ consists of clean speech $s(k)$ and background noise $n(k)$. Let $Z_i(\omega)$, $S_i(\omega)$, and $N_i(\omega)$ denote the short-time Fourier transforms of the $z(k)$, $s(k)$, and $n(k)$, respectively. In achieving NR based on STSA estimation, the short-time Fourier transform of the clean speech is estimated by $Z_i(\omega)$ multiplied by a gain $G_i(\omega)$ in the frequency domain as follows:

$$\hat{S}_i(\omega)=G_i(\omega)Z_i(\omega), \tag{12}$$

where $\hat{S}_i(\omega)$ is the estimate of $S_i(\omega)$. The Wiener-filtering-based [18] gain is calculated as

$$G_i(\omega)=\frac{|Z_i(\omega)|^2-E[|\hat{N}_i(\omega)|^2]}{|Z_i(\omega)|^2}, \tag{13}$$

where $E[\cdot]$ denotes the ensemble average, $|\cdot|^2$ denotes the absolute square, and $E[|\hat{N}_i(\omega)|^2]$ is the estimate of the noise level $E[|N_i(\omega)|^2]$. The noise level can be defined as follows:

$$E[|N_i(\omega)|^2]=\gamma_i(\omega)\cdot E[|Z_i(\omega)|^2], \tag{14}$$

where $\gamma_i(\omega)$ is a noise ratio defined by

$$\gamma_i(\omega)=\frac{E[|N_i(\omega)|^2]}{E[|Z_i(\omega)|^2]}. \tag{15}$$

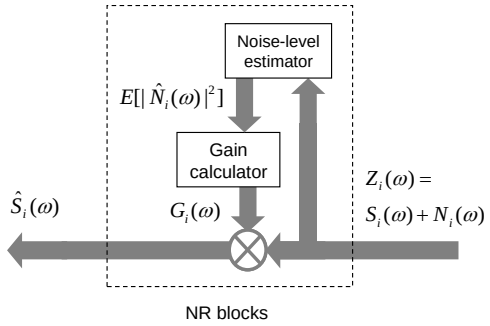


Fig. 5. Block diagram of NR.

This paper introduces a method for directly estimating the noise level from microphone input level $E[|Z_i(\omega)|^2]$, even during speech periods, by estimating the noise ratio [17]. However, the noise ratio cannot be estimated directly. Therefore, this method uses different variances of speech and noise signals. If the noise power $|N_i(\omega)|^2$ stays stationary whereas the speech power $|S_i(\omega)|^2$ is non-stationary for a finite period, the following assumptions hold:

$$(E[|N_i(\omega)|])^2 \approx E[|N_i(\omega)|^2], \tag{16}$$

$$(E[|S_i(\omega)|])^2 << E[|S_i(\omega)|^2], \tag{17}$$

i.e., the noise level $E[|N_i(\omega)|^2]$ can be approximated by

$$(E[|Z_i(\omega)|])^2 \approx E[|N_i(\omega)|^2]. \tag{18}$$

This method estimates the noise ratio using these assumptions as

$$\hat{\gamma}_i(\omega)=\Phi\left[\frac{(E[|Z_i(\omega)|])^2}{E[|Z_i(\omega)|^2]}\right], \tag{19}$$

where $\Phi[\cdot]$ denotes the noise-ratio-emphasis function designed to bring the noise ratio closer to zero or one defined by:

$$\Phi[A]=\begin{cases}1, & \text{if } A\le 0.8\\ 1.6667A-0.33334, & \text{if } 0.2<A<0.8.\\ 0, & \text{otherwise}\end{cases} \tag{20}$$

This emphasis reduces the estimation error of the noise ratio caused by a small margin of errors in the assumptions of Eqs. (16) and (17).

The new NR process can estimate noise level accurately, even during speech periods, and maintains natural near-end speech.

C. ER and VLIC processes

The ER process is based on the STSA estimation and suppresses the residual echo of the ADF by multiplying echo-reduction gains in the frequency domain. The ER process consists of an acoustic-coupling level (ACL) estimator, an echo-level estimator, and a gain calculator. The ACL estimator is based on a cross-spectrum method [8] and expanded to time and frequency spectral domains [19] to increase the tracking speed as follows:

$$|\hat{H}_i(\omega)|^2 = \left(\frac{\sum_r E[|X_i(\omega+r)||\hat{S}_i(\omega+r)|]}{\sum_r E[|X_i(\omega+r)|^2]} \right)^2, \quad (21)$$

where $|\hat{H}_i(\omega)|^2$ denotes the estimated ACL. The echo level estimator calculates the amount of residual echo by multiplying the reference signal by the estimated ACL as

$$|\hat{D}_i(\omega)|^2 = |\hat{H}_i(\omega)|^2 |X_i(\omega)|^2, \quad (22)$$

where $|\hat{D}_i(\omega)|^2$ is the estimated echo level. The gain calculator determines ER gains, which have to be calculated in a short period under nonstationary speech conditions as follows:

$$G_i(\omega) = \frac{|\hat{S}_i(\omega)|^2 - \tau_i(\omega)|\hat{D}_i(\omega)|^2}{|\hat{S}_i(\omega)|^2}, \quad (23)$$

where $G_i(\omega)$ denotes the ER gain and

$$\tau_i(\omega) = \frac{P\left[|\hat{D}_i(\omega)||\hat{S}_i(\omega)|\right]}{P\left[|\hat{D}_i(\omega)|^2\right]}. \quad (24)$$

The gain calculator is designed to calculate ER gains on the assumption that a slight correlation between echo and near-end speech signals remains in short-period processing [20]. An advantage of adopting this strategy is the ability to calculate the ER gains with high accuracy even in a short period, which contributes to sufficiently suppress the residual echo by the ER process resulting in natural near-end speech.

The VLIC process is used for howling cancellation. When the ACL is above 0 dB, the echo canceller begins howling immediately after it is turned on because there is no prior training. To prevent howling, variable losses are inserted into the system. When the far-end speech level is higher than the near-end speech level, the loss is inserted into the send side. On the contrary, when the near-end speech level is higher than the far-end speech level, the loss is inserted into the received side. The loss-insertion level is determined from the ACL. The variable-loss insertion also applies different losses to each frequency component to decrease the loss margin [21].

IV. Performance Evaluation

The performance of the new AENC was evaluated using the objective assessment methods. The sampling frequency of test signals was 16 kHz and their frequency range was 100-7000 Hz. The room reverberation time was about 300 ms.

A. Complexity evaluation of ADF process

The optimal step-size control schemes are often proposed as the robust frequency domain ADF methods [5]-[7]. In this paper the algorithm that can achieve the low computational complexity while maintaining the robustness of these schemes was added to these schemes. This complexity reduction algorithm makes effective use of buffer delays caused by A/D and D/A converters, as described in section III.A. This paper compared a difference in the computational complexity caused by the presence or absence of the complexity reduction algorithm. To keep the comparison fair, same level of optimization was applied both to the codes with and without the complexity reduction algorithm. The computational complexity was evaluated by measuring the processing speed by using a commercial 2.8-GHz central processing unit (CPU). In fact this is a different processor from the target processor thus the evaluation results could only be used to compare the methods. The test used a speech signal whose duration was 60 s. The experiment showed that the required processing time of the ADF with the complexity reduction algorithm was approximately 0.8 s and the required processing time of the ADF without its algorithm was approximately 2.5 s. These results suggest that the computational complexity was reduced to about 1/3 by using the complexity reduction algorithm.

B. Performance evaluation of NR process

To evaluate the performance of the proposed NR method, described in section III.B, noise reduction rates (NRRs) [22] for various types of noise, i.e. the airport, lobby and office noises, were calculated. The noise signals were selected from an environmental noise database of NTT Advanced Technology Corporation [23]. The signal-to-noise ratio (SNR) between the speech and background noise signals was about 6 dB. Voices of five males and five females were employed as the speech signals; these signals were recorded based on international standards ITU-T Recommendation P.800 [24]. The ordinary noise-level estimation method [13] that has almost the same complexity as that of the proposed method was used as the conventional method for the performance evaluation. The computational complexities of the conventional and proposed methods, which were measured under the same test condition as stated in section IV.A, were approximately 0.1 s and 0.12 s, respectively. Likewise, these evaluation results could only be used to compare the methods because the processor used was different from the target processor. These satisfy the requirements of the casing size and DSP-chip performance on videophones. The comparison results of the NR performance are shown in Table II. As shown in the table, NRR of the proposed method outperformed the conventional methods for all cases.

TABLE II
COMPARISON BY NRR

Noise Type	Conventional	Proposed
Airport noise	6.31 dB	6.64 dB
Lobby noise	5.58 dB	6.31 dB
Office noise	8.09 dB	8.22 dB

C. Comparison of Noise-level Estimation Accuracy

This paper verified the effects of the combination of the ADF with the noise robust step-size control and the noise-level estimation. The received and near-end speech signals were male and female English speeches, respectively. The background noise used in the experiment was a white noise. The microphone input signal included the echo, near-end speech, and background noise. The noise-level estimation accuracy was calculated from a segmental SNR of the target noise level and the estimated noise level.

The comparison results of the noise-level estimation accuracy are shown in Table III. The estimation accuracy was evaluated in both a single-talk situation (when the microphone picks up only the echo and the background noise) and a double-talk situation. "NR only" denotes that the ADF process was omitted. "Conventional ADF + NR" is the combination of the conventional ADF and NR; the conventional ADF calculates the step size based on the ordinary Gaussian-Gaussian mixture assumption [25]. "Proposed ADF + NR" is the combination of the ADF with the noise robust step-size control and NR. As Table III indicates, the noise-level estimation accuracy was improved by combining the proposed ADF with the NR compared with "NR only" and "Conventional ADF + NR". A better score was observed in both the single-talk and double-talk situations by combining the proposed ADF.

TABLE III
COMPARISON OF NOISE-LEVEL ESTIMATION ACCURACY

Category	Single-Talk Situation	Double-Talk Situation
NR only	12.63 dB	10.91 dB
Conventional ADF + NR	14.23 dB	11.90 dB
Proposed ADF + NR	14.41 dB	12.05 dB

D. Overall Performance Test

Finally, the overall performance of the proposed AENC was experimentally evaluated using measured data. The experimental arrangement conformed to ITU-T Recommendation P.340 [26]. The experimental conditions of the received and near-end speech signals included four different patterns consisting of the combinations: male-female, female-male, male-male and female-female. The language used in this test was Japanese.

Figs. 6 and 7 show an example of the received and near-end speech signals. The microphone input signal is shown in Fig. 8. During the entire measuring period, the microphone always picked up background noise. The figure also shows period "A", which represents a single-talk situation, and period "B", which represents a double-talk situation, where the microphone picks up the echo, near-end speech, and background noise. The back ground noise used in the experiment is the air conditioning noise shown in Fig. 9. The send signal after the AENC is shown in Fig. 10. This figure shows that the proposed AENC effectively suppresses echo and background noise and the near-end speech signal is hardly distorted.

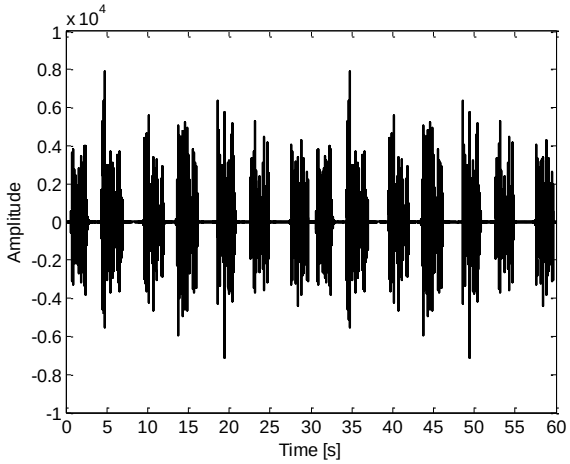


Fig. 6. Received speech signal (male).

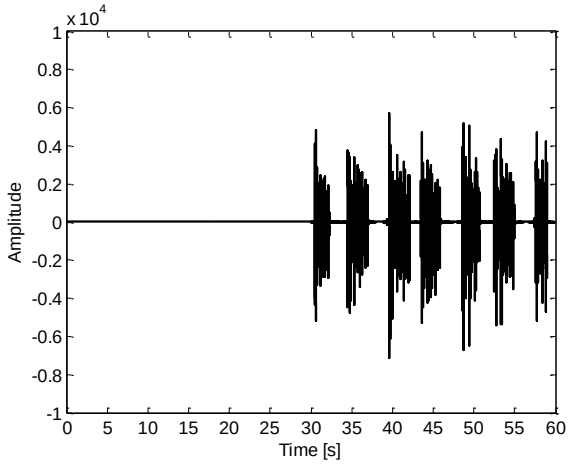


Fig. 7. Near-end speech signal (female).

The performance of the proposed AENC during the single-talk situation was evaluated using the echo suppressing level (ESL) and the noise suppressing level (NSL), respectively. These objective evaluation values were calculated from the difference of the signal levels before and after processing during the single-talk situation. The average ESL across the four speaker combination patterns was about 33 dB and the average of the NSL was about 20 dB, respectively.

The performance of the AENC during the double-talk situation was evaluated using the perceptual evaluation of speech quality (PESQ) [27]. The PESQ calculates the distance between the near-end speech signal and the send signal, and obtains a prediction value of the subjective mean opinion score (MOS) as the PESQ score. The PESQ score is mapped

from 1.0 (worst) up to 4.5 (best). The averages of PESQ score of the microphone input and send signals were 1.14 points and 1.88 points, respectively. These results show that the PESQ score was improved by suppressing echo and background noise while maintaining the quality of the near-end speech.

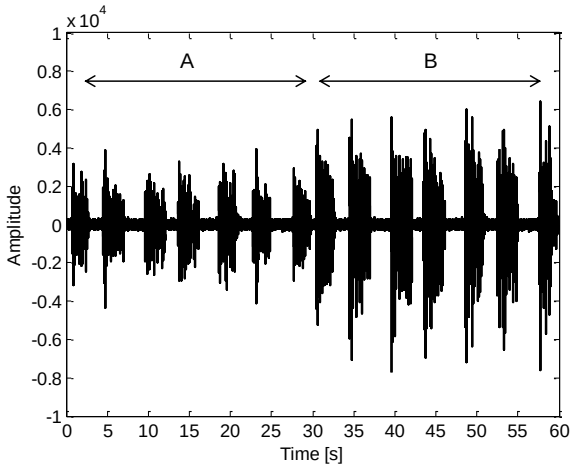


Fig. 8. Microphone input signal. Period A: echo and noise signals during single-talk situation, period B: double-talk situation.

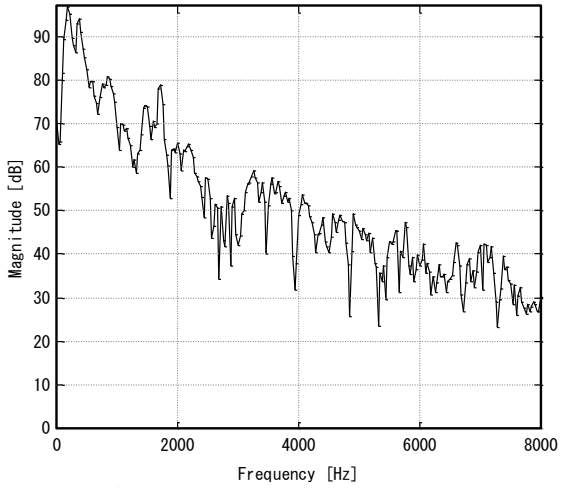


Fig. 9. Spectrum of background noise.

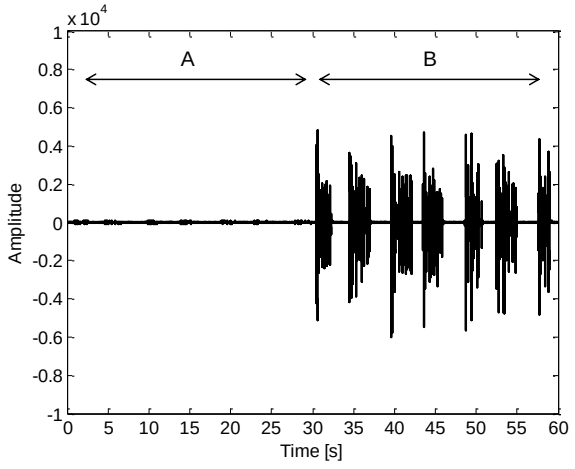


Fig. 10. Send signal. Period A: single-talk situation, period B: double-talk situation.

The overall performances of the proposed AENC under various noise conditions are summarized in Table IV. The results are averages across different speakers. The received and near-end speech signals are used by four different patterns consisting of the combinations: male-female, female-male, male-male and female-female; the language is Japanese. Three background noise signals, namely airport, lobby and office noises, were selected from the environmental noise database [23]. The SNR between the near-end speech and background noise signals was about 15 dB. The table shows that the proposed method sufficiently suppressed the echo and background noise irrespective of the noise type while keeping the near-end speech to be natural-sounding.

TABLE IV
OVERALL PERFORMANCES OF PROPOSED AENC

Noise Type	ESL	NSL	PESQ
Airport Noise	30.14 dB	20.68 dB	1.75 points
Lobby Noise	32.98 dB	20.39 dB	1.60 points
Office Noise	34.25 dB	21.07 dB	1.87 points

V. CONCLUSION

This paper implemented a robust frequency domain ADF method that uses a normalized residual echo enhancement; the filter calculation is based on the Gaussian-Laplacian mixture assumption for the signals normalized by the reference input signal amplitude; this method provides an optimal step-size control scheme for acoustic echo cancellation in a noisy environment. In addition, to reduce the computational complexity of the ADF, a three-frame echo replica is calculated at the same time without increasing the processing delay. This paper also introduced a new NR method that can emphasize the target near-end speech with low degradation. This method estimates the noise level for each frequency bin in the signal whether or not the background noise is superimposed by the speech. The proposed AENC was implemented in a hands-free videophone prototype. The experiments were conducted with the prototype and demonstrated that the proposed method suppresses undesired echo and noise, resulting in natural-sounding near-end speech.

REFERENCES

[1] 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," *IEEE Trans. Consumer Electron.*, vol. 60, no. 3, pp. 461-467, Aug. 2014.

[2] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Process.*, vol. 64, no. 1, pp. 21-32, Jan. 1998.

[3] S. Haykin, *Adaptive filter theory*, 3rd ed., Prentice-Hall, Inc.: New Jersey, USA, pp. 365-444, Dec. 1995.

[4] Y. Haneda, S. Makino, J. Kojima, and S. Shimauchi, "Implementation and evaluation of an acoustic echo canceller using duo-filter control system," *in Proc. European Signal Process. Conference*, Trieste, Italy, vol. 2, pp. 1115-1118, Sept. 1996.

[5] S. Shimauchi, Y. Haneda, and A.Kataoka, "Robust frequency domain acoustic echo cancellation filter employing normalized residual echo enhancement," *IEICE Trans. Fundamentals*, vol. E91-A, no. 6, pp. 1347-1356, Jun. 2008.

[6] T. S. Wada and B.-H. Juang, "Enhancement of residual echo for robust acoustic echo cancellation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 1, pp. 175-189, Jan. 2012.
[7] J. M. Gil-Cacho, T. V. Waterschoot, M. Moonen, and S. H. Jensen, "A frequency-domain adaptive filter (FDAF) prediction error method (PEM) framework for double-talk-robust acoustic echo cancellation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 12, pp. 2074-2086, Dec. 2014.
[8] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals," *Signal Process.*, vol. 64, no. 1, pp. 33-47, Jan. 1998.
[9] C. Avendano, "Acoustic echo suppression in the STFT domain**,"** *in Proc. IEEE Workshop Signal Process. to Audio and Acoustics*, 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, Sep. 2005.
[11] M. Fukui, S. Shimauchi, Y. Hioka, A. Nakagawa, and Y. Haneda, "Double-talk robust acoustic echo cancellation for CD-quality hands-free videoconferencing system," *IEEE Trans. Consumer Electron.*, vol. 60, no. 3, pp. 468-475, Aug. 2014.
[12] J. Li, Q.-J. Fu, H. Jiang, and M. Akagi, "Psychoacoustically-motivated adaptive β-order generalized spectral subtraction for cochlear implant patients," *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, pp. 4665-4668, Apr. 2009.
[13] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1383-1393, May 2012.
[14] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 6, pp. 1240-1250, Jun. 2013.
[15] Y. A. Huang, A. Luebs, J. Skoglund, and W. B. Kleijn, "Globally optimized least-squares post-filtering for microphone array speech enhancement," *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, pp. 380-384, Mar. 2016.
[16] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 27, no. 2, pp. 113-120, Apr. 1979.
[17] M. Fukui, S. Shimauchi, Y. Hioka, A. Nakagawa, Y. Haneda, H. Ohmuro, and A. Kataoka, "Noise-power estimation based on ratio of stationary noise to input signal for noise reduction," *Journal of Signal Process.*, vol. 18, no. 1, pp. 17-28, Jan. 2014 (in Japanese).
[18] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 32, no. 6, pp. 1109-1121, Jan. 2003.
[19] M. Fukui, S. Shimauchi, Y. Hioka, Y. Haneda, H. Ohmuro, and A. Kataoka, "Fast and accurate acoustic-coupling level estimation for echo reduction," *Journal of Signal Process.*, vol. 17, no. 5, pp. 167-177, Sep. 2013 (in Japanese).
[20] M. Fukui, S. Shimauchi, Y. Hioka, A. Nakagawa, Y. Haneda, A. Kataoka, and H. Ohmuro, "Wiener solution considering cross-spectral term between echo and near-end speech for acoustic echo reduction," *Acoustical Science and Technology*, vol. 35, no. 3, pp. 150-158, May 2014.
[21] K. Kobayashi, K. Furuya, Y. Haneda, and A. Kataoka, "Howling canceller based on sparseness of speech for hands-free system," *IEICE Technical Report*, vol. 107, no. 170, EA2007-35, pp. 1-6, Jul. 2007 (in Japanese).
[22] H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ICA and beamforming," *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, vol. 5, pp. 2733-2736, Mar. 2001.
[23] "Ambient noise database CD-ROM," NTT-AT.
[24] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," International Telecommunications Union, Geneva, Aug. 1996.
[25] K. Fujii and J. Ohga, "Optimum adjustment of step gain in learning identification algorithm," *IEICE Trans. Fundamentals* (Japanese Edition), vol. J75-A, no. 6, pp. 975-982, Jun. 1992.
[26] ITU-T Recommendation P.340, "Telephone transmission quality, telephone installations, local line networks," International Telecommunication Union, Geneva, May 2000.
[27] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunications Union, Geneva, Feb. 2001.

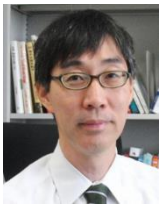
BIOGRAPHIES

Masahiro Fukui (M'09) received his BE degrees in information science from Ritsumeikan University, Shiga, Japan, in 2002. He received his ME 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 the International Conference on Consumer Electronics (ICCE) 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-SM'14) received his BE, ME, 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 BE, ME, 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), 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.