Doctoral Dissertation

Acoustic Echo Control for Improving Double-Talk Performance

December 2018

FUKUI Masahiro

Doctoral Dissertation Reviewed

by Ritsumeikan University

Acoustic Echo Control for Improving Double-Talk Performance

（同時通話性能を改善するための音響エコー制御）

December 2018

2018 年 12 月

FUKUI Masahiro

福井 勝宏

Principal referee: Professor YAMASHITA Yoichi

主査：山下 洋一教授

ABSTRUCT

This dissertation focuses on robustness against double talk for an acoustic echo canceller (AEC) and introduces high-performance acoustic echo control algorithms to improve quality of telecommunication systems.

On an acoustic echo control strategy, development of a double-talk-robust echo reduction (ER) process is one of the main targets. This study proposes a novel ER process robust against the double talk in estimations of echo-path power spectrum, echo-reduction gains, and late echo components. The echo-path power spectrum estimation algorithm is an echo-path-change robust algorithm that estimates the echo-path power spectrum in time and frequency spectral domains; this algorithm achieves the high tracking performance and accuracy of the echo-path power spectrum. The echo-reduction gain estimation algorithm is a novel estimation algorithm that solves a least square error problem of Wiener filtering (WF) method while taking into account cross-spectral term of the signals; thereby, this algorithm obtains a better echo-reduction gain than that of the conventional WF method. The late echo component estimation algorithm is a novel estimation algorithm that accurately estimates the echo power spectrum corresponding to early impulse response and late echo components resulting from reverberation beyond a length of fast Fourier transform (FFT) block; this algorithm estimates the echo power spectrum by assuming a finite nonnegative convolution model.

In addition, this study develops an AEC devices and an application software where the proposed estimation algorithms of the echo-path power spectrum and the echo-reduction gain are implemented. The developed AEC device is combined with a videoconferencing system and used in hands-free telecommunication; its frequency band of audio signal is supported up to compact disc (CD)-quality, i.e. 20-kHz wideband and delivered natural-sounding speech. The developed software for voice over internet protocol (VoIP) hands-free phone application on smartphone and tablet devices automatically tailor its performance to the acoustic characteristics of individual smartphone and tablet devices, and reduces the influence due to the difference in the acoustic characteristics of individual devices. The developed video phone employs noise-robust adaptive filter (ADF) and noise reduction (NR)

processes and maintains the acoustic echo and noise cancelling performance in a noisy environment such as an open-plan office.

内容概要

本論文では，音響エコーキャンセラ（AEC）の同時通話に対するロバスト性に重点を置いており，遠隔会議システムの品質を向上させる高性能音響エコー制御アルゴリズムについて述べる．

本論文の主な目標は，同時通話にロバストなエコーリダクション（ER）を開発することであり，エコー経路パワースペクトル，エコーリダクションゲイン，後部残響成分の推定における同時通話に対してロバストな ER を提案する．エコー経路パワースペクトル推定アルゴリズムは，時間および周波数スペクトル領域においてエコー経路のパワースペクトルを推定する，エコー経路の変化にロバストなアルゴリズムである．このアルゴリズムは，エコー経路パワースペクトルに対する高い追従性能と推定精度を達成する．エコーリダクションゲイン推定アルゴリズムは，信号のクロススペクトル項を考慮しながら，ウィーナフィルタ（WF）法の最小二乗誤差問題を解くアルゴリズムである．従来の WF 法に比べて推定誤差の少ないゲインを得ることができる．後部残響成分推定アルゴリズムは，高速フーリエ変換（FFT）ブロックの長さを超える残響成分に相当するエコーパワースペクトルを高精度に推定するアルゴリズムである．このアルゴリズムは，有限非負畳み込みモデルを仮定することによってエコーパワースペクトルを推定する．

提案するエコー経路パワースペクトル推定アルゴリズムとエコーリダクションゲイン推定アルゴリズムは， AEC 端末とアプリケーションソフトウェアに実装された．開発された AEC 端末は，ビデオ会議システムと組み合わせ，拡声通話で使用される．本端末は，周波数帯域を 20 kHz の広帯域までサポートし，自然な発話を提供する．スマートフォンやタブレット端末用の VoIP ハンズフリー電話アプリケーションとして開発されたソフトウェアは，個々のスマートフォンやタブレット端末の音響特性に自動的に適合し，端末ごとに異なる音響特性の違いによる性能低下を抑える．開発されたテレビ電話は，雑音高耐性の適応フィルタ（ADF）とノイズリダクション（NR）を採用し，オープンプランオフィスなどの騒がしい環境で高いエコー・ノイズキャンセル性能を達成する．

ACKNOWLEDGEMENTS

Over the course of this work and my research career at NTT Laboratories, a number of professors, managers, research leaders, colleagues, friends, and family made significant contributions, directly and indirectly. Without the wisdom and teachings of each one of these people, my undertaking of this work would be tremendously hindered.

I would especially like to express my sincere thanks to my supervisor, Professor Yoichi Yamashita of Ritsumeikan University for providing the academic hospitality, supportive professional guidance, and constructive discussions leading to this dissertation. I would also like to express my thanks to Professor Suehiro Shimauchi of Kanazawa Institute of Technology, my first supervisor at NTT Laboratories, for kindly directing me at the beginning of my research career at NTT. In addition, I obtained special research stimulation from Professor Kenichi Furuya of Oita University. Moreover, my special thanks go to Professor Yusuke Hioka of The University of Auckland for his precious discussions and advices.

I am grateful to my current research project manager at NTT Laboratories, Dr. Sumitaka Sakauchi, for providing me with the opportunity to carry out this research work to its completion. My current research group leader at NTT Laboratories, Dr. Noboru Harada kindly encouraged me to carry out the research work.

I am very thankful to my supervisors and colleagues of NTT Laboratories; the members who should be especially mentioned are Mr. Shigeaki Sasaki, Dr. Yusuke Hiwasaki, Mr. Akira Nakagawa, Dr. Satoru Emura, and Dr. Kazunori Kobayashi. I am also indebted to former research managers and leaders of NTT Laboratories; Mr. Hisashi Ohara, Mr. Akihiro Imamura, Dr. Hirohito Inagaki, Dr. Satoshi Takahashi, Dr. Akitoshi Kataoka, Dr. Yoichi Haneda, and Mr. Hitoshi Ohmuro have been very supportive of my research endeavor.

Finally, I appreciate to my parents, Mitsuyuki and Chieko, for their patients and supports. Special thanks also go to my wife, Emi, for understanding my ambitions and supporting me to cope with my work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1.

Introduction

1.1 *Background and Objective*

1.1.1 Proposed speech-quality improving algorithms

Due to the amazing growth in networking technologies, the recent broadband environments enable us to comfortably communicate with someone in a remote site in a variety of styles. A hands-free telecommunication is a very natural style because it provides us to talk with far-end people by using loudspeaker and microphone instead of holding a handset. Especially, the hands-free telecommunications are really convenient in the cases where many people on the same site want to communicate with remote site at the same time, e.g., a business teleconferencing situation. This style is also used where one wants to talk to the remote site while handling something else, e.g., car-driving situation.

In order to realize a natural hands-free telecommunication without a great stress, the acoustical problems that degrade the transmitted speech quality must be solved. Acoustic echo cancellation is one of fundamental techniques that eliminates the acoustic echo and howling caused by acoustic coupling between loudspeaker and microphone. The other fundamental techniques are noise reduction, beamforming and so on. Noise reduction technique suppresses ambient noise picked up by the microphone. Beamforming technique focuses on the talker's speech based on the microphone array processing. However, the most indispensable technique is acoustic echo cancellation; it is because once howling occurs,

conversation cannot be sustained.

Most acoustic echo cancellers (AECs) use in series an adaptive filter (ADF) [1] [2] [3] [4] that has a finite impulse response (FIR) filter structure, and an echo reduction (ER) [5] [6] [7] [8] [9] which is a nonlinear post-filter based on a short-time spectral amplitude (STSA) estimation [10]. The ADF process cancels out the echo signal by adaptively modeling an unknown acoustic echo path but some residual echo signal still remains in its output (often called *error signal*). Thus, the ER process follows the ADF process and reduces the residual echo signal [11] [12]. This process suppresses the residual echo by applying a multiplicative gain, which is called an echo-reduction gain, for each frequency component of the error signal; the process affects only the amplitude components of the signal and does not affect the phase components. The echo-reduction gain is calculated from the estimate of power spectrum of the echo-path impulse response (we call it *echo-path power spectrum* hereafter).

The echo-reduction performance depends on the estimation accuracy of echo-path power spectrum; and therefore, many methods for estimating the echo-path power spectrum have been proposed and applied so far. Two representative methods exist to estimate the echo-path power spectrum; one is a straightforward and simple method [5] [6] [12] [13] that measures the echo-path power spectrum when a near-end talker's voice (often called *near-end speech*) is detected to be absent; the other is a frequency-domain cross correlation (FDCC) method [7] [8] [9] using the correlation between received and error signals.

The literatures [5] [6], which are the above straightforward echo-path power estimation methods, employ a double-talk detector [14] [15] [16] [17]; herewith, the estimation process is suspended if the double-talk periods, which are periods when near-end and far-end talkers speak simultaneously, are detected; therefore, the disturbance of the estimation value is reduced. Instead of adopting double-talk detetor, the literature [12] calculates the echo-path power spectrum based on the local minimum value on the amplitude ratio between the spectra of the microphone-input and received signals across the past processing frames; herewith, this method can reduce an influence of the disturbance during the double-talk periods. However, in these methods suspending the echo-path power estimation during the double-talk periods, it is theoretically difficult to track the echo-path change in the double-talk periods.

On the other hand, a major advantage of FDCC is that it can be used to estimate the echo-

path power spectrum even during double-talk periods because this method can reduce the influence of the disturbance by using the cross-correlation between the microphone-input and received signals. Therefore, the FDCC approach is gradually becoming the mainstream regarding the echo-path power spectrum estimation. However, a problem with FDCC is its slow tracking speed. This method assumes that the received and near-end speech signals are statistically uncorrelated to estimate the echo-path power spectrum and to obtain the echo component. This assumption only holds true if the observation period (time period) of the signal is long enough. As a result, the FDCC approach might fail to accurately track immediate echo-path change in the residual echo level.

The echo-reduction gain, which decides the echo-suppression level of the residual echo, is obtained based on the estimate of the echo-path power spectrum; this gain is also decided according to the ratio of the near-end speech components included in the microphone input signal. Thereby, it is possible to suppress the residual echo components according to the echo-superimposing ratio even in the double-talk periods and to have small speech distortion although the ER approach is a nonlinear process if the high-accuracy echo-reduction gain can be obtained.

As the representative methods to derive the echo-reduction gain, there are spectral subtraction [10], Wiener filtering (WF) [18] [19], maximum likelihood (ML) estimation [20] [21], minimum mean-square error (MMSE) estimation [22] [23] and so on. The spectral subtraction method is a straightforward method of suppressing the residual echo components by subtracting the estimated echo power spectrum from the power spectrum of the echo-superimposing signal. The WF method is designed to minimize the mean square error between the estimated speech and the desired speech, and derives the echo-reduction gain with less speech distortion compared with the spectral subtraction method. The ML and MMSE estimation methods are statistical models using a general framework of the estimation theory such as Bayesian estimation; they derive the echo-reduction gain by regarding an echo-suppression problem as a problem of estimating and restoring the desired near-end speech spectrum.

The WF method has been widely utilized as the echo-reduction estimation approach that establishes practically high performance and low computational complexity; thereby, this study focuses on the WF method and addresses its practical problem. The WF method

estimates the echo-reduction gain based on the assumption that acoustic echo and near-end speech signals are statistically uncorrelated; and so a cross-spectral term of their signals is regarded as zero. However, in its application to the hands-free telecommunication system, the echo-reduction gain needs to be calculated from a very short period because most echo and near-end speech signals are usually nonstationary; therefore, the cross-spectral term between echo and near-end speech signals will not always become zero because the number of samples used in calculating the cross-spectral term is small. As a result, the ER process based on the conventional WF method might cause speech distortions during double-talk periods because the assumption does not hold in practical use.

In order to apply the AEC to the telecommunication system, the processing delay caused by the ADF and ER processes should be as short as possible so as not to hinder comfortable telecommunications with far-end talkers; therefore, the length of its processing frame must be as short as possible. The FDCC method, which is the estimation method of the echo-path power spectrum based on the cross-correlation approach, can estimate the echo-path power spectrum even during double-talk periods; however, this approach regards only a fraction of the echo-path impulse response corresponding to the early response that mainly consists of direct sound and early reflections as the echo-path power spectrum if calculating in the short-time processing frame. This is because the time length of one block where the fast Fourier transform (FFT) is applied is usually far shorter than that of the echo path impulse response [24]. Thereby, the late echo components that result from late reverberation are not considered.

To address this problem, a moving average (MA) model [24] [25] of the late reverberation for the echo-path impulse response is often used in conjunction with the FDCC method in order to offset the late echo components. In this model, the late echo components are compensated for by adding recursively the estimate of the echo power spectrum obtained with the FDCC method. The added amount of the late echo components is usually adjusted according to the reverberation time. An adaptive method to estimate the reverberation time was also proposed recently [24]. However, a problem with the MA model is to require the assumption that the spectral structure of the microphone input signal is stationary. This assumption does not necessarily hold true because a non-stationary signal such as speech is given by conversation on the telecommunications. As a result, the MA model often fails to

accurately estimate the echo power spectrum in practical reverberant environments, and this causes the perceptual degradation of sound quality after applying the ER process.

This dissertation focuses on the development of the ER algorithms that are suitable for the estimations of echo-path power spectrum, echo-reduction gain, and late echo components that result from late reverberation. Details for the ADF process is another interesting research topic in the AEC; and thereby, it is out of focus in the present dissertation.

For an estimation of the echo-path power spectrum, this dissertation proposes a new estimation algorithm in time and frequency spectral domains; this study aims to immediately deal with instantaneous echo-path change accompanying the switching between different microphones while presuming that each of talkers uses their own microphone during the teleconference. Of course, this kind of instantaneous echo-path change can be avoided if the AEC is introduced for each microphone. However, there is a problem of cost increase due to the increase in both used memory size and computational complexity when introducing the same number of AECs as microphones. Therefore, this study aims at improving the tracking performance and accuracy of the echo-path power spectrum, which are important in improving the echo-reduction performance, in order to realize the ER process which exhibits high performance even when the AEC is introduced after switching of microphones.

For an estimation of the echo-reduction gain, this dissertation derives a novel estimation algorithm that solves a least mean square error of the WF method by taking into account the cross-spectral term of the signals; thereby, this method obtains a better echo-reduction gain than that of the conventional WF method. While the conventional WF method assumes that the cross-spectral term of the echo and near-end signals is zero, the proposed method estimates the echo-reduction gain based on the assumption that the cross-spectral term of their signals is not zero because the time-sequence period is short. An advantage of this strategy is to be able to accurately calculate the echo-reduction gain even in a short period and to decrease speech distortions.

For an estimation of the late echo components, this dissertation derives a novel estimation algorithm that accurately estimates residual echo spectrum that includes the echo components corresponding to the early impulse response and the late echo components resulting from reverberation beyond a length of FFT block. This method estimates the echo power spectrum by not using the MA model, but by assuming a *finite nonnegative convolution model* [26];

this model convolutes each segment of echo-path impulse response with the received signal in the power spectral domain, taking into account the non-stationarity of the signal. To estimate the echo power spectrum using this model, this dissertation proposes an innovative algorithm to obtain the optimal solutions of the echo-path power spectra in all segments. The proposed algorithm is based on a simultaneous equation model for power spectra of microphone input signal, received signal, and each segment of echo-path impulse response. The solution for each segment is equal to the least squares solution between the microphone-input and estimated-echo-power spectra. With the proposed algorithm, the proper estimates of each segment can be obtained so that the difference between the power spectrum of residual echo and its estimate is minimized per frame.

1.1.2 Developed devices and application software

The estimation algorithm of the echo-path power spectrum, which is one of methods presented in this dissertation, was implemented into a newly-developed AEC unit; its unit is combined and used with a videoconferencing system; and the frequency band of the audio signal in this AEC unit is supported up to compact disc (CD)-quality, i.e. 20-kHz wideband.

Until now, the AEC consisting of the ADF process and the ER process, has been frequently applied to practical products because of its promising performance; an ordinary AEC 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; 1) the sound of the near-end speech is sometimes muffled in high frequency ranges during the double-talk periods because of the incomplete ER process caused by a low-accuracy estimation of the echo-path power spectrum; and 2) the amount of computation for the ADF increases exponentially as the sampling frequency becomes higher.

One of the objectives of this study is to improve the estimation accuracy of the echo-path power spectrum and to eliminate the muffled sound of near-end speech after applying ER. Another objective is 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 study

addresses the estimation-accuracy improvement of the echo-path power spectrum 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 the above-mentioned proposed method, but different in that the target is 20-kHz wideband. Then, to improve the computational efficiency, this study employs a low-complexity subband approach with a new subband filtering algorithm in the AEC. The 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 study evaluated the performance of the proposed method by implementing it using an AEC-unit prototype that has a digital signal processor (DSP) board.

This dissertation introduces a newly-developed AEC software for voice over internet protocol (VoIP) hands-free phone application on smartphone and tablet devices. Smartphones and tablets have rapidly become popular among internet users in recent years. Along with their popularization, VoIP phone applications that can run on such devices are also becoming popular, whose worldwide market size has been increasing [26] [27] [28]; in this situation, 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. However, three model-specific problems arise when an ordinary AEC software for providing the hands-free conversation is applied to various smartphones or tablets: 1) loudspeaker distortion, 2) microphone sensitivity variation, and 3) audio input/output delay variation [9]. Problem 1) causes non-linear distortions in the echo, and problems 2) and 3) 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.

This dissertation therefore proposes a new AEC method that automatically tailors its performance to the acoustic characteristics of individual devices. 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.

An above-mentioned study discussed an implementation of the AEC for VoIP phones on smartphone and tablet devices. On the other hand, this study addresses a new acoustic echo and noise canceller (AENC) [29] for videotelephony-enabled personal hands-free internet protocol (IP) phones. The popularity of videotelephony-enabled 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 input signals because personal videophones are usually used in noisy environments such as offices.

The videotelephony-enabled IP phone has to be practically integrated into a single fixed-point 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 the ADF process, the ER process, and a noise-reduction (NR) [30] [31] [32] [33] process. The NR process lowers the level of background noise by multiplicative gains in the frequency domain. This process is based on an STSA estimation as same as the ER process. In recent years, high-performance NRs with a microphone array that consists of multiple microphones placed at different spatial locations have been proposed [32] [33]. 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 [30] [31], the number of components of which is small and the computational complexity is relatively low, are still widely used in the videophones.

This study addresses two issues in using the AENC for personal videophones; 1) to maintain the AENC performance in a noisy environment such as the open-plan office, and 2) to attain the performance while reducing the computational complexity to make real-time

implementation possible. For the issue 1), this dissertation introduces two methods. One of them is controlling the step size of the ADF [34] [35] [36]. 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 [37], 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 in the double-talk periods.

For the issue 2), this study attempts to reduce a 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 process, which is employed in the above-described VoIP phone application on smartphone and tablet devices, updates and convolves the filter coefficients within each single frame. As a result of such modification, the computational complexity is drastically decreased.

1.2 *Organization of Dissertation*

This dissertation is organized as follows. Chapters 2, 3, and 4 present novel algorithms specialized for the improvements for the echo-reduction performance in the ER process. Chapter 5, 6, and 7 introduce methods, which improve speech quality and calculation efficiency against an acoustic echo and noise control, implemented in developed products such as the hands-free telecommunication devices and VoIP application software.

Chapter 2 presents a new estimation method of the echo-path power spectrum for the ER process based on Wiener filtering. This method employs two techniques: 1) a fast and robust echo-path power spectrum estimation using a short-time correlation between received and microphone input signal in time and frequency spectral domains and 2) compensation of the estimation error caused by an inaccurate correlation that occurs when applying an FFT of a short finite frame length; and the proposed method continuously updates the estimation value and immediately tracks a rapid echo-path change even during double-talk periods.

Chapter 3 presents an ER process based on a new Wiener-filtering method taking into account the cross-spectral term between the acoustic echo and the near-end speech. The goal of this study is to accurately calculate the echo-reduction gain to decrease the speech distortions produced by the ER process. The proposed method solves a least mean square error of the WF method by taking into account the cross-spectral term between the echo and the near-end speech to obtain a better echo-reduction gain.

Chapter 4 deals with a problem of estimating residual echo spectrum that results from reverberant component beyond a length of FFT block. This chapter introduces a finite nonnegative convolution method by which each segment of echo-impulse response is convoluted with a received signal in a power spectral domain. With the proposed method, the power spectra of each segment of echo-impulse response are collectively estimated by solving the least-mean-squares problem between the microphone-input and the estimated-residual-echo power spectra.

Chapter 5 introduces a new monaural AEC method developed for a 20-kHz wideband hands-free video-teleconferencing system. The method can effectively reduce undesired acoustic echo included in a signal arriving at a microphone from a loudspeaker and can emphasize the near-end speech in the double-talk periods. The method estimates the echo-

path power spectrum whether or not double-talk has occurred, then calculates an echo-reduction gain that effectively reduces the residual echo. In the echo cancellation processing, the computational 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.

Chapter 6 presents a new AEC method developed for VoIP phone applications on smartphone and tablet devices. This method can effectively reduce the residual echo and emphasize the near-end speech in the double-talk periods, 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.

Chapter 7 presents implementation and evaluation of a new AENC for a hands-free personal videophone. This canceller has the following features: noise-robust performance, low processing delay, and low computational complexity. The AENC employs a new ADF and NR methods that can effectively eliminate undesired acoustic echo and background noise included in a microphone input signal even in a noisy environment. The ADF method uses a 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 dissertation 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.

Chapter 8 concludes this dissertation and indicate some future works, which should follow this work to innovate the tele communication technologies.

Chapter 2.

Robust Improvement in Estimation of Echo-Path Power Spectrum

2.1 *Introduction*

An ER process is an ordinary process for reducing residual echo components remaining in an error signal after an ADF process; and the FDCC method employed in the ER process is widely used as an effective approach for estimating an echo-path power spectrum. This method uses the correlation between received and error signals in a frequency domain, reduces an influence of disturbance components, and estimates the echo-path power spectrum. The method exhibits two primary advantages over the implementation to the ER process in practice. One is not to need a double-talk detector. The other is to be able to estimate the echo-path power spectrum even during double-talk periods. However, an issue with the FDCC method is its slow tracking speed; it is because this method needs long time observation period of the signals in order to regard the disturbance components as uncorrelation components. This slow tracking speed makes difficult to correspond to the immediate echo-path change.

This chapter presents a conventional ER process employing the FDCC method, an WF method, and an MA model, and then develops a novel estimation method of the echo-path power spectrum; its estimation method is based on a concept of the FDCC approach. This chapter further shows comparisons between the conventional and proposed methods in terms

of the estimation accuracy of the echo-path power spectrum, and demonstrates their echo-reduction performances.

2.2 *Configuration for Ordinary ER Process*

This section explains the ER process in a short-time spectral amplitude (STSA) domain using the FDCC, WF, and MA approaches. The structure of the ER process is illustrated in Figure 2.1, and the adaptive filter (ADF) process is omitted to simplify the explanation. The microphone input signal $y(k)$ is expressed as

$$y(k) = d(k) + s(k), \tag{2.1}$$

where k is a sample number of discrete time; $d(k)$ and $s(k)$ are acoustic echo and near-end speech signals, respectively. $d(k)$ is represented by a convolution of an echo-path impulse response $h(k)$ and a received signal $x(k)$, i.e.,

$$d(k) = h(k) * x(k), \tag{2.2}$$

where $*$ denotes the convolution. The short-time spectrum of $y(k)$ is represented as

$$Y_i(\omega) = D_i(\omega) + S_i(\omega), \tag{2.3}$$

where ω is a discrete frequency index, i is a discrete time-frame index, and $D_i(\omega)$ and $S_i(\omega)$ are the short-time spectra of $d(k)$ and $s(k)$, respectively.

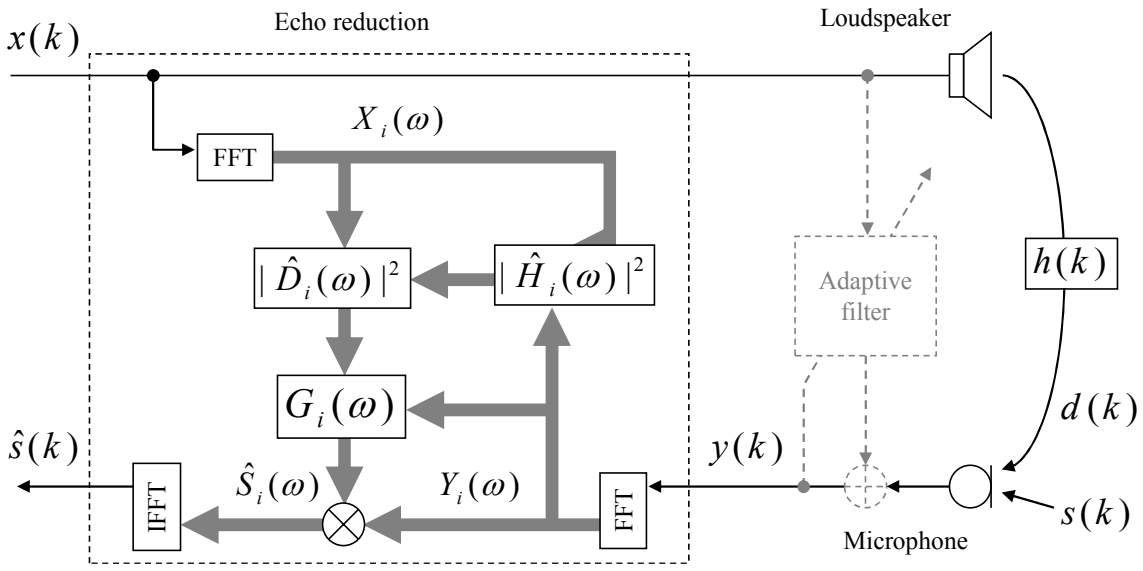


Figure 2.1. Structure of echo reduction process.

The echo reduction can be expressed using an echo-reduction gain $G_i(\omega)$ in a general form as follows:

$$\hat{S}_i(\omega) = G_i(\omega)Y_i(\omega), \tag{2.4}$$

where $\hat{S}_i(\omega)$ denotes the estimate of $S_i(\omega)$. $G_i(\omega)$ is for example calculated by an WF method obtained by the following equation:

$$G_i(\omega) = \frac{|Y_i(\omega)|^2 - |\hat{D}_i(\omega)|^2}{|Y_i(\omega)|^2}, \tag{2.5}$$

where $|\hat{D}_i(\omega)|^2$ is the estimate of echo power spectrum $|D_i(\omega)|^2$. The obtained estimate $\hat{S}_i(\omega)$ is transformed into the time domain signal $\hat{s}(k)$, which is the send signal, by an inverse fast Fourier transform (IFFT).

The echo power spectrum is estimated using the MA model as

$$|\hat{D}_i(\omega)|^2 = |\hat{H}_i(\omega)|^2 |X_i(\omega)|^2 + \alpha |\hat{D}_{i-1}(\omega)|^2, \tag{2.6}$$

where $|\hat{H}_i(\omega)|^2$ denotes the estimate of the echo-path power spectrum $|H_i(\omega)|^2$ and α is a design parameter to control the reverberation time and determine the amount of added late echo components; $0 \le \alpha \le 1$ is used according to the reverberation time; for instance, α is set at 0.7 when the frame shift size is 8 ms and the reverberation time with an RT60 [38] is 160 ms.

The echo-path power spectrum estimate $|\hat{H}_i(\omega)|^2$ is given by

$$|\hat{H}_i(\omega)|^2 = \left| \frac{\langle \mathbf{x}_i(\omega), \mathbf{y}_i(\omega) \rangle}{\|\mathbf{x}_i(\omega)\|^2} \right|^2, \tag{2.7}$$

where $\langle , \rangle$ and $\|\cdot\|$ are an inner product and a norm, respectively. Boldface denotes a time-sequence vector of a short-time spectrum: $\mathbf{p}_i(\omega) = [P_i(\omega), \cdots, P_{i-L+1}(\omega)]^T$. L is the number of frames, meaning the observation time parameter, and T is a transposition. The estimate of the echo-path power spectrum in the above equation is finally obtained from the following relationship:

$$\begin{aligned} |\hat{H}_i(\omega)|^2 &\approx \left| \frac{\langle \mathbf{x}_i(\omega), \mathbf{d}_i(\omega) \rangle + \langle \mathbf{x}_i(\omega), \mathbf{s}_i(\omega) \rangle}{\|\mathbf{x}_i(\omega)\|^2} \right|^2 \\ &\approx \left| \frac{\langle \mathbf{x}_i(\omega), \mathbf{d}_i(\omega) \rangle}{\|\mathbf{x}_i(\omega)\|^2} \right|^2 \\ &\approx \left| \frac{H_i(\omega) \langle \mathbf{x}_i(\omega), \mathbf{x}_i(\omega) \rangle}{\|\mathbf{x}_i(\omega)\|^2} \right|^2 \\ &= |H_i(\omega)|^2 \end{aligned}, \tag{2.8}$$

2.3 ***Problems with FDCC Method***

The FDCC method approximately simulates the echo-path power spectrum by the following three assumptions.

(i) The received signal and the near-end speech are uncorrelated.
(ii) The echo-path power spectrum is constant in the section of the number of frame L.
(iii) The received signal and the acoustic echo are highly correlated.

However, the assumptions (i), (ii), and (iii) have two significant problems as described below.

2.3.1 Problem resulting from assumptions (i) and (ii)

This subsection explains a problem on followability of the echo-path power spectrum estimation caused by the assumptions (i) and (ii).

The calculation range of the correlation is controlled by the number of frames L; namely, the influence of disturbance such as the near-end speech can be sufficiently eliminated by enough increasing L. However, when L is large, the assumption (ii) does not hold for a long time after the echo path changes; therefore, there is a problem that as the followability to the echo path degrades as L increases. On the other hand, if L is set small, the followability to the echo path improves; however, the influence of the disturbance cannot be sufficiently removed and so the assumption (i) does not hold. Therefore, the estimation of the echo-path power spectrum during the double-talk periods becomes difficult. Due to a trade-off relationship as mentioned above, there is a problem that it is difficult to achieve sufficient tracking performance by the conventional method.

2.3.2 *Problem resulting from assumption (iii)*

This subsection explains a problem on accuracy of the echo-path power spectrum estimation caused by the assumption (iii).

When an angle between time-sequence vectors of short-time spectral amplitudes of the received and echo signals is given by $\phi_i(\omega)$, the estimate of the echo-path power spectrum is represented by the following equation which is multiplied by the square of cosine similarity $|\cos\phi_i(\omega)|$:

$$\begin{aligned} |\hat{H}_i(\omega)|^2 &\approx \left| \frac{\langle \mathbf{x}_i(\omega), \mathbf{d}_i(\omega) \rangle + \langle \mathbf{x}_i(\omega), \mathbf{s}_i(\omega) \rangle}{\|\mathbf{x}_i(\omega)\|^2} \right|^2 \\ &\approx \left| \frac{\langle \mathbf{x}_i(\omega), \mathbf{d}_i(\omega) \rangle}{\|\mathbf{x}_i(\omega)\|^2} \right|^2 \\ &= \left| \frac{\|\mathbf{x}_i(\omega)\| \|\mathbf{d}_i(\omega)\| \cos\phi_i(\omega)}{\|\mathbf{x}_i(\omega)\|^2} \right|^2 \\ &= \frac{\|\mathbf{d}_i(\omega)\|^2}{\|\mathbf{x}_i(\omega)\|^2} |\cos\phi_i(\omega)|^2 \end{aligned} . \quad (2.9)$$

Namely, the estimate of the echo-path power spectrum is based on $|\cos\phi_i(\omega)| = 1$. However, the value of $|\cos\phi_i(\omega)|$ changes corresponding to the distance between the loudspeaker and microphone, and so on.

For instance, this subsection considers the changes of cosine similarity $|\cos\phi_i(\omega)|$ with the switching of two microphones when one speaker and two microphones are arranged as illustrated in Figure 2.2. And Figure 2.3 plots a comparison of cosine similarities when the reverberation time in this experiment is 300 ms and the distances of the microphone from the loudspeaker are 1 m and 2m. As these results show, the cosine similarities that the distances of the microphone from the loudspeaker are 1 m and 2m are significantly different.

Thereby, the error associated with the cosine similarity $|\cos\phi_i(\omega)|$ occurs regarding the estimate of the echo-path power spectrum because the assumption (iii) does not hold true.

Figure 2.4 plots the relationship of the distance of the loudspeaker and microphones and the estimation accuracy of the echo-path power spectrum. As this figure shows, the estimation accuracy at the distance of 1 m from the loudspeaker are significantly different from that of 2m.

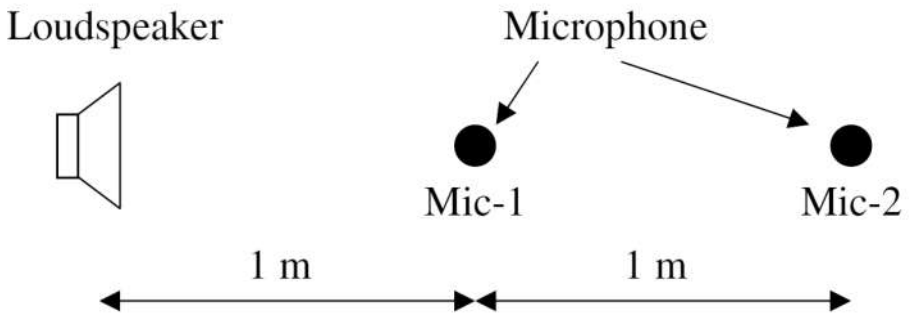


Figure 2.2. Locations of loudspeaker and microphones.

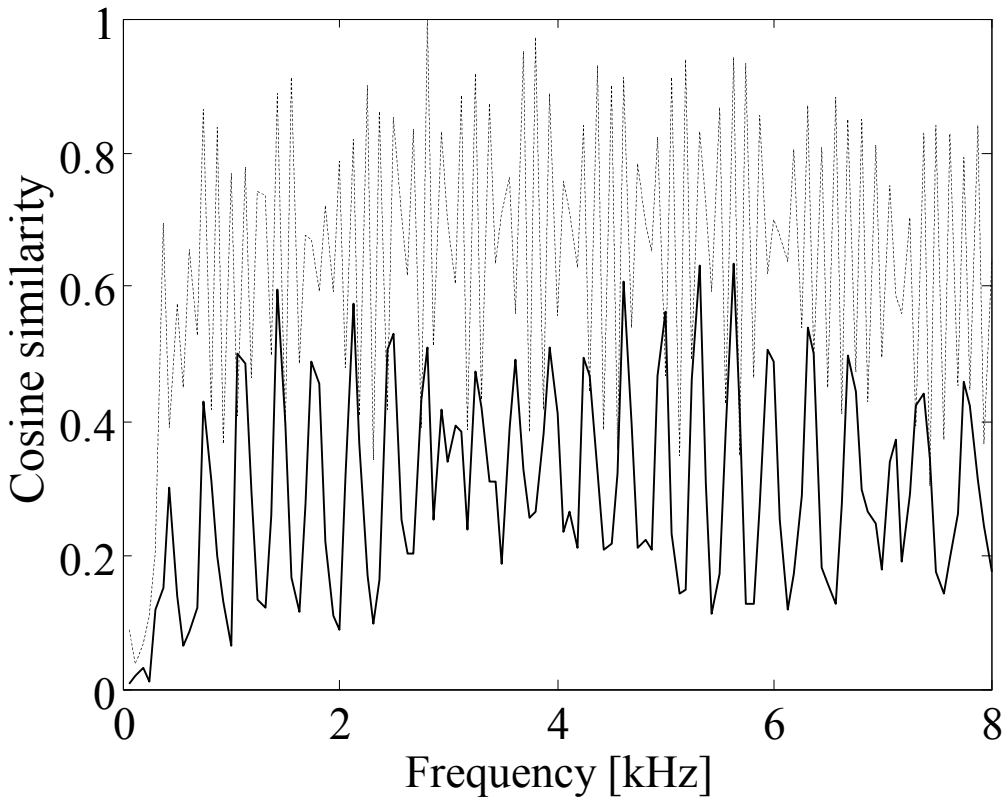


Figure 2.3. Relationship of cosine similarity and distance between loudspeaker and microphones (dotted line: 1 m, solid line: 2 m).

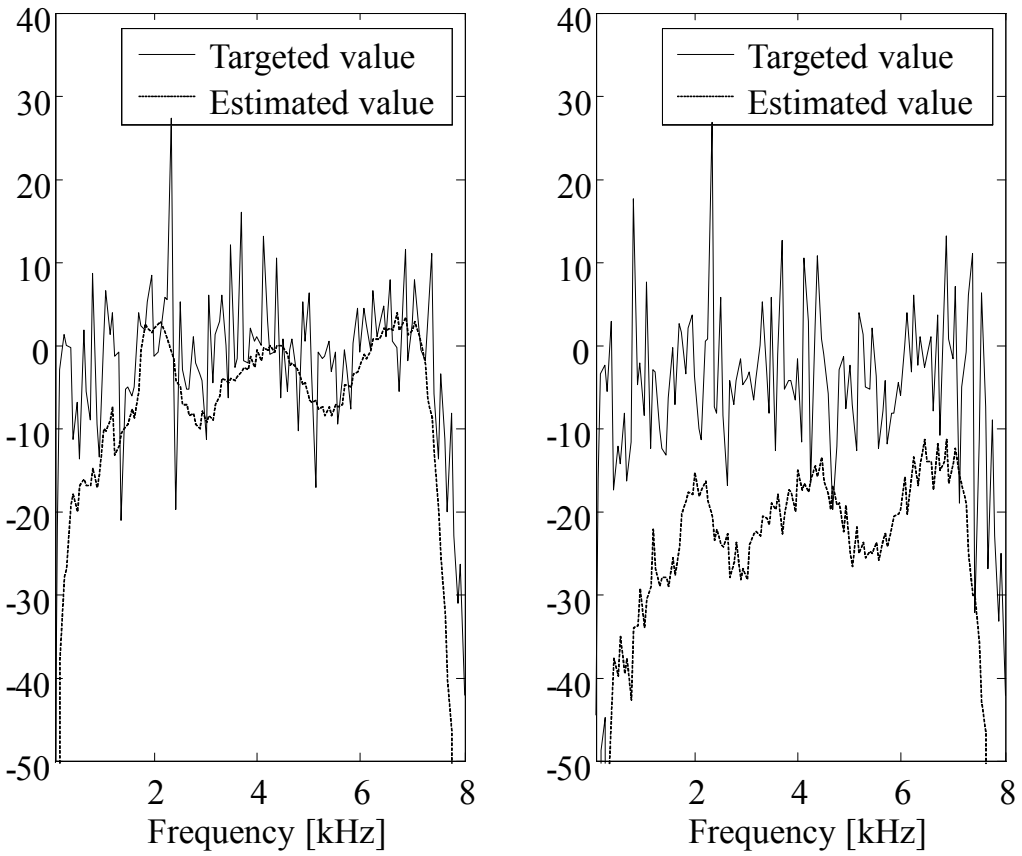


Figure 2.4. Relationship of distance between loudspeaker and microphones and estimation accuracy of echo-path power spectrum: left: 1 m, right: 2 m.

2.4 *Proposed Echo-Path Power Spectrum Estimation Method*

This section proposes a new method that solves the problem accompanied with conventional FDCC method by improving a followability while achieving a high-accuracy estimation of the echo-path power spectrum. The proposed method has two steps: (i) an improvement of tradeoff between the followability and accuracy controlled by the number of frames L : and (ii) an improvement of the estimation accuracy considering the influence of cosine similarity. The structure of the proposed method on the echo-path power spectrum estimation is illustrated in Figure 2.5.

2.4.1 Correlation calculation method for improving followability

The proposed method focuses on not only a time axis direction but also a frequency axis direction in the correlation calculation between received and microphone input signals in order to improve the followability to echo-path changes. This subsection describes the method for estimating the echo-path power spectrum by using the correlation in the frequency axis direction to the correlation in the time axis direction, and presents its validity.

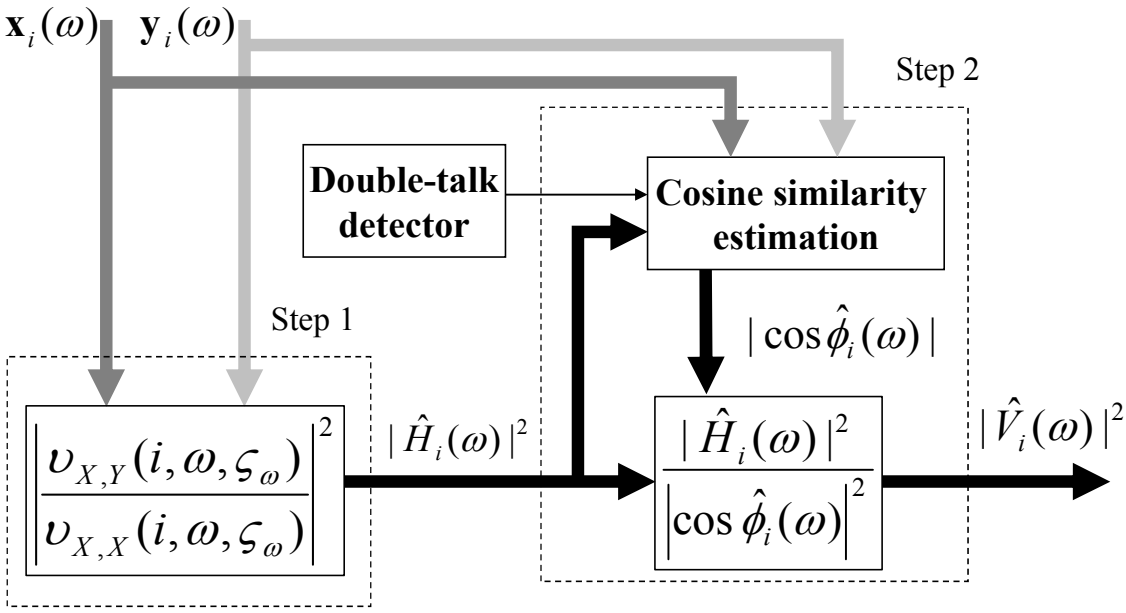


Figure 2.5. Structure of proposed echo-path power spectrum estimation.

A) Echo-path power spectrum estimation focusing on both time and frequency axis directions

The proposed echo-path power spectrum estimation method calculates the ratio of the averaged cross-spectrum between received and microphone input signals to the averaged-received-signal power spectrum in time and frequency spectral domains in order to sufficiently shorten the number of frames L as follows:

$$|\hat{H}_i(\omega)|^2 = \left|\frac{\upsilon_{X,Y}(i,\omega,\varsigma_\omega)}{\upsilon_{X,X}(i,\omega,\varsigma_\omega)}\right|^2, \tag{2.10}$$

where

$$\upsilon_{X,Y}(i,\omega,\varsigma_\omega)=\frac{1}{2\varsigma_\omega+1}\sum_{m=-\varsigma_\omega}^{\varsigma_\omega}\left\langle \mathbf{x}_i(\omega+m),\mathbf{y}_i(\omega+m)\right\rangle, \tag{2.11}$$

$$\upsilon_{X,X}(i,\omega,\varsigma_\omega)=\frac{1}{2\varsigma_\omega+1}\sum_{m=-\varsigma_\omega}^{\varsigma_\omega}\left\|\mathbf{x}_i(\omega+m)\right\|^2, \tag{2.12}$$

and ς_ω is a constant parameter for determining a bandwidth to be averaged in the frequency axis direction.

The proposed inner product, $\upsilon_{X,Y}(i,\omega,\varsigma_\omega)$, and norm, $\upsilon_{X,X}(i,\omega,\varsigma_\omega)$, are calculated in the expanded range including the peripheral frequency around frequency bin ω. This means that the calculation periods has been expanded from L of the conventional FDCC method to $L(2\varsigma_\omega+1)$; in the proposed method, a larger number of $2\varsigma_\omega+1$ expands the calculation period in the frequency direction in addition to the calculation in the time axis direction. From this fact, it can be expected that the number of frames L of the proposed method can be set shorter than that of the conventional method, and the followability to the echo-path changes is improved.

B) Relationship between bandwidth parameter ς_ω and accuracy of echo-path power spectrum estimation

This subsection verifies the accuracy of the echo-path power spectrum estimation when the correlation in the frequency axis direction is used. It is necessary to maintain the invariance of the echo-path power spectrum within the correlation calculation in order to estimate accurately the echo-path power spectrum. If this assumption holds true not only in the time axis direction but also in the frequency axis direction, the echo-path power spectrum can be approximated as follows:

$$
\begin{aligned}
|\hat{H}_i(\omega)|^2 &= \left|\frac{\upsilon_{X,D}(i,\omega,\varsigma_\omega)+\upsilon_{X,S}(i,\omega,\varsigma_\omega)}{\upsilon_{X,X}(i,\omega,\varsigma_\omega)}\right|^2 \\
&\approx \left|\frac{\upsilon_{X,D}(i,\omega,\varsigma_\omega)}{\upsilon_{X,X}(i,\omega,\varsigma_\omega)}\right|^2 \\
&\approx \frac{\sum_{m=-\varsigma_\omega}^{\varsigma_\omega}\left\|\mathbf{d}_i(\omega+m)\right\|^2}{\sum_{m=-\varsigma_\omega}^{\varsigma_\omega}\left\|\mathbf{x}_i(\omega+m)\right\|^2} \\
&\approx \frac{|D_i(\omega)|^2}{|X_i(\omega)|^2}
\end{aligned}
, \qquad (2.13)
$$

These approximation holds when the echo-path power spectrum is a constant value within the correlation calculation including the frequency axis direction. However, the assumption does not hold true because the value of the echo-path power spectrum is different in the frequency axis direction under an actual environment; as a result, an estimation error might occur. Particularly, the magnitude of the echo-path power spectrum of the frequency far away from the center frequency is significantly different from that of the center frequency by comparison with that of the frequency adjacent to the center frequency. Therefore, if the bandwidth parameter ς_ω is set to be too large, there is a possibility that the estimation accuracy is lowered.

Figure 2.6 shows the relationship between the bandwidth parameter ς_ω at 3 kHz and the estimation error of the echo-path power spectrum. The horizontal axis is the bandwidth $2\varsigma_\omega+1$ and displayed by the frequency; and the vertical axis is the RMSE (Root Mean Square Error) between the echo-path power spectrum $|H_i(\omega)|^2$ and its estimate $|\hat{H}_i(\omega)|^2$. This figure demonstrates that the estimation error minimizes at the bandwidth around 1.5 kHz, and the estimation error increases at 1.5 kHz or more.

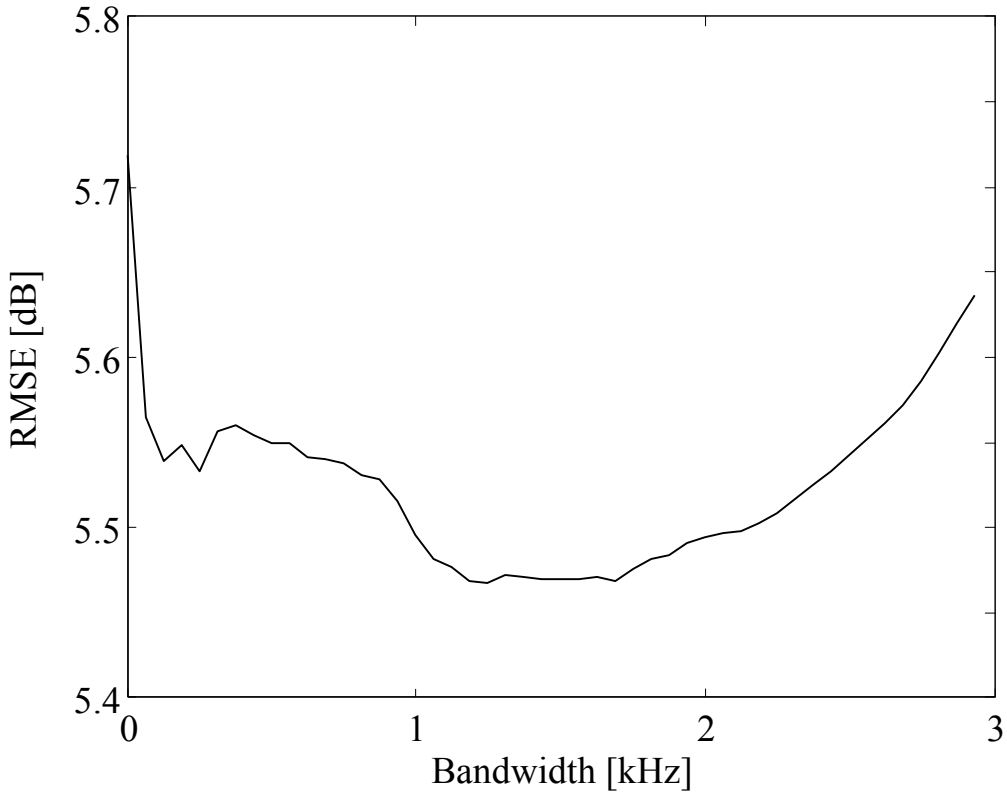


Figure 2.6. Changes of estimation error on echo-path power spectrum at 3 kHz.

2.4.2 Cosine similarity estimation method for improving accuracy of echo-path power spectrum estimation

If the direction of the received-signal time-sequence vector $\mathbf{x}_i(\omega)$ is different from the direction of the echo-signal vector $\mathbf{d}_i(\omega)$, the estimation error corresponding to the cosine similarity $\left|\cos\phi_i(\omega)\right|$ occurs in the estimate $|\hat{H}_i(\omega)|^2$ of the echo-path power spectrum. Therefore, in order to improve the estimation accuracy, this subsection proposes a method to

obtain the estimate $\left|\cos\hat{\phi}_i(\omega)\right|$ of cosine similarity, and corrects the estimate $|\hat{H}_i(\omega)|^2$ as follows:

$$|\hat{V}_i(\omega)|^2 = \frac{|\hat{H}_i(\omega)|^2}{\left|\cos\hat{\phi}_i(\omega)\right|^2}, \tag{2.14}$$

where $|\hat{V}_i(\omega)|^2$ is a corrected estimate of the echo-path power spectrum. The estimate $\left|\cos\hat{\phi}_i(\omega)\right|$ of the cosine similarity is calculated according to the talk situation, which is judged using the double-talk detector, as follows.

[CASE OF NO DOUBLE-TALK SITUATION]

The short-time spectrum of the microphone input signal can be considered as

$$Y_i(\omega) \approx D_i(\omega) \tag{2.15}$$

during the received single-talk periods that only the far-end talker speaks; because the microphone picks up only the received signal. Then, using the estimate in the echo-path power spectrum, the cosine similarity between the received and echo signals can readily be estimated as

$$\begin{aligned} \left|\cos\hat{\phi}_i(\omega)\right| &= |\hat{H}_i(\omega)| \frac{\left\|\mathbf{x}_i(\omega)\right\|}{\left\|\mathbf{y}_i(\omega)\right\|} \\ &\approx |\hat{H}_i(\omega)| \frac{\left\|\mathbf{x}_i(\omega)\right\|}{\left\|\mathbf{d}_i(\omega)\right\|} \end{aligned}. \tag{2.16}$$

On the other hand, when the received signal does not exist, the echo components are not contained in the microphone input signal; therefore, the echo-reduction gain in this period is set to

$$G_i(\omega) = 1 . \tag{2.17}$$

[CASE OF DOUBLE-TALK SITUATION]

The proposed method estimates the cosine similarity between the received and echo signals even during the double-talk periods only at frequency components without near-end speech components because the speech signal has a sparsity in the frequency domain [39]. This method gives an cosine similarity candidate $\rho_i(\omega)$ by

$$\rho_i(\omega) = | \hat{H}_i(\omega) | \frac{\|\mathbf{x}_i(\omega)\|}{\|\mathbf{y}_i(\omega)\|} . \tag{2.18}$$

The above equation readily finds that the candidate $\rho_i(\omega)$ is large when the near-end speech components do not exist in the microphone input vector $\|\mathbf{y}_i(\omega)\|$. Thus, the method calculates the estimate $|\cos\hat{\phi}_i(\omega)|$ of the cosine similarity in order to avoid the update at the frequency components including the near-end speech components as follows:

$$| \cos \hat{\phi}_i(\omega) | = \begin{cases} \rho_i(\omega), & \text{if } \rho_i(\omega) > | \cos \hat{\phi}_{i-1}(\omega) | \\ | \cos \hat{\phi}_{i-1}(\omega) |, & \text{otherwise} \end{cases} . \tag{2.19}$$

The above equation means that the estimate $|\cos\hat{\phi}_i(\omega)|$ of the cosine similarity is replaced by the candidate $\rho_i(\omega)$ of the cosine similarity when the candidate $\rho_i(\omega)$ is larger than the estimate $|\cos\hat{\phi}_{i-1}(\omega)|$ obtained before one frame.

2.5 ***Simulations***

In order to confirm the effectiveness of the proposed echo-path power spectrum estimation method, the proposed and conventional FDCC methods were applied to the ER process, respectively, and the performance comparison was performed. In the computer simulation, two omnidirectional microphones placing on the table were instantaneously switched in order to change the echo path; and the ADF process was omitted for simplifying the configuration, and the performance was verified only by the ER process.

The microphone input signal was calculated by adding the near-end speech signal to the echo signal obtained by convoluting the echo-path impulse response with the received signal. The length of the echo-path impulse response was set to 4096 samples.

2.5.1 Test conditions

The arrangement for a loudspeaker and microphones is shown in Figure 2.2. The echo-path impulse responses, $h(k)$, were measured while switching these microphones at the room of the reverberation time 300 ms. The sampling frequency is 16 kHz and its frequency band is from 100 Hz to 7000 Hz. The length of the echo-path impulse response was set to 4096 samples. The frame shift size is 128 samples and FFT points are 256 samples. The received and near-end speech signals, $x(k)$ and $s(k)$, are shown in Figure 2.7 (a) and (b), respectively. The microphone input signal, $y(k)$, was calculated by adding the near-end speech signal to the echo signal, $d(k)$, obtained by convoluting the echo-path impulse response with the received signal.

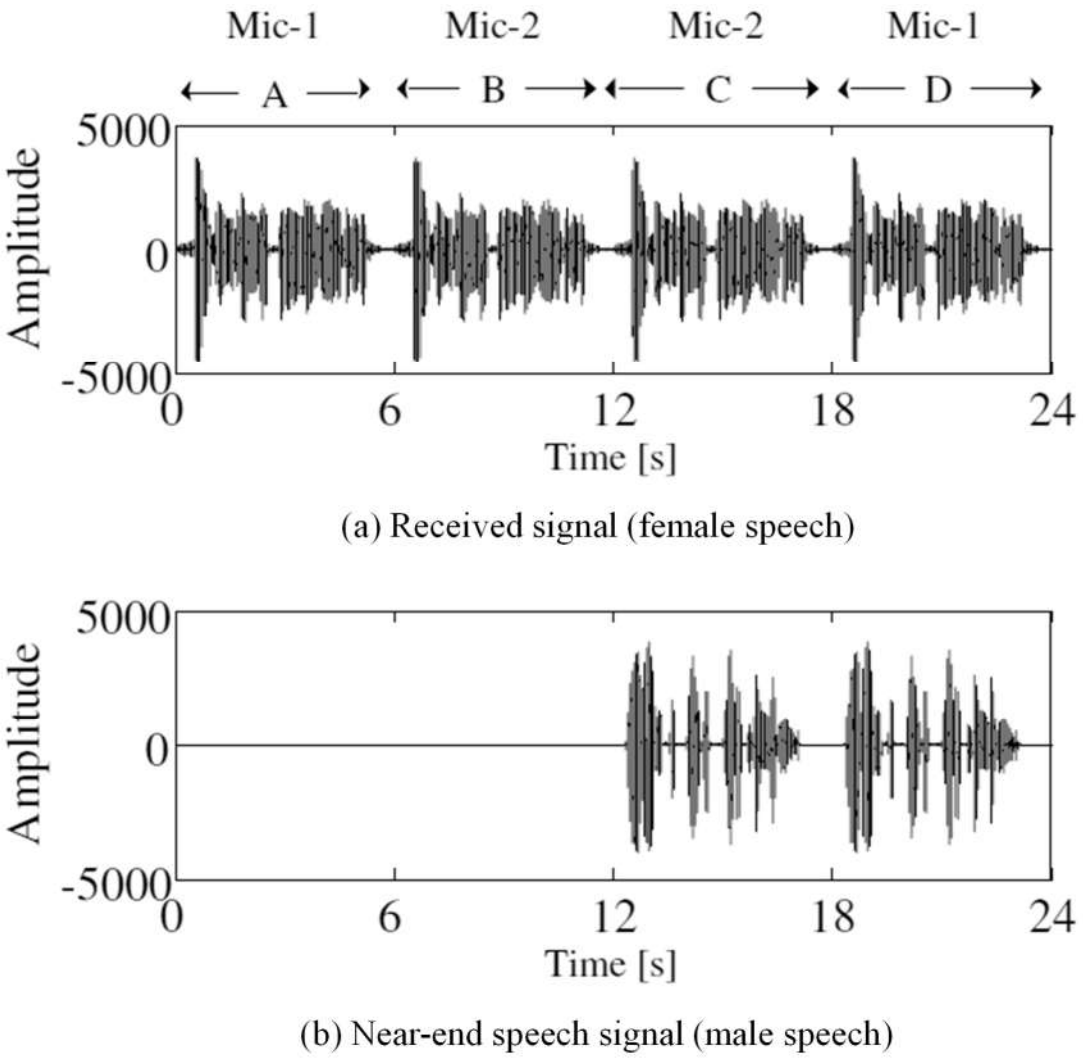


Figure 2.7. Received signal and near-end speech signal.

Table 2.1 shows test conditions for each talk situation. The period A (0 to 6 s) is the state of the received single-talk period (only the received signal is reproduced in this period) at Mic-1. The microphone is switched after the lapse of 6 s; thereafter, the period B (6 to 12 s) becomes the state of the received single-talk period at Mic-2. The period C (12 to 18 s) is the state of the double talk period in the case of Mic-2. The microphone is again switched after the lapse of 18 s; and thereafter, the period D (18 to 24 s) becomes the state of the double

talk period at the time of Mic-1.

Figure 2.8 shows the bandwidth $2\varsigma_\omega + 1$ that minimizes the estimation error of the echo-path power spectrum computed by the proposed method; and the regression line for this result is shown by the solid line. This regression line was obtained by calculating the bandwidth $2\varsigma_\omega + 1$ that minimizes the estimation error by using Equation (2.22) detailed later. In addition, in order to improve appropriateness of the results, the speech signals different from this experiment were used for calculating the bandwidth. Based on this simulation, the bandwidth parameter ς_ω is determined by

$$\varsigma_\omega = \frac{\left\lfloor \dfrac{(1.59\omega + 1009.9)}{\zeta} + 0.5 \right\rfloor - 1}{2}, \tag{2.20}$$

where ζ denotes a frequency resolution; $\zeta = 62.5$. The number of frames L were set to 625 in the conventional method and 62 in the proposed method, respectively; this means that the observation time for the correlation calculation of the proposed method is 1/10 in comparison with that of the conventional method.

In addition, Geigel algorithm [14], which is a simple double talk detection method, was employed in order to obtain the estimate $|\cos\hat{\phi}_i(\omega)|$ of the cosine similarity; the double talk period is detected when the magnitude relationship below is established.

$$\frac{\max\{|y(k)|, \cdots, |y(k-L_1+1)|\}}{\max\{|x(k)|, \cdots, |x(k-L_2+1)|\}} > T, \tag{2.21}$$

where $\max\{\cdot\}$ denotes a maximum selection. T is a threshold; this parameter was set to 0.5. L_1 and L_2 are constant parameters; these were adjusted to lengths matching the processing frame length and the reverberation time, respectively.

Table 2.1. Test condition of each period

Period	Microphone	Talk situation
A	Mic-1	Received single talk
B	Mic-2	Received single talk
C	Mic-2	Double talk
D	Mic-1	Double talk

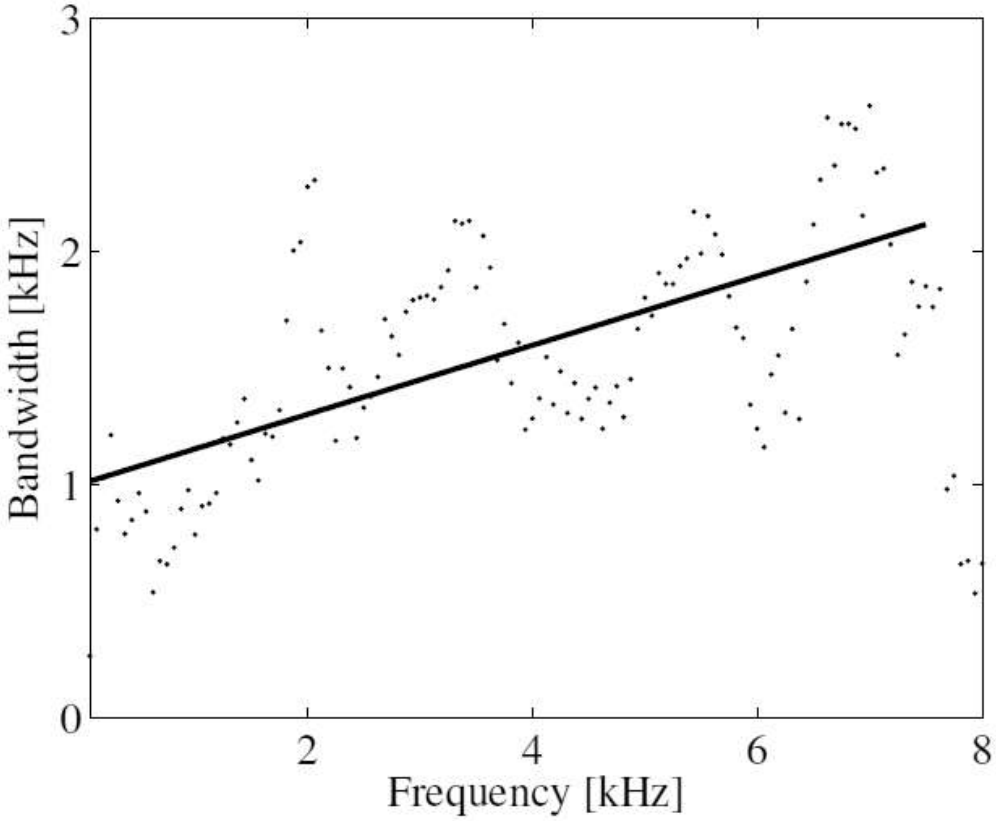


Figure 2.8. Time average (dots) and regression line (solid line) for minimal error bandwidth.

2.5.2 Performance evaluation on echo-path power spectrum estimation

This subsection compares the estimates of the echo-path power spectrum on the conventional FDCC method, the proposed method without the cosine similarity estimation, and the proposed method containing the cosine similarity estimation, respectively. As an evaluation measure, the estimation error $\varepsilon(\omega)$ of the echo-path power spectrum is defined by the following equation.

$$\varepsilon(\omega) = \left| 10 \log_{10} \frac{|\hat{H}_i(\omega)|^2}{|H_i(\omega)|^2} \right|. \tag{2.22}$$

Figure 2.9 shows the time transitions of the frequency-averaged estimation error of the echo-path power spectrum. The vertical axis is the frequency-averaged estimation error, and the horizontal axis is time. The dotted, solid, and bold lines denote the errors of the conventional method, of the proposed method without the cosine similarity estimation, and of the proposed method containing the cosine similarity estimation, respectively. The symbols $|\hat{H}_i(\omega)|^2_{\text{Conv}}$, $|\hat{H}_i(\omega)|^2_{\text{Prop}}$, and $|\hat{V}_i(\omega)|^2$ in the figure stand for the estimates of these methods.

As shown in Figure 2.9, the estimation accuracy of the proposed method using the cosine similarity estimation is superior to that of the conventional FDCC method. However, the estimation error of the conventional method became smaller than that of the proposed method immediately after the lapse of 6 s. This is because the conventional method that does not consider the cosine similarity already made small an estimate of the echo-path power spectrum during the period A.

The estimation errors during the received single-talk periods A and B of the conventional method were about 2 dB and 5 dB, respectively. The estimation error increased during the period B as compared with the period A; because the cosine similarity greatly changed as shown in Figure 2.3. Similarly, the estimate $|\hat{H}_i(\omega)|^2_{\text{Prop}}$ which does not consider the cosine

similarity also has the same problem. On the other hand, in the estimate $|\hat{V}_i(\omega)|^2$ employing the cosine similarity estimation, its error was less than 1 dB in the periods A and B. This result suggests that it is difficult to avoid the influence of the cosine similarity even if the conventional method compensates for the estimate of the echo-path power spectrum by multiplying the estimate with a constant parameter; it indicates the superiority of the proposed method considering the cosine similarity.

During the double-talk periods C and D, the error of the estimate $|\hat{H}_i(\omega)|^2_{\text{Prop}}$ is constantly smaller than that of the estimate $|\hat{H}_i(\omega)|^2_{\text{Conv}}$. This result indicates that in the estimate $|\hat{H}_i(\omega)|^2_{\text{Prop}}$, the accuracy degradation due to the influence of the near-end speech was sufficiently reduced by the effect of the correlation calculation using the frequency axis direction although the length of L is 1/10 of the conventional method. In addition, this figure shows that the estimate $|\hat{V}_i(\omega)|^2$ was the most accurate in the all methods; this result indicates that the cosine similarity was estimated without serious failure even during the double-talk periods.

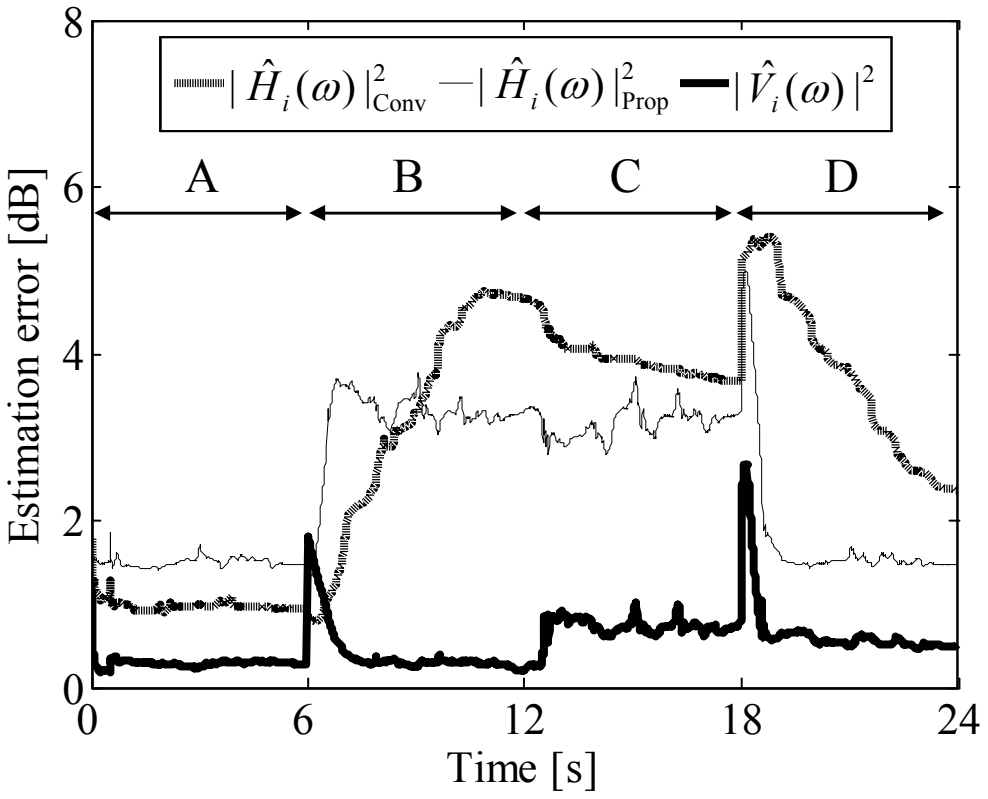


Figure 2.9. Time transitions in frequency-averaged estimation error of echo-path power spectrum.

2.5.3 Performance evaluation on echo reduction

This subsection evaluates the echo-reduction performance of the conventional FDCC and the proposed methods, using the ER process. The proposed method employs the cosine similarity estimation. The microphone input signal is shown in Figure 2.10 (a). The send signals after processing by conventional and proposed methods are shown in Figures 2.10 (b) and (c), respectively. As seen in these figures, the proposed and conventional methods sufficiently suppressed echo signals over the entire period. However, the conventional method suffers from near-end speech distortion during double-talk periods though the number of frames of

the conventional method is ten times as large as the number of frames of the proposed method. The near-end speech distortion was improved by using the proposed method, and the subjective quality was good.

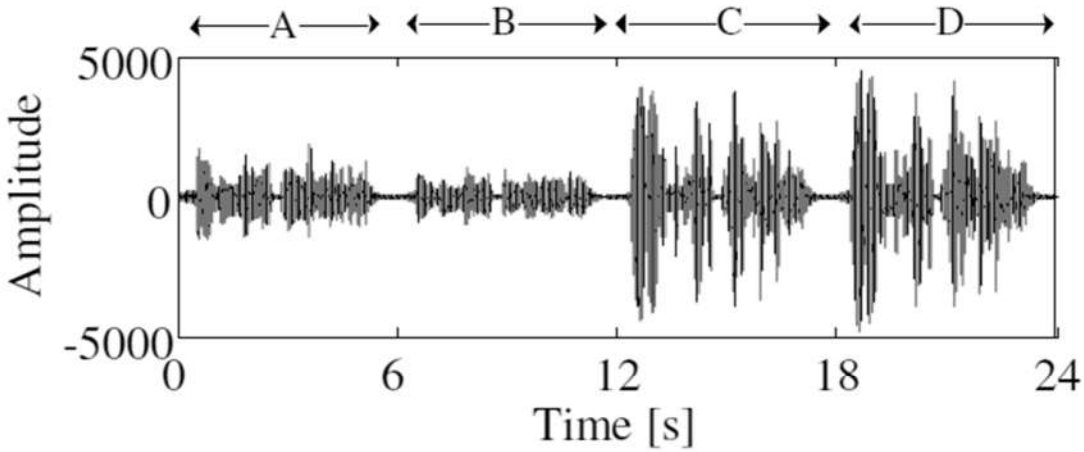
This subsection evaluates the echo-suppression level during the received single-talk periods A and B, and the distortion level of the send signal during the double-talk periods C and D, quantitatively. An echo return loss enhancement (ERLE) [40] was used in order to measure the echo-suppression level; the ERLE is defined as

$$\mathrm{ERLE} = 10\log_{10}\frac{p_{\mathrm{in}}}{p_{\mathrm{out}}}, \tag{2.23}$$

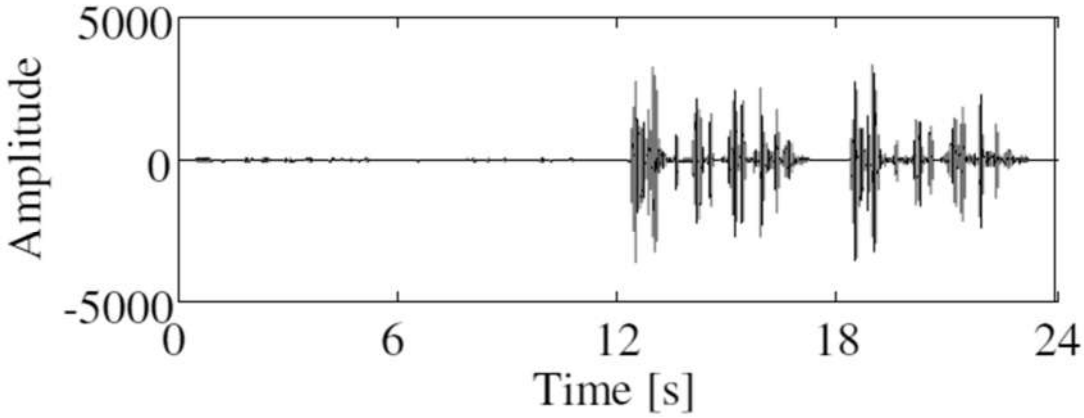
where p_{in} and p_{out} denote mean squares of the microphone input and send signals during the received single-talk periods A and B. The ERLEs are 36.2 dB in the conventional method and 37.0 dB in the proposed method, respectively. A signal-to-distortion ratio (SDR) [39] was used as the evaluation method of the speech quality after the nonlinear post-filtering; the SDR is defined as

$$\mathrm{SDR} = 10\log_{10}\frac{\sum_{\omega}|S_i(\omega)|^2}{\sum_{\omega}\max\left\{|S_i(\omega)|^2 - |\hat{S}_i(\omega)|^2, 0\right\}}. \tag{2.24}$$

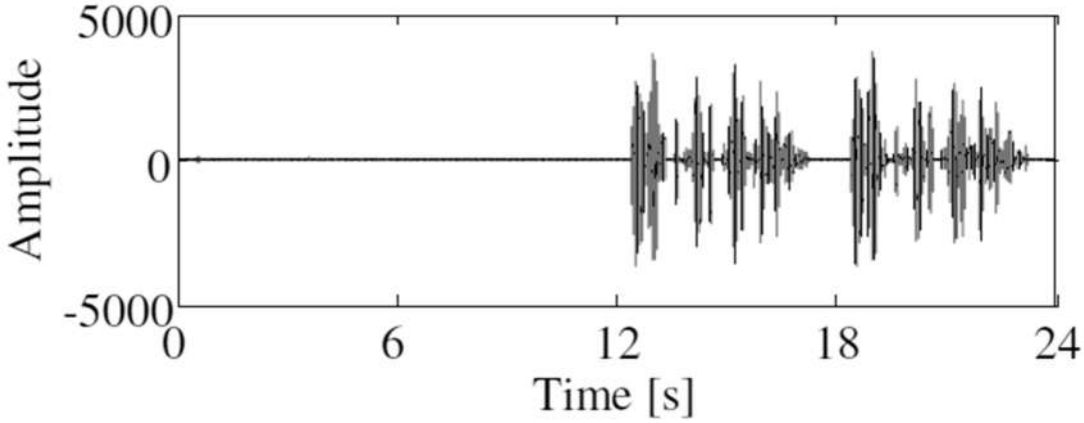
The SDRs are 9.4 dB in the conventional method and 18.7 dB in the proposed method, respectively. These results demonstrated that the proposed method can improve the speech distortion of the send signal about 10 dB compared with the conventional method while achieving the almost same echo-suppression level as the conventional method.



(a) Microphone input signal



(b) Send signal processed by conventional method



(c) Send signal processed by proposed method

Figure 2.10. Microphone input signal, send signal processed by conventional method, and send signal processed by proposed method.

2.6 *Conclusion*

An echo-path power spectrum estimation method for the ER process was proposed. The proposed method is focused on time and frequency spectral domains for estimating the echo-path power spectrum; and therefore, that method rapidly tracked echo-path changes. In addition, the proposed method improved the estimation accuracy by considering the cosine similarity between received and echo signals based on the talk situation. According to experimental results, this chapter confirmed that the proposed echo-path power spectrum estimation method achieved better echo reduction performance than that of the conventional method.

Chapter 3.

Wiener Solution Considering Cross-Spectral Term Between Echo and Near-End Speech

3.1 *Introduction*

This chapter introduces a frequency-domain acoustic echo reduction process based on a new WF method taking into account the cross-spectral term between the acoustic echo and the near-end speech. The conventional ER method based on Wiener filtering estimates the gain based on the assumption that the cross-spectral term of the echo and the near-end speech is zero because the acoustic echo and the near-end speech are statistically uncorrelated. However, this assumption does not always hold true in practice because the gain is estimated in a very short period where the amount of statistical data, which is used to calculate the ensemble averages of the observed signals, is insufficient. As a result, the conventional method occasionally causes the perceptual degradation of sound quality during a double-talk period; therefore, the performance is still not sufficient. Our goal was to accurately calculate the echo-reduction gain to decrease the speech distortions produced by the echo-reduction process. The proposed method solves a least mean square error of WF method by taking into account the cross-spectral term between the echo and the near-end speech to obtain a better echo-reduction gain.

3.2 *Wiener Filtering and Its Problem*

3.2.1 Echo-reduction gain estimation based on WF method

The echo-reduction gain based on the WF method can be approximately obtained by the following equation:

$$\begin{aligned} G_i(\omega) &= \frac{E[|S_i(\omega)|^2]}{E[|S_i(\omega)|^2]+E[|D_i(\omega)|^2]} \\ &\approx \frac{E[|Y_i(\omega)|^2]-E[|D_i(\omega)|^2]}{E[|Y_i(\omega)|^2]} \\ &\approx \frac{|Y_i(\omega)|^2-|\hat{D}_i(\omega)|^2}{|Y_i(\omega)|^2} \end{aligned}. \tag{3.1}$$

where $E[\cdot]$ denotes the expectation operation calculated over the signal frames. The WF method estimates the echo-reduction gain $G_i(\omega)$ by minimizing a squared error υ, which is given as follows:

$$\upsilon = \left\| \mathbf{S}_i(\omega) - G_i(\omega)\mathbf{Y}_i(\omega) \right\|^2. \tag{3.2}$$

where boldface denotes a time-sequence vector of a short-time spectral amplitude: $\mathbf{P}_i(\omega) = \left[|P_i(\omega)|, \cdots, |P_{i-L+1}(\omega)| \right]^T$. By solving the differential equation

$$\frac{\partial \upsilon}{\partial G_i(\omega)} = 2\left\{ G_i(\omega)\left\|\mathbf{Y}_i(\omega)\right\|^2 - \left\langle \mathbf{S}_i(\omega), \mathbf{Y}_i(\omega) \right\rangle \right\} \to 0, \tag{3.3}$$

the echo-reduction gain is then obtained as follows:

$$
\begin{aligned}
G_i(\omega) &= \frac{\langle \mathbf{S}_i(\omega), \mathbf{Y}_i(\omega) \rangle}{\|\mathbf{Y}_i(\omega)\|^2} \\
&= \frac{\langle \mathbf{Y}_i(\omega) - \mathbf{D}_i(\omega), \mathbf{Y}_i(\omega) \rangle + \delta_1}{\|\mathbf{Y}_i(\omega)\|^2} \\
&= \frac{\|\mathbf{Y}_i(\omega)\|^2 - \langle \mathbf{D}_i(\omega), \mathbf{D}_i(\omega) + \mathbf{S}_i(\omega) \rangle + \delta_1 + \delta_2}{\|\mathbf{Y}_i(\omega)\|^2} \\
&= \frac{\|\mathbf{Y}_i(\omega)\|^2 - \|\mathbf{D}_i(\omega)\|^2 - \langle \mathbf{D}_i(\omega), \mathbf{S}_i(\omega) \rangle + \delta_1 + \delta_2}{\|\mathbf{Y}_i(\omega)\|^2} \\
&= \frac{\|\mathbf{Y}_i(\omega)\|^2 - \|\mathbf{D}_i(\omega)\|^2}{\|\mathbf{Y}_i(\omega)\|^2} + \delta_W
\end{aligned}
, \tag{3.4}
$$

where

$$
\delta_W = \frac{\delta_1 + \delta_2 - \langle \mathbf{D}_i(\omega), \mathbf{S}_i(\omega) \rangle}{\|\mathbf{Y}_i(\omega)\|^2}, \tag{3.5}
$$

δ_1 and δ_2 are the errors caused by neglecting the phase components, respectively. If time-sequence vectors $\mathbf{D}_i(\omega)$ and $\mathbf{S}_i(\omega)$ are uncorrelated and $\delta_1 = \delta_2 = 0$, the approximation error of the WF method, δ_W, is regarded as zero; and $G_i(\omega)$ is given by\

$$
G_i(\omega) \approx \frac{\|\mathbf{Y}_i(\omega)\|^2 - \|\mathbf{D}_i(\omega)\|^2}{\|\mathbf{Y}_i(\omega)\|^2}, \tag{3.6}
$$

The WF-based gain is finally obtained by approximating the above equation for a very short period as follows:

$$
G_i(\omega) \approx \frac{|\, Y_i(\omega)\,|^2 - |\, D_i(\omega)\,|^2}{|\, Y_i(\omega)\,|^2}. \tag{3.7}
$$

3.2.2 Problem of Wiener filtering

It is assumed that the echo and near-end vectors are uncorrelated in the WF method; that means $\langle \mathbf{D}_i(\omega), \mathbf{S}_i(\omega) \rangle = 0$. This assumption holds only if a very long period of data is available because the statistical properties of data are used. However, the echo-reduction gain $G(\omega)$ needs to be calculated from a very short period in practice because most echo and near-end speech signals are usually nonstationary. In that case, the number of samples used in the cross-correlation calculation becomes insufficient, and so the inner product $\langle \mathbf{D}_i(\omega), \mathbf{S}_i(\omega) \rangle$ is not always zero. Therefore, the echo-reduction gain $G(\omega)$ is sometimes estimated to be smaller than the actual gain because the inner product between the echo and near-end vectors is ignored in the WF method. As a result, the ER process based on the WF method suffers from the speech distortions and quite often causes the perceptual degradation of the sound quality.

3.3 ***Proposed Echo-Reduction Gain Estimation Method***

3.3.1 Strategy for high-quality echo-reduction gain estimation

This subsection explains the concept of the proposed gain estimation method used in the ER process. As described above, the inner product between the echo and near-end vectors are not always zero in practice and so its inner product cannot be ignored with respect to the echo-reduction gain calculation. The proposed method derives a better ER gain $G(\omega)$ reconsidering the derivation of the conventional WF method as follows:

$$
\begin{aligned}
G_i(\omega) &= \frac{\langle \mathbf{S}_i(\omega), \mathbf{Y}_i(\omega) \rangle}{\|\mathbf{Y}_i(\omega)\|^2} \\
&= \frac{\langle \mathbf{Y}_i(\omega) - \mathbf{D}_i(\omega), \mathbf{Y}_i(\omega) \rangle + \delta_1}{\|\mathbf{Y}_i(\omega)\|^2} \\
&= \frac{\|\mathbf{Y}_i(\omega)\|^2 - \langle \mathbf{D}_i(\omega), \mathbf{Y}_i(\omega) \rangle + \delta_1}{\|\mathbf{Y}_i(\omega)\|^2} \\
&= \frac{\|\mathbf{Y}_i(\omega)\|^2 - \dfrac{\langle \mathbf{D}_i(\omega), \mathbf{Y}_i(\omega) \rangle}{\|\mathbf{D}_i(\omega)\|^2} \|\mathbf{D}_i(\omega)\|^2 + \delta_1}{\|\mathbf{Y}_i(\omega)\|^2}, \\
&= \frac{\|\mathbf{Y}_i(\omega)\|^2 - \gamma_i(\omega)\|\mathbf{D}_i(\omega)\|^2 + \delta_1}{\|\mathbf{Y}_i(\omega)\|^2} \\
&= \frac{\|\mathbf{Y}_i(\omega)\|^2 - \gamma_i(\omega)\|\mathbf{D}_i(\omega)\|^2}{\|\mathbf{Y}_i(\omega)\|^2} + \delta_P
\end{aligned}
\tag{3.8}
$$

where

$$
\gamma_i(\omega) = \frac{\langle \mathbf{D}_i(\omega), \mathbf{Y}_i(\omega) \rangle}{\|\mathbf{D}_i(\omega)\|^2} \tag{3.9}
$$

and

$$
\delta_P = \frac{\delta_1}{\|\mathbf{Y}_i(\omega)\|^2}. \tag{3.10}
$$

The parameter $\gamma_i(\omega)$ denotes the ratio of the inner product between echo and input vectors divided by the square norm of the echo vector. The parameter δ_P is the approximation error of the proposed method.

In the conventional WF method, the parameters $\gamma_i(\omega)$ and δ_P are set to constant value, $\gamma_i(\omega) = 1$, and the zero parameter, $\delta_P = 0$, respectively; namely, the echo-reduction gain obtained from these approximations becomes equivalent to the conventional echo-reduction

gain assuming $\delta_W = 0$. However, the assumption $\gamma_i(\omega) = 1$ cannot hold true during double-talk situations in practice due to $\langle \mathbf{D}_i(\omega), \mathbf{Y}_i(\omega) \rangle \neq \|\mathbf{D}_i(\omega)\|^2$ in a very short period.

3.3.2 Comparison of approximation accuracy

The approximation errors of the conventional and proposed methods δ_W and δ_P during the double-talk periods are shown in Figure 3.1. The approximation error is defined as the difference between the target and estimated echo-reduction gains, and this is the value that occurs by neglecting the unknown terms included in the estimation of echo-reduction gains. The vertical and horizontal axes are the approximation errors and the number of frame L, respectively. The simulation conditions are listed in Table 3.1. These approximation errors are calculated from average for all frames of two male signals and two female signals. As seen in Figure 3.1, with both cases the approximation errors decreases in proportion as increasing the amount of statistics which is determined by L. However, with the conventional method, the error δ_W significantly increases when L is small which is an amount of error that should not be ignored. On the other hand, in the proposed method, the error is smaller even if L is small. This shows that the proposed method works effectively for calculating the gain in a short period accurately.

This subsection also evaluated the time transitions of approximation errors δ_W and δ_P of when $L = 2$ and $L = 4$ during the double-talk periods, which are plotted in Figures 3.2 and 3.3, respectively. The proposed method showed the smaller approximation error over the entire period than that of the conventional method.

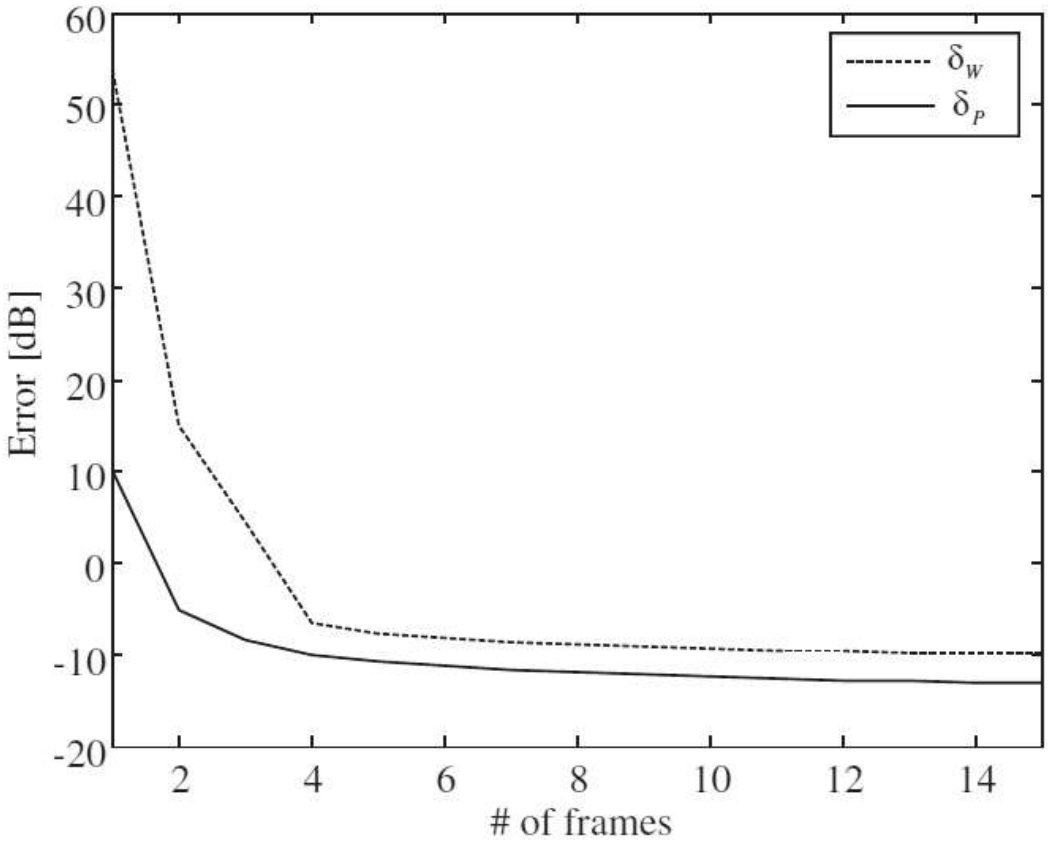


Figure 3.1. Comparison in approximation errors of conventional and proposed methods during double-talk periods.

Table 3.1. Simulation conditions

Sampling rate	16 kHz
Frame length	256 samples
Frame shift	128 samples
FFT points	256 samples
Sound item	Clean speech
Single length	4 s
# of signals	4 (2 males and 2 females)
Reverberation time	200 ms

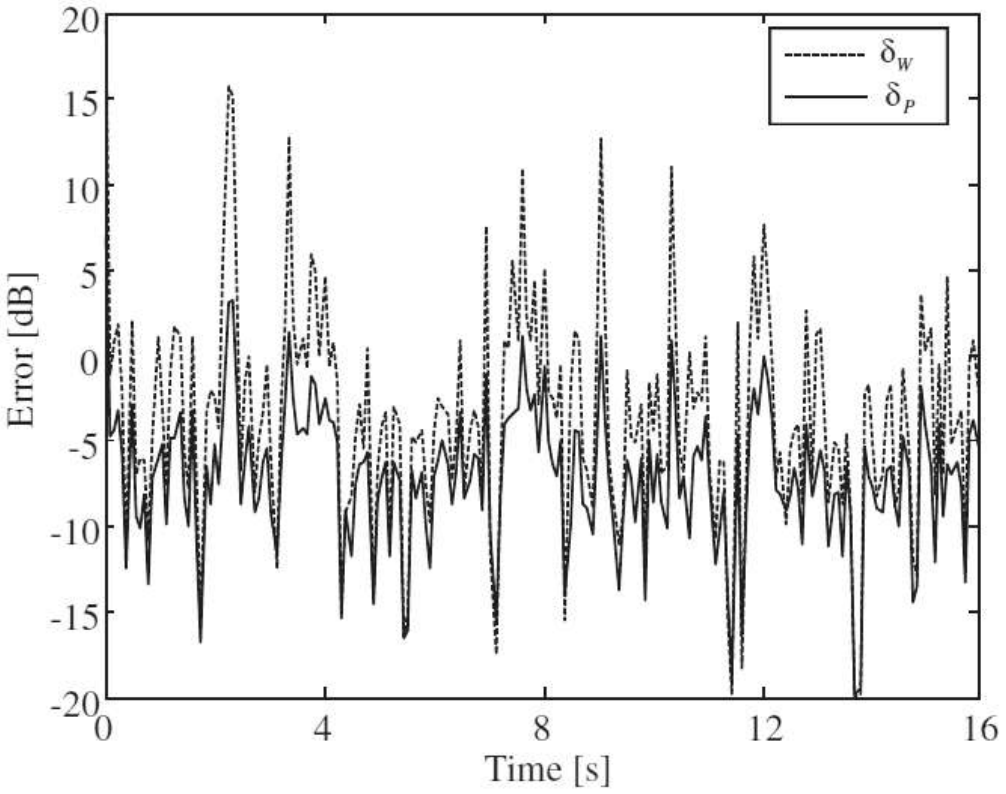


Figure 3.2. Time transition of approximation errors of conventional and proposed methods during double-talk situation ($L = 2$).

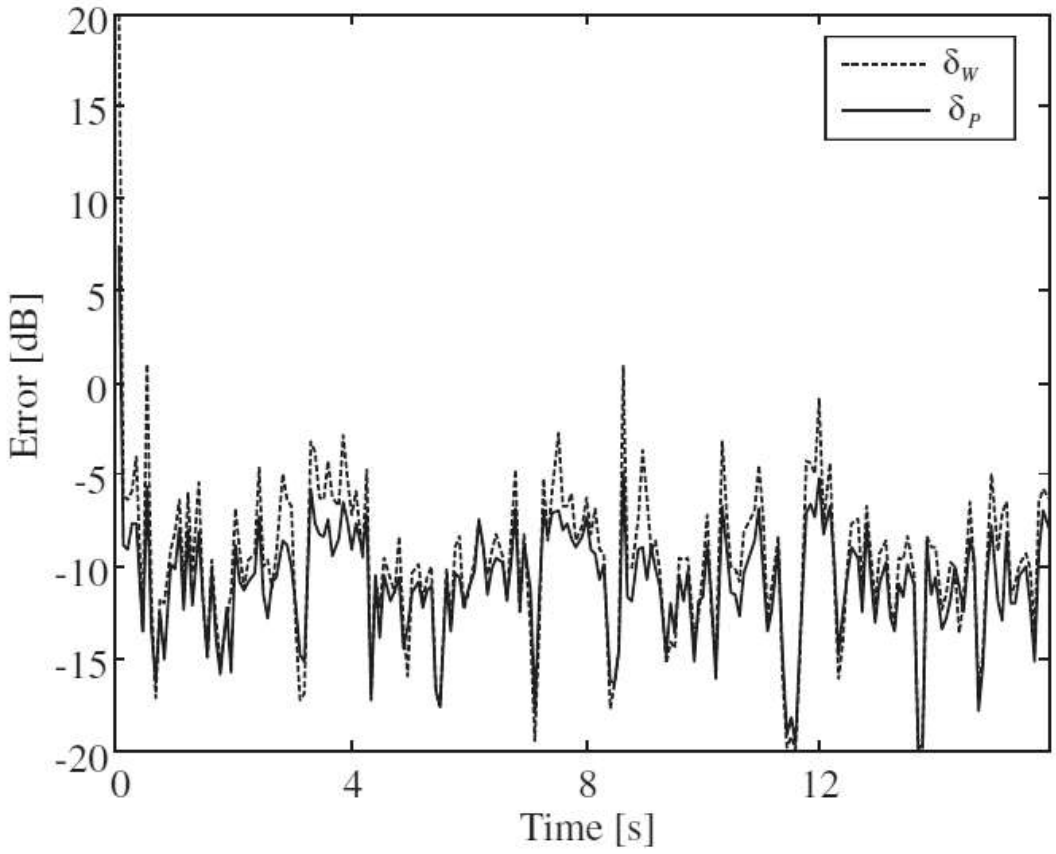


Figure 3.3. Time transition of approximation errors of conventional and proposed methods during double-talk situation ($L = 4$).

3.3.3 Calculation method of echo-reduction gain

Accurately estimating the correlation parameter $\gamma_i(\omega)$ is a key in solving the problem in which speech distortions are caused during the double-talk periods. To calculate the parameter $\gamma_i(\omega)$ in practice, the proposed method substitutes the unknown vector $\mathbf{D}_i(\omega)$ with $|\hat{D}_i(\omega)|^2 = |\hat{H}_i(\omega)|^2 |X_i(\omega)|^2$, i.e.,

$$\hat{\gamma}_i(\omega) = \frac{\left\langle \hat{\mathbf{D}}_i(\omega), \mathbf{Y}_i(\omega) \right\rangle}{\left\| \hat{\mathbf{D}}_i(\omega) \right\|^2}, \tag{3.11}$$

where

$$\hat{\mathbf{D}}_i(\omega) = \left[|\, \hat{D}_i(\omega)\,|, \cdots, |\, \hat{D}_{i-L+1}(\omega)\,| \right]^T. \tag{3.12}$$

The echo-reduction gain $G_i(\omega)$ is therefore represented using the estimated correlation parameter $\hat{\gamma}_i(\omega)$ as

$$G_i(\omega) = \frac{\left\| \mathbf{Y}_i(\omega) \right\|^2 - \hat{\gamma}_i(\omega) \left\| \hat{\mathbf{D}}_i(\omega) \right\|^2}{\left\| \mathbf{Y}_i(\omega) \right\|^2}. \tag{3.13}$$

The estimated correlation parameter $\hat{\gamma}_i(\omega)$ varies corresponding to the rate of the near-end speech component included in the microphone input signal. Equation (3.11) therefore takes a value closer to one during the single-talk periods whereas it takes a value larger than one during double-talk periods.

The echo-reduction gain is finally represented by using the estimated correlation parameter $\hat{\gamma}_i(\omega)$ and by replacing the norms $\left\| \hat{\mathbf{D}}_i(\omega) \right\|^2$ and $\left\| \mathbf{Y}_i(\omega) \right\|^2$ into instantaneous values $|\, \hat{D}_i(\omega)\,|^2$ and $|\, Y_i(\omega)\,|^2$ as follows:

$$G_i(\omega) = \frac{|\, Y_i(\omega)\,|^2 - \hat{\gamma}_i(\omega)\,|\, \hat{D}_i(\omega)\,|^2}{|\, Y_i(\omega)\,|^2}. \tag{3.14}$$

3.4 *Evaluation*

The performance of the new method was evaluated using both simulation and subjective listening tests. The conventional and proposed echo-reduction gain estimation methods were used to calculate the gain using Equations 3.14 and 3.7, respectively. Table 3.2 lists the experimental conditions. Figure 3.4 shows the frequency characteristics of impulse response used in the computer simulation. The numbers of frame L used to calculate the parameter $\hat{\gamma}_i(\omega)$ are set to four.

3.4.1 Simulation experiments

The subsection conducted simulations to compare the proposed echo-reduction gain estimation method to the conventional WF method. The received signal $x(k)$, the near-end signal $s(k)$, and the microphone input signal $y(k)$ are shown in Figures 3.5, 3.6, and 3.7, respectively. Period A is the received single talk; it means that only the far-end speaker is talking. Period B is the send single talk period; it means that only the near-end speaker is talking. The period C is the double-talk period; this period occurs when both the near-end and far-end speakers are talking concurrently.

Figures 3.8 and 3.9 plot the send signals after processing by the conventional and proposed echo-reduction gain estimation methods, respectively. The power envelopes of send signals of conventional and proposed methods in period C are plotted in Figure 3.10. As seen in Figures 3.7, 3.8, and 3.9, the conventional and proposed methods sufficiently suppressed echo signals over the entire period. The ERLEs that show the echo-suppression levels of the conventional and proposed methods during the period A were 33.79 dB with the conventional method and 33.72 dB with the proposed method, respectively.

Table 3.2. Experimental conditions

Sampling rate	16 kHz
Frame length	256 samples
Frame shift	128 samples
FFT points	256 samples
Sound item	Clean speech
Single length	12 s
Reverberation time	200 ms

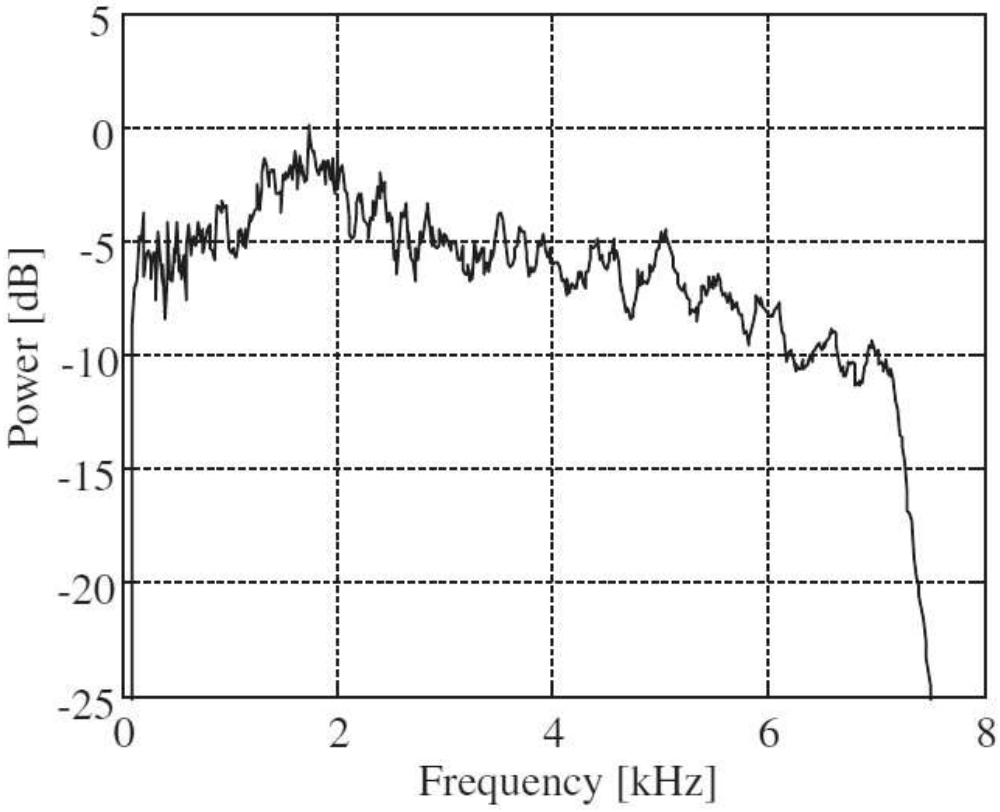


Figure 3.4. Frequency characteristics of impulse response used in computer simulation.

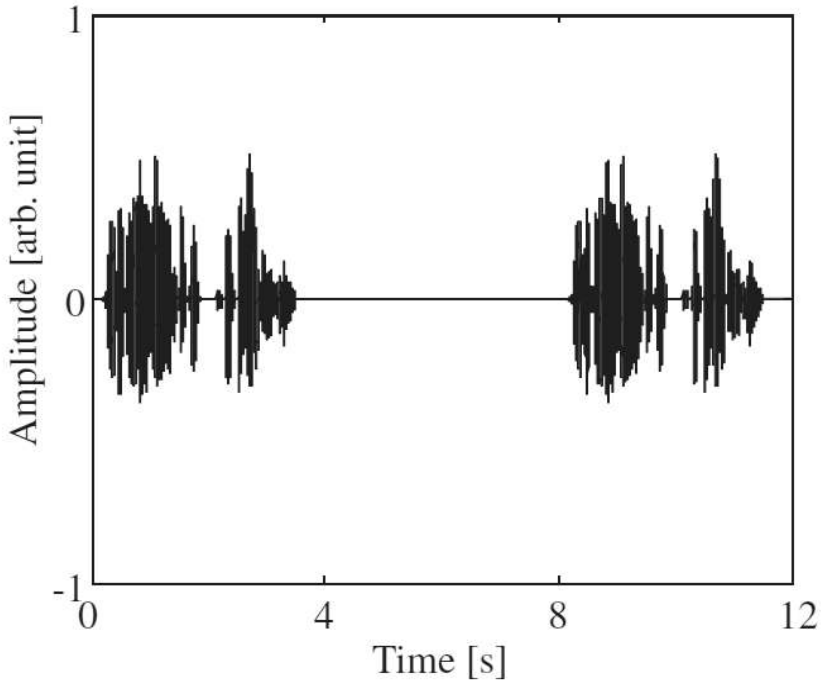


Figure 3.5. Received speech signal $x(k)$ (female speech).

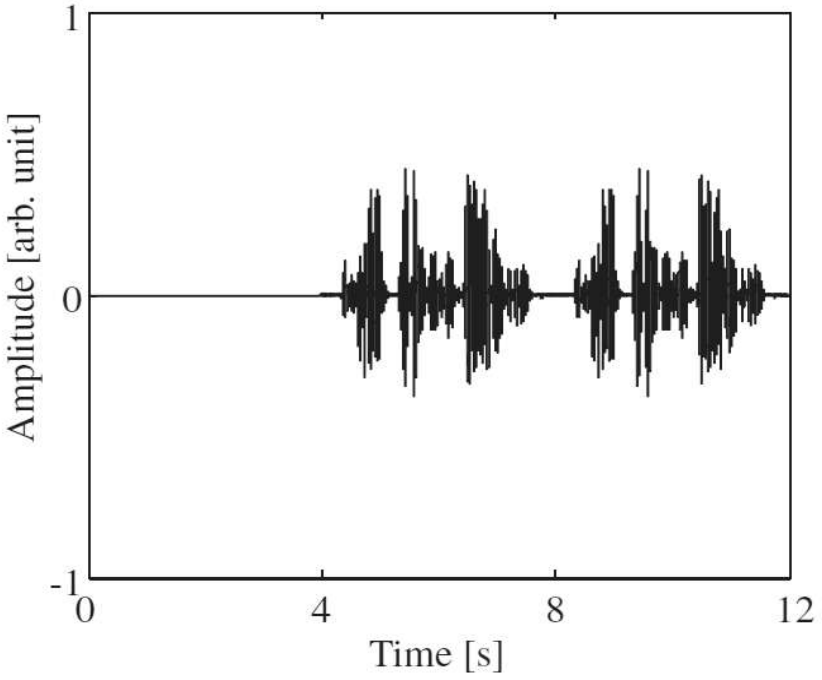


Figure 3.6. Near-end speech signal $s(k)$ (male speech).

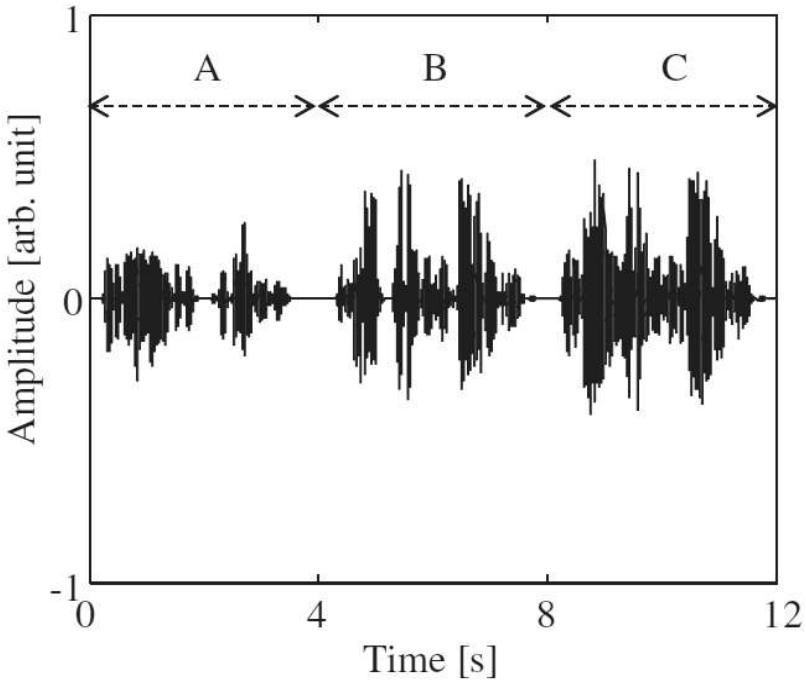


Figure 3.7. Microphone input signal $y(k)$.

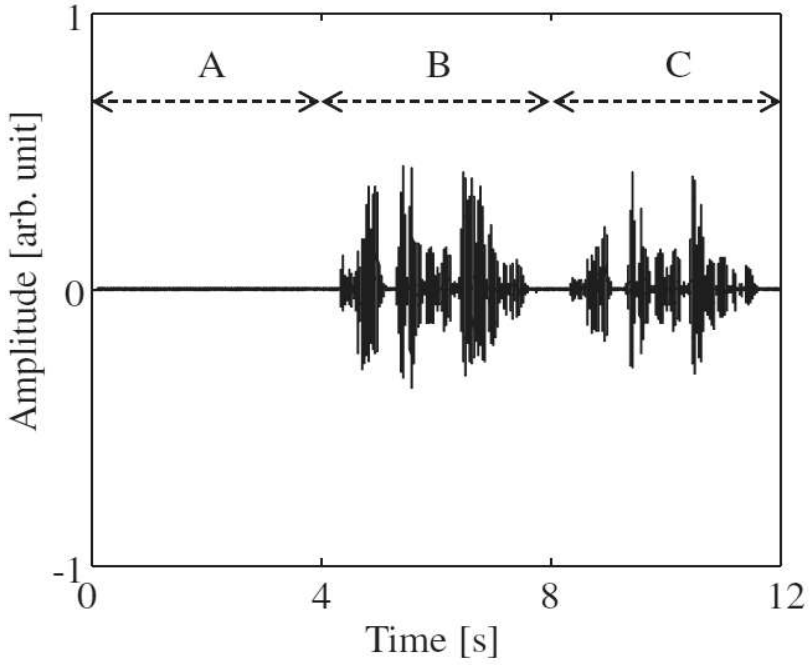


Figure 3.8. Send signal $\hat{s}(k)$ with conventional method.

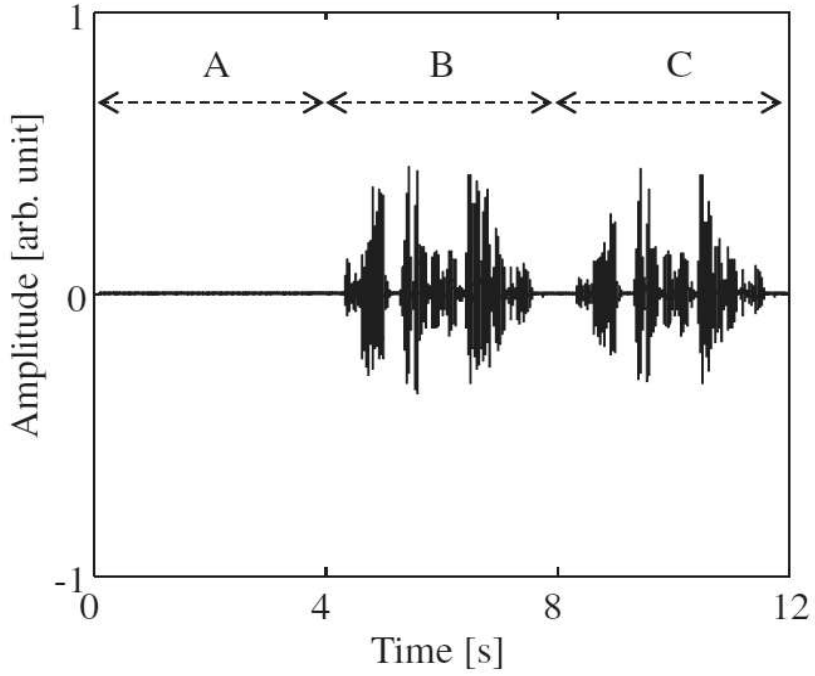


Figure 3.9. Send signal $\hat{s}(k)$ with proposed method.

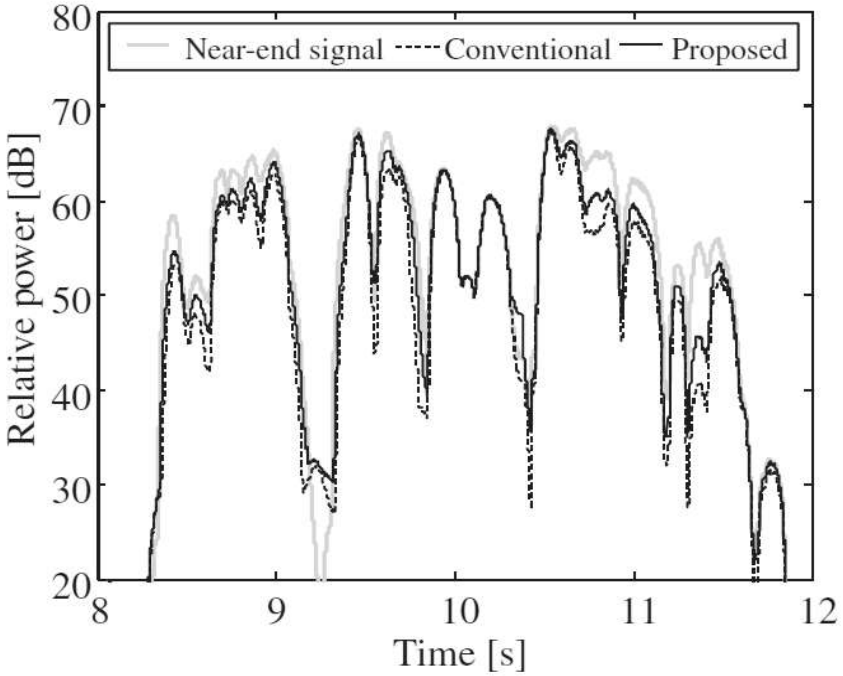


Figure 3.10. Each power envelope in each signal during double-talk period C.

However, as seen in Figure 3.10, the conventional echo-reduction gain estimation method seems to result in speech distortions during the double-talk period C compared with the proposed method. The amount of speech distortions during the double-talk period C were evaluated using a linear predictive coding (LPC) cepstral distance [41], which is computed as follows:

$$\mathrm{CD}(k) = \frac{10}{\log 10}\sqrt{2\sum_{m=1}^{16}\left[c(m,k) - \hat{c}(m,k)\right]^2}\,, \tag{3.15}$$

where $c(m,k)$ and $\hat{c}(m,k)$ are the k -th cepstral coefficients of near-end speech and send signals, respectively, and $\mathrm{CD}(k)$ is the LPC cepstral distance. The results from comparing the conventional and proposed methods using the LPC cepstral distance are shown in Figure 3.11. As these results indicate, the better scores in the LPC cepstral distance were observed than the conventional method over the entire period and the significant improvement was confirmed.

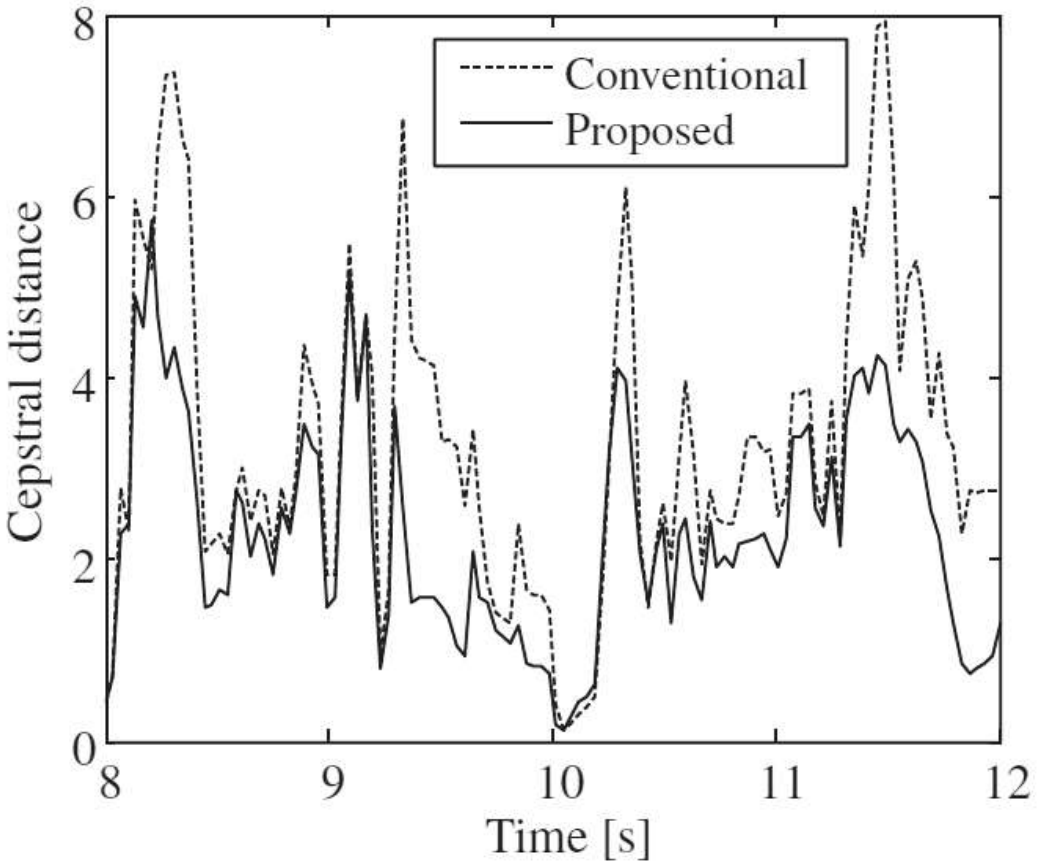


Figure 3.11. Comparison of LPC cepstrum distances during double-talk period C.

3.4.2 Subjective experiments

A multi-stimulus test with hidden reference and anchor (MUSHRA) using a 100-point scale, compliant with ITU-R BS.1534-1 [42], was used to test the quality of speech. All reference and evaluation signals were played to both ears with headphones (Sennheiser HD 280 Pro). Eight experienced listeners evaluated speech under six conditions: the near-end speech signal (c00: hidden reference), the near-end speech signal filtered by 3.5 kHz low-pass filter (c01: anchor A), the microphone input signal (c02: anchor B), the target signal of ER (c03: anchor C), and the send signals of conventional and proposed methods (c04: conventional method and c05: proposed method). The target signal $\hat{\mathrm{s}}_{\mathrm{T}}(k)$ denotes the limiting value of the ER process, which is simulated using the following equation:

$$\hat{\mathrm{s}}_{T}(k) = \mathrm{IFFT}\left[\,|\,\mathrm{S}_{\mathrm{i}}(\omega)\,|\,e^{j\theta_{Y}}\,\right], \tag{3.16}$$

where $\mathrm{IFFT}[\cdot]$ and θ_{Y} denotes the IFFT operation and the phase component of $Y_{i}(\omega)$, respectively.

The MUSHRA test results comparing the conventional and proposed methods during the double-talk situation of the period C are shown in Figure 3.12. The vertical lines in the figure denote a 95% confidence interval. For the double-talk period C, mean scores were awarded for four sound signals by eight listeners. As these results indicate, a better score was observed in the double-talk period by using the proposed method that calculates the echo-reduction gain considering the cross-spectral term between echo and near-end speech signals. The proposed method improved the sound quality by about six points on a 100-point scale compared with the conventional method, and a significant improvement was confirmed.

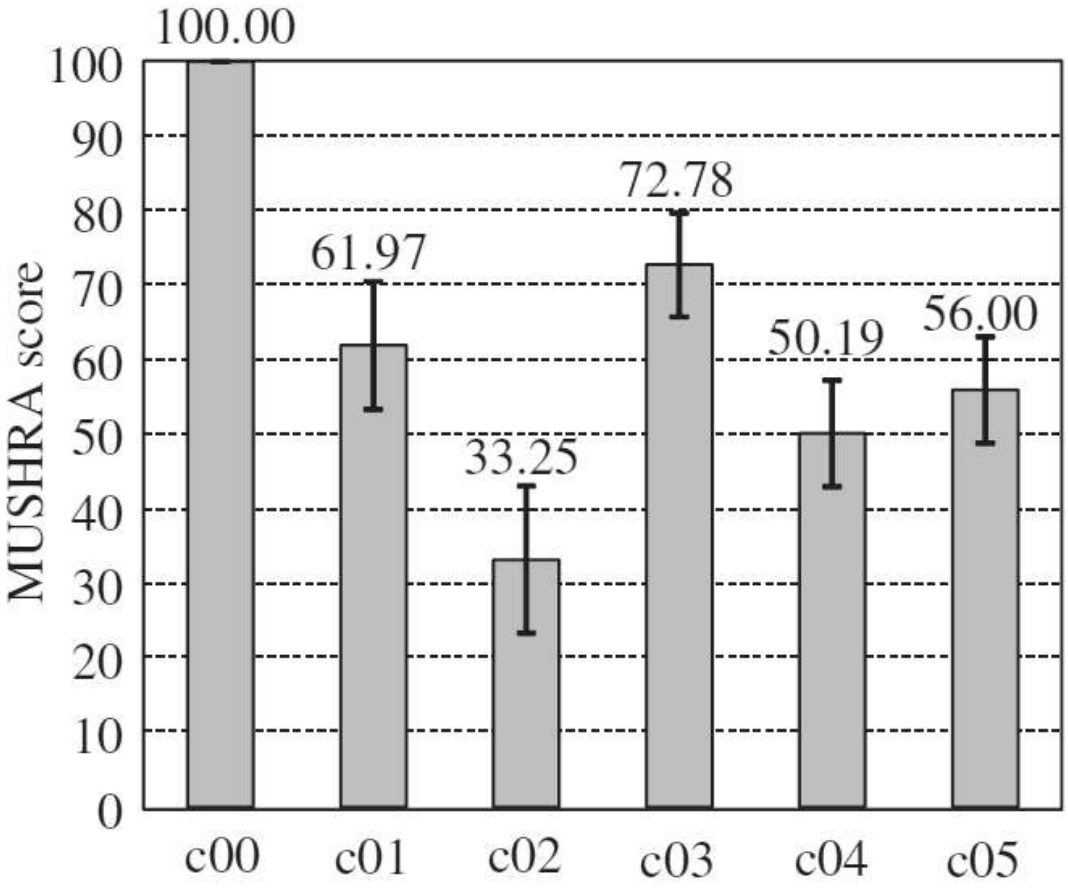


Figure 3.12. Double-talk quality assessments for period C.

3.5 *Conclusion*

This chapter proposed a novel modified gain-estimation method for the ER process. To reduce the speech distortion produced by the ER process, the proposed echo-reduction gain was calculated based on the assumption that the echo and near-end signals is uncorrelated but the cross-spectral term of their signals obtained in the short-time period is not zero. The experimental results showed that the proposed echo-reduction gain estimation method performed better than the conventional WF method by using the ER process, and significant improvement was confirmed.

Chapter 4.

Convolutive Echo Power Estimation for Addressing Long Reverberation-Time Problem

4.1 *Introduction*

This chapter deals with a problem of estimating the echo power spectrum that results from reverberant component beyond a length of FFT block. The ER process suppresses the residual echo signal by applying a multiplicative gain calculated from the estimated residual echo power spectrum. However, the estimated echo power spectrum reproduces only a fraction of the echo-path impulse response and so all the reverberant components are not considered. To address this problem, this chapter introduces a finite nonnegative convolution method by which each segment of echo-impulse response is convoluted with a received signal in a power spectral domain. With the proposed method, the power spectra of each segment of echo-impulse response are collectively estimated by solving the least-mean-squares problem between the microphone-input and the estimated-echo power spectra.

4.2 *Conventional Echo Power Estimation*

The echo spectrum, $D_i(\omega)$, which is the short-time spectrum of the echo signal $d(k)$, is expressed by the convolution of the echo path $h(k)$ and the received signal $x(k)$ in the frequency domain [43] as follows:

$$D_i(\omega) = \sum_{m=0}^{\infty} H_{m,i}(\omega) X_{i-m}(\omega), \tag{4.1}$$

where ω and m denote the frequency bin and the frame index, respectively. $H_{m,i}(\omega)$ is the short-time spectrum obtained from a fraction of the length of $h(k)$; $H_{0,i}(\omega)$ is the spectrum of the early part of $h(k)$ that includes the direct sound and early reflections (commonly called *early response*); the remaining portions ($H_{m,i}(\omega)$; $m > 0$) are the spectra of the late reverberation. The relationship between $h(k)$ and $H_{m,i}(\omega)$ is illustrated in Figure 4.1.

In the ER process, the echo power spectrum $|D_i(\omega)|^2$ is needed to calculate the echo-reduction gain $G_i(\omega)$; and its estimate $|\hat{D}_i(\omega)|^2$ is calculated using the MA model as

$$|\hat{D}_i(\omega)|^2 = |\hat{H}_{0,i}(\omega)|^2 |X_i(\omega)|^2 + \alpha |\hat{D}_{i-1}(\omega)|^2, \tag{4.2}$$

where α is a design parameter to control the reverberation time; the parameter α is set to the amount of added late echo components according to the reverberation time. $|\hat{H}_{0,i}(\omega)|^2$ denotes the power spectrum estimate of $H_{0,i}(\omega)$; it is given by

$$|\hat{H}_{0,i}(\omega)|^2 = \left| \frac{\langle \mathbf{X}_i(\omega), \mathbf{Y}_i(\omega) \rangle}{\|\mathbf{X}_i(\omega)\|^2} \right|^2. \tag{4.3}$$

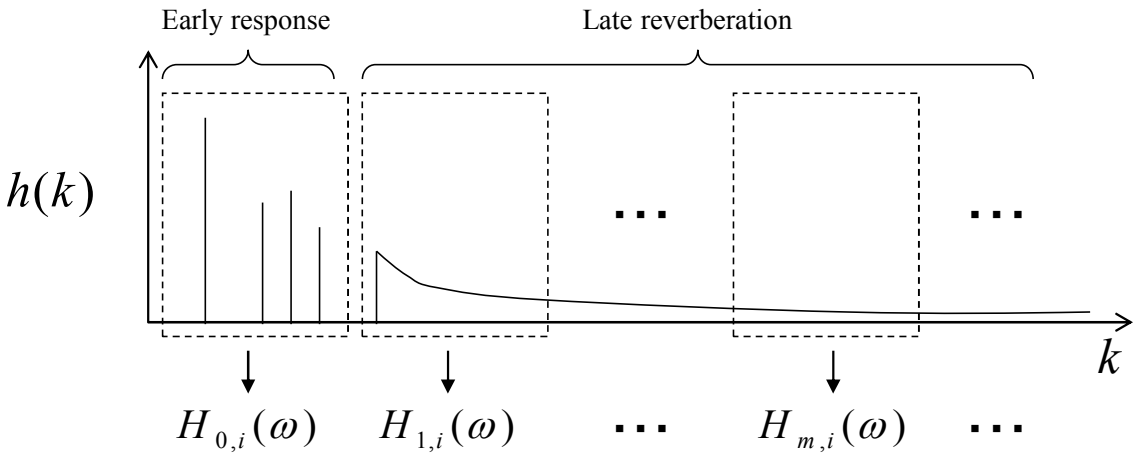


Figure 4.1. Relationship between echo-path impulse response and its short-time spectra.

4.3 *Problem with MA Model*

The MA model simulates the echo power spectrum by the following approximate expansions:

$$\begin{aligned} |\hat{D}_i(\omega)|^2 &= \left|\sum_{m=0}^{\infty} H_{m,i}(\omega) X_{i-m}(\omega)\right|^2 \\ &\approx \sum_{m=0}^{\infty} |H_{m,i}(\omega)|^2 |X_{i-m}(\omega)|^2 \\ &\approx \sum_{m=0}^{\infty} \alpha^m |H_{0,i}(\omega)|^2 |X_{i-m}(\omega)|^2 \\ &= |H_{0,i}(\omega)|^2 |X_{i-m}(\omega)|^2 + \alpha |D_{i-1}(\omega)|^2 \end{aligned} \quad , \qquad (4.4)$$

By the above estimate expansion: $|H_{m,i}(\omega)|^2 \approx \alpha^m |H_{0,i}(\omega)|^2$, the late echo components are modelled as $\alpha |D_{i-1}(\omega)|^2$. However, if this assumption holds true, the following relationships between $|H_{m,i}(\omega)|^2$ and $|H_{0,i}(\omega)|^2$ must hold:

$$
\begin{aligned}
|\hat{H}_{m,i}(\omega)|^2 &= \left|\frac{\langle \mathbf{X}_{i-m}(\omega), \mathbf{Y}_i(\omega)\rangle}{\|\mathbf{X}_{i-m}(\omega)\|^2}\right|^2 \\
&\approx \left|\frac{\left\langle \mathbf{X}_{i-m}(\omega), \alpha^{\frac{m}{2}}\mathbf{Y}_{i-m}(\omega)\right\rangle}{\|\mathbf{X}_{i-m}(\omega)\|^2}\right|^2, \\
&\approx \alpha^m \left|\frac{\langle \mathbf{X}_{i-m}(\omega), \mathbf{Y}_{i-m}(\omega)\rangle}{\|\mathbf{X}_{i-m}(\omega)\|^2}\right|^2 \\
&= \alpha^m |\hat{H}_{0,i}(\omega)|^2
\end{aligned}
\tag{4.5}
$$

where

$$
|\hat{H}_{0,i}(\omega)|^2 = |\hat{H}_{0,i-m}(\omega)|^2 \tag{4.6}
$$

if the echo-path impulse response remains unchanged.

The above equality expansions are based on the premise that the spectral structure of the microphone input signal is stationary: $\mathbf{Y}_i(\omega) = \alpha^{\frac{m}{2}}\mathbf{Y}_{i-m}(\omega)$. However, this is an impractical assumption because the non-stationary signal such as speech is received in practice. Consequently, the conventional approach suffers from the speech distortions because of the inaccurate estimate of the echo power spectrum which occasionally causes the perceptual degradation of the sound quality.

4.4 *Proposed Echo Power Estimation*

4.4.1 Strategy for accurate echo power estimation

This section proposes a method to improve the accuracy of conventional echo power spectral estimation. The proposed method introduces a finite nonnegative convolution model [44] in the power spectral domain from the following equation:

$$|\hat{D}_i(\omega)|^2 = \sum_{m=0}^{M-1} |\hat{H}_{m,i}(\omega)|^2 |X_{i-m}(\omega)|^2 , \tag{4.7}$$

where M is the number of frame determined by the reverberation time; M is for example set at 20 when the frame shift size is 8 ms and the reverberation time is 160 ms. This model is able to deal with the non-stationarity of the microphone input signal unlike the conventional MA approximation model.

The estimate $|\hat{H}_{m,i}(\omega)|^2$ in Equation (4.7) can be calculated from

$$|\hat{H}_{m,i}(\omega)| = \left| \frac{\langle \mathbf{X}_{i-m}(\omega), \mathbf{Y}_i(\omega) \rangle}{\|\mathbf{X}_{i-m}(\omega)\|^2} \right|^2 \tag{4.8}$$

which is included in Equation (4.5). However, this equation does not consider whether or not the estimated echo-path power spectra for each segment are optimal. Therefore, this study proposes the proper estimation method of the power spectra of early response and late reverberation for the finite nonnegative convolution model by solving the least squares problem between the microphone input power spectrum $|Y_i(\omega)|^2$ and the estimate $|\hat{D}_i(\omega)|^2$ of the echo power spectrum as follows:

$$\mathbf{W}_i(\omega) = \Phi_i^{-1}(\omega)\mathbf{R}_i^T(\omega)\tilde{\mathbf{Y}}_i(\omega), \tag{4.9}$$

where

$$\mathbf{W}_i(\omega) = \begin{bmatrix} |\hat{H}_{0,i}(\omega)|^2 \\ \vdots \\ |\hat{H}_{M-1,i}(\omega)|^2 \end{bmatrix}, \tag{4.10}$$

$$\Phi_i(\omega) = \mathbf{R}_i^T(\omega)\mathbf{R}_i(\omega), \tag{4.11}$$

$$\mathbf{R}_i(\omega) = \begin{bmatrix} |X_i(\omega)|^2 & \cdots & |X_{i-M+1}(\omega)|^2 \\ \vdots & \ddots & \vdots \\ |X_{i-L+1}(\omega)|^2 & \cdots & |X_{i-M-L+2}(\omega)|^2 \end{bmatrix}, \text{ and} \tag{4.12}$$

$$\widetilde{\mathbf{Y}}_i(\omega) = \begin{bmatrix} |Y_i(\omega)|^2 \\ \vdots \\ |Y_{i-L+1}(\omega)|^2 \end{bmatrix}. \tag{4.13}$$

Equation (4.9) is the solution for the following simultaneous equation model derived from Equation (4.7):

$$\widetilde{\mathbf{Y}}_i(\omega) = \mathbf{R}_i(\omega)\mathbf{W}_i(\omega), \tag{4.14}$$

and also corresponds to a least squares solution to the following equation:

$$\left\| \widetilde{\mathbf{Y}}_i(\omega) - \mathbf{R}_i(\omega)\mathbf{W}_i(\omega) \right\|^2 \to 0. \tag{4.15}$$

4.4.2 Practical calculation method

In practice, it is important to reduce the computational complexity of the method for real-time processing. This study proposes a method that improves the complexity of calculating the inverse of the matrix $\Phi_i(\omega)$ by applying the matrix inversion lemma to the following recursive equation:

$$\Phi_i(\omega) = \beta\Phi_{i-1}(\omega) + \dot{\mathbf{X}}_i(\omega)\dot{\mathbf{X}}_i^T(\omega), \tag{4.16}$$

where

$$\dot{\mathbf{X}}_i(\omega) = \begin{bmatrix} |X_i(\omega)|^2 \\ \vdots \\ |X_{i-M+1}(\omega)|^2 \end{bmatrix} \tag{4.17}$$

and β is the forgetting factor corresponding to time period L. The inverse of $\Phi_i(\omega)$ is for instance obtained using the recursive least-squares (RLS) algorithm [45] [46] as

$$\Phi_i^{-1}(\omega) = \beta^{-1}\Phi_{i-1}^{-1}(\omega) - \frac{\beta^{-2}\Phi_{i-1}^{-1}(\omega)\dot{\mathbf{X}}_i(\omega)\dot{\mathbf{X}}_i^T(\omega)\Phi_{i-1}^{-1}(\omega)}{1+\beta^{-1}\dot{\mathbf{X}}_i^T(\omega)\Phi_{i-1}^{-1}(\omega)\dot{\mathbf{X}}_i(\omega)}. \tag{4.18}$$

Also, the ER process often reconstructs the send signal by overlapping and adding the windowed output frames. If the FFT window overlap of 50% is adopted in the ER process, the proposed method allows us to omit some of the calculations and to estimate the echo power spectrum as follows:

$$|\hat{D}_i(\omega)|^2 = \sum_{m=0}^{M/2-1} |\hat{H}_{2m,i}(\omega)|^2 |X_{i-2m}(\omega)|^2 . \tag{4.19}$$

4.5 ***Simulations***

4.5.1 Comparison of estimation accuracy

The echo-power estimation errors of the conventional and proposed methods based on Equations (4.2) and (4.19) are shown in Figure 4.2. The estimation error is represented using the SDR as the difference between the target and estimated echo-power spectra in a received single-talk period. These sound qualities after processing were compared under reverberation times of 50, 100, 150, 250, 350, 500, and 750 ms. The female talker signal digitized at a sampling rate of 16 kHz was used as the sound source of the received speech signal. The language is English. The design parameter α was set at 0.3, 0.55, 0.7, 0.8, 0.85, 0.9, and 0.93, respectively, according to the reverberation time. The number of frame M was set at four. As Figure 4.2 indicates, the proposed echo-power estimation method improved the estimation accuracy under all these reverberation time conditions.

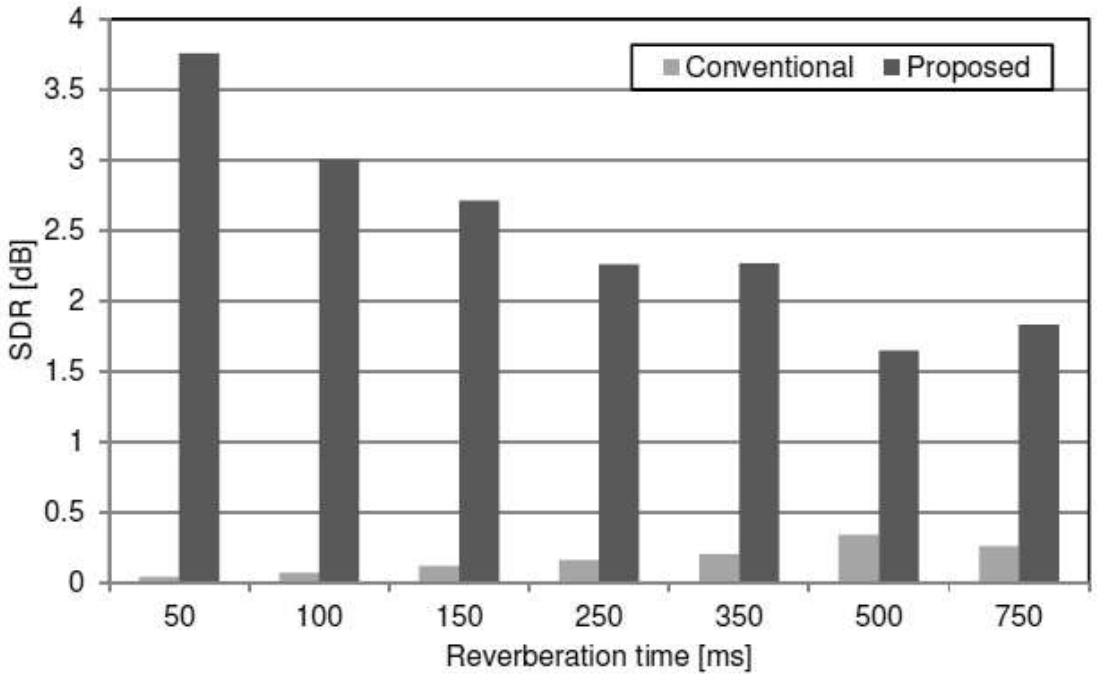


Figure 4.2. Comparison of SDR during received single-talk period.

4.5.2 Comparison of echo-reduction performance

This subsection evaluated the performance of the proposed echo power estimation method using simulation tests in order to compare it with the conventional method (MA model). The conventional and proposed estimation methods were applied to calculate the echo power spectrum using Equations (4.2) and (4.19), respectively. Table 4.1 lists the experimental conditions.

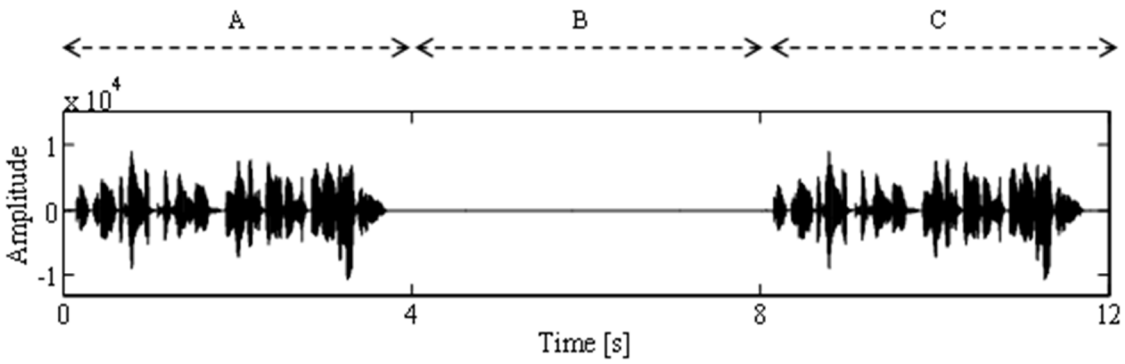
The received and near-end speech signals are shown in Figures 4.3 (a) and (b), respectively. Periods A and B are the received and send single-talk periods, respectively. A double-talk period occurs during a period C. The microphone input signal $y(k)$ is shown in Figure 4.4 (a). Figures 4.4 (b) and (c) plot the send signals $\hat{s}(k)$ after the conventional and proposed methods were respectively applied to them. As seen in Figures 4.4 (a), (b), and (c), the conventional and proposed methods sufficiently suppressed the echo signals in the

received single-talk period A. The ELREs of 31.05 dB and 31.38 dB were achieved as the echo-suppression levels using the conventional and proposed methods, respectively. However, these results show that speech distortions seemed to occur during the double-talk period C when the conventional method was applied.

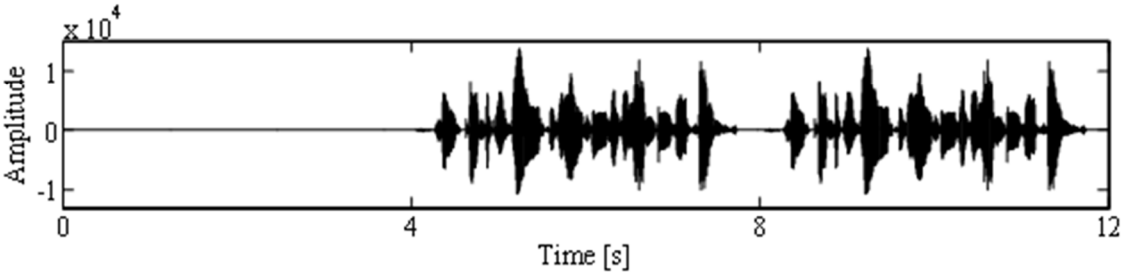
The amount of speech distortion during the double-talk period C was evaluated using the LPC cepstral distance. The results of comparing the conventional and proposed methods are shown in Figure 4.5. The average cepstral distance scores of the conventional and proposed methods were 5.42 points and 2.19 points, respectively. In particular, the proposed method improved sound quality by a maximum of 9.72 points in a period of about 11 s over the conventional method. Better scores were observed over almost the entire period, and significant improvement was confirmed.

Table 4.1. Experimental conditions

Sampling rate	16 kHz
Frame length	256 samples
Frame shift	128 samples
FFT points	256 samples
Sound item	Clean speech
Single length	12 s
Reverberation time	160 ms

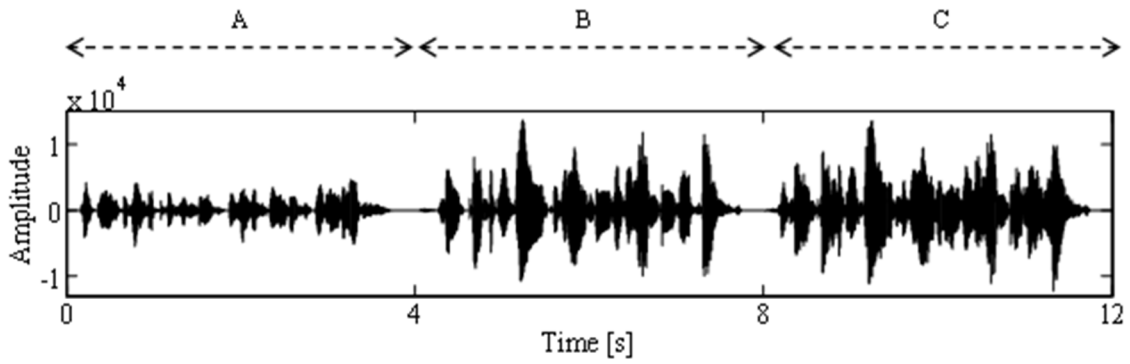


(a) Received speech signal (female speech)

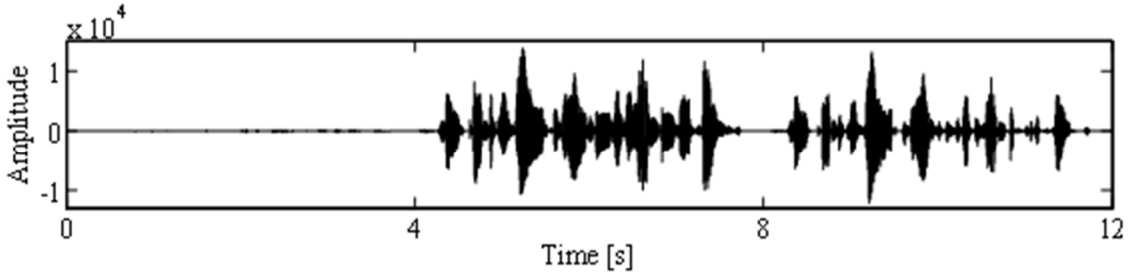


(b) Near-end speech signal (male speech)

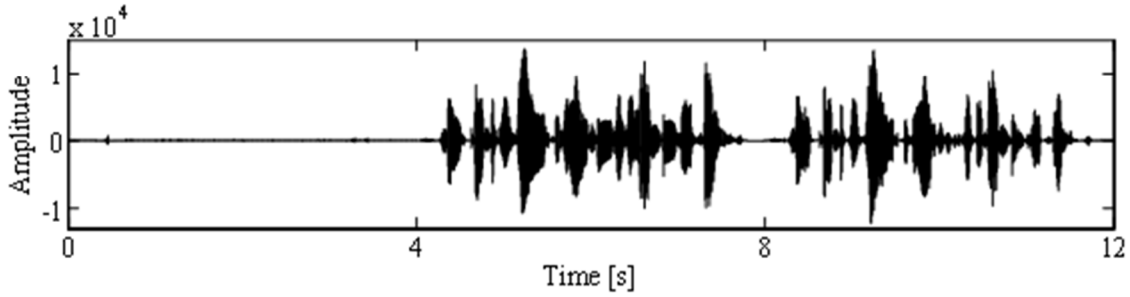
Figure 4.3. Received speech signal and near-end speech signal.



(a) Microphone input signal



(b) Send signal processed by conventional method



(c) Send signal processed by proposed method

Figure 4.4. Microphone input signal, send signal with conventional method, and send signal with proposed method.

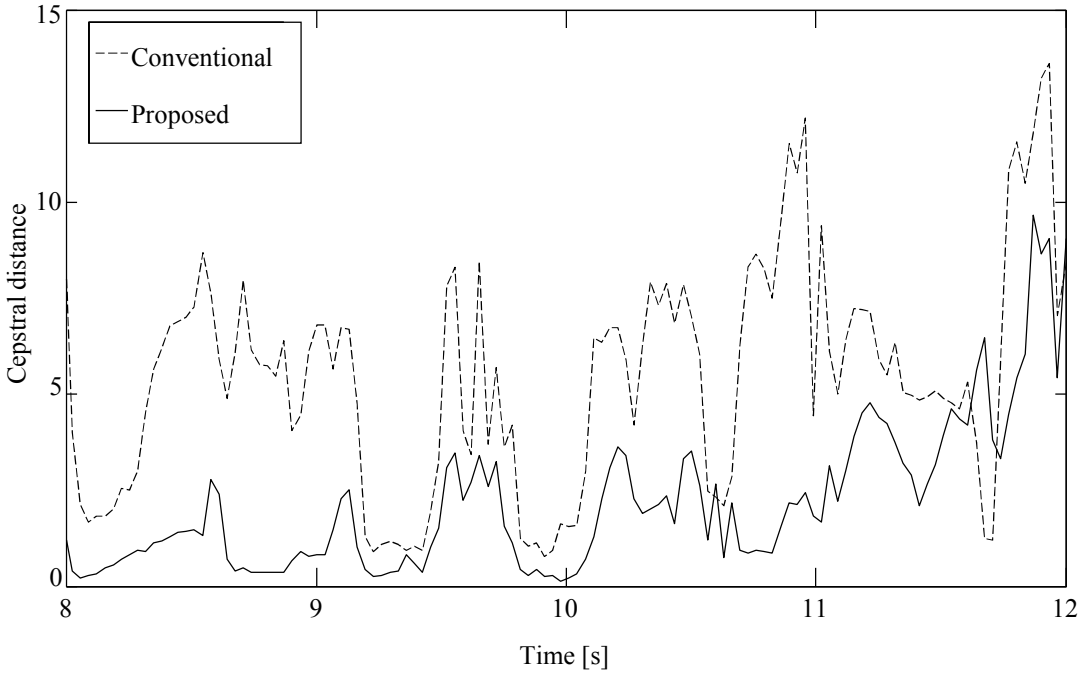


Figure 4.5. Comparison of LPC cepstrum distances during double-talk period C.

4.6 *Conclusion*

A novel method to estimate the echo power spectrum for the ER process was proposed. The echo power spectrum was calculated using the finite convolution between the echo path and the received signal in the power spectral domain. The power spectra of early response and late reverberation were jointly estimated by solving the least squares problem for the early response and the late reverberation in the power spectral domain. The proposed method achieves the proper residual echo suppression and marginal speech distortion in the send signal during the double-talk period. The experimental results revealed that the proposed method outperformed the conventional MA-model approach, and improved the speech quality by about three points in the average score of the LPC cepstral distance during the double-talk period.

Chapter 5.

Acoustic Echo Cancellation for CD-quality Hands-free Videoconferencing System

5.1 *Introduction*

This chapter describes a monaural AEC unit newly developed for a 20-kHz wideband hands-free video- or teleconferencing system. The unit can effectively reduce undesired acoustic echo that occurs in the system, and can emphasize the target talker's voice during the double-talk period. The algorithm implemented to this unit estimates an echo-path power spectrum whether or not double-talk has occurred, then calculates a post-filter that effectively reduces the undesired echo. In the echo cancellation processing, the computational 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.

This study evaluated the echo-cancellation performance by implementing it using an AEC-unit prototype that has a DSP board. Experiments were conducted to examine the performance of the developed AEC unit; the results indicated that the developed unit delivered natural-sounding near-end speech even during the double-talk periods while sufficiently suppressing the undesired echo.

5.2 *Videoconferencing System with AEC Unit*

Figure 5.1 shows a videoconferencing system with an AEC unit; the flow of the audio signal processing of the AEC is illustrated in Figure 5.2. 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 input signal in order to transmit only the near-end speech to the far-end.

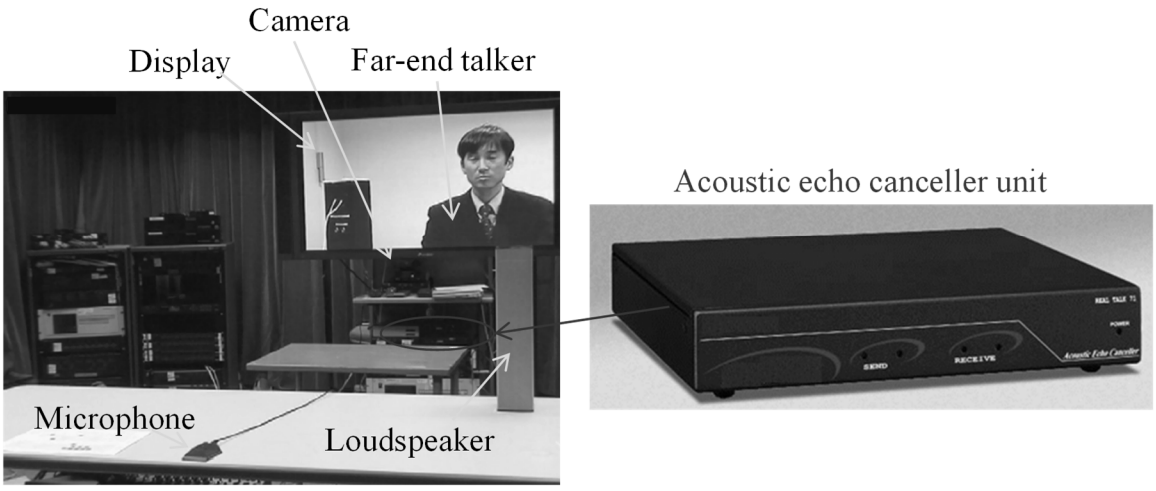


Figure 5.1. Left: hands-free video conference system, right: AEC unit.

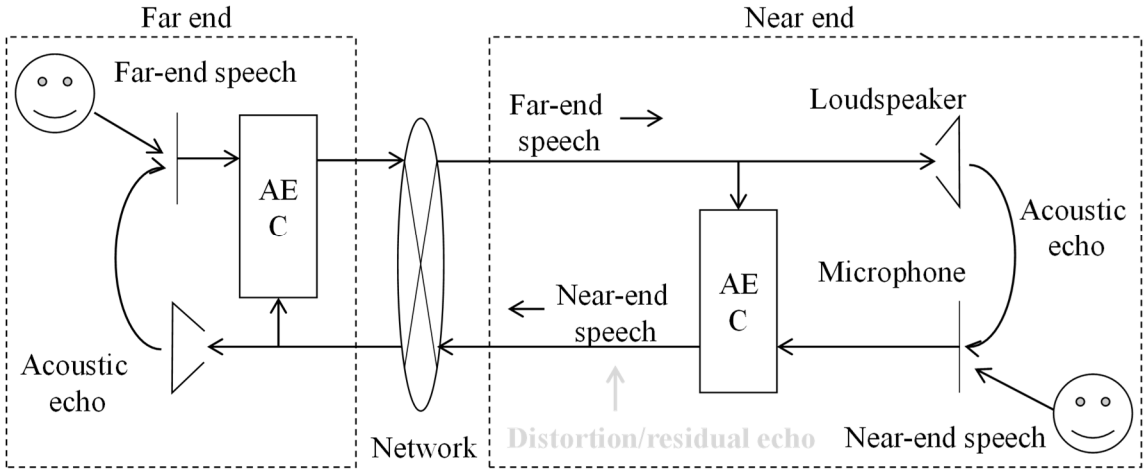


Figure 5.2. Flow of AEC.

5.3 *Algorithm Improving Double-Talk Performance*

The problem of near-end speech degradation is mainly caused by the low estimation accuracy of the echo-path power spectrum used in the ER process. The conventional echo-path power spectrum estimation method [12] sometimes over-estimates the echo-path power spectrum during the double-talk periods; this overestimation causes near-end speech in double-talk being muffled after ER process is applied. To solve this problem, this study employs the echo-path power spectrum estimation method similar to the method described in Chapter 2.

The AEC receives the signal $x(k)$ from the far-end; and this signal is picked up as an echo signal $d(k)$ by the microphone after passing through the room echo path that has an impulse response modeled as $\mathbf{h}(k)=[h_1(k),\ldots,h_{L_f}(k)]^T$, where L_f is the filter length. Given the reference input vector, $\mathbf{x}(k)=[x(k),\ldots,x(k-L_f+1)]^T$, and the adaptive filter vector, $\mathbf{w}(k)=[w_1(k),\ldots,w_{L_f}(k)]^T$, the output signal of ADF, $y(k)$, can be written in terms of the reference input vector, $\mathbf{x}(k)$, which is convoluted by the impulse response between reference and ADF output signals (*residual echo path*), $\mathbf{h}'(k)=[h'_1(k),\ldots,h'_{L_f}(k)]^T$, including the near-end speech signal $s(k)$ as

$$\begin{aligned} y(k) &= d(k)+s(k) \\ &= \mathbf{x}^T(k)\{\mathbf{h}(k)-\mathbf{w}(k-1)\}+s(k), \\ &= \mathbf{x}^T(k)\mathbf{h}'(k)+s(k) \end{aligned} \tag{5.1}$$

The output signal $y(k)$ is transformed to the frequency as follows:

$$Y_i(\omega)=D_i(\omega)+S_i(\omega), \tag{5.2}$$

The output of the ER is obtained as

$$\hat{S}_i(\omega)=G_i(\omega)Y_i(\omega). \tag{5.3}$$

Here, $G_i(\omega)$ is calculated by the Wiener filtering method

$$G_i(\omega) = \frac{|Y_i(\omega)|^2 - |\hat{D}_i(\omega)|^2}{|Y_i(\omega)|^2}, \tag{5.4}$$

where

$$|\hat{D}_i(\omega)|^2 = |\hat{H}_i^{'}(\omega)|^2 |X_i(\omega)|^2 . \tag{5.5}$$

The proposed method obtains an estimate of the echo-path power spectrum using the following equation:

$$|\hat{H}_i^{'}(\omega)|^2 = \left(\frac{\sum_{r=-\varsigma_\omega}^{\varsigma_\omega} E[|X_i(\omega+r)||Y_i(\omega+r)|]}{\sum_{r=-\varsigma_\omega}^{\varsigma_\omega} E[|X_i(\omega+r)|^2]} \right)^2, \tag{5.6}$$

Equation (5.6) is a new simplified equation for the proposed technique in Chapter 2. In Equation (5.6), 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 components can be removed by calculating the amplitude correlation between both signals, and Equation (5.6) is then approximated to the following equation:

$$
\begin{aligned}
|\hat{H}_i^{'}(\omega)|^2 &= \left(\frac{\sum_{r=-\varsigma_\omega}^{\varsigma_\omega} E[|X_i(\omega+r)\,\|\,D_i(\omega+r)+S_i(\omega+r)|]}{\sum_{r=-\varsigma_\omega}^{\varsigma_\omega} E[|X_i(\omega+r)|^2]} \right)^2 \\
&\approx \left(\frac{\sum_{r=-\varsigma_\omega}^{\varsigma_\omega} E[|X_i(\omega+r)\,\|\,D_i(\omega+r)|]}{\sum_{r=-\varsigma_\omega}^{\varsigma_\omega} E[|X_i(\omega+r)|^2]} \right)^2 \\
&\approx \frac{\sum_{r=-\varsigma_\omega}^{\varsigma_\omega} E[|D_i(\omega+r)|^2]}{\sum_{r=-\varsigma_\omega}^{\varsigma_\omega} E[|X_i(\omega+r)|^2]}
\end{aligned}
\quad . \qquad (5.7)
$$

This means the ER is able to directly estimate the echo-path power spectrum even during double-talk periods.

5.4 *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 sutudy 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.

5.4.1 Specifications

The proposed AEC algorithm was utilized for a hands-free voice communication unit. A circuit board of the unit is shown in Figure 5.3, and the specifications are listed in Table 5.1. The AEC was implemented on a single fixed-point DSP chip. The sampling frequency of the analog-to-digital (A/D) or digital-to-analog (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.

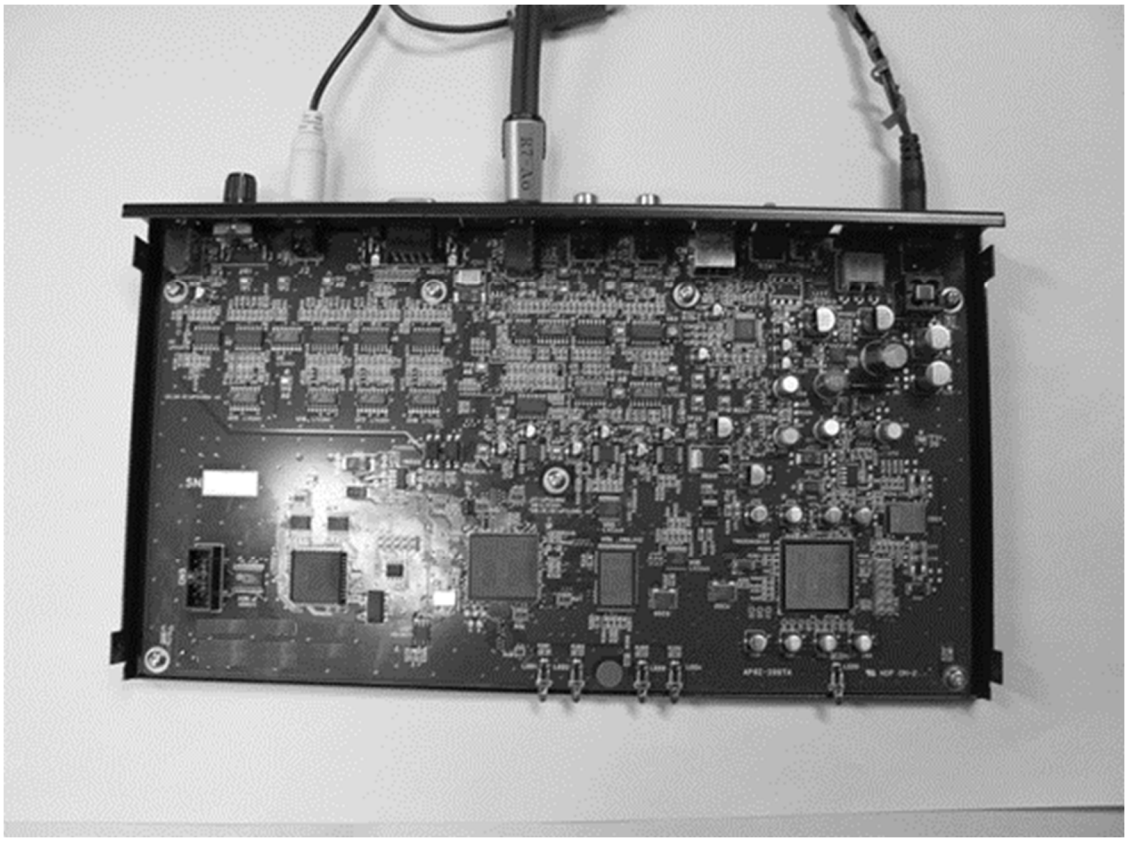


Figure 5.3. Circuit board of AEC unit.

Table 5.1. Specifications of AEC 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

5.4.2 Subband approach

In the AEC, the ADF process has a significant impact on the overall complexity because the degree of computational complexity required in the ADF process 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 process 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 process to achieve a real-time CD-quality wideband AEC.

This study 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 5.4 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 Figure 5.5.

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 process was reduced by 75% by quartering the band. Moreover, if the ADF process is not processed for frequencies above 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 components above 12 kHz can be sufficiently suppressed without the ADF process but only by the ER process. In total, the computational complexity required in the ADF process was reduced by about 90% using this configuration.

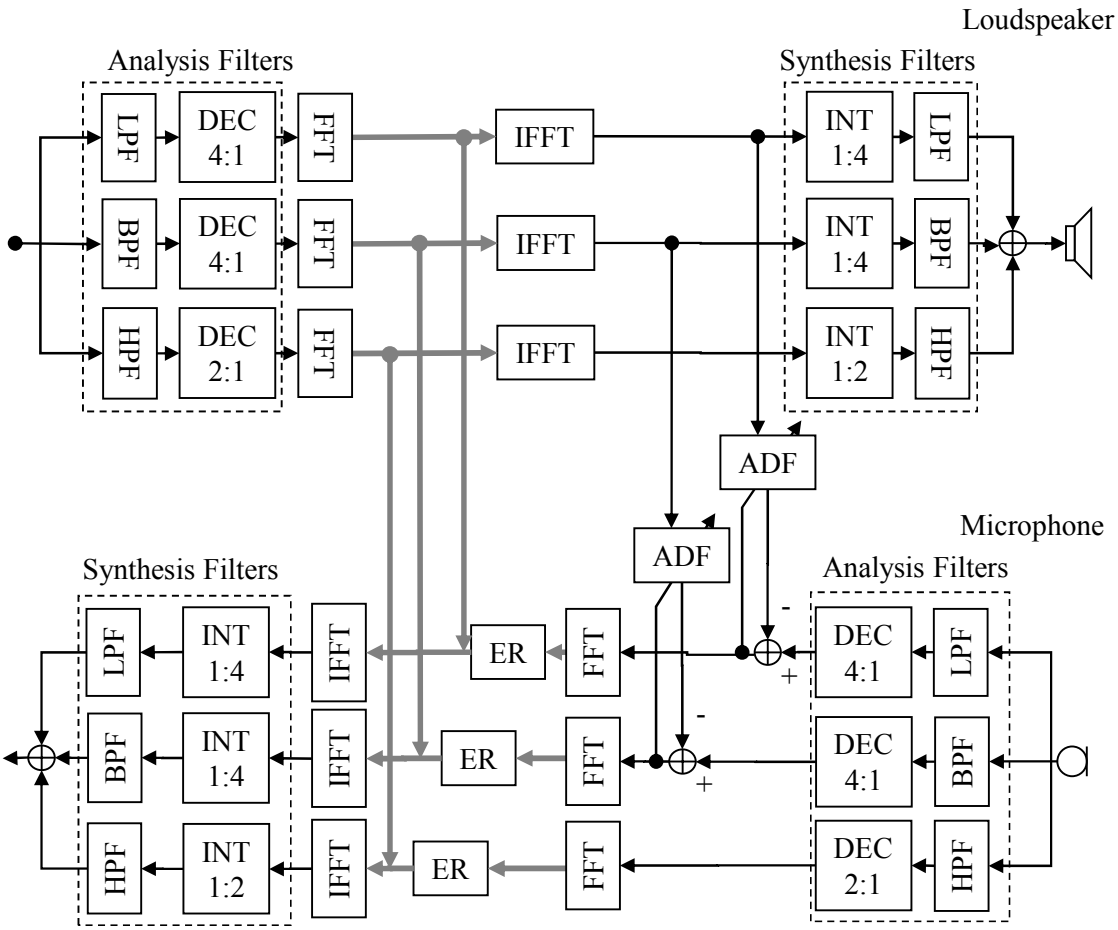


Figure 5.4. Signal flow of implemented AEC.

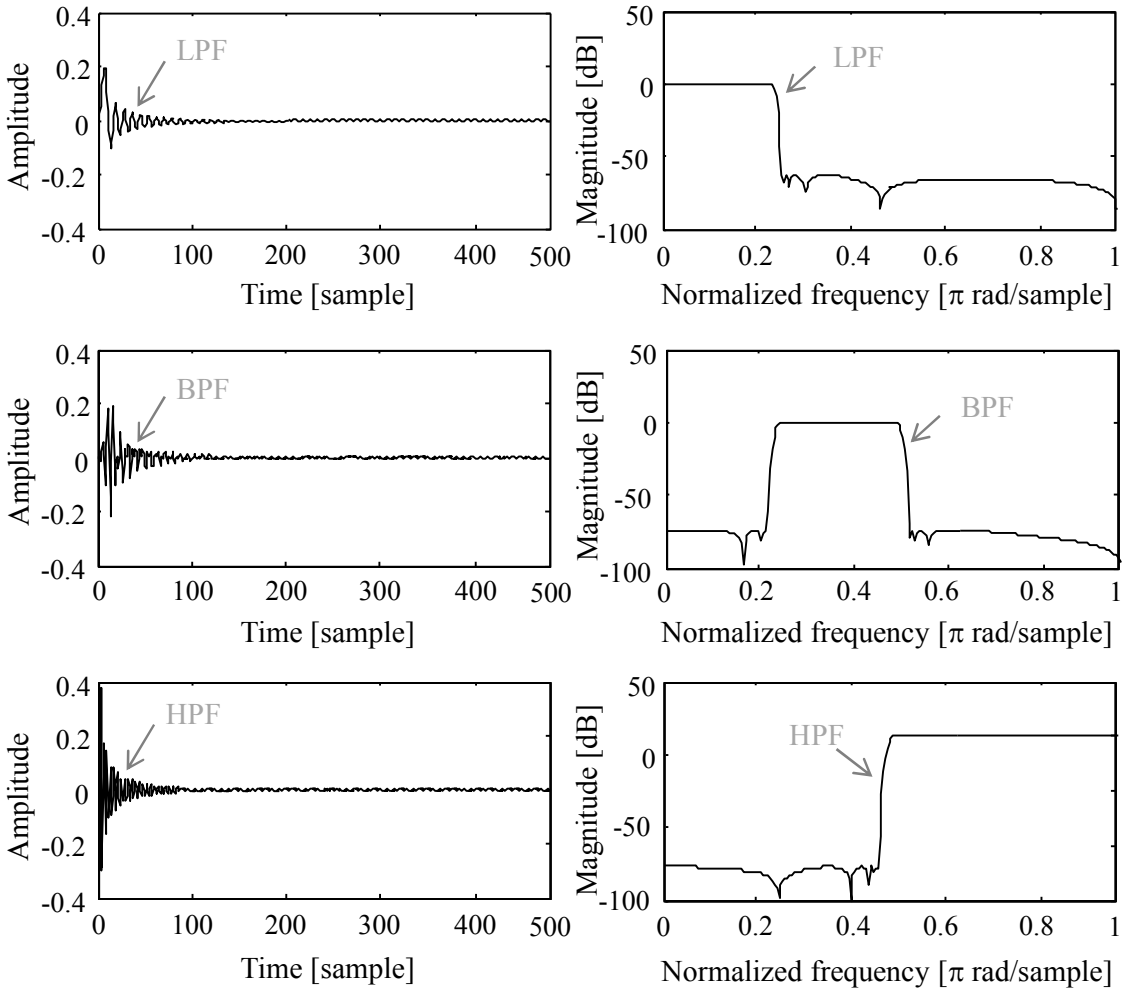


Figure 5.5. Impulse responses and frequency responses of low-pass, band-pass, and high-pass filters.

5.4.3 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 study 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 Figure 5.6. The frequency ranges of the LPF, BPF, and HPF are 0–6, 6–12, and 12–24 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

$$Z_i(e^{j\omega\tau}) = U_i(e^{j\omega\tau})I_i(e^{j\omega\tau}), \tag{5.8}$$

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:

$$Z_{\mathrm{i}}^{\mathrm{D}}(e^{j\omega\tau'}) = \frac{1}{M}\sum_{k=0}^{M-1} Z\left(e^{j\frac{\omega\tau'-2\pi k}{M}}\right), \tag{5.9}$$

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

Figure 5.6. For example, if the DEC ratio M is two, Eq. (10) is convertible as follows:

$$\begin{aligned} Z_{\mathrm{i}}^{\mathrm{D}}(e^{j\omega\tau'}) &= \frac{1}{2}\sum_{k=0}^{2-1} Z\left(e^{j\frac{\omega\tau'-2\pi k}{2}}\right). \\ &= Z\left(e^{j\omega\tau}\right)+Z\left(e^{j\omega\tau-\pi}\right) \end{aligned} \tag{5.10}$$

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 Figure 5.4. 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

$$I_i^U(e^{j\omega\tau}) = I_i(e^{j\frac{\omega\tau'}{M}}). \tag{5.11}$$

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 Figure 5.7.

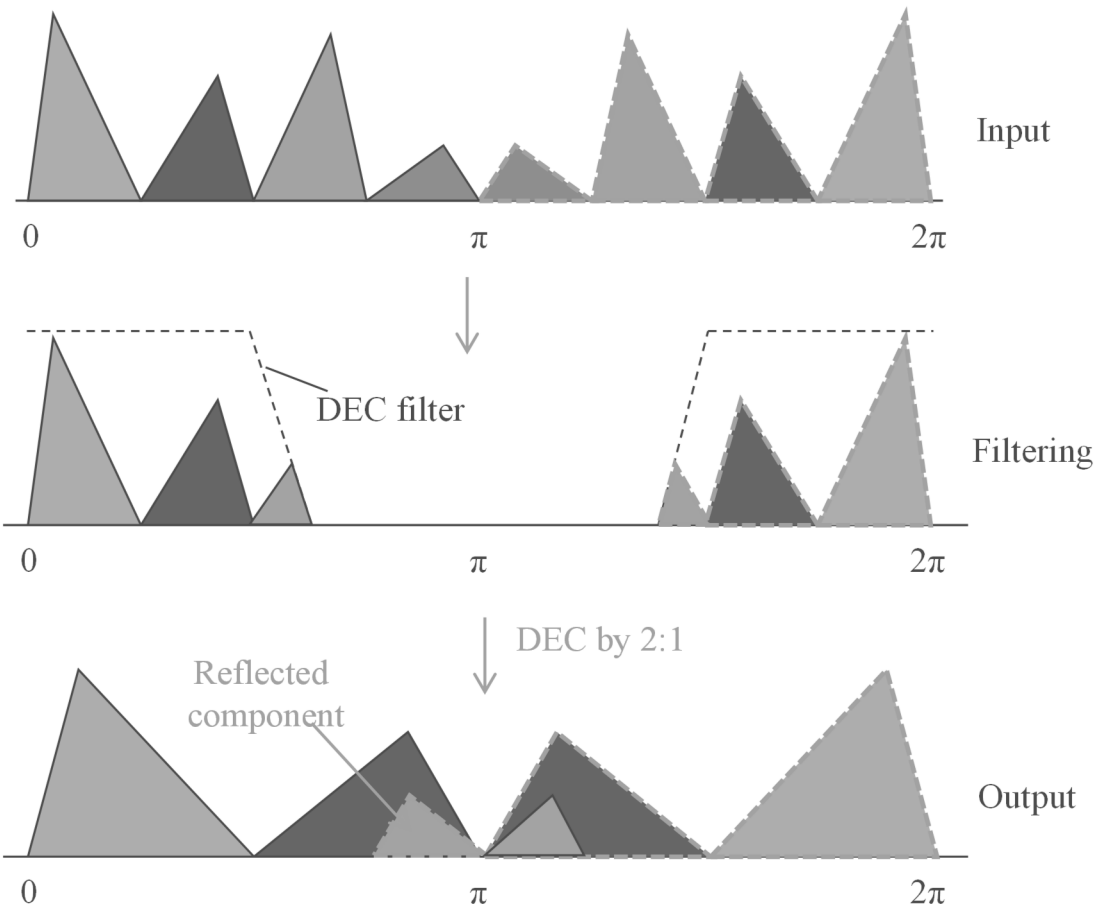


Figure 5.6. Illustration showing how decimation can be performed in frequency domain.

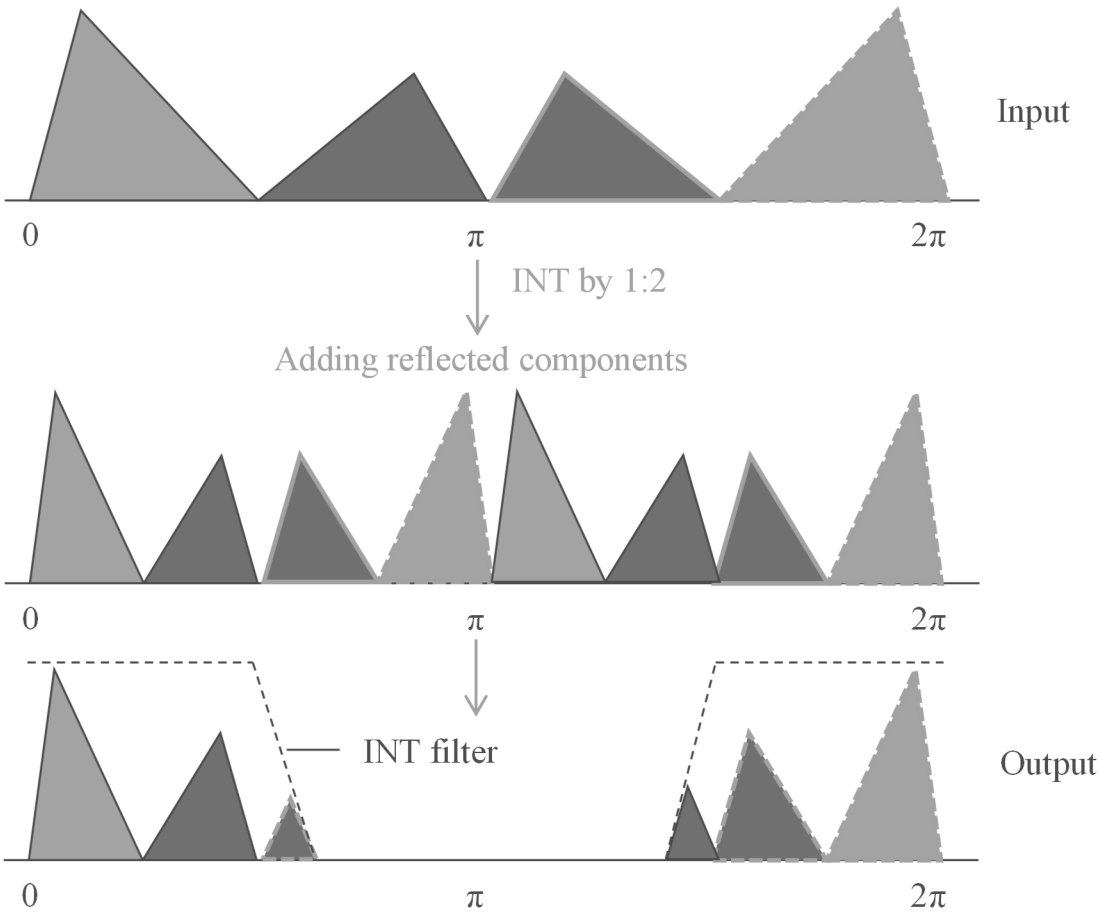


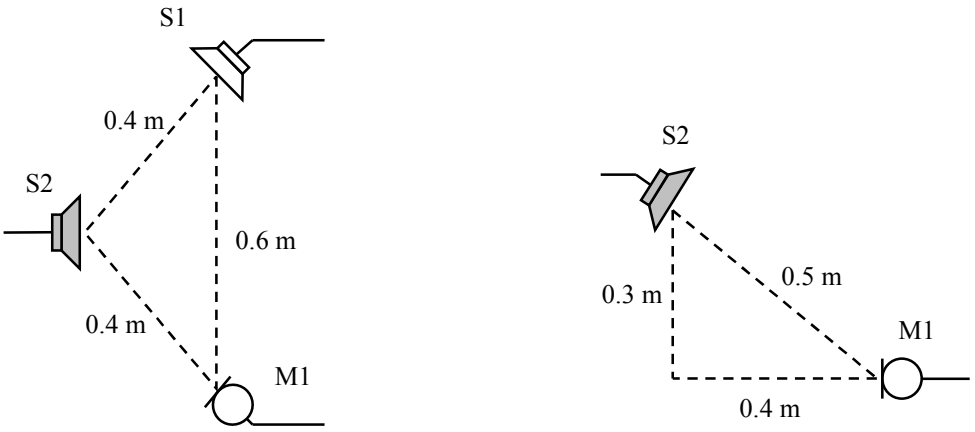
Figure 5.7. Illustration showing how interpolation can be performed in frequency domain.

5.5 ***Performance Evaluation***

The performance of the proposed method implemented in the AEC-unit prototype was evaluated in a practical environment. This study examined how the near-end ambient sound is processed and transmitted under various conversation states, particularly during the double-talk periods. If the 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 echo-path power spectrum estimation method [12].

5.5.1 Test conditions

The arrangement of the microphones and sound sources used in the test is shown in Figure 5.8. 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 [47]. 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 Figure 5.9.



(a) Top view (b) Side view

Figure 5.8. Test arrangements for objective measurements.

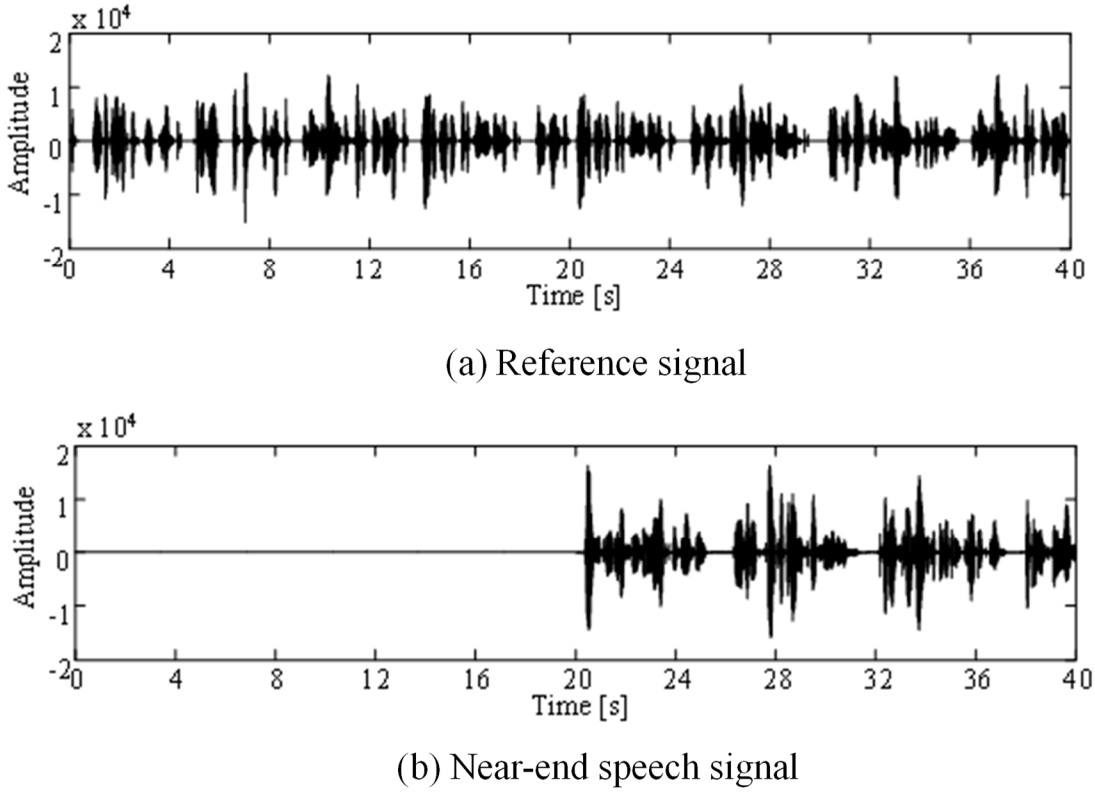


Figure 5.9. Reference signal and near-end speech signal.

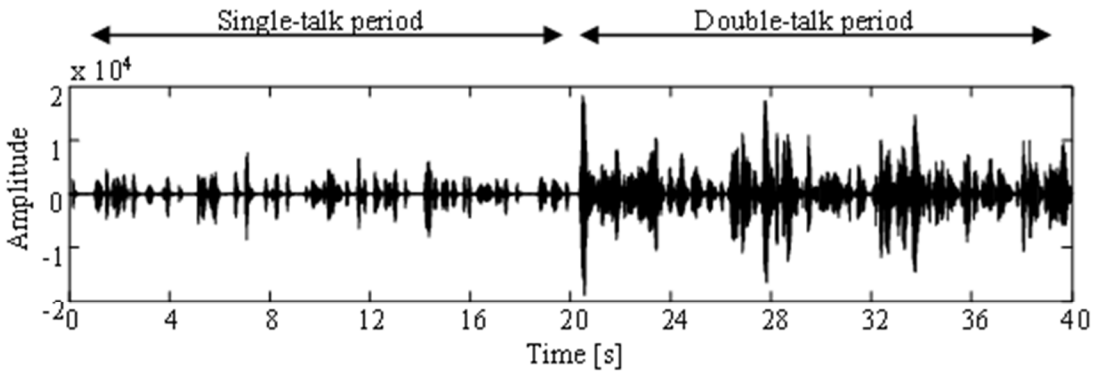
5.5.2 Experimental results

The microphone input signal and the transmitted signals received after processing by the conventional and proposed AECs are shown in Figure 5.10, 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 Figure 5.10.

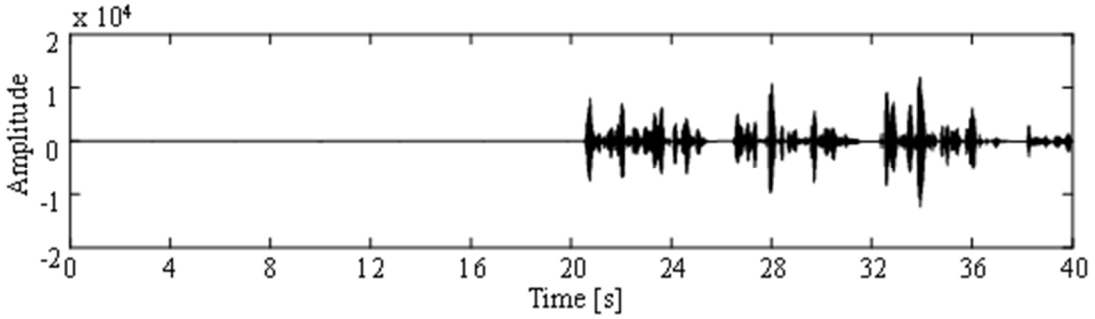
For the evaluation of the echo-suppression level, the ERLE measure, which is defined by ITU-T Recommendation G.168 [48], was used. The ERLE gives large values when the echo component is well suppressed by the AEC. The average ERLEs of the conventional and proposed methods are 36.1 dB and 37.6 dB in the single-talk period, respectively.

The levels of transmitted signals processed with the conventional and proposed methods during a double-talk period are shown in Figures 5.11 and 5.12, 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, as seen from these figures. This means that the proposed method achieved much better accuracy in estimating the echo-path power spectrum compared to 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.

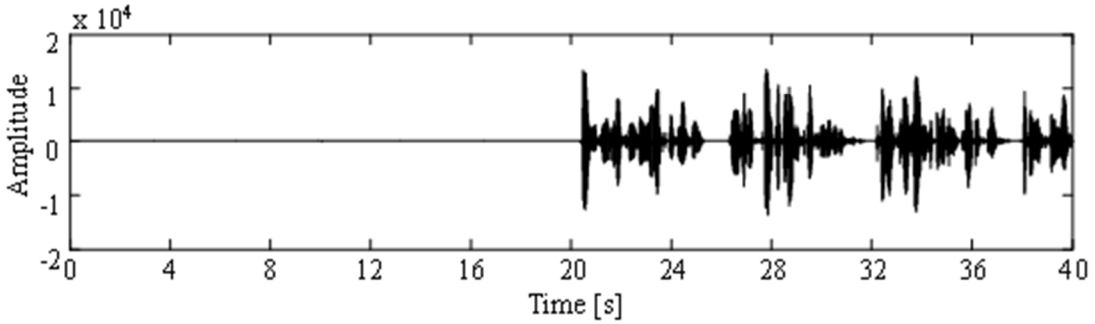
This study also evaluated the amount of speech distortion during the double-talk period using the LPC cepstral distance between the near-end speech and transmitted signals. The LPC cepstral distance gives a small value when the transmitted signal does not suffer a loss of the near-end speech and when the echo is well suppressed. The time transitions of the LPC cepstral distance of the conventional and proposed methods are shown in Figure 5.13. 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.



(a) Microphone input signal



(b) Transmitted signal processed by conventional method



(c) Transmitted signal processed by proposed method

Figure 5.10. Microphone input signal, transmitted signal processed using conventional method, and transmitted signal processed using proposed method.

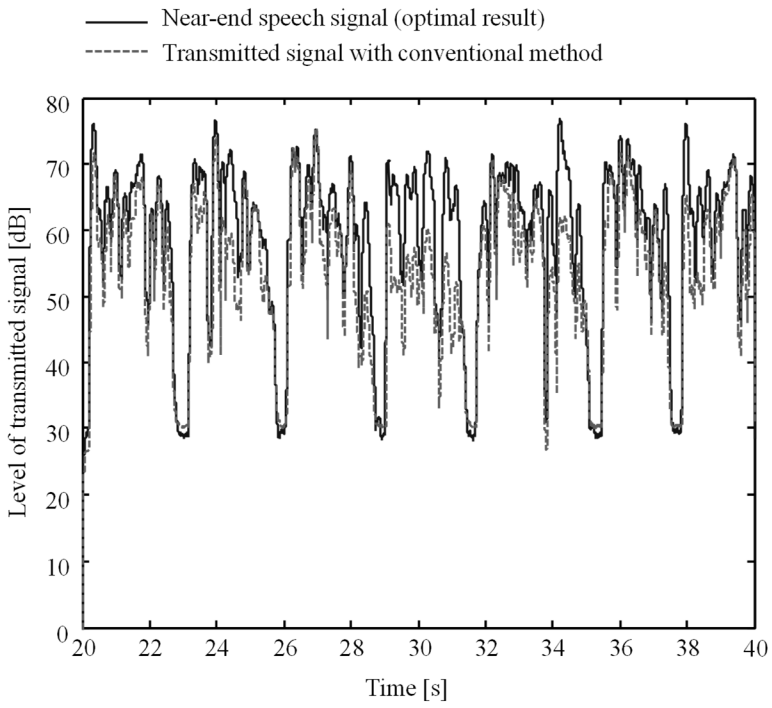


Figure 5.11. Level of transmitted signal processed with conventional method.

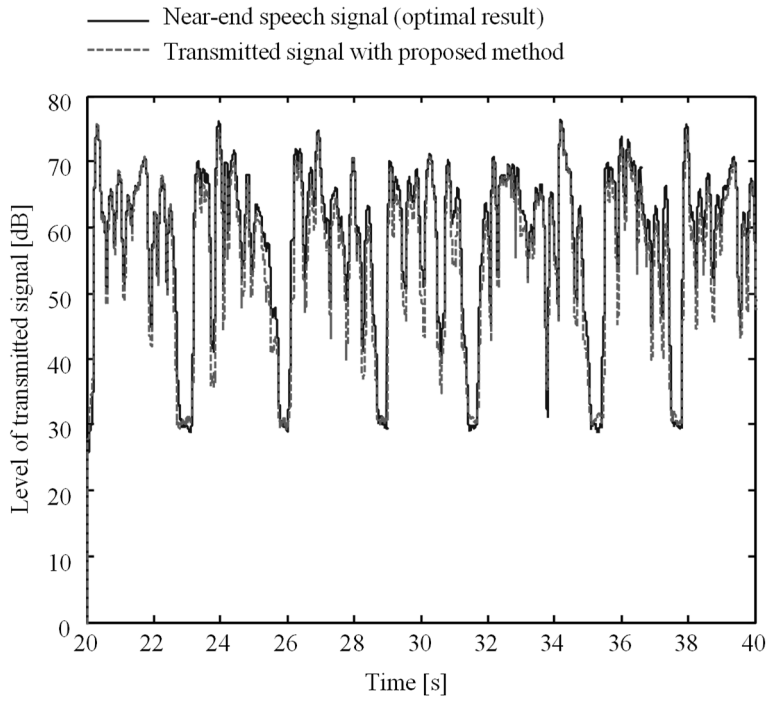


Figure 5.12. Level of transmitted signal processed with proposed method.

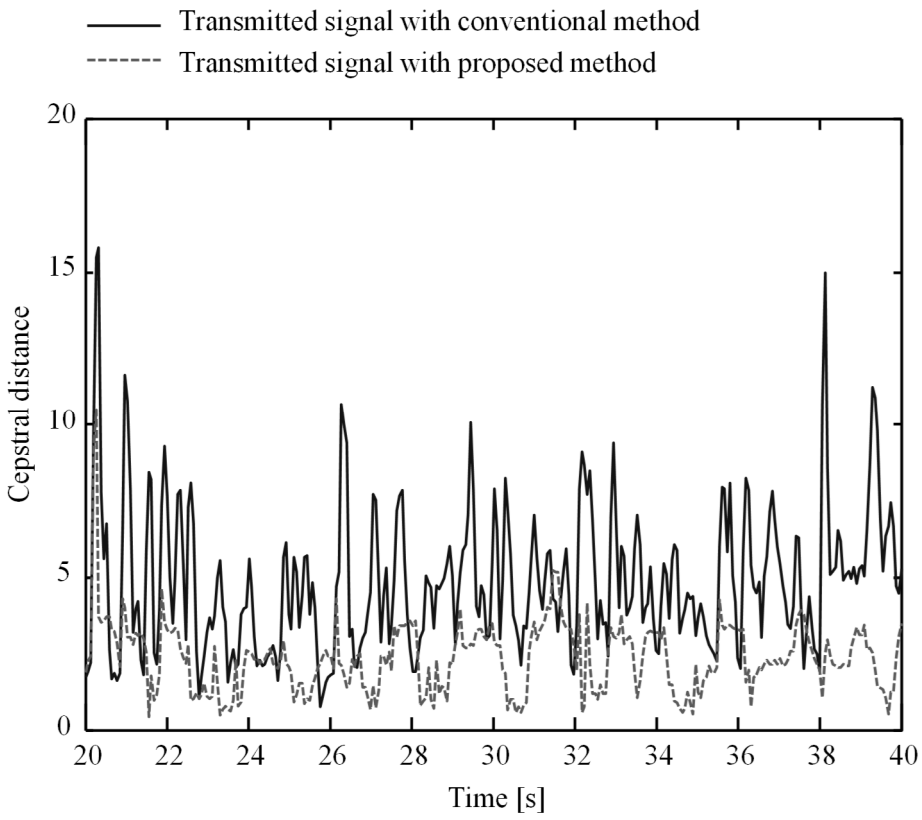


Figure 5.13. LPC cepstral distances of transmitted signals processed with conventional and proposed methods during double-talk period.

5.6 *Conclusion*

This chapter employed the echo-path power spectrum estimation method similar to the method described in Chapter 2, and presented an AEC method that can emphasize the target near-end speech with low degradation during the double-talk periods while reducing undesired acoustic echo. The method estimates the echo-path power spectrum between reference and ADF-output signals whether or not double-talk has occurred and calculates the post-filter based on the estimated echo-path power spectrum 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 computational complexity, the echo-path power spectrum 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.

Chapter 6.

AEC Software for VoIP Hands-free Application on Smartphone and Tablet Devices

6.1 *Introduction*

An AEC method developed for VoIP hands-free applications on smartphone and tablet devices is proposed. This method can effectively reduce undesired acoustic echo arriving at a microphone from a loudspeaker and emphasize the target talker's voice during the double-talk periods, irrespective of smartphone/tablet device models. This method mainly involves cancellation of the non-linear acoustic echo caused by the loudspeaker distortion, the residual echo reduction robust against echo-path change, and the 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 the undesired echo for every smartphone or tablet used for the evaluation. This result indicates that the performance of the proposed AEC method does not depend on the difference in the acoustic characteristics of individual devices.

6.2 *AEC Approach for VoIP Application on Smartphones and Tablets*

A block diagram of the developed AEC algorithm is illustrated in Figure 6.1. 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 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.

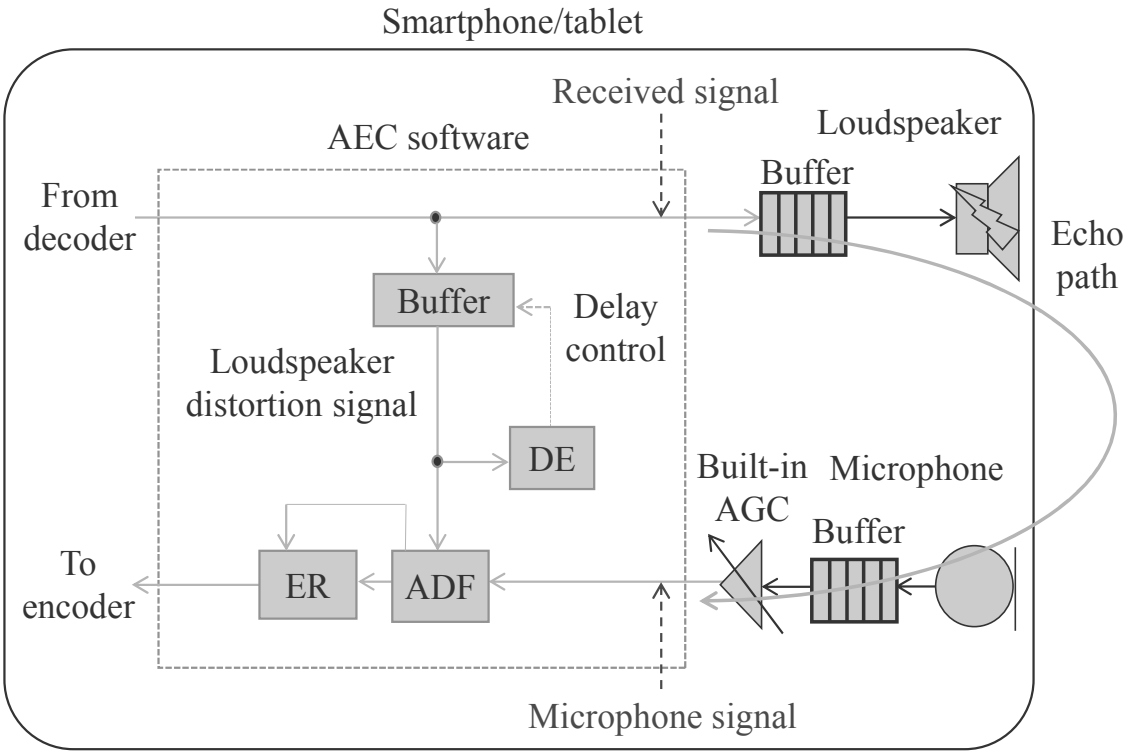


Figure 6.1. Block diagram depicting proposed AEC method developed for smartphone and tablet devices.

6.2.1 Nonlinear ADF

The cascade adaptive filtering scheme [49] [50] is used in the proposed AEC method, in which the adaptive hard clipping function is followed by the adaptive 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 $u(k)$ as follows:

$$u(k)=\begin{cases} a(k) & \left(x(k)>a(k)\right) \\ x(k) & \left(\left|x(k)\right|\le a(k)\right) \\ -a(k) & \left(x(k)<-a(k)\right) \end{cases}, \tag{6.1}$$

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

$$V(\omega)=W(\omega)U(\omega), \tag{6.2}$$

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:

$$a(k)\leftarrow a(k)+\mu\Delta a(k), \tag{6.3}$$

$$W(\omega)\leftarrow W(\omega)+\mu\frac{U(\omega)^{*}}{|U(\omega)|^{2}}\left[Y(\omega)-\Delta a(k)W(\omega)U'(\omega)\right], \tag{6.4}$$

where

$$\Delta a(k)=\text{real}\left[\frac{\sum_{\omega=0}^{\varsigma-1}\frac{\left(W(\omega)U'(\omega)\right)^{*}Y(\omega)}{\left|U(\omega)\right|^{2}}}{\chi+\sum_{\omega=0}^{\varsigma-1}\frac{\left|W(\omega)U'(\omega)\right|^{2}}{\left|U(\omega)\right|^{2}}}\right], \tag{6.5}$$

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

Unlike the original cascade scheme [49] [50], the adaptive FIR filter is implemented in the frequency domain [51] for the computational efficiency. To do this, the parameter updates shown in Equations (6.3) and (6.4) are also differently formulated from those of the original scheme in order to jointly estimate the time domain threshold parameter $a(k)$ and the filter coefficient of each frequency bin $W(\omega)$ by commonly evaluating the frequency domain error $Y(\omega)$.

6.2.2 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 [52]. 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

$$|\hat{D}_i^{'}(\omega)|^2 = \hat{g}_i |\hat{H}_{m,i}^{'}(\omega)|^2 |X_i(\omega)|^2, \tag{6.6}$$

where

$$\hat{g}_i = \max\left[\frac{\sum_{q=0}^{Q-1}\sum_{\omega=0}^{\varsigma-1}|Y_{i-q}(\omega)|^2 |\hat{H}_{m,i-q}^{'}(\omega)|^2 |X_{i-q}(\omega)|^2}{\sum_{q=0}^{Q-1}\sum_{\omega=0}^{\varsigma-1}\left(|\hat{H}_{m,i-q}^{'}(\omega)|^2 |X_{i-q}(\omega)|^2\right)^2}, 1\right], \tag{6.7}$$

$|\hat{H}_{m,i}^{'}(\omega)|^2$ 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.

6.2.3 *Delay estimation*

A block diagram of the DE technique is illustrated in Figure 6.2. 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 $H''_{0,i}(\omega),\ldots,H''_{L_d-1,i}(\omega)$ by using a generalized cross correlation (GCC) method [53] consisting of a multi-delay filter (MDF) [54] as follows:

$$\begin{bmatrix} H''_{0,i}(\omega) \\ \vdots \\ H''_{L_d-1,i}(\omega) \end{bmatrix} \approx \frac{\begin{bmatrix} X_0^*(\omega) & 0 & 0 \\ \vdots & \ddots & 0 \\ X^*_{M_d-1}(\omega) & \ddots & X_0^*(\omega) \\ 0 & \ddots & \vdots \\ 0 & 0 & X^*_{M_d-1}(\omega) \end{bmatrix}^{\mathrm{H}} \begin{bmatrix} Z_0(\omega) \\ \vdots \\ Z_{M_d+L_d-2}(\omega) \end{bmatrix}}{\sum_{i=0}^{M_d} X_i^*(\omega)X_i(\omega)}, \tag{6.8}$$

where $Z_i(\omega)$ indicates the frequency-domain microphone input signal, H is the complex conjugate transposition, and L_d and M_d are the numbers of the multi-delay filters and delay search frames, respectively. In the GCC method, the frame number i of $H''_{m,i}(\omega)$ 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.

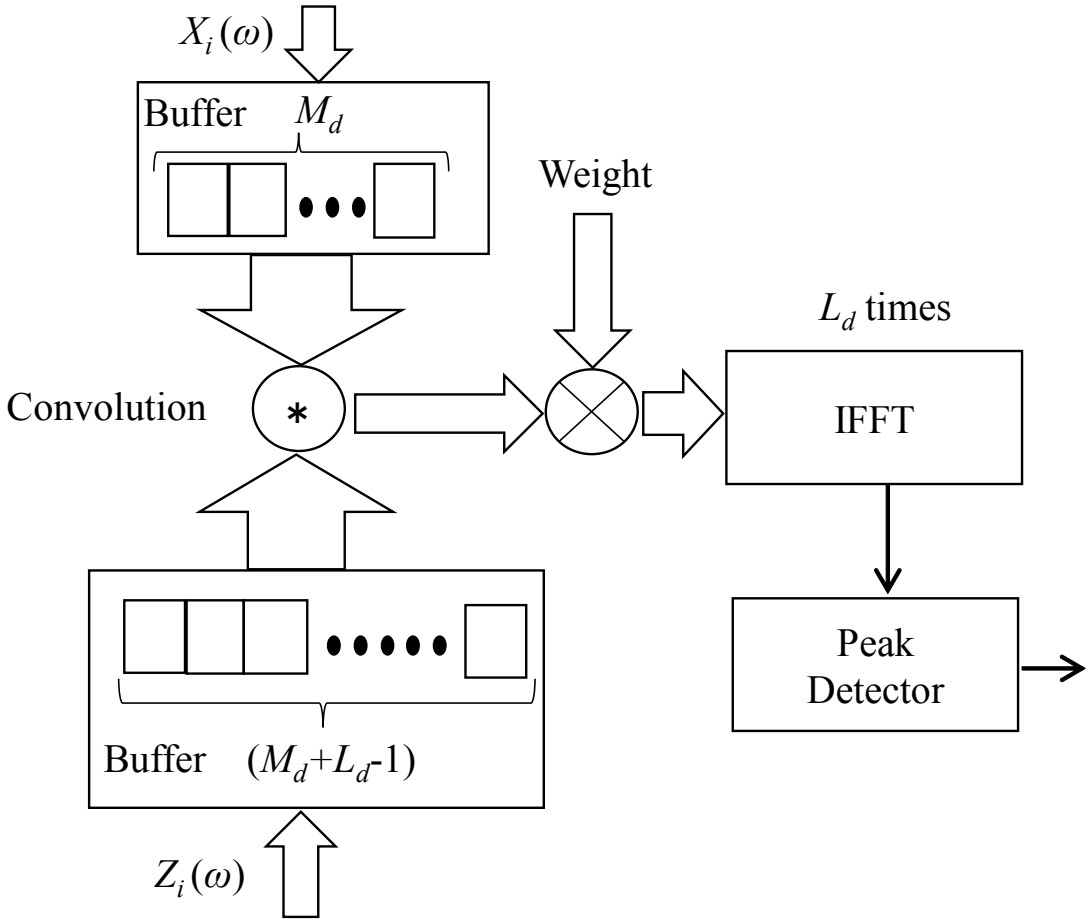


Figure 6.2. Block diagram depicting proposed AEC method developed for smartphone and tablet devices.

6.3 *Prototype Overview*

Photographs of a VoIP-phone prototype implemented with the proposed AEC method are shown in Figure 6.3. 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 [55], G.711.1 [56], and G.711.1 Annex D [57]. The sampling frequencies of these codecs 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 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 Figure 6.4. 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.

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.

Figure 6.3. Photograph of VoIP phone prototype equipped with proposed AEC method.

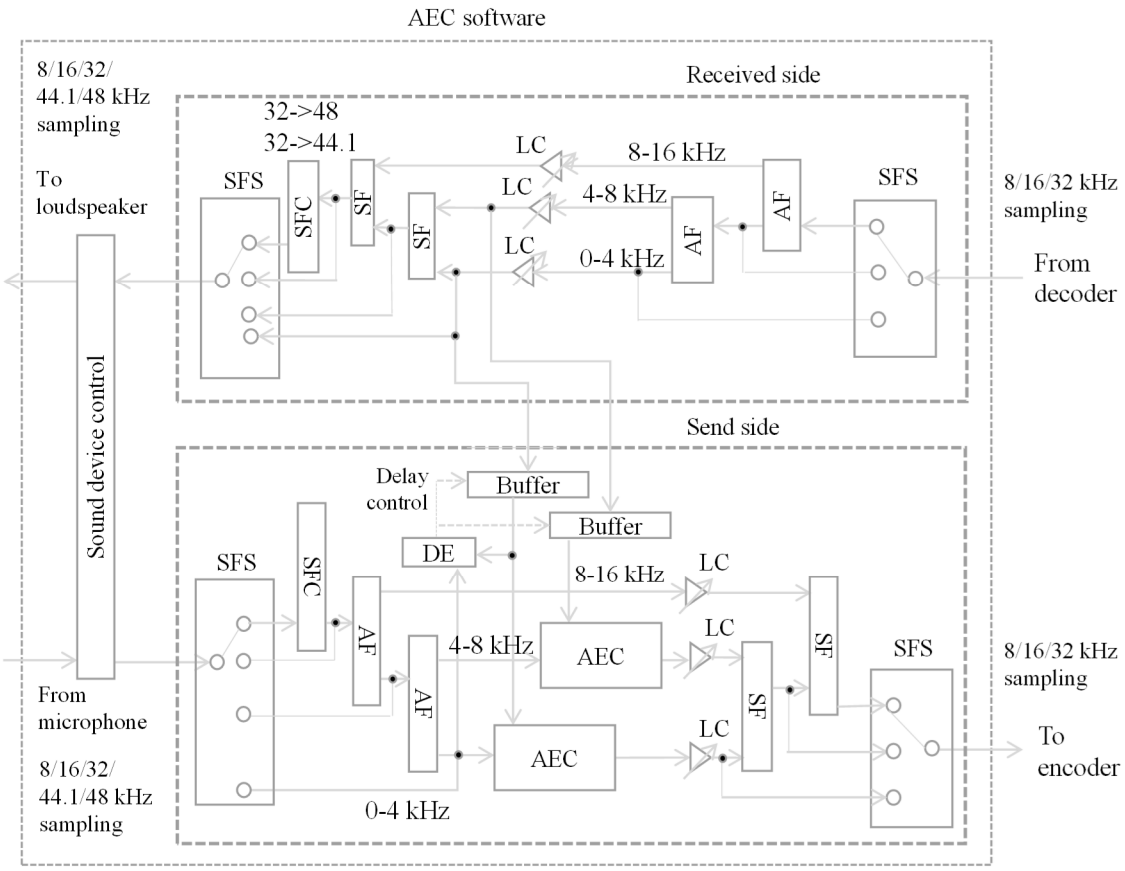


Figure 6.4. Block diagram of AEC software implemented in VoIP application.

6.4 ***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 is that in which the ADF and ER techniques were methods described in Chapter 5, and the DE technique is not employed.

6.4.1 Test conditions

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 is that in which the ADF and ER techniques were used, as described in Section 5.3, without a DE technique. The arrangement of the smartphone/tablet device and sound source is shown in Figure 6.5. 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 [58]. The tests were based on specific test signals as prescribed in ITU-T Recommendation P.501 [59]. 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.

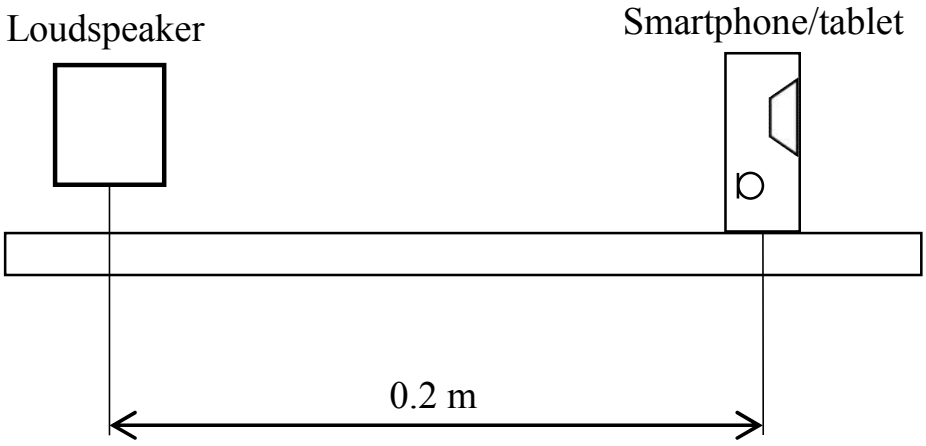


Figure 6.5. Test arrangements for objective measurements.

6.4.2 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:

$$\mathrm{SER}_{\mathrm{ST}} = 10\log_{10}\frac{\sum_{k=0}^{\vartheta}|\hat{s}_S(k)|^2}{\sum_{k=0}^{\vartheta}|\hat{s}(k)|^2}, \text{ and} \tag{6.9}$$

$$\mathrm{SER}_{\mathrm{DT}} = 10\log_{10}\frac{\sum_{k=0}^{\vartheta}|\hat{s}_S(k)|^2}{\sum_{k=0}^{\vartheta}|\hat{s}(k)-\hat{s}_S(k)|^2}, \tag{6.10}$$

where ϑ is the signal length; $\hat{s}(k)$ is the transmitted signal, and $\hat{s}_S(k)$ is the transmitted signal observed when only the near-end speaker is talking.

These results are shown in Figures 6.6 to 6.11. Figures 6.6 to 6.8 are the single-talk cases with and without background noise, and Figures 6.9 to 6.11 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.

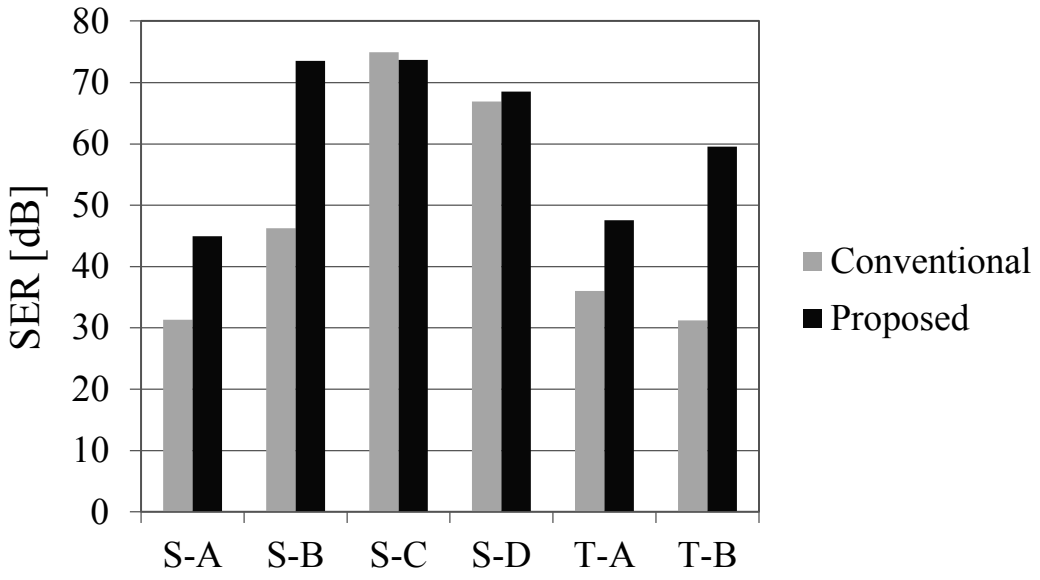


Figure 6.6. Comparison of echo-reduction performance by SER (single talk without background noise).

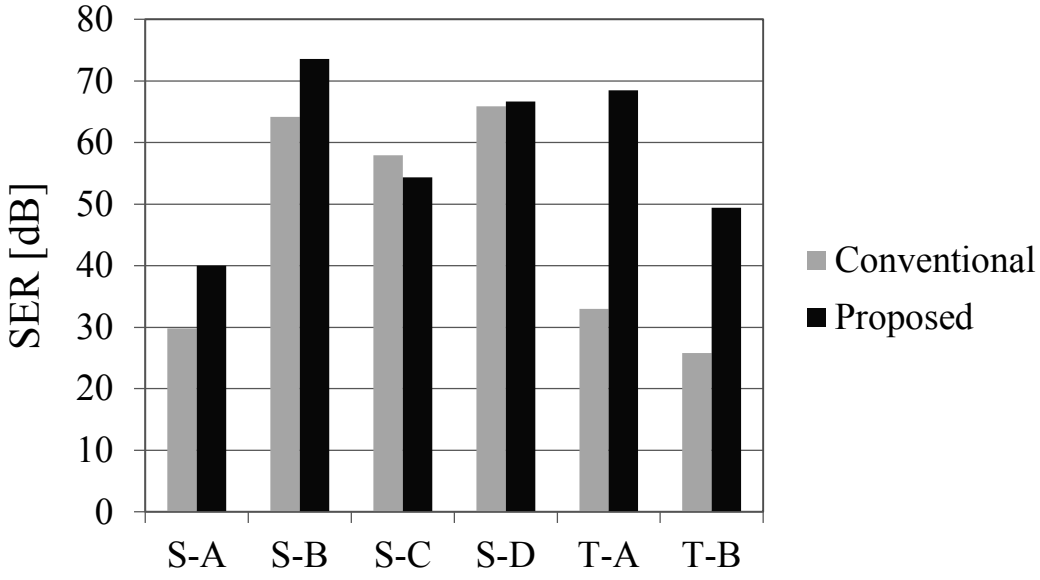


Figure 6.7. Comparison of echo-reduction performance by SER (single talk with pink noise at 20 dB SNR).

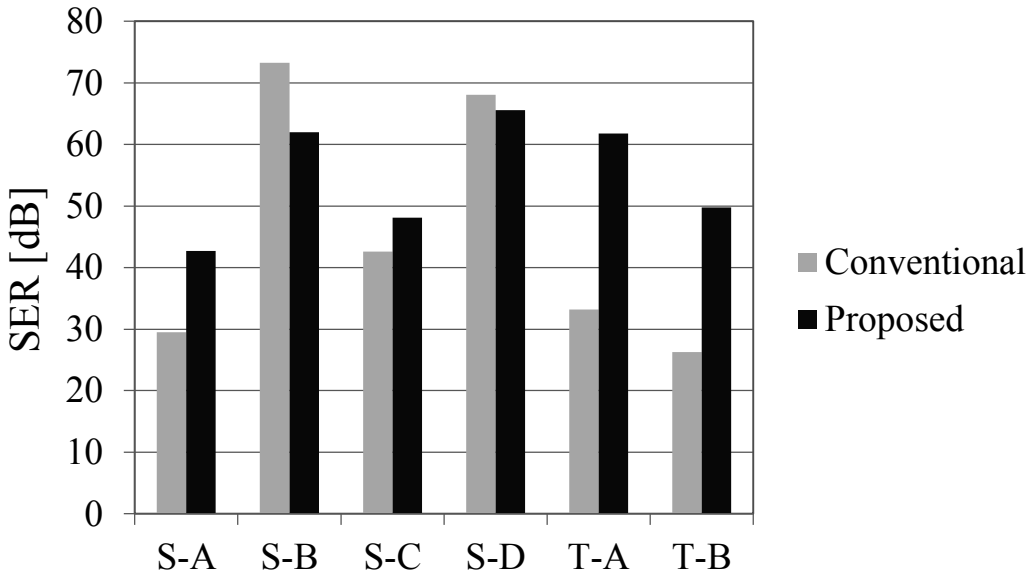


Figure 6.8. Comparison of echo reduction-performance by SER (single talk with office noise at 15 dB SNR).

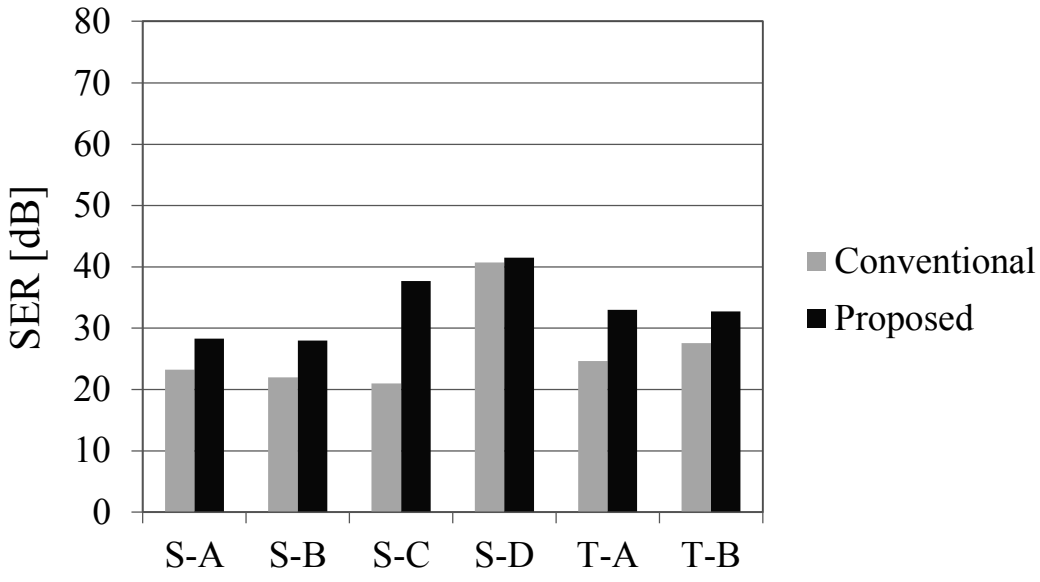


Figure 6.9. Comparison of echo reduction-performance by SER (double talk without background noise).

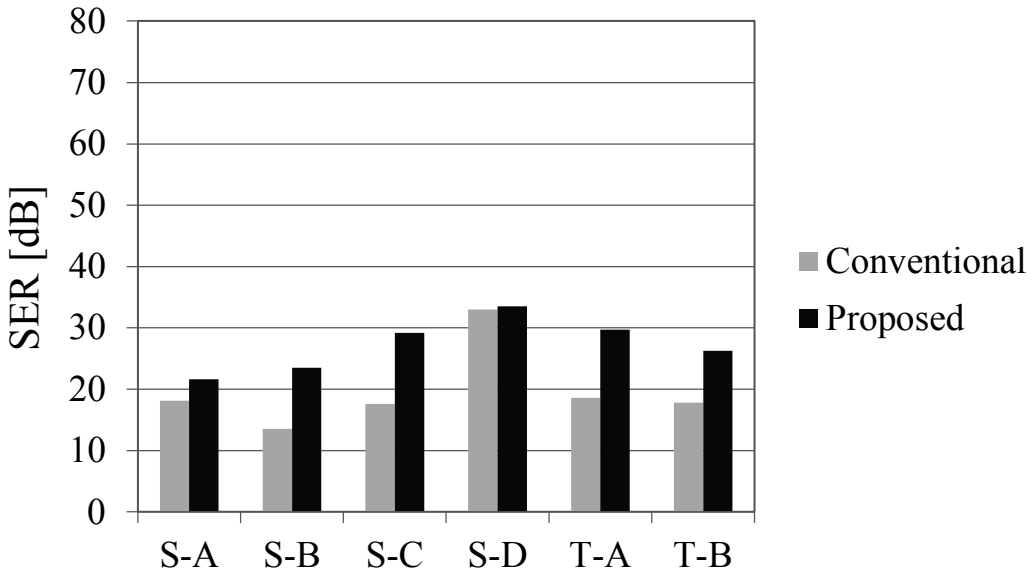


Figure 6.10. Comparison of echo reduction-performance by SER (double talk with pink noise at 20 dB SNR).

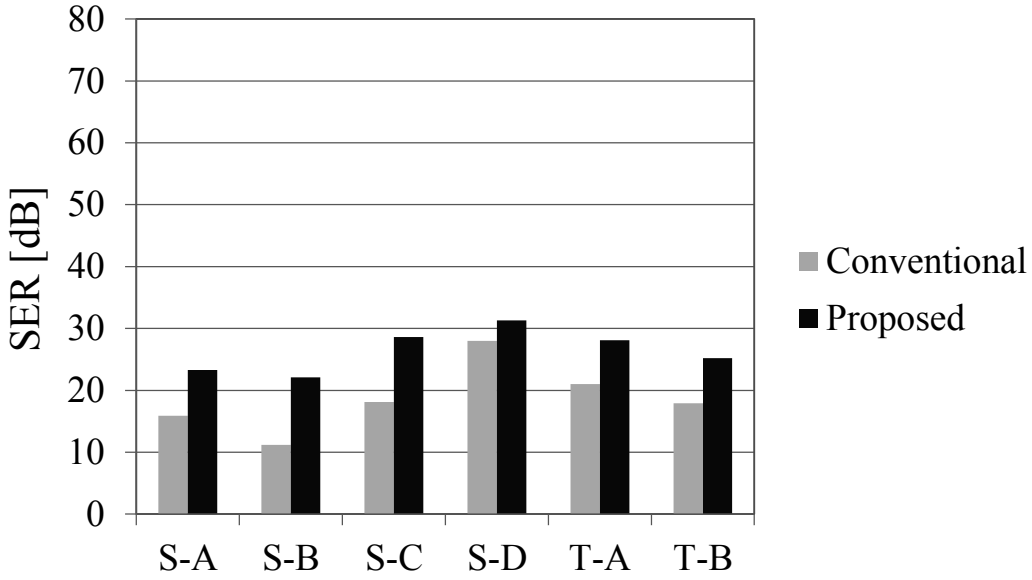


Figure 6.11. Comparison of echo reduction performance by SER (double talk with office noise at 15 dB SNR).

6.5 *Performance Evaluation*

An AEC method for VoIP hands-free application on smartphones and tablets was proposed. The proposed method can reduce the undesired acoustic echo and emphasize the target near-end speech during the double-talk periods, irrespective of the smartphone and tablet models. This method can estimate the non-linear acoustic echo caused by the loudspeaker distortion, the instantaneous residual echo variation caused by the echo-path change, and the 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 the 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 that does not consider the model-specific problems due to the difference in the acoustic characteristics of individual devices.

Chapter 7.

AENC for Videotelephony-Enabled Personal Hands-Free IP Phone

7.1 *Introduction*

This chapter presents implementation and evaluation of a proposed AENC for personal hands-free Video IP phones. This canceller has the following features: noise-robust performance, low processing delay, and low computational complexity. The AENC employs the ADF and NR methods that can effectively eliminate undesired acoustic echo and background noise included in a microphone input 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 the 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 chapter 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 the echo and noise sufficiently in the noisy environment; thus, resulting in the natural-sounding speech.

7.2 *Specifications*

An external view of a personal hands-free videophone prototype equipped with the AENC is shown in Figure 7.1. 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 Figure 7.2. The circuit uses a fixed-point DSP. Specifications of the prototype are listed in Table 7.1. 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 300-3400 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.

Figure 7.1. External view of videophone prototype.

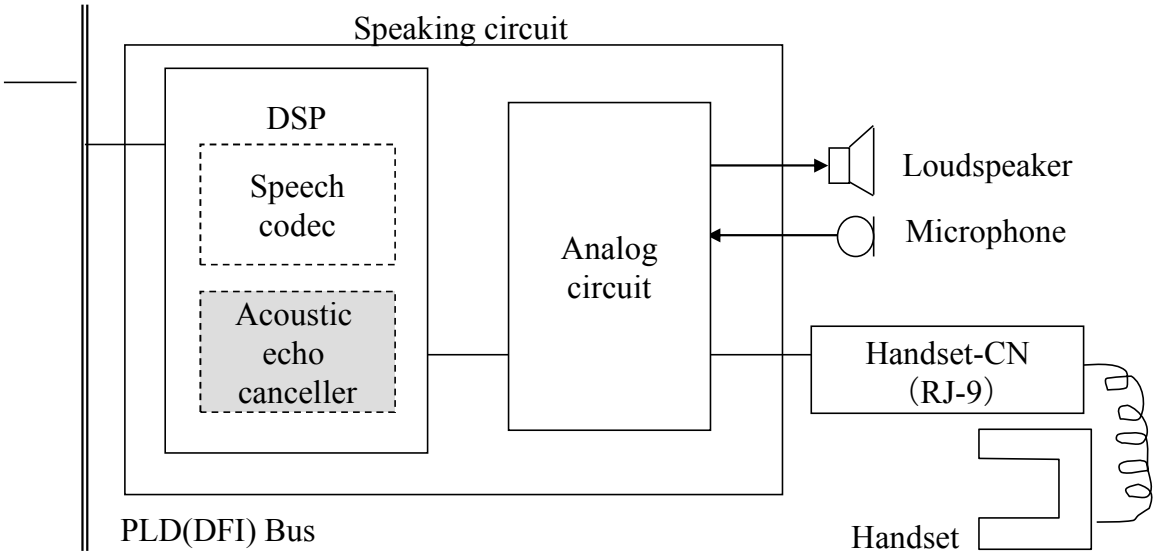


Figure 7.2. Block diagram of speaking circuit.

Table 7.1. 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)

7.3 *System Description of AENC*

A block diagram of the proposed AENC that employs the ADF and NR methods robust against the noisy environment is shown in Figure 7.3. It consists of four blocks with the following functions: ADF process, NR process, ER process, and variable-loss insertion control (VLIC) [60] process. These processes are carried out in the frequency domain. In this section, the detailed features of the ADF and NR processes are described first then the outlines of the ER and VLIC processes are summarized.

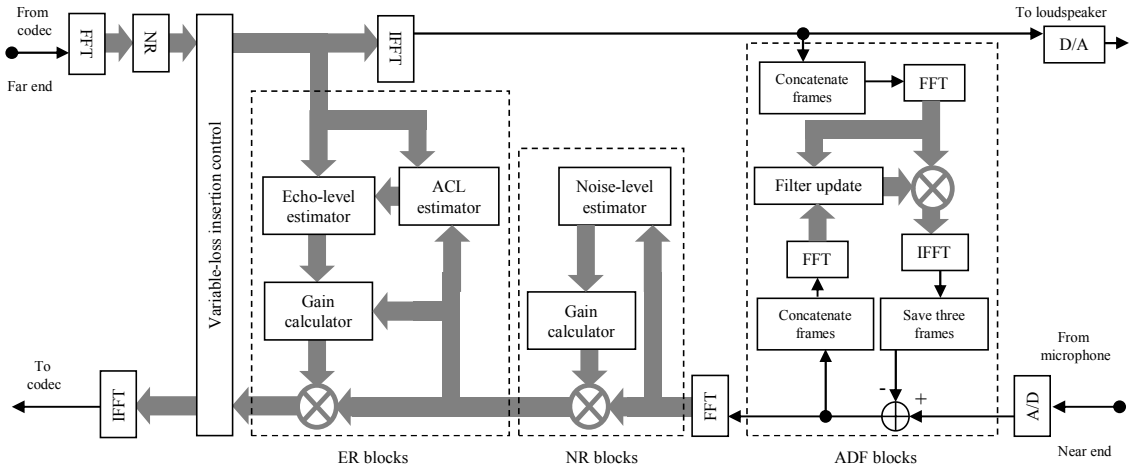


Figure 7.3. Block diagram of new AENC.

7.3.1 ADF process

A block diagram of the proposed ADF is shown in Figure 7.4. 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 $\mathbf{h}_{L_f} = [h_1, \ldots, h_{L_f}]^T$. The microphone input signal $y(k)$ is expressed as

$$y(k) = d(k) + \upsilon(k),, \tag{7.1}$$

where $\upsilon(k)$ is the signal, called the outlier, which includes the near-end speech, background noise, and so on.

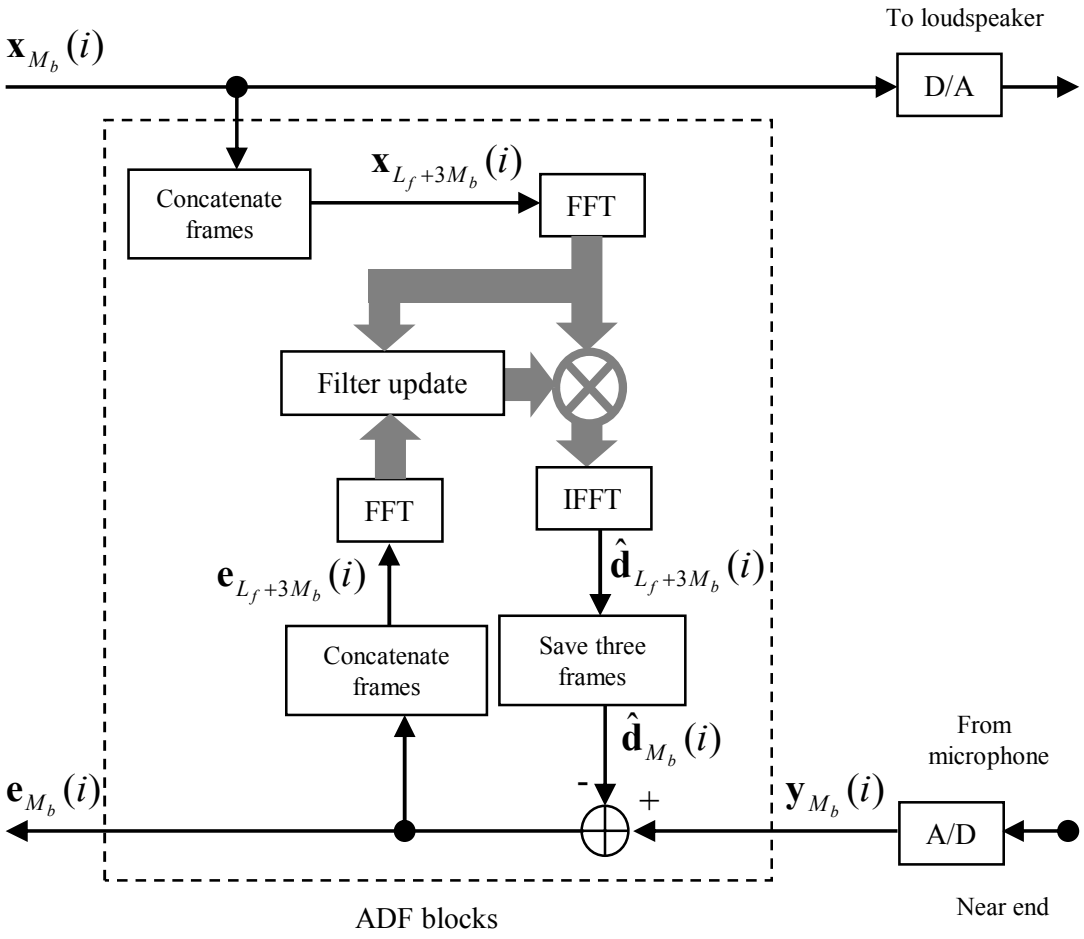


Figure 7.4. Block diagram of ADF.

In the D/A converter, the received speech signal $x(k)$ is stored as the received speech vector $\mathbf{x}_{M_b}(i) = [x(iM_b - M_b + 1), \ldots, x(iM_b)]^T$, where i is the frame index and M_b denotes the buffer size of the signal. The microphone input signal is also stored as vector $\mathbf{y}_{M_b}(i)$ with length M_b. The ADF calculates the echo-replica vector $\hat{\mathbf{d}}_{M_b}(i)$ corresponding to vector $\mathbf{y}_{M_b}(i)$ and achieves block echo cancellation as

$$\mathbf{e}_{M_b}(i) = \mathbf{y}_{M_b}(i) - \hat{\mathbf{d}}_{M_b}(i), \tag{7.2}$$

where

$$\hat{\mathbf{d}}_{M_b}(i) = \begin{bmatrix} x(iM_b - M_b + 1) & \cdots & x(iM_b - M_b - L_f + 2) \\ \vdots & \ddots & \vdots \\ x(iM_b) & \cdots & x(iM_b - L_f + 1) \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_{L_f} \end{bmatrix} \tag{7.3}$$

and $\mathbf{e}_{M_b}(i)$ is a residual signal vector and is fed back to the ADF to update the filter coefficients $w_1, \cdots, w_i, \cdots, w_{L_f}$. The filter-coefficient update is based on a frequency-domain adaptive filter algorithm [2] given by

$$W_{i+1}(\omega) = W_i(\omega) + \mu \frac{\varepsilon_i(\omega) X_i^*(\omega)}{P\left[\left|X_i(\omega)\right|^2\right]}, \tag{7.4}$$

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 μ is the step size, $*$ denotes the conjugate and $P[\cdot]$ is the smoothing function [2].

The filter-coefficient accuracy becomes important when using the ADF process. 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 process adaptively calculates the step size based on the Gaussian-Laplacian mixture assumption for the signals normalized by the reference input signal amplitude [34] 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:

$$\mu = \alpha \cdot \Psi\left(\frac{|\varepsilon_i(\omega)|}{|X_i(\omega)|}, \frac{\sigma_\omega^2}{\lambda_\omega}\right)\frac{|X_i(\omega)|}{|\varepsilon_i(\omega)| + \delta}, \tag{7.5}$$

where

$$\Psi(b, b_{\text{th}}) = \begin{cases} |b| & \text{if } |b| \le b_{\text{th}} \\ b_{\text{th}} & \text{otherwise} \end{cases}, \tag{7.6}$$

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

$$\text{Gaussian}\left(R_i(\omega)\right) = \frac{1}{\sqrt{2\pi\sigma_\omega^2}} e^{-\frac{|R_i(\omega)|^2}{\sigma_\omega^2}}, \tag{7.7}$$

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

$$\text{Laplacian}\left(V_i(\omega)\right) = \frac{1}{2\lambda_\omega} e^{-\frac{|V_i(\omega)|}{\lambda_\omega}}, \tag{7.8}$$

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

$$\frac{\varepsilon_i(\omega)}{|X_i(\omega)|} = R_i(\omega) + V_i(\omega). \tag{7.9}$$

The normalized residual echo spectra are estimated by the maximum a posteriori (MAP) estimate maximizing the posteriori probability $p\left(R_i(\omega)\left|\varepsilon_i(\omega)\right|X_i(\omega)\right|^{-1}\right)$, and used to derive the optimal step size in Equation (7.5). 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, $L_f + M_b$ 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 $\mathbf{x}_{M_b}(i-2)$, $\mathbf{x}_{M_b}(i-1)$ 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 $\hat{\mathbf{d}}_{M_b}(i)$ but also future echo-replica vectors $\hat{\mathbf{d}}_{M_b}(i+1)$ and $\hat{\mathbf{d}}_{M_b}(i+2)$ simultaneously.

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

$$\hat{\mathbf{d}}_{L_f+3M_b}(i) = \begin{bmatrix} \mathbf{c}_{L_f}(i) \\ \hat{\mathbf{d}}_{M_b}(i) \\ \hat{\mathbf{d}}_{M_b}(i+1) \\ \hat{\mathbf{d}}_{M_b}(i+2) \end{bmatrix}, \tag{7.10}$$

where $\hat{\mathbf{d}}_{M_b}(i)$, $\hat{\mathbf{d}}_{M_b}(i+1)$, and $\hat{\mathbf{d}}_{M_b}(i+2)$ correspond to the microphone input signal frames

$\mathbf{y}_{M_b}(i)$, $\mathbf{y}_{M_b}(i+1)$, and $\mathbf{y}_{M_b}(i+2)$, respectively, and $L_f \times 1$ vector $\mathbf{c}_{L_f}(i)$ is unimportant because it is not used here. Thus, by waiting for $\mathbf{y}_{M_b}(i+2)$ until the $(i+2)$-th frame, the concatenated error vector

$$\mathbf{e}_{L_f+3M_b}(i) = \begin{bmatrix} \mathbf{0}_{L_f} \\ \mathbf{e}_{M_b}(i) \\ \mathbf{e}_{M_b}(i+1) \\ \mathbf{e}_{M_b}(i+2) \end{bmatrix} \tag{7.11}$$

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.

7.3.2 NR process

A block diagram of the proposed NR process is shown in Figure 7.5. 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:

$$\hat{S}_i(\omega) = G_i(\omega) Z_i(\omega), \tag{7.12}$$

where $\hat{S}_i(\omega)$ is the estimate of $S_i(\omega)$. The Wiener-filtering-based [22] gain is calculated as

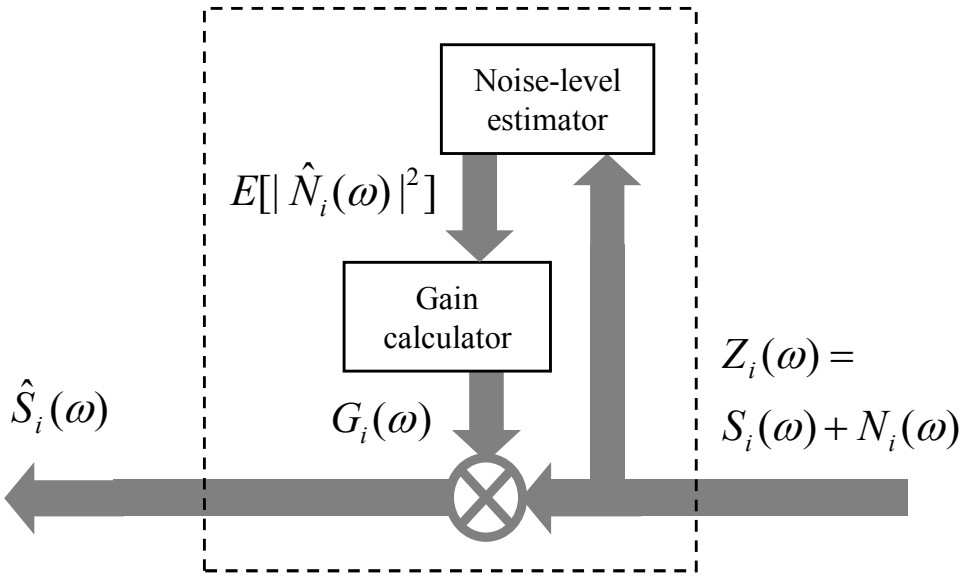


Figure 7.5. Block diagram of NR process.

$$G_i(\omega) = \frac{|Z_i(\omega)|^2 - E[|\hat{N}_i(\omega)|^2]}{|Z_i(\omega)|^2}, \tag{7.13}$$

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

$$E[|N_i(\omega)|^2] = \eta_i(\omega) \cdot E[|Z_i(\omega)|^2], \tag{7.14}$$

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

$$\eta_i(\omega) = \frac{E[|N_i(\omega)|^2]}{E[|Z_i(\omega)|^2]}. \tag{7.15}$$

This study 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:

$$(E[|N_i(\omega)|])^2 \approx E[|N_i(\omega)|^2], \tag{7.16}$$

$$(E[|S_i(\omega)|])^2 << E[|S_i(\omega)|^2], \tag{7.17}$$

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

$$(E[|Z_i(\omega)|])^2 \approx E[|N_i(\omega)|^2]. \tag{7.18}$$

This method estimates the noise ratio using these assumptions as

$$\hat{\eta}_i(\omega) = \Theta\left[\frac{(E[|Z_i(\omega)|])^2}{E[|Z_i(\omega)|^2]}\right] \tag{7.19}$$

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

$$\Theta[A] = \begin{cases} 1 & \text{if } A \le 0.8 \\ 1.6667A - 0.3334 & \text{if } 0.2 < A < 0.8 \\ 0 & \text{otherwise} \end{cases} \tag{7.20}$$

This emphasis reduces the estimation error of the noise ratio caused by a small margin of errors in the assumptions of Equations (7.16) and (7.17).

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

7.3.3 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 calculate the echo-path power spectrum, $|\hat{H}_i(\omega)|^2$, based on a cross-spectrum method in Chapter 5. The echo level estimator calculates the amount of residual echo, $|\hat{D}_i(\omega)|^2$, by multiplying the reference signal by the estimate $|\hat{H}_i(\omega)|^2$. The gain calculator determines the echo-reduction gains, $G_i(\omega)$, which have to be calculated in a short period under nonstationary speech conditions as follows:

$$G_i(\omega) = \frac{|\hat{S}_i(\omega)|^2 - \tau_i(\omega)|\hat{D}_i(\omega)|^2}{|\hat{S}_i(\omega)|^2}, \tag{7.21}$$

where

$$\tau_i(\omega) = \frac{P\left[|\hat{D}_i(\omega)||\hat{S}_i(\omega)|\right]}{P\left[|\hat{D}_i(\omega)|^2\right]}. \tag{7.22}$$

The gain calculator is designed to calculate the echo-reduction gains on the assumption that a slight correlation between echo and near-end speech signals remains in short-period processing; Equation (7.19) is a simplified equation for the proposed technique in Chapter 3. An advantage of adopting this strategy is the ability to calculate the echo-reduction 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 [61].

7.4 ***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.

7.4.1 Complexity evaluation of ADF process

The optimal step-size control schemes are often proposed as the robust frequency domain ADF methods [34] [35] [36]. In this study, 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 7.3.1. This study 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 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.

7.4.2 Performance evaluation of NR process

To evaluate the performance of the proposed NR method, described in Section 7.3.2, noise reduction rates (NRRs) [62] 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 [63]. The 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 [64]. The ordinary noise-level estimation method [31] 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 7.4.1, 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 7.2. As shown in the table, NRR of the proposed method outperformed the conventional methods for all cases.

Table 7.2. 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

7.4.3 Comparison of noise-level estimation accuracy

This study verified the effects of the combination of the ADF process 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 7.3. 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 processes; the conventional ADF calculates the step size based on the ordinary Gaussian-Gaussian mixture assumption [65]. “Proposed ADF + NR” is the combination of the ADF process with the noise robust step-size control and NR process. As Table 7.3 indicates, the noise-level estimation accuracy was improved by combining the proposed ADF process with the NR process 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 process.

Table 7.3. 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

7.4.4 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 [58]. 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.

Figures 7.6 and 7.7 show an example of the received and near-end speech signals. The microphone input signal is shown in Figure 7.8. 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 Figure 7.9. The send signal after the AENC is shown in Figure 7.10. This figure shows that the proposed AENC effectively suppresses echo and background noise and the near-end speech signal is hardly distorted.

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) [66]. 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.

The overall performances of the proposed AENC under various noise conditions are summarized in Table 7.4. 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 [63]. 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.

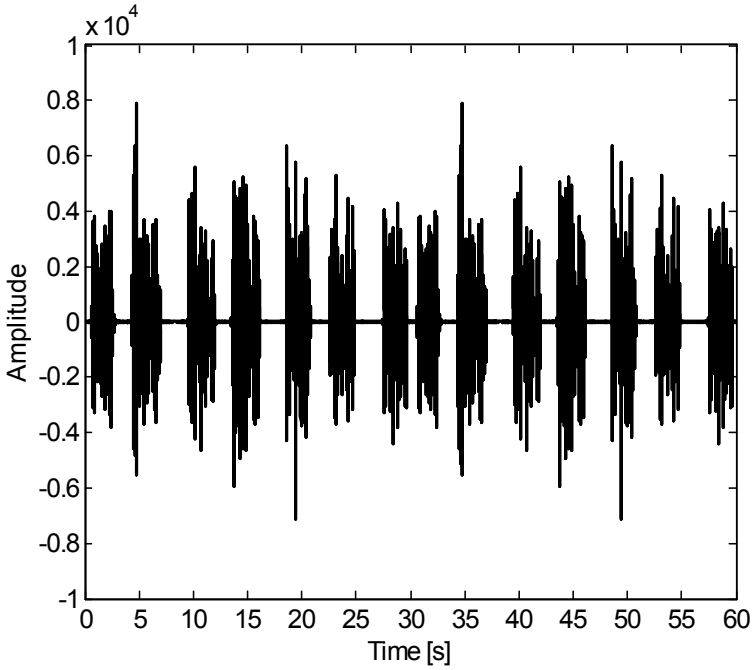


Figure 7.6. Received speech signal (male).

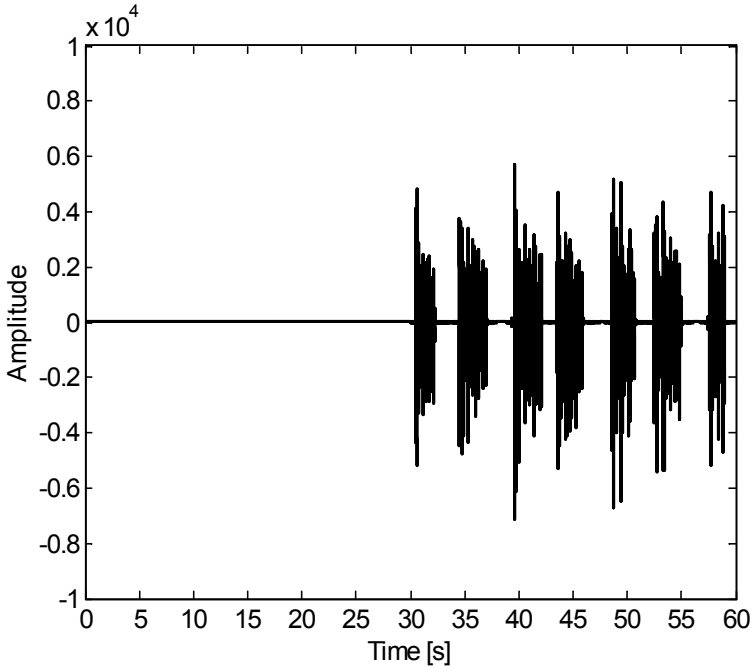


Figure 7.7. Near-end speech signal (female).

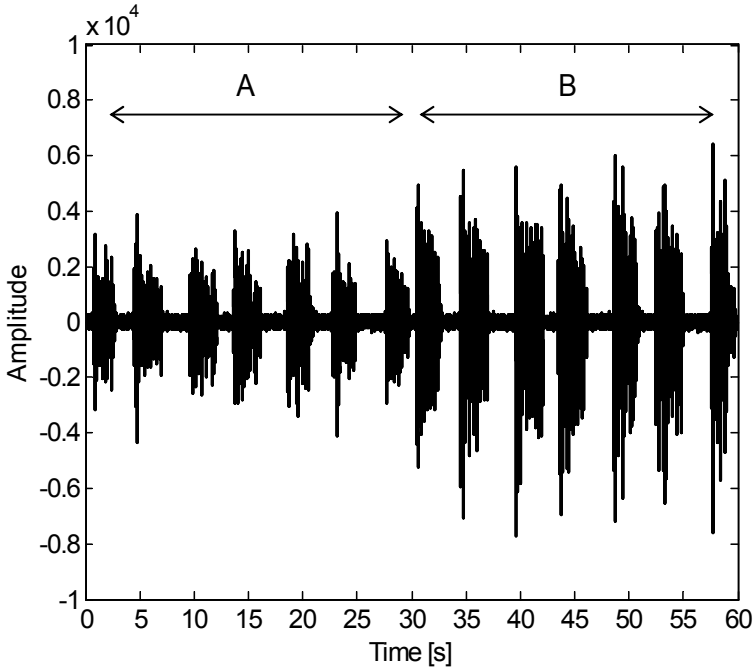


Figure 7.8. Microphone input signal. Period A: echo and noise signals during single-talk situation, period B: double-talk situation.

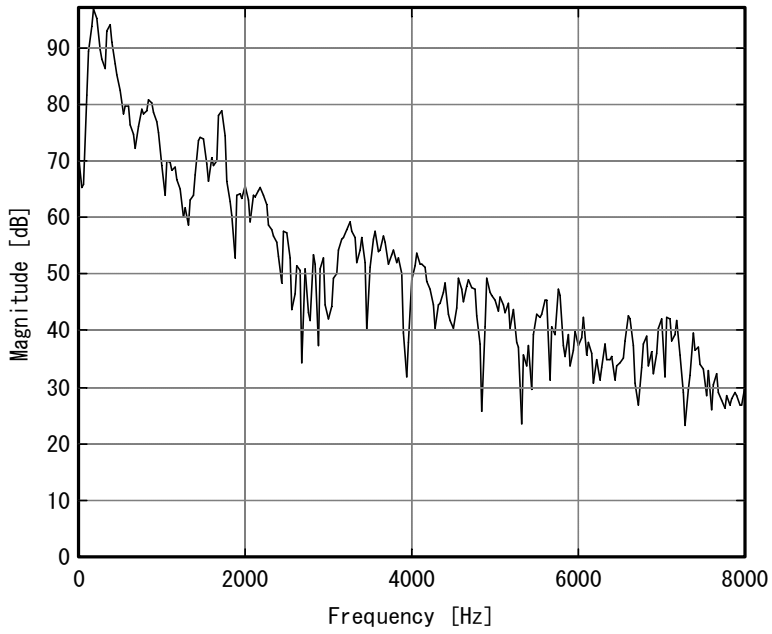


Figure 7.9. Spectrum of background noise.

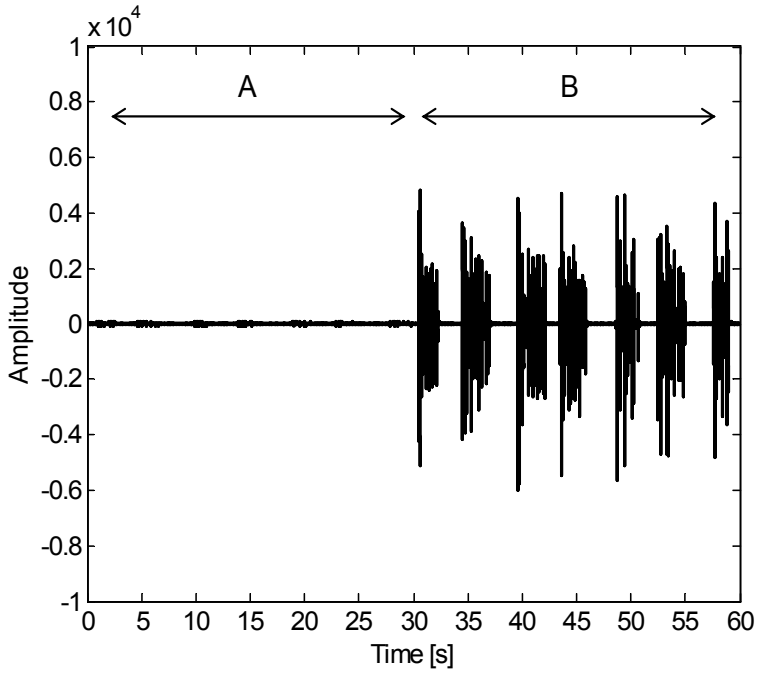


Figure 7.10. Send signal. Period A: single-talk situation, period B: double-talk situation.

Table 7.4. Overall performance 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

7.5 *Conclusion*

This study 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 study 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.

Chapter 8.

Conclusions

This dissertation focused on the robustness against the double talk for the acoustic echo canceller (AEC) and introduced high-performance acoustic echo control algorithms to improve the quality of telecommunication systems.

On the acoustic echo control strategy, the development of the double-talk-robust echo reduction (ER) process was one of the main targets. This study could develop the novel ER process robust against the double talk in the estimations of the echo-path power spectrum, the echo-reduction gain, and the late echo components. The echo-path power spectrum estimation algorithm was an echo-path-change robust algorithm that estimates the echo-path power spectrum in time and frequency spectral domains; this algorithm could achieved the high tracking performance and accuracy of the echo-path power spectrum. The echo-reduction gain estimation algorithm was a novel estimation algorithm that solves a least mean square error of the Wiener filtering (WF) method while taking into account the cross-spectral term of the signals; thereby, this algorithm obtained a better echo-reduction gain than that of the conventional WF method. The late echo component estimation algorithm was a novel estimation algorithm that accurately estimated the echo power spectrum corresponding to the early impulse response and the late echo components resulting from reverberation beyond a length of fast Fourier transform (FFT) block; this algorithm estimated the echo power spectrum by assuming a finite nonnegative convolution model.

This study developed the AEC devices and the application software where the proposed estimation algorithms of the echo-path power spectrum and the echo-reduction gain were

implemented. The developed AEC unit was combined with a videoconferencing system and used in hands-free telecommunication; its frequency band of the audio signal in this AEC unit was supported up to compact disc (CD)-quality, i.e. 20-kHz wideband and delivered natural-sounding speech. The developed software for voice over internet protocol (VoIP) hands-free phone application on smartphone and tablet devices automatically could tailor its performance to the acoustic characteristics of individual smartphone and tablet devices, and reduced the influence due to the difference in the acoustic characteristics of individual devices. The developed video phone employed the noise-robust adaptive filter (ADF) and noise reduction (NR) processes and could maintain the acoustic echo and noise cancelling performance in a noisy environment such as the open-plan office.

Future works include the following interesting studies. The concept of the double-talk robust ER process can be extend to multi-channel algorithm such as stereo AEC. Influence of disturbance such as the near-end speech on the estimation accuracy of the echo power spectrum may be investigated more analytically by specifying statistical models of the signal. The problem of the residual echo power estimation may be extended as the estimation problem of the optimal solution including nonlinear components such as the loudspeaker distortion by using reasonable approximations.

List of Publications and Presentations

Journal Article

1. *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 Processing (in Japanese), vol. 17, no. 5, pp. 167-177, Sep. 2013.
2. M. Fukui, K. Tsutsumi, S. Sasaki, Y. Hiwasaki, and Y. Haneda, "Multilayer coding using selection of modes based on sound source characteristics," Journal of the Acoustical Society of Japan (in Japanese), vol. 69, no. 12, pp. 623-631, June 2013.
3. 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 Processing (in Japanese), vol. 18, no. 1, pp. 17-28, Jan. 2014.
4. M. Fukui, S. Sasaki, Y. Hiwasaki, K. Tsutsumi, S. Kurihara, H. Ohmuro, and Y. Haneda, "Adaptive spectral masking of AVQ coding and sparseness detection for ITU-T G.711.1 Annex D and G.722 Annex B standards," IEICE Trans. Information and Systems, vol. E97-D, no. 5, May 2014.
5. *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.

6. *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.
7. *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.
8. M. Fukui, K. Kobayashi, S. Shimauchi, Y. Hioka, and H. Ohmuro, "Low-complexity dereverberation for hands-free audio conferencing unit," IEEE Trans. Consumer Electron., vol. 61, no. 4, pp. 539-545, Nov. 2015.
9. *M. Fukui, S. Shimauchi, Y. Hioka, A. Nakagawa, and Y. Haneda, "Acoustic echo and noise canceller for personal hands-free video IP phone," IEEE Trans. Consumer Electron., vol. 62, no. 4, pp. 454-462, Nov. 2016.
10. M. Fukui, T. Watanabe, and M. Kanazawa, "Sound source separation for plural passenger speech recognition in smart mobility system," IEEE Trans. Consumer Electron., vol. 64, no. 3, pp. 399-405, Aug. 2018.

Reviews

1. M. Fukui, S. Sasaki, Y. Hiwasaki, and S. Kurihara, "International standard for 14 kHz band speech codec: ITU-T G.711.1 Annex D," NTT Technical Journal (in Japanese), vol. 24, no. 9, pp. 74-77, Sep. 2012.

International Conferences

1. M. Fukui, Y. Yamashita, H. Saruwatari, and K. Shikano, "Pitch recognition for successive musical cords using short-term estimation method for peak frequencies," Proc. International Symposium on Musical Acoustics, 3-S2-9, pp. 285-288, Nara, Japan, Mar. 2004.

2. 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, pp. 1-4, Seattle, USA, Sept. 2008.
3. M. Fukui, A. Nakagawa, S. Shimauchi, Y. Haneda, and A. Kataoka, "20-kHz frequency-range acoustic echo canceller for high-quality TV conferencing," Proc. IEEE International Conference on Consumer Electronics, 5.1-2, pp. 1-2, Las Vegas, USA, Jan. 2009.
4. M. Fukui, A. Nakagawa, S. Shimauchi, Y. Haneda, T. Somei, H. Fuyuki, and A. Kataoka, "New acoustic echo canceller for videotelephony-enabled wideband business phone," Proc. IEEE 13th International Symposium on Consumer Electronics, pp. 228-232, Las Kyoto, Japan, May 2009.
5. M. Fukui, S Sasaki, Y Hiwasaki, K Sachiko, and Y Haneda, "Dual-mode AVQ coding based on spectral masking and sparseness detection for ITU-T G. 711.1/G. 722 super-wideband extensions," Proc. 12th Annual Conference of the International Speech Communication Associations, pp. 2525-2528, Florence, Italy, Aug. 2011.
6. L. Miao, Z. Liu, C. Hu, V. Eksler, S. Ragot, C. Lamblin, B. Kövesi, J. Sung, M. Fukui, S. Sasaki, and Y. Hiwasaki, "G.711.1 Annex D and G.722 Annex B - new ITU-T superwideband codecs," Proc. IEEE International Conference on Acoust., Speech and Signal Processing, pp. 5232-5235, Prague, Czech Republic, May 2011.
7. M. Fukui, A. Nakagawa, S. Shimauchi, Y. Haneda, and A. Kataoka, "Echo reduction using Wiener gains considering short-time correlation between echo and near-end speech," Proc. International Workshop on Acoustic Signal Enhancement, pp. 1-4, Aachen, Germany, Sep. 2012.
8. *M. Fukui, S. Shimauchi, Y. Hioka, H. Ohmuro, and Y. Haneda, "Accurate acoustic echo reduction with residual echo power estimation for long reverberation," Proc. 134th Convention of Audio Engineering Society, pp. 132–139, Rome, Italy, May 2013.
9. M. Fukui, S. Shimauchi, Y. Hioka, H. Ohmuro, Y. Haneda, "Acoustic echo reduction robust against echo-path change with instant echo-power-level adjustment," Proc. European Signal Processing Conference, pp. 1-5, Marrakech, Morocco, Sep, 2013.
10. 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, pp.133-134, Las Vegas, USA, Jan. 2014.

11. M. Fukui, K. Kobayashi, S. Shimauchi, Y. Hioka, and H. Ohmuro, "Hands-free audio conferencing unit with low-complexity dereverberation," Proc. IEEE International Conference on Consumer Electronics, pp.128-129, Las Vegas, USA, Jan. 2015.
12. M. Fukui, Y. Wakisaka, T. Watanabe, and M. Kanazawa, "Sound source separation for plural passenger speech recognition in smart mobility system," Proc. IEEE International Conference on Consumer Electronics, pp.1-2, Las Vegas, USA, Jan. 2018.

Technical Reports

1. M. Fukui, Y. Yamashita, H. Saruwatari, and K. Shikano, "Large pitch extraction for successive musical cord using short-term peak frequency estimation method," Technical Meeting of ASJ Musical Acoust. (in Japanese), MA2004-6, pp. 7-12, Kyoto, Japan, June 2004.
2. M. Fukui, S. Shimauchi, Y. Haneda, and A. Kataoka, "Echo path change robust acoustic coupling estimation for echo suppression," 21st SIP Symposium (in Japanese), C8-2, Kyoto, Japan, Nov. 2006.
3. M. Fukui, S. Shimauchi, A. Nakagawa, Y. Haneda, and A. Kataoka, "Noise power estimation based on noise ratio in signal and its noise reduction performance," IEICE Technical Report (in Japanese), vol. 107, no. 532, EA2007-126, pp. 85-90, Tokyo, Japan, Mar. 2008.
4. M. Fukui, K. Tsutsumi, S. Sasaki, Y. Hiwasaki, and Y. Haneda, " Multi-layer Speech Coding Using Combination of Modes Based on Sparseness of Input Spectrum—ITU-T G.722/G.711.1 Super Wideband Extension Candidate—," IEICE Technical Report (in Japanese), vol. 109, no. 375, EA2009-137, pp. 279-284, Kyoto, Japan, Jan. 2010.
5. S. Kurihara, S. Shimauchi, M. Fukui, and N. Harada, "Quality of experience assessment in hands-free communications—Study on subjective evaluation method consistent with PESQ measure—," IEICE Technical Report (in Japanese), vol. 117, no. 386, CQ2017-96, pp. 63-68, Tokyo, Japan, Jan. 2018.

6. M. Fukui, S. Shimauchi, and Y. Hioka, "Convolutive residual echo power estimation for addressing long reverberation-time problem," IEICE Technical Report (in Japanese), vol. 117, no. 515, EA2017-148, pp. 255-260, Okinawa, Japan, Mar. 2018.

Domestic Conferences

1. M. Fukui and Y. Yamashita, "High temporal resolution pitch recognition of musical sounds based on defference informantion on spectrum peak between adjuacent windows," Proc. Spring Meeting of the Acoust. Society of Japan (in Japanese), 3-7-1, pp. 833-834, Mar. 2003.
2. M. Fukui, Y. Yamashita, H. Saruwatari, and K. Shikano, "Pitch recognition for successive musical cords using shor-term estimation method for peak frequencies and peak clustering," Proc. Spring Meeting of the Acoust. Society of Japan (in Japanese), 2-9-5, pp. 693-694, Mar. 2004.
3. M. Fukui, S. Shimauchi, Y. Haneda, and A. Kataoka, "A multi-channel echo reduction by send signal selection in frequency domain," Proc. Spring Meeting of the Acoust. Society of Japan (in Japanese), 3-Q-3, pp. 511-512, Mar. 2005.
4. M. Fukui, S. Shimauchi, Y. Haneda, and A. Kataoka, "Evaluation of time-frequency domain acoustic coupling estimation," Proc. Spring Meeting of the Acoust. Society of Japan (in Japanese), 3-Q-33, pp. 715-716, Mar. 2007.
5. S. Terauchi, M. Fukui, and Y. Yamashita, "Pitch recognition of musical data based on peak selection by a decision tree," Proc. Autummn Meeting of the Acoust. Society of Japan (in Japanese), 1-1-10, pp. 831-832, Sep. 2007.
6. M. Fukui, S. Shimauchi, Y. Haneda, and A. Kataoka, "Accurate noise estimation based on time variability in input signal," Proc. Autumn Meeting of the Acoust. Society of Japan (in Japanese), 3-7-5, pp. 725-726, Sep. 2007.
7. M. Fukui, A. Nakagawa, Y. Haneda, and A. Kataoka, "Gain estimation for echo reduction with less speech distortion based on input-output cross correlation," Proc. Spring Meeting of the Acoust. Society of Japan (in Japanese), 3-P-7, pp. 851-852, Mar. 2008.

8. S. Terauchi, M. Fukui, and Y. Yamashita, "Pitch recognition of musical data based on a decision tree technique using spectrum and cepstrum features," Proc. Autumn Meeting of the Acoust. Society of Japan (in Japanese), 1-9-19, pp. 893-896, Sep. 2008.
9. M. Fukui, A. Nakagawa, S. Shimauchi, Y. Haneda, and A. Kataoka, "Accurate echo power estimation for echo reduction," Proc. IEICE General Conference (in Japanese), A-4-18, pp. 122, Mar. 2009.
10. S. Saito, M. Fukui, A. Nakagawa, and Y. Haneda, "Nonlinear echo canceller with filter-switching algorithm depending on reference signal vectors," Proc. Spring Meeting of the Acoust. Society of Japan (in Japanese), 1-4-1, pp. 609-612, Mar. 2009.
11. K. Tsutsumi, M. Fukui, S. Kurihara, S. Sasaki, Y. Hiwasaki, and Y. Haneda, "On performance of a candidate algorithm for joint super wideband extension to ITU-T G.722/G.711.1," Proc. Autummn Meeting of the Acoust. Society of Japan (in Japanese), 1-2-18, pp. 275-276, Sep. 2009.
12. M. Fukui, K. Tsutsumi, S. Sasaki, Y. Hiwasaki, and Y. Haneda, "8-14 kHz band speech codec based on mode selection considering characteristics of sound source—ITU-T G.722/G.711.1 super wideband extension candidate—," Proc. Autummn Meeting of the Acoust. Society of Japan (in Japanese), 1-2-22, pp. 283-284, Sep. 2009.
13. K. Tsutsumi, M. Fukui, S. Kurihara, S. Sasaki, Y. Hiwasaki, and Y. Haneda, "A method for stereo speech rendering on multi point teleconference : Partial-mixing with scalable speech codecs," Proc. Spring Meeting of the Acoust. Society of Japan (in Japanese), 3-7-14, pp. 387-388, Mar. 2010.
14. M. Fukui, S. Kurihara, S. Sasaki, Y. Hiwasaki, and Y. Haneda, "8-14 kHz band speech codec using mode selection in ITU-T G.722/G.711.1 super wideband extension candidate," Proc. Autummn Meeting of the Acoust. Society of Japan (in Japanese), 1-1-6, pp. 217-218, Sep. 2010.
15. K. Kobayashi, S. Shimauchi, M. Fukui, and H. Ohmuro, "A study of delay estimation method with low complexity and high accuracy for echo canceller," Proc. Spring Meeting of the Acoust. Society of Japan (in Japanese), 1-P-6, pp. 831-834, Mar. 2013.
16. S. Kurihara, S. Shimauchi, M. Fukui, and N. Harada, "Study on quality of experience assessment in hands-free communications," Proc. IEICE General Conference (in Japanese), B-11-7, Mar. 2018.

Bibliography

[1] J. Nagumo and A. Noda, "A learning method for system identification," IEEE Trans. Autom. Control, vol. 12, no. 3, pp. 282-297, June 1967.

[2] S. Haykin, "Adaptive filter theory," Prentice-Hall, Inc., pp. 365-444, New Jersey, USA, Dec. 1995.

[3] C. Breining, P. Dreiscitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt and J. Tilp, "Acoustic echo scontrol: an application of very-high-order adaptive filters," IEEE Signal Processing Magazine, vol. 16, no. 4, pp. 42–69, July 1999.

[4] 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.

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

[6] C. Avendano, "Acoustic echo suppression in the STFT domain," Proc. IEEEWorkshop on Applications of Signal Processing to Audio and Acoustics, vol. 21, no. 24, pp. 175–178, New York, USA, Oct. 2001.

[7] C. Faller and C. Tournery, "Robust acoustic echo control using a simple echo path model," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 281–284, Toulouse, France, May 2006.

[8] Y. S. Park and J. H. Chang,, "Frequency domain acoustic echo suppression based on soft decision," IEEE Signal Processing Letters, vol. 16, no. 1, pp. 53–56, Jan. 2009.

[9] Y. Tong and Y. Gu, "A modified a priori SER for acoustic echo suppression using wiener filter," Proc. IEEE International Workshop on Acoustic Signal Enhancement, pp. 1–4, Xi'an, China, Oct. 2016.

[10] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal, Processing, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

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

[12] S. Sakauchi, A. Nakagawa, Y. Haneda and A. Kataoka, "Implementing and evaluating of an audio teleconferencing terminal with noise and echo reduction," Proc. IEEE International Workshop on Acoustic Echo and Noise Control pp. 191–194, Kyoto, Japan, Sep. 2003.

[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] D. L. Duttweiler, "A twelve-channel digital echo canceler," IEEE Trans. Commun., vol. COM-26, no. 5, pp. 647-653, May 1978.

[15] H. Ye and B.-X. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," IEEE Trans. Commun., vol. 39, no. 11, pp. 1542-1545, Nov. 1991.

[16] T. Gansler, M. Hansson, C.-J. Ivarsson and G. Salomonsson, "A double-talk detector based on coherence," IEEE Trans. Commun., vol. 44, no. 11, pp. 1421-1427, Nov. 1996.

[17] J. Benesty, D. R. Morgan and J. H. Cho, "A new class of doubletalk detectors based on cross-correlation," IEEE Trans. Speech and Audio, vol. 8, no. 2, pp. 168-172, Mar. 2000.

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

[19] R. L. B. Jeannes, P. Scalart, G. Fauco'n and C. Beaugeant, "Combined noise and echo reduction in handsfree systems: a survey," IEEE Trans. Speech Audio Processing, vol. 9, no. 8, pp. 808–820, Nov. 2001.

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

[21] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 629-632, Atlanta, USA, May 1996.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-32, no. 6, Dec. 1984.

[23] J.-H. Chang, H.-G. Kim and S. Kang, "Residual echo reduction based on MMSE estimator in acoustic echo canceller," IEICE Elecronics Express, vol. 4, no. 24, pp. 762-767, Dec. 2007.

[24] A. Favrot, C. Faller and F. Kuech, "Modeling late reverberation in acoustic echo suppression," Proc. IEEE International Workshop on Acoustic Signal Enhancement, pp. 1–4, Aachen, Germany, Sep. 2012.

[25] S. Leglaive, R. Badeau and G. Richard, "Autoregressive moving average modeling of late reverberation in the frequency domain," Proc. European Signal Processing Conference, pp. 1478–1482, Budapest, Hungary, Aug. 2016.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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, pp. 4665-4668, Taipei, Taiwan, Apr. 2009.

[31] 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 Processing, vol. 20, no. 4, pp. 1383-1393, May 2012.

[32] 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 Processing, vol. 21, no. 6, pp. 1240-1250, June 2013.

[33] 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, pp. 380-384, Shanghai, China, Mar. 2016.

[34] 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, June 2008.

[35] T. S. Wada and B.-H. Juang, "Enhancement of residual echo for robust acoustic echo cancellation," IEEE Trans. Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 175-189, Jan. 2012.

[36] J. M. G.-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 Processing, vol. 22, no. 12, pp. 2074-2086, Dec. 2014.

[37] 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 Processing (in Japanese), vol. 18, no. 1, pp. 17-28, Jan. 2014.

[38] ISO 3382-2, "Acoustics — measurement of room acoustic parameters — part 2: reverberation time in ordinary rooms," International Organization for Standardization, Geneva, Switzerland, June 2008.

[39] M. Aoki, K. Furuya and A. Kataoka, "Improvement of "SAFIA" source separation method under reverberant conditions," IEICE Trans. Fundamentals (in Japanese), vol. J87-A, no. 9, pp. 1171-1186, Sep. 2004.

[40] S. Sakauchi, Y. Haneda and A. Kataoka, "Gain emphasis method for echo reduction based on a short-time spectral amplitude estimation," IEICE Trans. Fundamentals (in Japanese), vol. J88-A, no. 6, pp. 695-703, June 2005.

[41] A. H. Gray Jr. and J. D. Markel, "Distance measures for speech processing," IEEE Trans. Acoust. Speech Signal Processing, vol. 24, no. 5, pp. 380-391, Oct. 1976.

[42] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," International Telecommunications Union, Geneva, Switzerland, Jan. 2003.

[43] H. Kanai, T. Hori, N. Chubachi and T. Ono, "Delayed block transfer function in the frequency domain," IEEE Trans. Signal Processing, vol. 42, no. 7, pp. 1669-1684, July 1994.

[44] M. Fukui, K. Kobayashi, S. Suehiro, Y. Hioka and H. Ohmuro, "Low-complexity dereverberation for hands-free audio conferencing unit," IEEE Trans. Consumer Electron., vol. 61, no. 4, pp. 539-545, Nov. 2015.

[45] S. Qiao, "Fast adaptive RLS algorithms: a generalized inverse approach and analysis," IEEE Trans. Signal Processing, vol. 39, no. 6, pp. 1455-1459, June 1991.

[46] S. Makino and Y. Kaneda, "A new RLS algorithm based on the variation characteristics of a room impulse response," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. 373-376, New York, USA, Apr. 1994.

[47] ITU-T Recommendation P.34, "Transmission performance of hands-free telephones," International Telecommunication Union, Geneva, Switzerland, Mar. 1993.

[48] ITU-T Recommendation G.168, "Digital network echo cancellers," International Telecommunications Union, Geneva, Switzerland, Apr. 1997.

[49] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," Signal Processing, vol. 80, no.9, pp.1747-1760, Sep. 2000.

[50] 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, pp. 1-4, Aachen, German, Sep. 2012.

[51] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," IEEE Signal Processing Magazine, vol. 9, no. 1, pp. 14-37, Jan. 1992.

[52] 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, pp. 1–5, Marrakech, Morocco, Sep. 2013.

[53] C. H. Knapp and C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[54] J.-S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 38, no. 2, pp. 373-376, Feb. 1990.

[55] ITU-T Recommendation G.711, "Pulse code modulation (PCM) of voice frequencies," International Telecommunications Union, Geneva, Switzerland, Nov. 1988.

[56] ITU-T Recommendation G.711.1, "Wideband embedded extension for G.711 pulse code modulation," International Telecommunications Union, Geneva, Switzerland, Mar. 2008.

[57] ITU-T Recommendation G.711.1 Annex D, "New Annex D with superwideband extension," International Telecommunications Union, Geneva, Switzerland, Nov. 2010.

[58] ITU-T Recommendation P.340, "Transmission characteristics and speech quality parameters of hands-free terminals," International Telecommunications Union, Geneva, Switzerland, May 2000.

[59] ITU-T Recommendation P.501, "Test signals for use in telephonometry," International Telecommunications Union, Geneva, Switzerland, Aug. 1996.

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

[61] K. Kobayashi, K. Furuya, Y. Haneda and A. Kataoka, "Howling canceller based on sparseness of speech for hands-free system," IEICE Technical Report (in Japanese), vol. 107, no. 170, EA2007-35, pp. 1-6, July 2007.

[62] H. Saruwatari, S. Kurita and K. Takeda, "Blind source separation combining frequency-domain ICA and beamforming," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 2733-2736, Salt Lake City, USA, Mar. 2001.

[63] NTT-AT Database, "Ambient noise database CD-ROM," NTT Advanced Technology Corporation, Kanagawa, Japan.

[64] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," International Telecommunications Union, Geneva, Switzerland, Aug. 1996.

[65] K. Fujii and J. Ohga, "Optimum adjustment of step gain in learning identification algorithm," IEICE Trans. Fundamentals (in Japanese), vol. J75-A, no. 6, pp. 975-982, June 1992.

[66] 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, Switzerland, Feb. 2001.