공학박사 학위논문

Enhanced Acoustic Echo Suppression Techniques Based on Spectro-Temporal Correlations

주파수 및 시간적 상관관계에 기반한 음향학적 에코 억제 기법

2016년 8월

서울대학교 대학원

전기·컴퓨터공학부

이 철 민

Abstract

In the past decades, a number of approaches have been dedicated to acoustic echo cancellation and suppression which reduce the negative effects of acoustic echo, namely the acoustic coupling between the loudspeaker and microphone in a room. In particular, the increasing use of full-duplex telecommunication systems has led to the requirement of faster and more reliable acoustic echo cancellation algorithms. The solutions have been based on adaptive filters, but the length of these filters has to be long enough to consider most of the echo signal and linear filtering in these algorithms may be limited to remove the echo signal in various environments.

In this thesis, a novel stereophonic acoustic echo suppression (SAES) technique based on spectral and temporal correlations is proposed in the short-time Fourier transform (STFT) domain. Unlike traditional stereophonic acoustic echo cancellation, the proposed algorithm estimates the echo spectra in the STFT domain and uses a Wiener filter to suppress echo without performing any explicit double-talk detection. The proposed approach takes account of interdependencies among components in adjacent time frames and frequency bins, which enables more accurate estimation of the echo signals.

Due to the limitations of power amplifiers or loudspeakers, the echo signals captured in the microphones are not in a linear relationship with the far-end signals

even when the echo path is perfectly linear. The nonlinear components of the echo cannot be successfully removed by a linear acoustic echo canceller. The remaining echo components in the output of acoustic echo suppression (AES) can be further suppressed by applying residual echo suppression (RES) algorithms. In this thesis, we propose an optimal RES gain estimation based on deep neural network (DNN) exploiting both the far-end and the AES output signals in all frequency bins. A DNN structure is introduced as a regression function representing the complex nonlinear mapping from these signals to the optimal RES gain. Because of the capability of the DNN, the spectro-temporal correlations in the full-band can be considered while finding the nonlinear function. The proposed method does not require any explicit double-talk detectors to deal with single-talk and double-talk situations.

One of the well-known approaches for nonlinear acoustic echo cancellation is an adaptive Volterra filtering and various algorithms based on the Volterra filter were proposed to describe the characteristics of nonlinear echo and showed the better performance than the conventional linear filtering. However, the performance might be not satisfied since these algorithms could not consider the full correlation for the nonlinear relationship between the input signal and far-end signal in time-frequency domain. In this thesis, we propose a novel DNN-based approach for nonlinear acoustic echo suppression (NAES), extending the proposed RES algorithm. Instead of estimating the residual gain for suppressing the nonlinear echo components, the proposed algorithm straightforwardly recovers the near-end speech signal through the direct gain estimation obtained from DNN frameworks on the input and far-end signal. For echo aware training, *a priori* and *a posteriori* signal-to-echo ratio (SER) are introduced as additional inputs of the DNN for tracking the change of the echo signal. In addition, the multi-task learning (MTL) to the DNN-based NAES

is combined to the DNN incorporating echo aware training for robustness. In the proposed system, an additional task of double-talk detection is jointly trained with the primary task of the gain estimation for NAES. The DNN can learn the good representations which can suppress more in single-talk periods and improve the gain estimates in double-talk periods through the MTL framework. Besides, the proposed NAES using echo aware training and MTL with double-talk detection makes the DNN be more robust in various conditions.

The proposed techniques show significantly better performance than the conventional AES methods in both single- and double-talk periods. As a pre-processing of various applications such as speech recognition and speech enhancement, these approaches can help to transmit the clean speech and provide an acceptable communication in full-duplex real environments.

Keywords: Acoustic echo cancellation, acoustic echo suppression, signal-to-echo ratio, spectro-temporal correlations, stereophonic acoustic echo suppression, residual echo suppression, nonlinear echo, deep neural networks, optimal gain regression, adaptive filtering, nonlinear acoustic echo suppression, echo aware training, multi-task learning

Student number: 2009-20876

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In full-duplex hands-free telecommunication systems such as mobile phones, speakerphones, and teleconferencing system, acoustic echo is easily generated from the acoustic coupling between a loudspeaker and a microphone. Since even a small acoustic echo picked up by the microphone in a receiving room may be very annoying and significantly deteriorate the quality of speech signal, various algorithms for acoustic echo cancellation (AEC) or suppression (AES) are required to remove the echo components and overcome serious conversation trouble. In the last decades, a number of algorithms have been proposed to solve the acoustic echo problem and produced some successful results in telecommunication systems [1]–[5].

Traditionally, single channel AEC has been achieved by identifying the echo path with respect to room environment and the positions of the microphone and the loudspeaker and deducting the echo estimates from the input signal. This process can be viewed as a system identification and the echo path can be generally modeled as

a finite impulse response filter. The solutions for AEC mostly are based on adaptive filters which are the well-known approaches such as normalized least mean square, recursive least squares, affine projection and so on [6]. For the reasonable performance of the acoustic echo cancellation algorithms, the length of the adaptive filters should be long enough to consider most of the echo signal. However, these long filters demand the computational complexity of the algorithms, so the modified techniques in the time or frequency domain have been introduced for fast computation [6].

As an alternative, AES algorithms based on speech enhancement framework like spectral modification or Wiener filtering have been researched to solve the echo issue [7]–[11]. These approaches can help to enhance the perceptual quality of the near-end speech without post-processing for suppressing residual echo and may be robust to echo path changes. Additionally, these techniques may be computationally more efficient than the AEC methods based on adaptive filters.

For spatial sound reproduction, the multi-channel AEC has been also researched over the last decade and most of the traditional stereophonic AEC algorithms are based on an adaptive filters for tracking several echo paths [12]–[14]. However, because of the strong cross-correlation between the stereo signals, these approaches require various de-correlation pre-processes which demand substantial complexity and cause distortion of the reproduced signal [4], [12]. To avoid the disadvantages of the de-correlation methods, a stereophonic AES algorithms was presented recently [15]. This method estimates echo spectra and utilizes them to obtain *a priori* and *a posteriori* signal-to-echo ratio (SER) information which are exploited by the single channel AES methods.

Although AEC or AES algorithms with linear filtering have been proven to remove echo successfully, a certain amount of residual echo remains at the output of

these methods possibly due to the inherent nonlinearity of the loudspeakers and power amplifiers, nonlinear acoustic transfer function of the echo path, or imperfection of the algorithms. First, to alleviate the nonlinearity in the output of AEC or AES, several residual echo suppression (RES) methods have been introduced. The authors in [9] and [16] proposed an RES gain function based on the SER estimated in a decision-directed manner [17]. Recently, RES based on artificial neural network (ANN) was proposed to model the mapping from the far-end to the residual echo signal [18]. Second, for direct nonlinear AEC or AES without post-filtering, the adaptive Volterra filters have been widely used because the filter structure can be viewed as a generalization of linear adaptive filters [19]–[25]. Also, other approaches based on the tap-delayed neural networks (TDNN) [26], kernel modification in AP algorithm [27] were proposed to attenuate the nonlinear echo.

1.2 Scope of thesis

In this thesis, we propose three approaches incorporating spectro-temporal correlations for acoustic echo suppression.

First, we propose an enhanced stereophonic AES (SAES) algorithm based on spectral and temporal correlations among adjacent time frames and frequency bins to improve the echo estimation performance of the conventional SAES method. Since linear systems can be accurately represented by cross-band filtering in the short-time Fourier transform (STFT) domain. The augmented vectors considering the continuity in the time-frequency domain are introduced in order to estimate the stereo echo more precisely, and calculate the extended power spectral density (PSD) matrices and cross-PSD vectors combining adjacent components in the STFT domain. In

various simulated conditions, experimental results showed better performances than that of the conventional SAES technique.

Second, a new residual echo suppression using deep neural networks (DNNs) is proposed in the single channel case. The DNN system estimates the optimal RES gain based on both the far-end and the output signals of AES in all frequency bins. We expect that the architecture can accommodate to model a nonlinear regression function from these signals to optimal RES gain based on DNN training using multi-condition data even though the room impulse responses (RIRs) used in the training do not match the RIRs for the test. The proposed system can consider spectro-temporal correlations which may come from harmonic distortion, insufficient frequency resolution or nonlinear echo path without use of any explicit double-talk detectors since the training data include both of the situations. The overall results obtained in matched and mismatched conditions for various RIRs, SER, clipping type, and level of nonlinearity in loudspeaker show that the proposed RES outperforms the conventional ANN-based RES method in terms of various objective measures.

Finally, extending the DNN-based RES technique, we propose a novel approach in DNN framework for nonlinear acoustic echo suppression (NAES). The proposed algorithm tries to directly recover the near-end speech signal by applying the optimal gain estimation based on DNNs. However, the structures have not dynamic, but fixed networks, so it may be impossible to track the nonlinear echo signal effectively in the various environments or room impulse responses (RIRs) compared to adaptive filtering. In order to overcome the issue, we use the echo information such as *a priori* and *a posteriori* signal-to-echo ratio (SER) as additional inputs of the DNN. These SER features may have appropriate information for tracking the change of the echo

signal. This is called echo aware training. Furthermore, we introduce the multi-task learning (MTL) to the DNN-based NAES incorporating echo aware training for robust NAES. In the proposed technique, the primary task of the gain estimation for NAES is jointly trained with an additional task of double-talk detection. The network can learn the good representations which can suppress more in single-talk periods and improve the gain estimates in double-talk periods through the double-talk detection task. Therefore, the proposed method makes the DNN be more robust in various conditions. Experimental results evaluated under real environments show that the proposed method is superior to the conventional one, especially in double-talk situations.

The rest of the thesis is organized as follows: Conventional AEC and AES approaches are briefly reviewed in Chapter 2. In Chapter 3, a novel stereophonic AES algorithm incorporating spectro-temporal correlations is proposed in the STFT domain. The correlation between adjacent time frames and frequency bins is helpful to suppress the echo signals and to preserve the near-end speech effectively. Chapter 4 proposes a new residual echo suppression based on deep neural networks. The DNN is employed to find the complex regression function among the optimal RES gain, the AES output, and the far-end signal in nonlinear environments which are originated from cheap loudspeakers and power amplifiers. In Chapter 5, a DNN-based nonlinear acoustic echo suppression using echo aware training and multi-task learning is proposed for effective nonlinear echo suppression. In views of tracking the echo and robustness to various conditions, this approach can achieve good performance. Finally, we conclude this dissertation in Chapter 6.

Chapter 2

Conventional Approaches for Acoustic Echo Suppression

In the last few decades, many conventional approaches have been dedicated to acoustic echo cancellation and suppression which reduce the negative effects of acoustic echo, namely the acoustic coupling between the loudspeaker and microphone in a room. In particular, the increasing use of teleconferencing systems has led to the requirement of faster and more reliable acoustic echo cancellation algorithms. In the chapter, a few conventional approaches for acoustic echo suppression are briefly introduced. First, classical single channel acoustic echo cancellation based on adaptive filters and acoustic echo suppression using speech enhancement techniques are reviewed. Second, stereophonic acoustic echo cancellation in a full-duplex stereophonic system for multichannel acoustic echo cancellation is introduced. Lastly, the nonlinear residual echo suppression technique based on artificial neural network is presented. The remarkable studies in this chapter may be helpful to understand the details in the following chapters of the thesis.

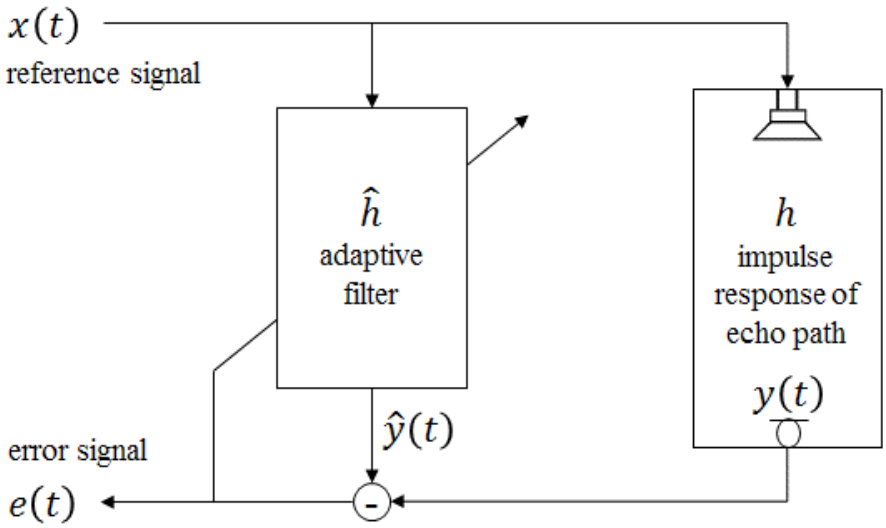

Figure 2.1: Schematic diagram of an adaptive acoustic echo canceller.

2.1 Single Channel Acoustic Echo Cancellation and Suppression

2.1.1 Single Channel Acoustic Echo Cancellation

In full-duplex hands-free telecommunication systems, single channel acoustic echo cancellation (AEC) and suppression have been researched to eliminate the undesired acoustic echo, which is the coupling between a loudspeaker and a microphone [1]–[5]. Historically, the echo cancellation process is achieved by identifying the echo path and subtracting an estimate of the echo signal from the microphone signal. In Fig. 2.1, a typical AEC is represented. The far-end signal $x(n)$ which is called the reference signal is played in a receiving room and the echo path generated from the signal can be modeled as a finite impulse response (FIR) filter. The echo

signal adds to the microphone signal $y(n)$ together with the near-end speech $s(n)$. Thus, the microphone signal $y(n)$ can be modeled as

$$y(n) = \mathbf{h}^T \mathbf{x}(n) + s(n) \tag{2.1}$$

where

$$\mathbf{x}(n) = [x(n), x(n-1), ..., x(n-N+1)]^T$$
$$\mathbf{h} = [h_0, h_1, ..., h_{N-1}]^T$$

N is the length of the echo path impulse response, and ^{T} denotes the transpose operation. To remove the echo signal, the echo estimate is calculated by identifying the coefficients of an FIR filter,

$$\hat{\mathbf{h}} = [\hat{h}_0, \hat{h}_1, ..., \hat{h}_{L-1}]^T. \tag{2.2}$$

The error signal $e(n)$ can be yield by subtracting the echo estimate from the microphone signal,

$$\begin{aligned} e(n) &= y(n) - \hat{y}(n) \\ &= [\mathbf{h} - \hat{\mathbf{h}}]^T \mathbf{x}(n) + s(n). \end{aligned} \tag{2.3}$$

For an optimal error criterion, the mean square error (MSE) can be used as follows,

$$E\{e^2(n)\} = E\{[(\mathbf{h} - \hat{\mathbf{h}})^T \mathbf{x}(n)]^2\} + E\{s^2(n)\} \tag{2.4}$$

where $E\{\cdot\}$ denotes the expectation operation and the echo signal and the near-end speech signal are assumted to be uncorrelated. In other words, $E\{s^2(n)\}$ is not affected by estimating the echo path. Thus, minimizing the MSE criterion means that $E\{[(\mathbf{h} - \hat{\mathbf{h}})^T \mathbf{x}(n)]^2\}$ is minimized and it is the objective of AEC to suppress the echo.

2.1.2 Adaptive Filters for Acoustic Echo Cancellation

In order to identify the optimum echo path, a number of adaptive techniques have been already addressed. Mostly, well-known approaches are based on normalized least mean square (NLMS), recursive least squares (RLS), affine projection (AP), and so on [6]. The invention of the least mean square (LMS) algorithm can be seen as the most crucial development for adaptive filtering. The potential of the LMS algorithm for acoustic echo cancellation and suppression was recognized for several researches. However, compared with line echoes on long distance transmission lines, suppressing acoustic echoes requires advanced adaptive filters which are extremely demanding with respect to signal processing power. In the past decades, many experiments and simulations affirmed that the weakness of the LMS algorithm with respect to correlated signals like speech. Although several modified LMS algorithms in the time and frequency domain were proposed for improving the performance and efficiency, these results included a certain amount of residual echo. In contrast to LMS algorithm, the recursive least squares (RLS) algorithms for acoustic echo processing can handle correlated signals very well since it has a built in decorrelation facility. However, this needs the inversion of the short-term correlation matrix of the input signal. Specially, the matrix may become singular by the characteristics of the input signal or the estimation procedure. As a result, the RLS algorithm frequently becomes instable for echo processing. To stabilize the RLS algorithm, it may be necessary to revise the technique with a long memory, but this might cause the tracking problem when the system to be identified changes. The affine projection (AP) algorithm [28] can provide a compromise method between the LMS and RLS algorithms. Like the RLS algorithm, the AP method also requires the matrix

inversion, but its numerical complexity is lower than that of the RLS algorithm and the convergence speed almost reaches that of the RLS processing for speech signal.

When the signal-to-noise ratio (SNR) is high and the near-end speech is absent, the estimated echo path coefficients can converge and the echo is suppressed well. However, when the near-end speech is active in the echo process, the adaptive FIR filter can be diverged by the presence of the near-end signal. To alleviate this problem, the process of double-talk detection is needed [5], [29], [30]. Whenever double-talk periods are detected, the echo process for estimating the echo path is stopped and the filter coefficients are not updated. The double-talk detection based on normalized correlation coefficients (NCCs) is the most well-known approach to find the double-talk intervals [29].

2.1.3 Acoustic Echo Suppression Based on Spectral Modification

To achieve the reasonable performance of the echo canceller, the length of the adaptive filter should be long enough for considering most of the echo. However, the long filters cause that the computational cost is very high and implementing the filter in the frequency domain can reduce the complexity compared to that in the time domain. Instead of estimating the echo path directly, acoustic echo suppression (AES) based on speech enhancement techniques like spectral modification and Wiener filtering has been developed to reduce the echo effect [7]–[11]. AES approaches are easily incorporated and can enhance the perceptual speech quality. In addition, mostly, the AES process may be robust to echo path changes and operate well without post-processing for suppressing residual echo. If necessary, the AES and RES methods can be easily combined. Lastly, AES algorithms may be computationally more efficient than the conventional AEC methods based on adaptive

filtering.

The signal model given in 2.1 can be written in a vector form in the short-time Fourier transform (STFT) as

$$Y(k) = U(k) + S(k) \tag{2.5}$$

where $U(k)$ is the result after taking STFT on the echo signal $\mathbf{u}(n) = \mathbf{h}^T\mathbf{x}$. The echo cancellation can be seen as an estimation problem to calculate $S(k)$ from the microphone signal $Y(k)$. By obtaining the estimate of $U(K)$, the near-end signal $S(K)$ can be recovered. Thus, this problem is equivalent to the design of two signal estimators which are the spectral magnitude and the phase component. Fortunately, it has been proven that the relationship between the phase distortion and human perception is more insensitive than expected [17], [31], [32]. The phase component of the microphone signal can be used as an estimate of the echo signal for echo suppression. This serves as the basis for the echo suppression. In this framework, given $Y(k)$, $|S(k)|$ is estimated using spectral modification [11]. It is assumed that the microphone and the echo signals are uncorrelated. Based on this assumption, the instantaneous power spectrum of the microphone signal $U(k)$ can be approximated as follows,

$$|Y(k)|^2 \approx |U(k)|^2 + |S(k)|^2. \tag{2.6}$$

To recover the near-end signal, $|S(K)|^2$ can be estimated by subtracting $|U(K)^2|$ from $|Y(k)|^2$, and the corresponding spectral magnitude of the near-end speech is

calculated as

$$|S(k)| = \sqrt{|\hat{S(k)}|^2} \tag{2.7}$$

$$= \left[\frac{|Y(k)|^2 - |\hat{U}(k)|^2}{|Y(k)|^2} \right]^{1/2} \tag{2.8}$$

$$= G(k)|Y(k)| \tag{2.9}$$

where $G(k)$ is called a gain filter. If necessary, additional parameters can be combined to control the amount of echo in case it is under- or over-estimated. It has been widely adopted for the purpose of additive noise suppression and speech enhancement. This framework removes the echo in the time-frequency domain on a frame-by-frame basis. Finally, the estimated near-end speech or suppressed signal is generated by applying the overlap-add method with inverse STFT. However, the spectral modification often causes the musical noise which makes annoying phenomenon due to the isolated spectral peaks resulting from the nonlinear gain estimation.

2.2 Residual Echo Suppression

Acoustic echo cancellation (AEC) or suppression (AES) is a technique to reduce the echo originated from acoustic coupling between loudspeakers and microphones [7], [11], [33], [34]. Although there have been many techniques which are prove to suppress the echo successfully, there still exists some amount of residual echo at the outputs of these methods. One of the reasons for which the AEC or AES suffer is that the echo signal is not a linear function of the far-end digital signal even when the echo path is perfectly linear. The power amplifiers and loudspeakers, especially cheap and small ones, can be the sources of this nonlinearity. To overcome this problem, several residual echo suppression (RES) filters have been applied to the

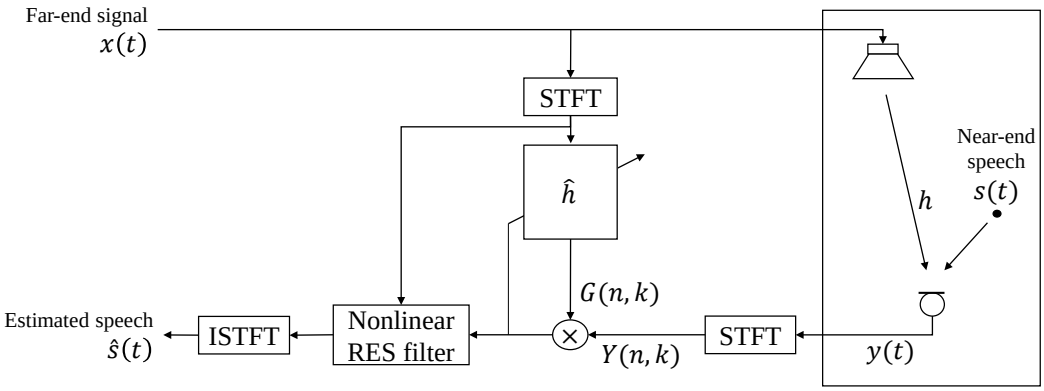

Figure 2.2: Schematic diagram of the stereophonic acoustic echo cancellation.

output of the AEC or AES to suppress remaining echo. The authors in [9] and [16] proposed RES methods to estimate the signal-to-echo ratio (SER) and then apply Wiener filters or spectral subtraction in the frequency domain. In [35], subband filtering based on the spectral subtraction was combined with a truncated Taylor series expansion of acoustic echo path for the estimation of power spectral density of the echo.

A single-channel AES system with a post filter is depicted in Figure 2.2. The far-end signal $x(t)$ at time index t is generated by the source signal through the acoustic impulse response in the transmission room. Let $y(t)$ be the input signal including near-end speech $s(t)$ in the receiving room and $Y(n,k)$ is the short-time Fourier Transform (STFT) coefficient of $y(t)$ for k-th frequency bin at the n-th frame. The spectral gain function to suppress the echo, $G(n,k)$, is obtained from the Wiener filtering or spectral subtraction in each frequency bin. However, due to limitations of linear echo modeling, the echo component may still remain in the output of AES including a considerable amount of nonlinear echo degrading the quality of the

near-end speech. To improve the output of AES, additional nonlinear residual echo suppression (RES) filter can be applied to the remaining signal. Using the residual echo suppression gain $G_{res}(n,k)$, the final estimated speech in the frequency domain, $\hat{S}(n,k)$ is given by,

$$\hat{S}(n,k) = \{G(n,k) \cdot G_{res}(n,k)\}Y(n,k). \tag{2.10}$$

When the power amplifiers and loudspeakers introduce severe nonlinearity, it is very important to calculate $\hat{G}_{res}(n,k)$ accurately in accordance with the nonlinearity of residual echo.

2.2.1 Spectral Feature-based Nonlinear Residual Echo Suppression

Recently, for the estimation of the residual echo magnitude spectrum in the nonlinear environment, a RES algorithm was proposed using artificial neural networks (ANNs) [18]. Due to the computational complexity of the physical processes leading to distortion artifacts, the approach using the spectral features from the far-end signal was attempted instead of modeling RES directly. Thus, by training a multiple-input regression model and realizing as ANNs, the RES processing is achieved. The MMSE-optimal suppression gain G_{res} can be estimated based on the estimated magnitude spectrum of the residual echo $u(t) = y(t) - \hat{y}(t)$ and the AEC or AES output magnitude $|\tilde{S}(n,k)|$,

$$G_{res}(n,k) = \max\left(G_{min}, 1 - \mu\frac{|\hat{U}(n,k)|^2}{|\tilde{S}(n,k)|^2}\right), \tag{2.11}$$

where μ is the overestimation factor and G_{min} is the mininum gain. Thus, this gain estimation is achieved in the Wiener filter framework.

In Fig. 2.3, an artificial neural network based on spectral features extracted from the far-end signal magnitude spectrum in the individual frequency bin is illustrated.

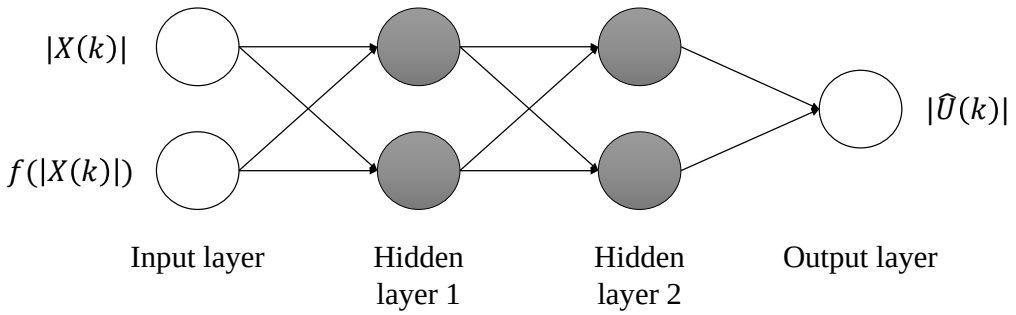

Figure 2.3: An ANN structure for spectral feature-based nonlinear residual echo suppression

This structure consists of an input layer, two hidden layers and an output layer. Each hidden layer node represents a weighted sum of the hidden layers outputs and bias values. For generating the feature, a feature extraction function is used as follows,

$$f(|X(n,k)|) = \frac{1}{k/2} \sum_{n=1}^{k/2} |X(n,k)|. \tag{2.12}$$

This feature is the average over all subbands up to half of the current subband k, motivated by the observation that nonlinear components in subband k are likely to generate from input magnitudes in subbands $k/2$ and less if they represent higher-order harmonics. This model can be thought as a generalization of the linear relationship between the reference signal and the residual echo magnitude, and the sparse coupling signal, which can be made by using the magnitudes of other subbands as input features. However, the training of a feedforward network is a non-convex optimization problem, so an additional online adaptation is needed to be feasible estimation. The initial residual echo magnitude estimate $|\hat{U}^*(n,k)|$ from the offline-trained net-

work, a scalar factor $a(n,k)$ is adopted in the online processing,

$$|\hat{U}(n,k)| = a(n,k)|\hat{U}^*(n,k)|. \tag{2.13}$$

The assumption behind this combination is that the nonlinear characteristics such as the ratio between the linear and the nonlinear components per each frequency band are not strongly dependent on the acoustic environment. Thus, the effect of time-varying acoustic characteristics can be modeled by the adaptive filtering like LMS algorithm [6]. This can be operated by estimating the weights in single-talk periods using the adaptive update.

2.3 Stereophonic Acoustic Echo Cancellation

In the last decades, AEC on a single full-duplex audio channel was the major research topic to remove undesired echoes that result from between a microphone and a loudspeaker. However, for effective audio communication between groups of people or multi-speaker conditions, AEC on multichannel environment is necessary and the stereophonic channel is the minimum case of the environment [4], [12], [36]–[39]. In Fig. 2.4, a schematic diagram of the stereophonic acoustic echo cancellation in the teleconferencing is illustrated. Basically, conventional AEC techniques based on adaptive filters estimate the echo using a FIR filter with adjustable coefficients to model the acoustic impulse response of the echo path. In other words, two adaptive FIR filters $\hat{h}_1$ and $\hat{h}_2$ are used to model the two echo paths in the receiving room.

In order to develop the LMS algorithm for stereo AEC, the echo signal can be expressed as,

$$y(n) = h_1(n) * x_1(n) + h_2(n) * x_2(n) \tag{2.14}$$

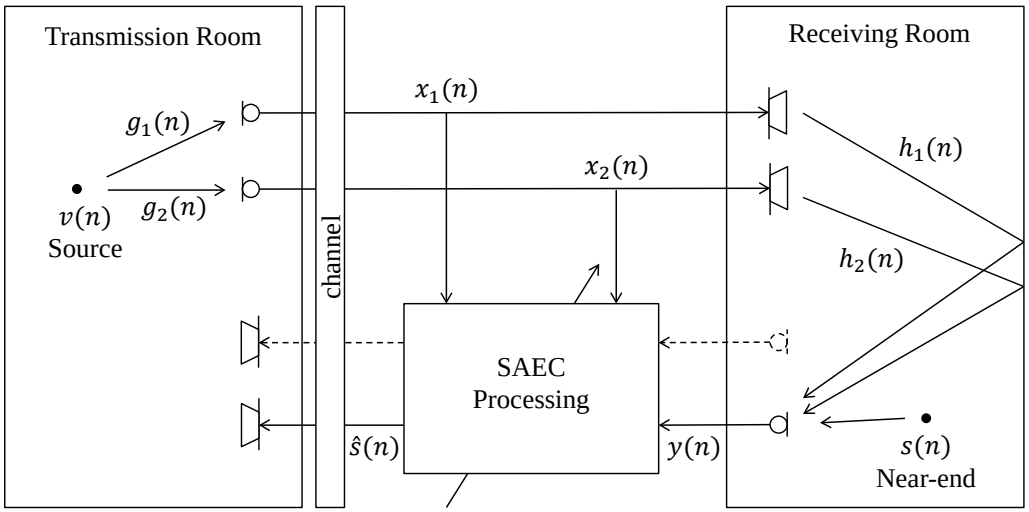

Figure 2.4: Schematic diagram of the stereophonic acoustic echo cancellation.

where h_1 and h_2 are the loudspeaker-to-microphone impulse responses in the receiving room, and * denotes the convolution operation. Thus, the error signal is formulated as,

$$e(n) = y(n) - \hat{\mathbf{h}}_1^T \mathbf{x}_1(n) - \hat{\mathbf{h}}_2^T \mathbf{x}_2(n) \tag{2.15}$$

where

$$\mathbf{x}_i(n) = [x_i(n), x_i(n-1), ..., x_i(n-N+1)]^T$$

$$\mathbf{h}_i = [h_i(0), h_i(1), ..., h_i(N)]^T (i = 1, 2).$$

Specially, the convergence issue of $\hat{h}_i$ becomes even more important in the stereo case and multichannel AEC processing [36]. Setting aside the important aspect of how convergence is achieved, it is assumed that the error signal can be ideally zero. (In this case, the conventional single channel echo cancellation methods are simply

extended to the stereo echo cancellation process.) As a result, it follows that

$$\tilde{h}_1 * x_1 + \tilde{h}_2 * x_2 = 0 \tag{2.16}$$

where $\tilde{h}_i = h_i - \hat{h}_i (i = 1, 2)$ is the misalignment vector. When the near-end speech $s(n)$ is talking in this case, it implies that

$$[\tilde{h}_1 * g_1 + \tilde{h}_2 * g_2] * s(n) = 0 \tag{2.17}$$

where g_1 and g_2 are the acoustic impulse responses in the transmission room, respectively. In the frequency domain, it becomes

$$[\tilde{H}_1(k)G_1(k) + \tilde{H}_2(k)G_2(k)]S(k) = 0 \tag{2.18}$$

where k is the k-th frequency bin. Considering the single channel case with $G_2 = 0$, the complete alignment ($\hat{\mathbf{h}}_1 = \mathbf{h}_1$) is achieved since $G_1 S$ does not vanish at any frequency. On the other hand, in the stereophonic case, the best choice is

$$\tilde{H}_1 G_1 + \tilde{H}_2 G_2 = 0 \tag{2.19}$$

even if S has no zeros in the frequency range. However this equation does not guarantee $\tilde{H}_1 = \tilde{H}_2 = 0$, which is the complete alignment condition. This issue is the most crucial problem to remove the echoes in the stereophonic AEC. Although h_1 and h_2 are fixed in the receiving room, any change in G_1 or G_2 requires adjustment of $\tilde{H}_1$ and $\tilde{H}_1$ except $\tilde{H}_1 = \tilde{H}_2 = 0$. Therefore, the adaptation algorithm must track not only variations in the receiving room, but also variations in the transmission room. The changes in the room are very hard to follow. For example, if one speaker stops talking and another speaker starts talking at a different location, the impulse responses g_1 and g_2 change suddenly. This difficult problem is to propose the new technique for the convergence independent on the variations in the transmission

room. So, this problem is called the nonuniqueness problem in the stereophonic AEC [4].

To solve this problem, several signal decorrelaion methods were proposed such as addition of random noise, decorrelation filters, interchannel frequency shifting and interleaving comb filters [4], [36]. Unfortunately, these techniques might not be satisfactory to obtain both the reduction of the misalignment and the perceptual improvement. Also, because of the strong cross-correlation between the stereo signals, most of the traditional stereo acoustic echo cancellation approaches based on an adaptive filter require some form of various de-correlation techniques, but the decorrelation processes demand substantial complexity as the pre-processing procedure and cause distortion of the reproduced signal.

2.4 Wiener Filtering for Stereophonic Acoustic Echo Suppression

Stereophonic acoustic echo cancellation (SAEC) has a fundamental issue due to the non-uniqueness problem which does not happen in the single channel case [4]. The traditional adaptive filter algorithms can not solve this problem well due to the strong correlation be tween the stereo signals. To alleviate the problem, several de-correlation algorithms have been proposed and tried. However, most of these preprocessing techniques may affect the negative stereo perception and require the computational complexity [38], [39]. In the past, a low complexity AES algorithm was proposed [10]. It is based on spectral modification method which is widely used in the speech enhancement area. A multichannel AES method was also presented with the assumption that an user use a reasonably symmetric loudspeaker and microphone

setup [11]. However, this assumption may be not realized in practice.

Recently, an open-loop stereophonic acoustic echo suppression (SAES) algorithm without preprocessing was proposed for teleconferencing systems, where the Wiener filter in the STFT domain is incorporated [15]. By using two weighting functions, the stereo echo spectrum can be estimated from the stereo signals. In other words, this approach does not identify the echo path impulse responses with adaptive filters, so it can avoid the non-uniqueness problem in the stereophonic case. The undesired echo can be suppressed by applying spectral modification techniques which are proposed for speech enhancement like noise reduction or speech dereverberation [17]. In addition, for real-time operations, signal-to-echo ratio (SER) based Wiener filter is employed as the echo suppression gain function to consider a trade-off between musical noise reduction and computational complexity.

Taking the STFT on both sides of (2.14) with the near-end speech $s(n)$ in Fig. 2.4 can yield as follows,

$$Y(k) = H_1(k)X_1(k) + H_2(k)X_2(k) + S(k). \tag{2.20}$$

In Fig. 2.5, the SAES algorithms using the two weighting functions based on the Wiener filter is represented. The stereo echo spectra consists of two parts. The first one $D_1(k) = G_1(k)X_1(k)$ is correlated with the spectrum of the far-end signal $x_1(n)$ and the other one $D_2(k) = G_2(k)X_2(k)$ is correlated with that of the second far-end signal $x_2(n)$ but uncorrelated with that of the first signal. By minimizing the mean-square error (MSE) J_1 and J_2, the two weighting functions $G_1(k)$ and $G_2(k)$ can be formulated as follows,

$$J_1 = E\big[|Y(k) - G_1(k)X_1(k)|^2\big], \tag{2.21}$$

$$J_2 = E\big[\big|Y_1(k) - G_2(k)X_2(k)|^2\big]. \tag{2.22}$$

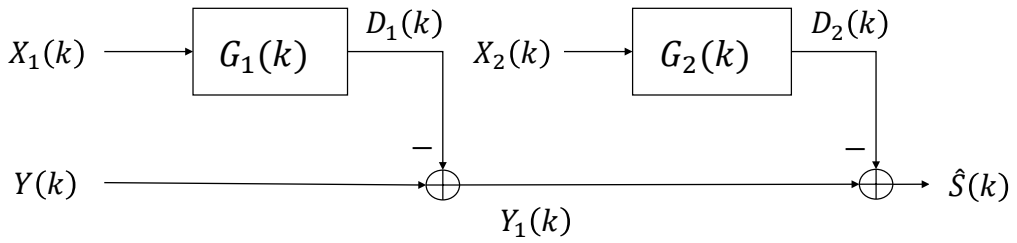

Figure 2.5: SAES algorithm via two weighting functions.

Actually, these cost functions may be modified to another MSE like

$$J = E\big[|Y(k) - G_1(k)X_1(k) - G_2(k)X_2(k)|^2\big]. \tag{2.23}$$

Minimizing the error with respect to $G(k) = [G_1(k)G_2(k)]^T$ makes the following estimator,

$$G(k) = \mathbf{\Phi}_{XX}^{-1}\mathbf{\Phi}_{XY} \tag{2.24}$$

where

$$\mathbf{\Phi}_{XX} = \begin{bmatrix} \Phi_{X_1X_1}(k) & \Phi_{X_1X_2}(k) \\ \Phi_{X_2X_1}(k) & \Phi_{X_2X_2}(k) \end{bmatrix},$$
$$\mathbf{\Phi}_{XY} = [\Phi_{X_1Y}(k)\ \Phi_{X_2Y}(k)]^T.$$

Unfortunately, the matrix $\mathbf{\Phi}_{XX}$ may be ill-conditioned since the far-end signals are strongly correlated. As a result, the matrix may cause inaccurate gain estimates and be not suitable for the echo suppression. Therefore, by minimizing (2.21) and (2.22)

with respect to $G_1(k)$ and $G_2(k)$, respectively, the following solutions are obtained,

$$G_1(k) = \frac{\mathbf{\Phi}_{X_1Y}(k)}{\mathbf{\Phi}_{X_1X_1}(k)}, \tag{2.25}$$

$$G_2(k) = \frac{\mathbf{\Phi}_{X_2Y_1}(k)}{\mathbf{\Phi}_{X_2X_2}(k)} \tag{2.26}$$

where $\mathbf{\Phi}_{XY}(k)$ denotes the cross power spectral density (PSD) between $x(n)$ and $y(n)$

$$\begin{aligned}\mathbf{\Phi}_{XY}(k) &= \sum_{\infty}^{\tau=\infty} E[x(n)y(n-\tau)]\exp(-jk\tau), \\ &= \sum_{\infty}^{\tau=\infty} R_{XY}\exp(-jk\tau).\end{aligned}$$

From the criteria (2.21) and (2.22), it is obtained that $E[D_1^*(k)Y_1(k)] = 0$ and $E[D_2^*(k)S(k)] = 0$, where $*$ denotes the operation for complex conjugation. Thus, the PSD of the near-end speech can be approximately estimated by

$$\begin{aligned}|\hat{S}(k)|^2 &\approx |Y(k)|^2 - |D_1(k)|^2 - |D_2(k)|^2 \\ &= |Y(k)|^2 - |G_1(k)X_1(k)|^2 - |G_2(k)X_2(k)|^2.\end{aligned} \tag{2.27}$$

Thus, the magnitude spectrum $|\hat{S}(k)|$ is calculated as $|\hat{S}(k)| = G(k)|Y(k)|$, where $G(k)$ is the gain function that can be estimated through spectral modification method like decision-directed approach [17]. In this framework, *a priori* signal-to-echo ratio (SER) based Wiener filter can be applied in the STFT domain. The *a posteriori* and *a priori* SER can be defined as

$$\gamma(n,k) \triangleq \frac{|Y(n,k)|^2}{\lambda_D(n,k)}, \tag{2.28}$$

$$\xi(n,k) \triangleq \frac{\lambda_S(n,k)}{\lambda_D(n,k)} \tag{2.29}$$

where n is the time index and $\lambda_D(n,k)$ and $\lambda_S(n,k)$ denote the PSD of the stereo echo and the near-end speech, respectively. To incorporate the Wiener filter for the

echo suppression, it is assumed that the echo and near-end spectra are mutually uncorrelated. Then, the gain function can be obtained as follows,

$$G(n,k) = \frac{\xi(n,k)}{1+\xi(n,k)}. \tag{2.30}$$

Even though the echo suppression algorithm is represented only for one microphone signal, this approach can be simply extended to another microphone signal.

Chapter 3

Stereophonic Acoustic Echo Suppression Incorporating Spectro-Temporal Correlations

3.1 Introduction

Acoustic echo cancellation techniques have been developed to overcome serious conversation trouble due to the acoustic coupling between microphones and loudspeakers [10]–[12], [16], [33]. Especially for spatial sound reproduction, the multi-channel acoustic echo cancellation problem has been researched over the last decade. Unlike single-channel echo cancellation, de-correlation algorithms are usually required to resolve the non-uniqueness problem, which results in a reconvergence issue [12]. However, these strategies are likely to distort signals reproduced by loudspeakers and demand a significant amount of computation.

Recently, inspired by several single-channel echo suppression methods [10], [11],

a stereophonic acoustic echo suppression (SAES) technique [15] was proposed. This approach estimates echo spectra in the short-time Fourier transform (STFT) domain without pre-processing by introducing an *a priori* signal-to-echo ratio (SER) and an *a posteriori* SER [33] under the Wiener filtering framework. This algorithm has been found to operate well during double-talk periods in spite of the fact that it does not apply any explicit double-talk detector.

In this chapter, to improve the estimation performance of the SAES method presented in [15], we propose an enhanced SAES (ESAES) algorithm that incorporates spectral and temporal correlations among adjacent time frames and frequency bins, based on the observation that linear systems can be accurately represented by cross-band filtering in the STFT domain [40]. We introduce augmented vectors considering the continuity in the time-frequency domain in order to estimate the stereo echo more precisely, and calculate the extended power spectral density (PSD) matrices and cross-PSD vectors incorporating adjacent components in the STFT domain. The performance of the proposed algorithm is evaluated by echo return loss enhancement (ERLE) and the ITU-T Recommendation P. 862 perceptual evaluation of speech quality (PESQ) [41] measures. Experimental results showed improved performances in terms of ERLE and PESQ compared with the conventional SAES technique.

3.2 Linear Time-Invariant Systems in the STFT Domain with Crossband Filtering

In this section, we briefly review the representation of linear time-invariant (LTI) systems in the STFT domain with crossband filtering [40]. Specially, the motivation

of the proposed echo suppression algorithm based on spectro-temporal correlations can be found in this framework. To identify LTI system, the STFT representation of a signal $x(n)$ is given by

$$x_{p,k} = \sum_{m} x(m)\psi^*_{p,k}(m) \tag{3.1}$$

where

$$\psi_{p,k}(n) \triangleq \psi(n - pL)\exp(j\frac{2\pi}{N}k(n - pL)). \tag{3.2}$$

$\psi(n)$ denotes an analysis window of length N, p is the frame index k, k represents the frequency-band index, L is the discrete-time shift, and * denotes complex conjugation operation. For the reconstruction of $x(n)$ from its STFT representation $x_{p,k}$, the inverse STFT is given by

$$x(n) = \sum_{p}\sum_{k=0}^{N-1} x_{p,k}\tilde{\psi}_{p,k}(n) \tag{3.3}$$

where

$$\tilde{\psi}_{p,k}(n) \triangleq \tilde{\psi}(n - pL)\exp(j\frac{2\pi}{N}k(n - pL)) \tag{3.4}$$

and $\tilde{\psi}$ denotes a synthesis window of length N.

Then, an STFT representation of LTI systems can be formulated using $h(n)$ and $d(n)$ which are an impulse response of an LTI system with length Q and a output signal, respectively, as follows,

$$\begin{aligned} d(n) &= \sum_{i=0}^{Q-1} h(i)x(n - i) \\ &= \sum_{m,l} h(l)x(m - l)\psi^*_{p,k}(m) \end{aligned} \tag{3.5}$$

which is obtained by using (3.1) and (3.2). Substituting (3.3) into (3.5), the output signal can be rewritten as

$$d_{p,k} = \sum_{k'=0}^{N-1}\sum_{p'} x_{p',k'}h_{p,k,p',k'} \tag{3.6}$$

where

$$h_{p,k,p',k'} = \sum_{m,l} h(l)\tilde{\psi}_{p',k'}(m-l)\psi^*_{p,k}(m) \tag{3.7}$$

may be interpreted as the STFT of $h(n)$ using a composite analysis window. Using (3.2) and (3.4), the equation (3.7) can be reformulated as

$$\begin{aligned} h_{p,k,p',k'} &= \sum_l h(l) \sum_m \psi(m) e^{(-j\frac{2\pi}{N}km)} \tilde{\psi}((p-p')L - l + m) e^{(j\frac{2\pi}{N}k'((p-p')L-l+m))} \\ &= \{h(n) * \phi_{k,k'}(n)\}|_{n=(p-p')L} \\ &\triangleq h_{p-p',k,k'} \end{aligned} \tag{3.8}$$

where $*$ denotes convolution with respect to the time index n and

$$\phi \triangleq e^{j\frac{2\pi}{N}k'n} \sum_m \psi(m)\tilde{\psi}(n+m)e^{-j\frac{2\pi}{N}m(k-k')}. \tag{3.9}$$

In other words, $h_{p,k,p',k'}$ depends on $p-p'$ rather than p and p' separately. Substituting (3.8) in (3.6), the final expression of the output signal can be obtained as

$$\begin{aligned} d_{p,k} &= \sum_{k'=0}^{N-1} \sum_{p'} x_{p',k'} h_{p-p',k,k'} \\ &= \sum_{k'=0}^{N-1} \sum_{p'} x_{p-p'} h_{p',k,k'}. \end{aligned} \tag{3.10}$$

Let N_h be the length of the cross-band filters and then the STFT of the output signal $d_{p,k}$ can be written as

$$d_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'=0}^{N_h-1} x_{p-p',k'} h_{p',k,k'}. \tag{3.11}$$

Let $\hat{d}_{p,k}$ be the resulting estimate of $d_{p,k}$ using only $2K$ cross-band filters around the frequency-band k and $\hat{h}_{p',k,k'}$ be an estimate of the cross-band filter $h_{p',k,k'}$. Then,

the estimated output signal $\hat{d}_{p,k}$ is obtained as follows,

$$\hat{d}_{p,k} = \sum_{k'=k-K}^{k+K} \sum_{p'=0}^{N_h-1} \hat{h}_{p',k,(k'modN)} x_{p-p',(k'modN)} \tag{3.12}$$

where the periodicity of the frequency-bands is exploited.

Therefore, the time domain convolution is not equivalent to the STFT domain multiplication any longer since finite length analysis windows are employed in the usual implementations. To solve this issue, crossband filtering is needed to perfectly represent an LSI system in the STFT domain. For successful echo suppression in adverse acoustic environments, system identification based on crossband filtering and spectral correlations has to be employed.

3.3 Enhanced SAES (ESAES) Utilizing Spectro-Temporal Correlations

A typical stereophonic acoustic echo scenario is illustrated in Fig. 3.1. The far-end signals $x_1(n)$ and $x_2(n)$ at time index n are generated by the source signal $v(n)$ through the acoustic impulse responses $g_1(n)$ and $g_2(n)$ in the transmission room. Let $y(n)$ be the signal picked up by one of the microphones in the receiving room. This signal can be modeled as

$$y(n) = \sum_{i=1}^{2} h_i(n) * x_i(n) + s(n) \tag{3.13}$$

where $h_i(n)$ represents the acoustic echo path from the ith loudspeaker to the microphone and $s(n)$ is the near-end signal. In this work, we focus on only one of the microphones to describe the stereophonic acoustic echo problem because we can apply the same approach to the other microphone.

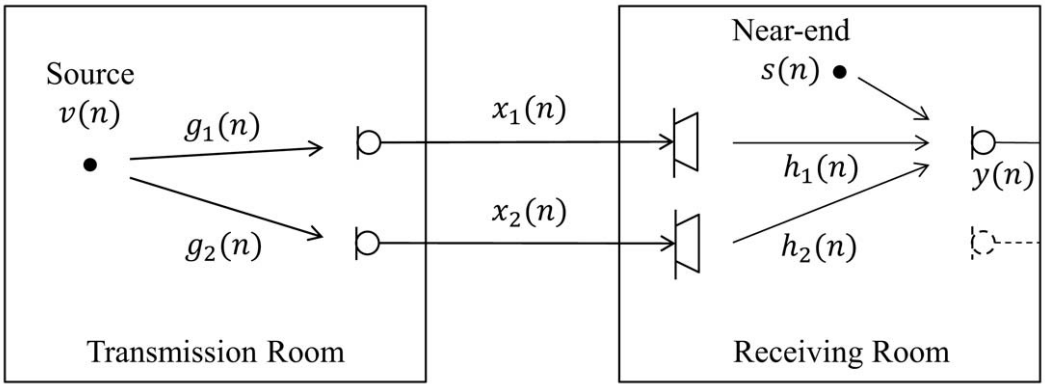

Figure 3.1: Schematic diagram of the stereophonic acoustic echo scenario.

The proposed ESAES algorithm extends the SAES method in [15] by taking account of correlations among adjacent time frames and frequency bins in the STFT domain. According to [40], linear systems can be more accurately represented by crossband filtering due to the effect of finite windows. Moreover, it is shown in [40] that considering a few neighboring bins was enough although all the frequency bins need to be taken into considerations theoretically. In order to combine this theory with the SAES algorithm, we introduce the following augmented vector (Type 1) for each far-end signal:

$$\begin{aligned}\underline{\mathbf{X}}_i^1(n,k) = [&X_i(n-T,k-K)\ \ldots\ X_i(n-T,k+K)\\ &X_i(n-T+1,k-K)\ \ldots\ X_i(n-T+1,k+K)\\ &\ldots\ X_i(n,k-K)\ \ldots\ X_i(n,k+K)]^T\ (i=1,2)\end{aligned} \tag{3.14}$$

where $X_i(n,k)$ is the STFT coefficient of the far-end signal $x_i(n)$ for the kth frequency bin at the nth frame. The augmented vector defined in (3.14) consists not only of the $(2K+1)$ adjacent frequency bins from the current nth frame, but also of

the previous T frames of $(2K+1)$ adjacent frequency bins. Thus, the dimension of this augmented vector becomes $M_1 = (T+1) \times (2K+1)$. Alternatively, by considering only adjacent frequencies of the current and previous frames of given frequency, we can reduce the dimension of the augmented vector (Type 2) as follows:

$$\begin{aligned}\underline{\mathbf{X}}_i^2(n,k) = [&X_i(n-T,k)\ X_i(n-T+1,k)\\ &\dots\ X_i(n-1,k)\ X_i(n,k-K)\\ &\dots\ X_i(n,k+K)]^T \quad (i=1,2).\end{aligned} \tag{3.15}$$

The augmented vector in (3.15) is made of the $(2K+1)$ frequency bins at the current nth frame and the kth frequency bin from frames $(n-T)$ to $(n-1)$, and its dimension becomes $M_2 = T + 2K + 1$. The components included in the two types of augmented vectors with $T = 2$ and $K = 2$ are illustrated in Fig. 3.2. In the remaining part of this work, for simplicity, we will use the notation $\underline{\mathbf{X}}_i(n,k)$ which represents the augmented vector shown either in (3.14) or (3.15), and use M to denote the dimension of this augmented vector.

3.3.1 Problem Formulation

Let $Y(n,k)$, $\underline{\mathbf{X}}_1(n,k)$, $\underline{\mathbf{X}}_2(n,k)$ denote the STFT coefficients of $y(n)$ and the augmented vectors corresponding to $x_1(n)$ and $x_2(n)$, respectively. Crossband convolutive filters are denoted by $\underline{\mathbf{H}}_1(n,k)$ and $\underline{\mathbf{H}}_2(n,k)$ that represent the acoustic paths relating $\underline{\mathbf{X}}_1(n,k)$ and $\underline{\mathbf{X}}_2(n,k)$ to $Y(n,k)$, respectively [40]. Then $Y(n,k)$ can be described as

$$Y(n,k) = \sum_{i=1}^{2} \underline{\mathbf{H}}_i^H(n,k)\underline{\mathbf{X}}_i(n,k) + S(n,k) \tag{3.16}$$

where $S(n,k)$ is the STFT coefficient of the near-end signal $s(n)$, including near-end speech and noise, and superscript ^{H} denotes conjugate transpose.

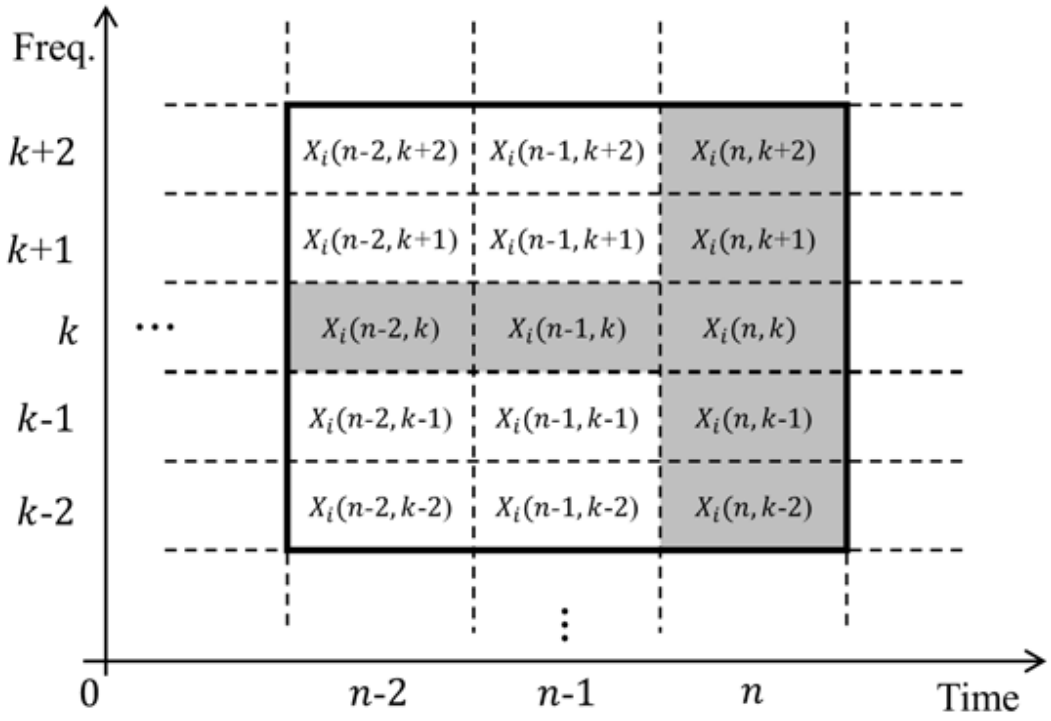

Figure 3.2: Two types of augmented vectors with $T = 2$, $K = 2$. The augmented vector (3.14) consists of 15 adjacent components in the bold square, and the augmented vector (3.15) is made of 7 adjacent component in the shaded region.

As in the conventional SAES, let us denote the STFT of the echo component due to $x_1(n)$ by $D_1(n,k)$, and likewise for $x_2(n)$ by $D_2(n,k)$. Then,

$$\begin{aligned} D_1(n,k) &= \underline{\mathbf{H}}_1^H(n,k)\underline{\mathbf{X}}_1(n,k), \\ D_2(n,k) &= \underline{\mathbf{H}}_2^H(n,k)\underline{\mathbf{X}}_2(n,k) \end{aligned} \tag{3.17}$$

assuming that $D_1(n,k)$ is correlated with $x_1(n)$ and $D_2(n,k)$ is correlated with $x_2(n)$ but uncorrelated with $x_1(n)$. In general, we obtain the optimal weight vectors $\widehat{\underline{\mathbf{H}}}_1(n,k)$ and $\widehat{\underline{\mathbf{H}}}_2(n,k)$ according to the minimum mean-square error (MMSE)

criterion, which jointly minimize

$$J_1 = E[|Y(n,k) - \underline{\mathbf{H}}_1^H(n,k)\underline{\mathbf{X}}_1(n,k)|^2], \tag{3.18}$$

$$J_2 = E[|Y_1(n,k) - \underline{\mathbf{H}}_2^H(n,k)\underline{\mathbf{X}}_2(n,k)|^2] \tag{3.19}$$

where $Y_1(n,k) = Y(n,k) - D_1(n,k)$ and $E[\cdot]$ denotes expectation. By minimizing (3.18) and (3.19) with respect to $\underline{\mathbf{H}}_1(n,k)$ and $\underline{\mathbf{H}}_2(n,k)$, we are led to the acoustic path estimates

$$\widehat{\underline{\mathbf{H}}}_1(n,k) = \mathbf{\Phi}^{-1}_{\underline{\mathbf{X}}_1\underline{\mathbf{X}}_1}(n,k)\mathbf{\Phi}_{\underline{\mathbf{X}}_1 Y}(n,k), \tag{3.20}$$

$$\widehat{\underline{\mathbf{H}}}_2(n,k) = \mathbf{\Phi}^{-1}_{\underline{\mathbf{X}}_2\underline{\mathbf{X}}_2}(n,k)\mathbf{\Phi}_{\underline{\mathbf{X}}_2 Y_1}(n,k) \tag{3.21}$$

where $\mathbf{\Phi}_{\underline{\mathbf{X}}\underline{\mathbf{X}}}(n,k)$ and $\mathbf{\Phi}_{\underline{\mathbf{X}}Y}(n,k)$ denote the extended PSD matrix and cross-PSD vector defined by

$$\mathbf{\Phi}_{\underline{\mathbf{X}}\underline{\mathbf{X}}}(n,k) = E[\underline{\mathbf{X}}(n,k)\underline{\mathbf{X}}^H(n,k)], \tag{3.22}$$

$$\mathbf{\Phi}_{\underline{\mathbf{X}}Y}(n,k) = E[\underline{\mathbf{X}}(n,k)Y^*(n,k)] \tag{3.23}$$

with superscript * denoting complex conjugation.

Given the estimated echo spectra, the near-end signal in the STFT domain, $S(n,k)$, can be estimated by means of the Wiener gain $G(n,k)$ as follows:

$$\widehat{S}(n,k) = G(n,k)Y(n,k) \tag{3.24}$$

under the assumption that the near-end signal and echo signal are uncorrelated. Details on the estimation of the extended PSD matrices, echo spectra, and gain function are described in the following subsection.

3.3.2 Estimation of Extended PSD Matrices, Echo Spectra, and Gain Function

The extended PSD matrix and cross-PSD vector related to $\underline{\mathbf{X}}_1(n,k)$ can be obtained by first-order recursive averaging in the following way:

$$\begin{aligned}\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_1\underline{\mathbf{X}}_1}(n,k) &= \alpha_{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_1\underline{\mathbf{X}}_1}(n-1,k) \\ &\quad + (1-\alpha_{\mathbf{\Phi}})\underline{\mathbf{X}}_1(n,k)\underline{\mathbf{X}}_1^H(n,k), \qquad (3.25) \\ \widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_1 Y}(n,k) &= \alpha_{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_1 Y}(n-1,k) \\ &\quad + (1-\alpha_{\mathbf{\Phi}})\underline{\mathbf{X}}_1(n,k)Y^*(n,k) \qquad (3.26)\end{aligned}$$

where $0 < \alpha_{\mathbf{\Phi}} < 1$ is a smoothing factor. With $\widehat{\underline{\mathbf{H}}}_1$ obtained by applying (3.25) and (3.26) to (3.20), the estimate of $D_1(n,k)$ can be calculated by introducing an additional overestimation control-factor matrix, $\mathbf{B}_1$, which is an extension of the echo suppression level control factor in the conventional SAES:

$$\widehat{D}_1(n,k) = |\widehat{\underline{\mathbf{H}}}_1^H(n,k)\mathbf{B}_1\underline{\mathbf{X}}_1(n,k)| \qquad (3.27)$$

where $\mathbf{B}_1$ is a diagonal matrix whereby the diagonal elements corresponding to $X_1(n,k)$ are emphasized over the other elements. After deriving $\widehat{D}_1(n,k)$, the spectral subtraction method in [10] is used to get the estimate of $Y_1(n,k)$.

In a similar way, $D_2(n,k)$ can be estimated by performing the following proce-

dures:

$$\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_2\underline{\mathbf{X}}_2}(n,k) = \alpha_{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_2\underline{\mathbf{X}}_2}(n-1,k), \\ + (1-\alpha_{\mathbf{\Phi}})\underline{\mathbf{X}}_2(n,k)\underline{\mathbf{X}}_2^H(n,k), \tag{3.28}$$

$$\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_2 Y_1}(n,k) = \alpha_{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_2 Y_1}(n-1,k) \\ + (1-\alpha_{\mathbf{\Phi}})\underline{\mathbf{X}}_2(n,k)Y_1^*(n,k), \tag{3.29}$$

$$\widehat{D}_2(n,k) = |\widehat{\underline{\mathbf{H}}}_2^H(n,k)\mathbf{B}_2\underline{\mathbf{X}}_2(n,k)| \tag{3.30}$$

where $\mathbf{B}_2$ is also an overestimation control-factor matrix. The overestimation control-factor matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ are applied to further reduce the residual echo. Let

$$\mathbf{B}_i = \text{diag}\{b_1 \ \ldots \ b_M\} \quad (i=1,2) \tag{3.31}$$

where b_m weights $X_i(n_m,k_m)$, which represents the mth element of $\underline{\mathbf{X}}_i(n,k)$. In this work, we choose each b_m as follows:

$$b_m = \alpha_{B_i}\exp\left(-\beta_{\text{t},i}|n-n_m| - \beta_{\text{f},i}|k-k_m|\right) \\ (i=1,2,\ m=1,\ldots,M) \tag{3.32}$$

in which the parameters α_{B_i}, $\beta_{\text{t},i}$, and $\beta_{\text{f},i}$ are determined experimentally.

In order to obtain the gain function $G(n,k)$, we introduce the *a priori* SER $\xi(n,k)$ and *a posteriori* SER $\gamma(n,k)$ as in [33],

$$\xi(n,k) \triangleq \frac{\lambda_S(n,k)}{\lambda_D(n,k)},\ \gamma(n,k) \triangleq \frac{|Y(n,k)|^2}{\lambda_D(n,k)} \tag{3.33}$$

where $\lambda_S(n,k)$ and $\lambda_D(n,k)$ denote the PSDs of the near-end signal and composite echo, respectively. Estimates of $\lambda_D(n,k)$, $\gamma(n,k)$, and $\xi(n,k)$ are formed and updated

as [15]

$$\widehat{\lambda}_D(n,k) = \alpha_D \widehat{\lambda}_D(n-1,k) + (1-\alpha_D)(|\widehat{D}_1(n,k)|^2 + |\widehat{D}_2(n,k)|^2), \tag{3.34}$$

$$\widehat{\gamma}(n,k) = \frac{|Y(n,k)|^2}{\widehat{\lambda}_D(n,k)}, \tag{3.35}$$

$$\widehat{\xi}(n,k) = \alpha_{DD}\widehat{\gamma}(n-1,k)G^2(n-1,k) + (1-\alpha_{DD})\max(\widehat{\gamma}(n,k)-1,0) \tag{3.36}$$

where $0 < \alpha_D < 1$ and $0 < \alpha_{DD} < 1$ are smoothing factors that have to be much smaller than conventional values for SNR estimation in which the noise is assumed to be stationary, because the echo signals are highly nonstationary in most cases. Finally, according to the Wiener estimator theory, the gain function is given by

$$G(n,k) = \frac{\widehat{\xi}(n,k)}{1+\widehat{\xi}(n,k)} \tag{3.37}$$

and the estimated near-end $\widehat{S}(n,k)$ is obtained from (3.24).

3.3.3 Complexity of the Proposed ESAES Algorithm

We investigate the computational complexity of the proposed ESAES algorithm with half-overlapping windows and compare it with that of the conventional SAES in [15], where 7/8-overlapping windows were applied. Considering the matrix inversions in (3.20) and (3.21) and assuming the use of the divide-and-conquer algorithm [42], the proposed technique requires a total of $(8M^3 + 24M^2 + 104M + 52 + 24\log_2 N)$ real-valued multiplications, $4M$ complex-valued divisions, 12 real-valued divisions, and 20 square root calculations per frequency bin to obtain N samples in the time domain, considering the frame overlaps where M and N are the dimension of the

augmented vector and FFT size, respectively. On the other hand, the conventional method needs $(656 + 92\log_2 N)$ real-valued multiplications, 80 real-valued divisions, and 80 square root calculations per frequency bin to produce the same number of samples. As we choose appropriate values for M and N (e.g., $M \leq 4$, $N = 2048$), the complexity of the proposed algorithm can be kept lower than that of the conventional SAES algorithm.

3.4 Experimental Results

To evaluate the performance of the proposed ESAES method, we conducted computer simulations under various conditions. For performance assessment, we created 20 data sets from the TIMIT database such that each set consists of a source signal $v(n)$ and near-end signal $s(n)$. The data sets were sampled at 16 kHz. The length of each data set ranges from 10 s to 18 s and the total length of the data is 302 s. The duration of the double-talk interval is between 5 s and 10 s. Both the transmission room and the receiving room were designed to simulate a small office room of a size $4\,\mathrm{m} \times 4\,\mathrm{m} \times 3\,\mathrm{m}$. All of the room impulse responses (RIRs) were generated with reverberation time $T_{60} = 200\,\mathrm{ms}$ by means of the image method [43]. The length of the RIRs was set to 512. The echo level measured at the input microphone is on average 3.5 dB lower than that of the near-end speech. A white noise was added to the microphone signals such that SNR $= 30, 20$, and 10 dB. We applied a Hamming window of length 2048, which is half-overlapped for taking the STFT. In the experiments, the parameter values were set as follows: $\alpha_{\mathbf{\Phi}} = 0.999$, $\alpha_D = 0.001$, $\alpha_{DD} = 0.001$, $\alpha_{B_1} = 1.35$, $\alpha_{B_2} = 1.2$, $\beta_{\mathrm{t},1} = \beta_{\mathrm{t},2} = 0$, $\beta_{\mathrm{f},1} = 0.3$, $\beta_{\mathrm{f},2} = 0.32$, and $N = 2048$. It is noted that the window size was rather long, but the smoothing factors, α_D and α_{DD}

Table 3.1: ERLE and PESQ scores of proposed ESAES algorithm in noiseless conditions with different values of T and K

Augmented Vector	T \ K	ERLE (dB)			PESQ		
		0	1	2	0	1	2
—	0	12.23	20.71	24.51	2.52	2.81	2.88
Type 1	1	22.93	34.39	34.98	2.84	2.90	2.78
	2	23.82	35.91	36.04	2.85	2.80	2.51
Type 2	1	22.93	29.55	31.98	2.84	2.92	2.93
	2	23.82	30.95	33.43	2.85	2.91	2.91

were quite small.

To verify the performance of the proposed ESAES, we evaluated the PESQ score [41] and the ERLE measure which is defined by [16]

$$\text{ERLE}(n) = 10\log_{10}\left[\frac{E[y^2(n)]}{E[\hat{s}^2(n)]}\right] \text{(dB)} \tag{3.38}$$

where $\hat{s}(n)$ denotes the residual echo signal at time index n after suppressing far-end echoes in the single-talk case.

The overall results of the ERLE and PESQ scores obtained in noiseless conditions are shown in Table 3.1 for different values of T and K. Type 1 and Type 2 in Table 3.1 indicate the two ways of constructing the augmented vectors, given in (3.14) and (3.15), respectively. It is noted that the ESAES algorithm with ($T = 0$, $K = 0$) is equivalent to the conventional SAES with 50% window overlap. From the

Table 3.2: ERLE and PESQ scores of proposed ESAES, compared to SAES (Yang) in different SNR conditions

SNR		30 dB	20 dB	10 dB
ERLE	ESAES (Type 2, $T = 1$, $K = 1$)	28.66	21.32	11.22
	SAES	19.48	17.85	11.00
PESQ	ESAES (Type 2, $T = 1$, $K = 1$)	2.87	2.62	1.93
	SAES	2.71	2.52	1.91

whole results, we can observe that as more correlations among adjacent components are taken into account, the higher the ERLE performance becomes. This means that the correlation between adjacent time frames and frequency bins is helpful to suppress the echo signals effectively. On the other hand, the PESQ scores of the ESAES algorithm could not always be improved with increasing number of adjacent components. It is found that the ESAES algorithm with the augmented vectors of Type 2 is capable of maintaining the near-end signal more faithfully than that with the augmented vectors of Type 1. Furthermore, when we consider the adjacent components (Type 2, $T = 1, K = 2$), the best PESQ score is obtained. In other words, adjacent components other than these may not be beneficial to estimate the stereo echo accurately without distorting the near-end signal.

In Table 3.2, the performance of the ESAES algorithm, using Type 2 augmented vectors with ($T = 1$, $K = 1$) and half-overlapping windows is compared to that of the conventional SAES with 7/8-overlapping windows, under various SNR conditions. The augmented vector of Type 2 with ($T = 1$, $K = 1$) was chosen as it provides a

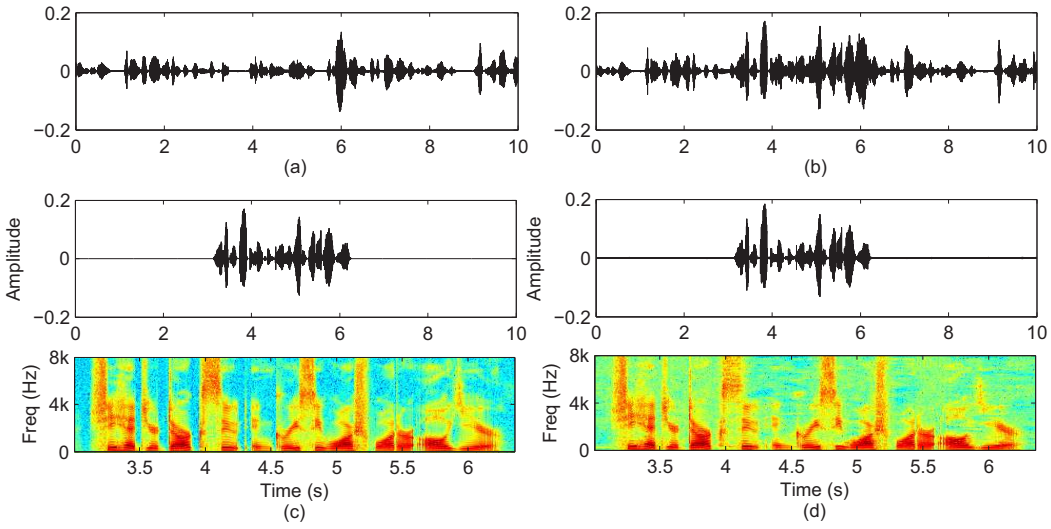

Figure 3.3: Waveforms and spectrograms for the double-talk case with 30 dB SNR. (a) one of the far-end signals, (b) microphone signal, (c) near-end speech, and (d) output of the ESAES.

good trade-off between the performance and computational complexity. We used the same parameters values as in [15] ($\beta_1 = 1.35,\ \beta_2 = 1.2,\ \alpha_\lambda = 0.6,\ \alpha_{DD} = 0.6,\ \alpha_\phi = 0.975,\ N = 2048$). In all the tested SNR conditions, the proposed approach outperformed the conventional SAES. In particular, it was found that the resulting signals of SAES had a significant level of residual echo compared to those of ESAES. Also, it could be seen that the ESAES preserved the near-end signal better as seen from the comparison of the PESQ scores or in Fig. 3.3, which illustrates double-talk performance through the waveforms and spectrograms of $x_1(n)$, $y(n)$, $s(n)$, and $\hat{s}(n)$.

We also investigated the tracking performance and convergence speed of the proposed ESAES and conventional SAES algorithms in the single-talk condition as displayed in Fig. 3.4. Fig. 3.4(a) shows the microphone signal $y(n)$ in the receiving

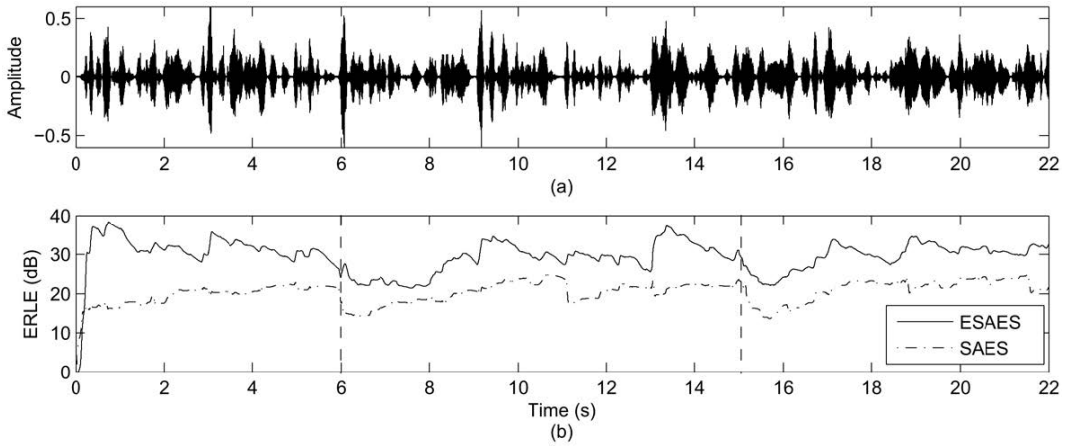

Figure 3.4: Comparison of tracking performance and convergence speed between the proposed ESAES and the SAES algorithms in the single-talk case. At 6 s, the source location in the transmission room was changed and at 15 s, the microphone in the receiving room moved, SNR $= 30$ dB and $T_{60} = 200$ ms. (a) $y(n)$ in the receiving room. (b) Temporal variation of ERLE.

room. In this experiment, the source location in the transmission room was changed at 6 s and the microphone in the receiving room changed its location at 15 s. Fig. 3.4(b) shows the variation of ERLE over time. From these results, we can observe that the proposed ESAES algorithm did not show any significant tracking difficulty in the dynamic environment and always outperformed the conventional SAES method.

3.5 Summary

In this chapter, we have proposed the ESAES algorithm using augmented vectors in order to incorporate spectral and temporal correlations. The approach takes

advantage of the correlations among components in adjacent time frames and frequency bins in the STFT domain. To estimate the stereo echo signal, the extended PSD matrices and cross-PSD vectors are derived from the signal statistics. Experimental results demonstrated that the proposed ESAES method is superior to conventional SAES in terms of both ERLE and PESQ.

Chapter 4

Nonlinear Residual Echo Suppression Based on Deep Neural Network

4.1 Introduction

Various acoustic echo cancellation (AEC) and suppression (AES) techniques have been proposed to reduce the echo components from the microphone signals when there exists acoustic coupling between loudspeakers and microphones [7], [8], [11], [33], [34], [44]. However, in most of the cases, a certain amount of residual echo remains at the output of these methods possibly due to the inherent nonlinearity of the loudspeakers and power amplifiers, nonlinear acoustic transfer function of the echo path, or imperfection of AEC and AES.

Several residual echo suppression (RES) post-filtering techniques have been introduced to further attenuate the remaining echo in the output of AEC or AES [9],

[16], [18], [35], [45], [46]. [9] and [16] proposed an RES gain function similar to that of the spectral subtraction or Wiener filters based on the signal-to-echo ratio (SER) estimated in a decision-directed manner. A subband filtering technique utilizing spectral subtraction was also developed for which the power spectral density of the residual echo was estimated using truncated Taylor series expansion [35]. In [45], the magnitude of the residual echo in a subband was approximated by a linear function of both the current and previous spectra of the far-end signal in the same band, while it was modeled as a function of the harmonic frequency components at the current frame of the far-end signal in [46]. Recently, RES based on artificial neural network (ANN) was proposed to model the mapping from the far-end to the residual echo signal [18]. The inputs to this ANN are the magnitude spectrum of the given frequency bin in the far-end signal and the sum of the spectra that may affect that frequency bin. Though some improvements have been observed, these methods are not considered to fully exploit the nonlinear relationship among the residual echo, far-end signal and AES output in all the frequency bins.

In this chapter, we propose a novel approach to nonlinear RES using deep neural network (DNN) which estimates the optimal RES gain based on both the far-end signal and the AES output. It is beneficial to utilize the DNN for describing the highly complex mapping between the RES gain and the relevant signals considering all frequency bins for several consecutive frames jointly since DNNs have been successfully applied to automatic speech recognition and speech enhancement [47]–[51] due to their capability in learning complicated mappings among various data. We have evaluated the overall performance of the proposed technique not only in matched but also in mismatched conditions with various RIRs, signal-to-noise ratios (SNRs), SERs, and amplifier characteristics. As measures of performance, we use echo return

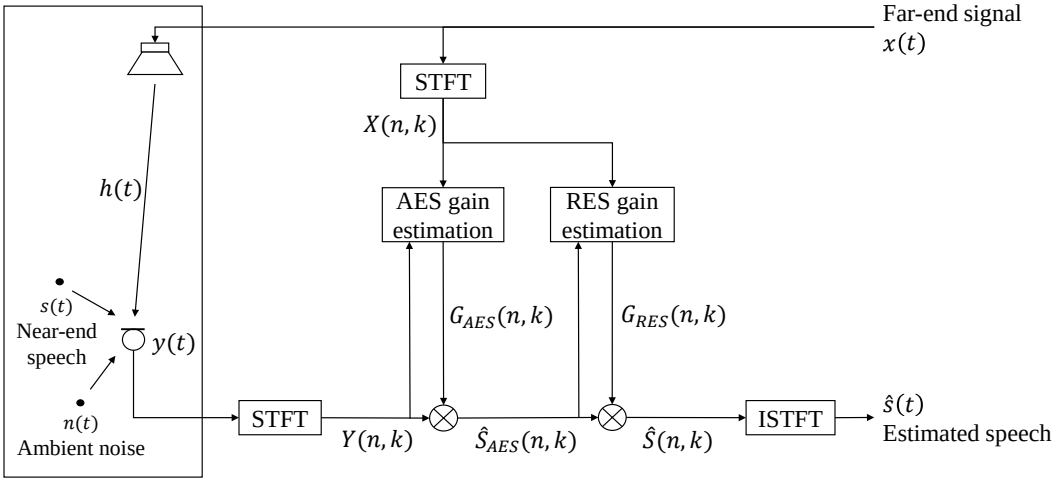

Figure 4.1: Schematic diagram of AES system with RES post-filter.

loss enhancement (ERLE) for single-talk periods and ERLE, signal-to-distortion ratio (SDR) [52], the ITU-T Recommendation P. 862 perceptual evaluation of speech quality (PESQ) scores [41] for double-talk periods. Experimental results show that the proposed method achieves improved speech quality and echo suppression compared to the conventional algorithm with ANN-based residual echo estimation and Wiener filtering [18].

4.2 A Brief Review on RES

AES [7], [8], [11], [33], [34], [44] modifies the spectra of the microphone signal aiming at attenuating the acoustic echo for hands-free communication or teleconference. A single-channel AES system is illustrated in Fig. 4.1. Let $x(t)$ denote the far-end signal. The microphone signal, $y(t)$, is composed of the echo signal, $h(t)*x(t)$, the near-end speech $s(t)$, and the ambient noise $n(t)$ where $h(t)$ represents the im-

pulse response of the echo path and $*$ is the convolution operation. Let $X(n,k)$ and $Y(n,k)$ be the short-time Fourier Transform (STFT) coefficients of $x(t)$ and $y(t)$ for the k-th frequency bin at the n-th frame, respectively. The spectral gain function of AES, $G_{AES}(n,k)$ is derived similarly to spectral subtraction or Wiener filtering. Due to various factors such as the limitations of power amplifiers or loudspeakers, the output of AES, $\hat{S}_{AES}(n,k)$ still possesses signal components caused by echo. To further suppress these components, an additional nonlinear RES filter can be applied to the AES output. With the RES gain $G_{RES}(n,k)$, the final estimated speech spectrum, $\hat{S}(n,k)$ is given by

$$\begin{aligned}\hat{S}(n,k) &= \{G_{RES}(n,k) \cdot G_{AES}(n,k)\}Y(n,k) \\ &= G_{RES}(n,k)\hat{S}_{AES}(n,k). \end{aligned} \tag{4.1}$$

Finally, the estimated speech signal $\hat{s}(t)$ is computed by taking the inverse STFT (ISTFT) with overlap-add.

4.3 Deep Neural Networks

In this section, we briefly introduce deep neural networks (DNNs) which is a conventional multilayer perceptron (MLP) with many hidden layers (≥ 2). Fig. 4.2 represents a DNN which consists of an input layer, hidden layers, and an output layer. For the sake of notation simplicity, let the input layer be layer 0 and the output layer be layer L for an $(L+1)$-layer DNN. The representation of the DNN at the l-th layer is formulated as follows

$$\mathbf{v}^l = \sigma(\mathbf{z}) = \sigma(\mathbf{W}^l\mathbf{v}^{l-1} + \mathbf{b}^l),\ (0 < l < L) \tag{4.2}$$

where $\mathbf{z} = \mathbf{W}^l \mathbf{v}^{l-1} + \mathbf{b}^l \in \Re^{N_l \times 1}$, $\mathbf{v}^l \in \Re^{N_l \times 1}$, $\mathbb{W}^l \in \Re^{N_l \times N_{l-1}}$, $\mathbf{b}^l \in \Re^{N_l \times 1}$, and N_l are the excitation vector, the activation vector, the weight matrix, the bias vector, and the number of neurons at l-th layer. The observation vector as input feature is $\mathbf{v}^0$ and its dimension is N_0. For the activation function $\sigma(\cdot)$, the sigmoid function

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \tag{4.3}$$

or the hyperbolic tangent function

$$tanh(z) = \frac{\exp(z) - \exp(-z)}{\exp(z) + \exp(-z)} \tag{4.4}$$

or the rectified linear unit (ReLU) function

$$\text{ReLU}(z) = \max(0, z) \tag{4.5}$$

is employed in the DNN framework and the sigmoid function is the most popular activation function in most applications, so it is assumed that this function is used unless noted.

The target data type at the output layer is chosen based on the tasks. For the regression tasks, the values in the output layer can be generated through a linear layer $\mathbf{v}^L = \mathbf{z} = \mathbf{W}^l \mathbf{v}^{L-1} + \mathbf{b}^L$ or a layer with sigmoid output functions $\mathbf{v}^L = \sigma(\mathbf{z}) = \sigma(\mathbf{W}^L \mathbf{v}^{L-1} + \mathbf{b}^L)$. The output vector $\mathbf{V}^L \in \Re^{N_L}$ and N_L denotes the output dimension. For the multi-class classification tasks, each output neuron represents each class $\{1, ..., i, ..., N_L\}$. The i-th value in the output layer indicates the probability $P(i|\mathbf{v}^0)$ where the observation vector $\mathbf{v}^0$ belongs to i-th class. To calculate a multinomial probability distribution, the output vector $\mathbf{v}^L$ has to be normalized. This process can be successfully done by applying a softmax function

$$v_i^L = softmax_i(\mathbf{z}^L) = \frac{\exp(z_i^L)}{\sum_{j=1}^{N_L} \exp(z_j^L)} \tag{4.6}$$

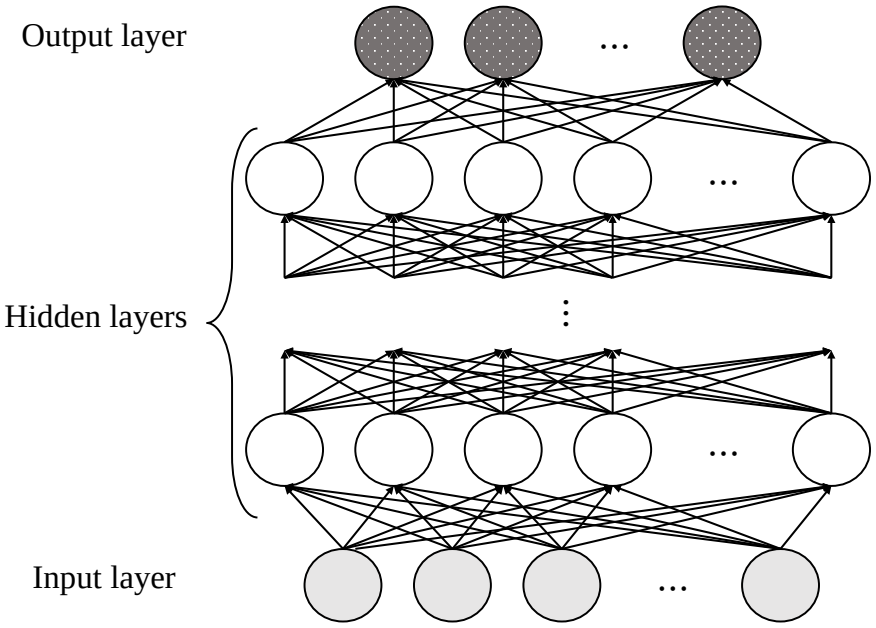

Figure 4.2: An example of deep neural network.

where z_i^L and N_L denote the i-th element of the excitation vector $\mathbf{z}^L$ and the number of classes at the output layer, respectively.

The DNN model parameters are unknown values and should be estimated from the training samples which are already labeled as the target values. Let the training set be $\{(\mathbf{o}_n,\ \mathbf{y}_n)|\ 0 \leq n < N\}$ where $\mathbf{o}$ and $\mathbf{y}$ are the n-th observation vector and the corresponding vector as the desired target output. This process is called the parameter estimation or the training process.

For the initialization of the DNN parameters, restricted Boltzmann machines (RBMs) can be utilized as a pre-train process [49]. After that, to train the DNN model, it is the most important to decide the training criterion because the goal

of the task and these are strongly correlated. The two popular training criteria are widely used for the practical applications. The first one for the regression tasks is the mean square error (MSE) criterion

$$J_{MSE}(\mathbf{W}, \mathbf{b}|\mathbf{o}, \mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} \sum_{n'=1}^{N_L} (y_{n,n'} - v_{n,n'}^{L})^2 \tag{4.7}$$

where n' denotes the n'-th element of $\mathbf{y}_n$. For the classification tasks, the second one is the cross-entropy (CE) criterion

$$J_{CE}(\mathbf{W}, \mathbf{b}|\mathbf{o}, \mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} \sum_{n'=1}^{N_L} -y_{n,n'} \log(v_{n,n'}^{L}) \tag{4.8}$$

where $\mathbf{y}$ is a probability distribution. For the supervised training process, the DNN model parameters can be learned with the error backpropagation (BP) algorithm which is based on gradient algorithm. This is called the fine-tuning process.

4.4 Nonlinear RES using Deep Neural Network

In this section, we propose an approach to estimate the RES gain based on DNN for signal-channel case. Here, the DNN system is employed to find the highly complex mapping between the RES gain and the relevant signals such as far-end signal and AES output. There are several reasons why we choose the RES gain as the target output of the DNN instead of the clean near-end speech or the residual echo signal. First, since the RES gain is confined to the finite range $(0, 1)$, it straightforwardly fits the output of a sigmoid function which is used as the activation function in the output layer. Also, additional gain modifications such as the application of minimum and maximum gains and temporal smoothing may be easily applied to the gain if necessary.

A DNN system for the proposed method is illustrated in Fig. 4.3 where $\boldsymbol{X}(n)$, $\hat{\boldsymbol{S}}_{AES}(n)$ and $\boldsymbol{G}_{RES}(n)$ are defined as follows:

$$\boldsymbol{X}(n) = [X(n,1) \dots X(n,K)]^T, \tag{4.9}$$

$$\hat{\boldsymbol{S}}_{AES}(n) = [\hat{S}_{AES}(n,1) \dots \hat{S}_{AES}(n,K)]^T, \tag{4.10}$$

$$\boldsymbol{G}_{RES}(n) = [G_{RES}(n,1) \dots G_{RES}(n,K)]^T, \tag{4.11}$$

where $K = N/2+1$ when taking N-point STFT and ^{T} denotes the transpose operation. The DNN is a feed-forward neural network which includes an input layer, three hidden layers and an output layer. The magnitude spectra of the far-end signal and the AES output in all frequency bins over T successive frames are fed to the input layer, which makes the number of the input units $2 \times K \times T$. The input features are normalized to have zero mean and unit variance. The output is the RES gain vector in the current frame, corresponding to K output units. Each hidden layer consists of binary units and the logistic sigmoid function is applied as nonlinear activation of the units. With this network structure, the proposed system can consider spectro-temporal correlations which may come from harmonic distortion, insufficient frequency resolution or nonlinear echo path.

To initialize the DNN parameters, we pre-train a model built by stacking restricted Boltzmann machines (RBMs) [49]. Taking account of the real-valued input feature, the first RBM is a Gaussian-Bernoulli RBM and two Bernoulli-Bernoulli RBMs can be stacked on top of the first RBM. These RBMs can be trained layer-by-layer in an unsupervised greedy fashion using contrastive divergence (CD). After this pre-training, we run the supervised fine-tuning stage where the back-propagation algorithm with the minimum mean squared error (MMSE) between the estimated RES gain, $G_{RES}(n,k)$ and the optimal gain, $G_{RES,opt}(n,k)$, is employed to train the

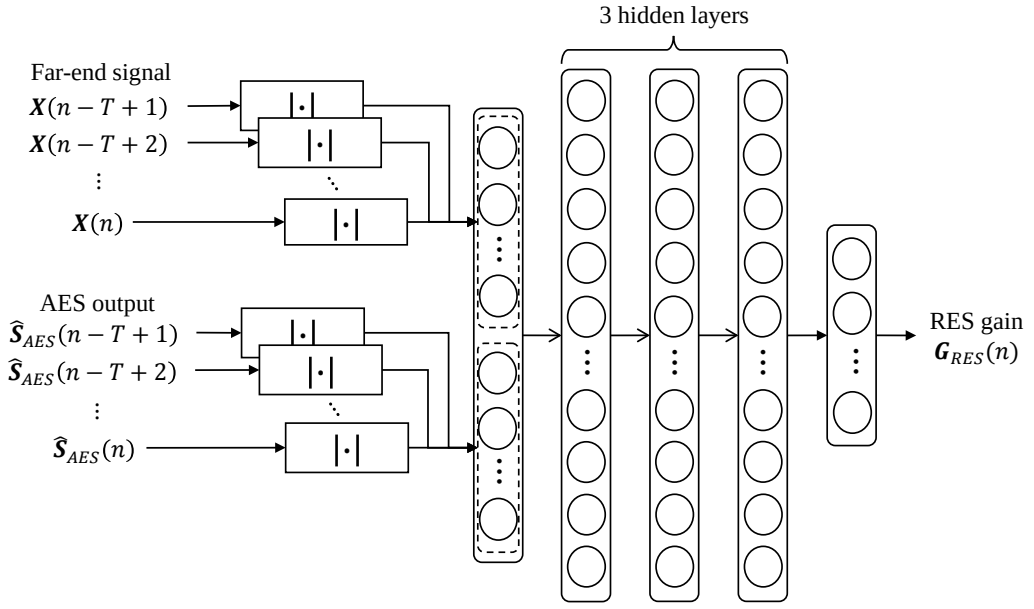

Figure 4.3: A DNN system for the proposed residual echo suppression.

DNN. The optimal RES gain $G_{RES,opt}(n,k)$ is defined as follows:

$$G_{RES,opt}(n,k) = \max\left\{G_{min}, \min\left(\frac{|S(n,k)|}{|\hat{S}_{AES}(n,k)|}, 1\right)\right\}, \tag{4.12}$$

where $S(n,k)$ is the STFT coefficient of the clean near-end speech and $G_{min} = 10^{-4}$ is introduced to reduce musical artifacts. A stochastic gradient descent algorithm is performed in mini-batches to improve learning convergence with the objective function,

$$\text{MMSE} = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}(G_{RES,opt}(m,k) - G_{RES}(m,k))^2 \tag{4.13}$$

where M is the mini-batch size. The detailed procedures for pre-training and fine-tuning are described in [49], [50].

As for the training data, multi-condition data with various levels of SNR and echo paths including both single- and double-talk cases are utilized. It is noted that considering only a few different echo paths was enough to train the mapping as the linear echo which is heavily dependent on the echo path was suppressed in the AES output. Moreover, the proposed deep structure can accommodate the mapping for the both single- and double-talk cases without use of any explicit double-talk detectors since the training data include both of the situations. The proposed method was the first attempt to suppress residual echo using DNN and the DNN structure can be easily used as a pre-processing of speech applications in a DNN framework.

4.5 Experimental Results

To evaluate the performance of the proposed RES technique, we conducted several experiments under various conditions. From the TIMIT database, we created 500 files (4444 s) as the far-end signals of which 400 files (3576 s) were used for training while the other 100 files (868 s) were used for the test. These files were sampled at 16 kHz. To simulate realistic nonlinear echo signal captured by the microphone, we performed three processes on the far-end signals: clipping, application of a model for a nonlinear loudspeaker, and convolution with RIRs. As for the artificial clipping mimicking amplifier characteristics, both the hard and soft clippings were considered. If $x(n)$ denotes the far-end input signal, the outputs of the hard and soft

clippings, x_{hard} and x_{soft} can be obtained as [53]

$$x_{hard}(n) = \begin{cases} -x_{max}, & x(n) < -x_{max} \\ x(n), & |x(n)| \leq x_{max} \\ x_{max}, & x(n) > x_{max} \end{cases} \tag{4.14}$$

and

$$x_{soft}(n) = \frac{x_{max}x(n)}{\sqrt[\rho]{|x_{max}|^{\rho} + |x(n)|^{\rho}}}, \tag{4.15}$$

respectively, where x_{max} is the maximum value of the output signal. For soft clipping, the value of ρ was set to 2. The output of the hard or soft clipping was then processed by a memoryless sigmoidal function simulating a nonlinear loudspeaker as follows [54] :

$$x_{NL}(n) = \gamma\Big(\frac{1}{1 + \exp\left(-a \cdot b(n)\right)} - \frac{1}{2}\Big) \tag{4.16}$$

where

$$b(n) = \frac{3}{2}x(n) - \frac{3}{10}x(n)^2, \tag{4.17}$$

in which $x(n)$ and $x_{NL}(n)$ are the input and output signals of the loudspeaker. The parameter γ is the sigmoid gain which was set to $\gamma = 2$. The sigmoid slope value a was chosen as $a = 4$ if $b(n) > 0$ and $a = 1/2$ otherwise. A receiving room was designed as a small office room with dimensions $4 \times 4 \times 3\,\mathrm{m}^3$. By using the image method [43], the RIRs from 9 loudspeaker locations to the microphone in the receiving room were generated with reverberation time $T_{60} = 200\,\mathrm{ms}$. The locations of the loudspeakers and the microphone are given in Fig. 4.4, and the length of the RIRs was set to 512. The echo level measured at the microphone was on average 3.5 dB lower than that of the near-end speech. For performance evaluation, the ERLE and PESQ scores [41]

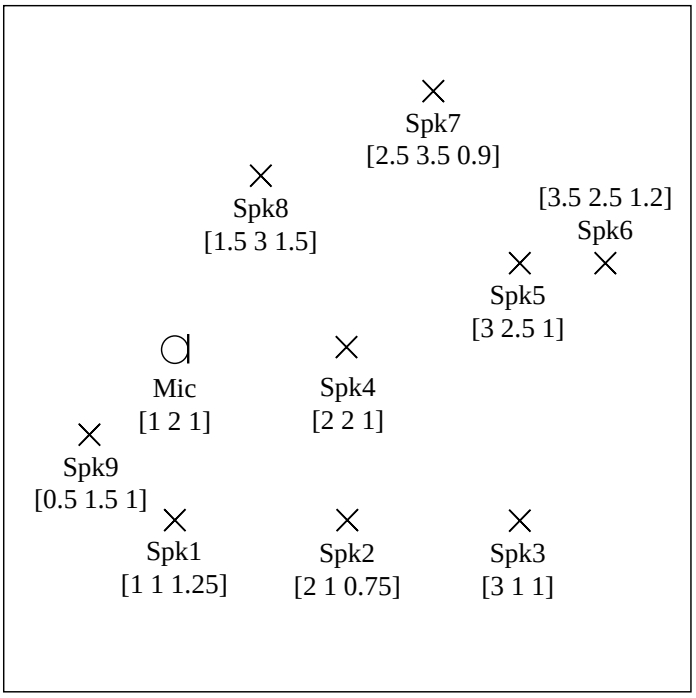

Figure 4.4: Locations of one microphone and 9 loudspeakers in a simulated receiving room of $4 \times 4 \times 3\,\mathrm{m}^3$ for echo DB.

were used as objective measures. The ERLE is defined by

$$\mathrm{ERLE}(t) = 10 \log_{10} \left[\frac{E[y^2(t)]}{E[\hat{s}^2(t)]} \right] (\mathrm{dB}). \tag{4.18}$$

First, the conventional AES technique [34] was applied to the whole data set. The AES algorithm was slightly modified so that it fitted to single channel AES by eliminating the second-channel echo estimation. The parameters for the AES were set to the values given in [34]. The average ERLE for the tested data was approximately 9 dB due to the severe nonlinear distortion even though the AES algorithm was shown to remove the linear echo to a certain extent.

Second, we implemented the ANN-based RES method using spectral features [18] to compare with the proposed RES technique. The 128-point STFT was applied with 75% overlap. The estimator in each frequency bin for the residual echo was a feed-forward network with two hidden layers of two log-sigmoid units. The magnitude spectra of the far-end signal in the current band and the power over all subbands up to the half of the current band were used as the inputs. The learning rate for the online adaptive adjustment of the residual echo implemented with LMS algorithm was set to 0.008, which resulted in the best performance empirically. The training was performed on 30 files (267 s) of the residual echo applying RIRs from the locations of Spk1, Spk2, and Spk3 to that of the Mic in Fig. 4.4. The parameters were set as follows: the smoothing parameter $\lambda = 0.95$ and the echo suppression factor $\mu = 5.0$. The information of double-talk periods was manually marked and applied to the method instead of applying a double-talk detector. We also tried training with larger DB or taking 256-point STFT, but neither of them could bring about performance improvement.

For the training of the proposed technique, the 4800 files (11.92 hours) established with a hard clipping at 80% of the maximum amplitude for clean and 3 SNR levels (30, 20, and 10 dB) at the locations of Spk1, Spk2, and Spk3 were used to train the DNN. The additive noise was white Gaussian noise. The frame length was set to 256 samples with 50% overlap, and a 256-point STFT was applied to each frame. Each hidden and the output layer had 2048 and 129 units, respectively. The input vector included the current frame and the previous two frames, which made it a 774-dimensional vector. In the pre-training, the number of epochs for the RBM in each layer was 20 and the learning rate was 0.0005. In the fine-tuning, the learning rate was set to 0.1 for the first 10 epochs, then decreased by 10% after each epoch.

Total iteration number was 50 and the mini-batch size M was set to 256. For the tests, we generated two copies of 100 far-end speech files (868 s) for each of the 9 locations of the loudspeaker. One copy was mixed with another 100 files of near-end speech selected from TIMIT DB to evaluated double-talk performance, while the other one was used to assess the performance for single-talk periods.

In Table 4.1, the overall results of the ERLEs for single-talk periods and PESQ scores for double-talk periods without additive noise are shown, where the test data were constructed with a hard clipping at the 80% of the maximum amplitude of the input signal. The result demonstrates that the proposed method based on DNN outperformed the conventional RES [18]. Although the proposed method were trained with only a few of the RIRs, the performance for the matched and mismatched loudspeaker positions did not show significant differences. It may support our assumption that the mapping from both the far-end signal and the AES output to the RES gain would not be substantially affected by the acoustic environment.

Table 4.2 shows the performance at the fifth loudspeaker's position under various SNR conditions with a hard clipping at 80% of the maximum amplitude. Both the ERLEs and PESQ scores of the proposed method were improved compared with those of the conventional one. The performance of the proposed method with different numbers of units in each hidden layer is also demonstrated which shows a trade-off between the computational complexity and the performance of the RES algorithm, although the performance with 256 units was still significantly better than that of the conventional method.

To examine the effects of other factors such as signal-to-echo ratios (SERs), clipping types and amounts of clipping on the RES algorithms, we additionally tested several cases corresponding to other mismatches at the location of Spk5 without

additive noise, of which the result is given in Table 4.3. SER 0 dB means that the near-end speech level was adjusted so that the near-end speech to echo ratio was on average 0 dB. HC (l%) and SC (l%) indicate the hard and soft clipping at l% of the maximum amplitude of the input signal, respectively. In all 4 cases, the proposed method outperformed the conventional RES.

Through the PESQ comparison between the proposed and the conventional methods, the performance of the proposed algorithm based on DNN is more improved than the ANN-based. To validate the detail performance in double-talk periods, additional measures were adopted. The first measure is ERLE in double-talk (DT) and another one is speech-to-distortion ratio (SDR) which is defined as [52]

$$SDR \triangleq 10 \log 10 \frac{|s_{target}|^2}{|e_{interf} + e_{noise} + e_{artif}|^2} \text{ [dB]} \tag{4.19}$$

where s_{target} is a version of the true source modified by an allowed distortion and e_{interf}, e_{noise}, e_{artif} are respectively the interferences, noise and artifacts error terms. In Table. 4.4, the overall performance of the ERLEs and SDR for double-talk periods without noise is shown. The clipping condition is the same to the environment represented in Table. 4.4. As expected, the performance result shows that the proposed technique outperformed the conventional RES. Specially, compared with the ANN-based RES, ERLE in double-talk improves more than 2 dB on average. Consequently, it if found that the proposed method has the notable capacity for residual echo suppression in both single-talk and double-talk periods. Since the DNN-based algorithm can continuously track the residual echo by using the trained information in double-talk periods, the RES performance may be kept regardless of the existence of the near-end speech. On the other hand, the ANN-based method could not track the residual echo in double-talk periods because it is frozen or stopped to avoid the

divergence problem when the double-talk situation is detected. The SDR performance of the proposed method is also superior to the conventional one. While the ANN-based method may degrade the speech components, it can be seen that the proposed one recovers these components from the SDR comparison. In other words, we conclude that the DNN-based RES can suppress the residual echo including the nonlinearity components and preserve the near-end speech effectively.

Considering mismatched cases in different SNR conditions at the location of fifth speaker, the performance of both two methods is shown in Table. 4.5. Actually, the DNN structure of the proposed method was already trained on the various noise conditions, so both the ERLE and SDR of the proposed one were improved compared with the ANN-based algorithm. Therefore, the DNN-based technique may achieve not only residual echo suppression but also slight noise reduction. Whenever the size of each layer is larger, SDR performance is higher but ERLE results in DT are almost same. Thus, the DNN-based technique may have the capability for robust speech recovery when the residual echo and noise occur simultaneously.

We checked the effects of other factors (SER, clipping types and amounts of the clipping) as done in Tabel. 4.3 and the result is given in In Table. 4.6. Similar to the previous cases (Table. 4.3), the proposed method outperformed the conventional ANN-based RES in both terms of ERLE and SDR. The proposed method may be more robust to sereral mismatch environment than the conventional RES since the DNN can train the nonlinear characteristics of the residual echo by using a number of neurons and deep layers.

Fig. 4.5 shows the evolution of the ERLE over time in conjunction with the corresponding unprocessed echo waveform. Again, we could confirm the proposed algorithm attenuated the residual echo components more effectively than the con-

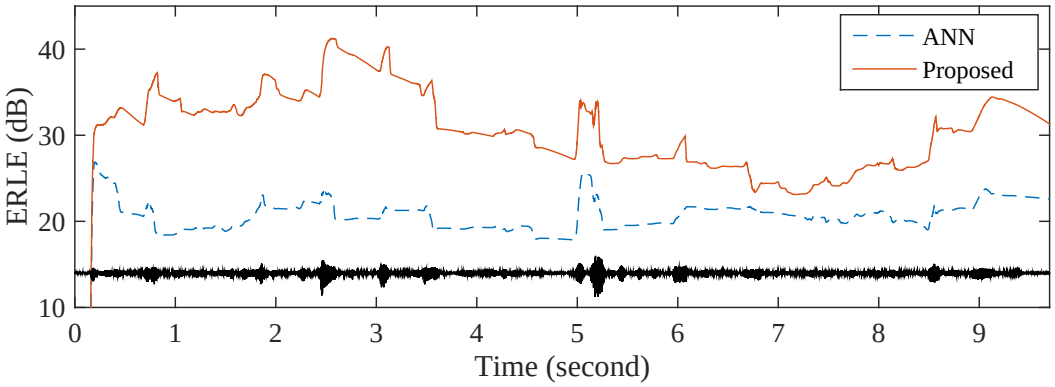

Figure 4.5: Comparison of ERLE at the location of Spk5 in a single-talk situation with SNR = 30 dB and T_{60} = 200 ms.

ventional RES.

4.5.1 Combination with Stereophonic Acoustic Echo Suppression

In this subsection, we evaluated the performance of the proposed method in the stereophonic case. We conducted several experiments under various conditions. From the TIMIT database, we created 500 files (4459 s) as the far-end signals of which 400 files (3591 s) were used for training while the other 100 files (868 s) were used for the test. These files were sampled at 16 kHz. The RIR simulation by image method [43] was also the same condition to that in [34]. The clipping condition was a hard clipping at 80 % of the maximum amplitude and the nonlinear function used in Section 4.5 were applied on the experiments. In the data, 4 SNR conditions were considered, which are clean, 30 dB, 20 dB, and 10 dB. The additive noise was white Gaussian noise. Considering the stereophonic AES, the method ($TK = 11, Type2$) [34] which was proposed in the previous chapter was used and the parameter setting was the

same.

In the proposed method, the number of the hidden layers in the DNN was 2. Each hidden and the output layer had 2048 and 129 units, respectively. The input vector included the current frame and the previous frame. The pre-training and the fine-tuning were operated through the same parameters and procedures in Section 4.5. These evaluations can be seen as tests in the matched conditions since the training and test environments were equivalent.

For objective evaluation, ERLE and PESQ were used in the various environments and the results are illustrated in Table. 4.7. Applying the proposed RES technique on the SAES case, the overall performance was much improved compared with only following SAES. It is found that the ERLE and PESQ score increased at least 20 dB and 0.3 point, respectively. Therefore, the proposed algorithm can also suppress the nonlinear echo in the stereophonic case.

For subjective evaluation, we conducted the mean opinion score (MOS) test in the various SNR conditions in verify the perceptual quality improvement. For this, 14 professional listeners decided the subjective score. The results are graded using a five-point score [55]. In Table. 4.8, we concluded that the proposed algorithm is superior in suppressing the nonlinear echo.

Last, we check the real time factor (RTFs) for complexity of the proposed RES. The RTFs for CPU were measured using one core of Intel Xeon E5-2620 2.4 Ghz processor. The regression process of the proposed algorithm was done with Kaldi platform. Since the process time in the system was 0.086 RTF, we think that the proposed RES can be applied to several applications based on DNN like speech recognition.

4.6 Summary

In this chapter, we have proposed an optimal residual echo suppression gain regression employing DNN. The DNN could represent the complicated mapping from the AES output and far-end signal in the whole frequency bins to RES gains. Furthermore, the proposed method does not need any explicit double-talk detectors as the DNN can accommodate the mapping for both single-talk and double-talk cases. The proposed RES algorithm outperformed the conventional one in terms of ERLE for single-talk situations and ERLE, SDR and PESQ scores for double-talk situations.

Table 4.1: ERLE and PESQ scores obtained with the matched and mismatched RIRs.

Measure	Condition		None	ANN [18]	Proposed
ERLE	Matched	Spk1	9.04	21.73	34.73
		Spk2	9.74	22.82	36.57
		Spk3	9.01	21.85	35.94
	Mismatched	Spk4	9.26	22.19	33.62
		Spk5	9.51	21.91	33.85
		Spk6	8.63	20.84	35.58
		Spk7	9.89	23.21	34.12
		Spk8	9.68	21.68	32.87
		Spk9	10.20	22.57	32.80
PESQ	Matched	Spk1	2.58	2.67	2.95
		Spk2	2.64	2.68	3.01
		Spk3	2.64	2.70	3.04
	Mismatched	Spk4	2.64	2.67	2.99
		Spk5	2.62	2.69	2.98
		Spk6	2.57	2.63	2.92
		Spk7	2.67	2.70	3.00
		Spk8	2.68	2.72	3.02
		Spk9	2.67	2.71	3.00

Table 4.2: ERLE and PESQ scores in different SNR conditions at the location of Spk5.

Measure	SNR	None	ANN [18]	Proposed		
				Size of each layer		
				256	512	2048
ERLE	Clean	9.51	21.91	31.06	31.89	33.85
	30 dB	9.48	21.90	31.18	31.83	33.84
	20 dB	9.23	21.57	29.36	29.91	31.92
	10 dB	7.38	18.73	25.62	25.96	27.37
PESQ	Clean	2.62	2.69	2.83	2.86	2.98
	30 dB	2.59	2.69	2.82	2.85	2.97
	20 dB	2.46	2.63	2.60	2.64	2.78
	10 dB	2.07	2.35	2.45	2.47	2.57

Table 4.3: ERLE and PESQ scores in the various mismatched conditions at the location of Spk5.

Measure	Condition	None	ANN [18]	Proposed
ERLE	SER 0 dB	9.57	21.91	32.71
	HC (70%)	9.50	21.84	34.43
	SC (80%)	9.47	21.61	33.48
	SC (70%)	9.46	21.53	33.10
PESQ	SER 0 dB	2.40	2.48	2.76
	HC (70%)	2.61	2.68	2.98
	SC (80%)	2.52	2.60	2.89
	SC (70%)	2.50	2.58	2.87

Table 4.4: ERLE in double-talk (DT) and SDR obtained with the matched and mismatched RIRs.

Measure	Condition		None	ANN [18]	Proposed
ERLE in DT	Matched	Spk1	0.91	1.90	3.80
		Spk2	1.11	2.10	4.29
		Spk3	0.83	1.75	4.02
	Mismatched	Spk4	0.86	1.79	3.66
		Spk5	0.95	1.96	4.04
		Spk6	1.10	2.22	4.06
		Spk7	0.85	1.75	3.84
		Spk8	0.84	1.83	3.77
		Spk9	1.02	1.94	3.94
SDR	Matched	Spk1	13.23	12.71	15.83
		Spk2	13.16	12.38	15.60
		Spk3	13.56	12.73	15.21
	Mismatched	Spk4	13.71	12.92	15.97
		Spk5	12.71	12.49	15.52
		Spk6	13.30	12.44	15.65
		Spk7	14.17	13.02	16.19
		Spk8	13.65	12.39	16.00
		Spk9	13.87	12.45	15.86

Table 4.5: ERLE in DT and SDR in different SNR conditions at the location of Spk5.

Measure	SNR	None	ANN [18]	Proposed		
				Size of each layer		
				256	512	2048
ERLE in DT	Clean	0.88	1.79	3.79	3.98	3.66
	30 dB	0.84	1.79	3.75	3.87	3.75
	20 dB	0.77	1.80	2.60	2.65	2.63
	10 dB	0.57	2.22	3.01	2.88	3.09
SDR	Clean	13.71	12.92	15.70	15.82	15.97
	30 dB	13.36	12.89	15.62	15.74	15.91
	20 dB	13.08	12.65	15.04	15.17	15.36
	10 dB	11.41	11.01	12.91	13.08	13.16

Table 4.6: ERLE in DT and SDR in the various mismatched conditions at the location of Spk5.

Measure	Condition	None	ANN [18]	Proposed
ERLE in DT	SER 0 dB	1.51	2.59	4.10
	HC (70%)	0.90	1.83	3.30
	SC (80%)	1.09	2.10	3.46
	SC (70%)	1.15	2.17	3.52
SDR	SER 0 dB	11.10	11.25	14.41
	HC (70%)	13.58	12.84	15.94
	SC (80%)	12.75	12.34	15.36
	SC (70%)	12.53	12.21	15.21

Table 4.7: ERLE and PESQ in the various SNR conditions.

Measure	Method	Clean	SNR 30 dB	SNR 20 dB	SNR 10 dB
ERLE	SAES	12.99	12.94	12.42	9.15
	SAES + proposed	39.76	39.03	32.85	30.09
PESQ	SAES	2.768	2.690	2.560	2.280
	SAES + proposed	3.100	2.988	2.875	2.589

Table 4.8: MOS results for subjective test in the various SNR conditions.

Method	Clean	SNR 30 dB	SNR 20 dB	SNR 10 dB
None	2.74	2.74	2.37	1.97
SAES	3.80	3.37	2.97	2.40
SAES + proposed	4.53	4.30	3.89	2.76

Chapter 5

Enhanced Deep Learning Frameworks for Nonlinear Acoustic Echo Suppression

5.1 Introduction

Due to the popularity of mobile phones and hands-free devices, nonlinear acoustic echo cancellation has become important and been developed over the last decades. Specially, cheap amplifiers and loudspeakers used in these devices mostly generate significant nonlinearities in echo signal which is not a linear relationship with the far-end signal any more even when the echo path is perfectly linear. These components cannot be easily removed by the linear echo cancellation algorithms which are kinds of adaptive filter methods based on gradient theory such as the least mean square (LMS), the recursive least squares (RLS), and affine projection (AP) algorithms.

To alleviate the nonlinear acoustic echo problem, some methods have been stud-

ied. In [26], based on the tap-delayed neural networks (TDNN), the nonlinear portion was first estimated and adaptively updated with the LMS scheme. Also, the adaptive Volterra filters have been widely used and revised since the structure of these filters can be seen as a straightforward generalization of linear adaptive filters. In [19], a memoryless polynomial Hammerstein model or its cascaded model with a linear finite impulse response (FIR) filter was exploited to describe the nonlinear characteristic. These filters can be considered as a subclass of the Volterra series filter. Similarly, a cascaded structure which consists of polynomial Volterra filters for the nonlinear loudspeaker and the normalized LMS (NLMS) algorithm for the linear property was proposed [20]. By Kuech and Kellermann, an approach based on adaptive second-order Volterra filter was proposed [21]. This method is regarded as an extension of partitioned block algorithm for efficient computation. Also, to approximate generic N-th order Volterra filters, the combinations of a linear kernel and quadratic kernels were introduced [22]. Based on the linear-to-nonlinear power ratio, this method can control the amount of nonlinear echo which has to be estimated from the quadratic kernel. For more efficient Volterra alternative, another kernel combination technique was integrated in the framework [23]. In this way, each kernel of a single Volterra filter was replaced by a combination of kernels for lower complexity. In [24], nonlinear echo power estimation using the second-order Volterra filter was adopted for acoustic echo suppression in the frequency-domain. For the precise estimation, a soft decision scheme with a *priori* probability of near-end speech absence was incorporated in the AES method. Besides, power filters as approximations of nonlinear acoustic echo paths that can be modeled by the cascade of a linear filter were proposed to improve the convergence speed [25]. In addition, a kernel which consists in a weighted sum of the linear and the Gaussian kernels

was exploited in the kernel based AP algorithm [27] or a multichannel structure for modeling a nonlinear echo path was proposed [53]. A functional link adaptive filtering was also used for capturing the nonlinearity in echo signal [54]. Recently, deep neural networks (DNNs) were adopted to suppress the nonlinear echo components in residual echo [56]. A DNN architecture, which is suitable to model a complicated nonlinear mapping between high-dimensional vectors, was employed as a regression function from these signals to the optimal RES gain. This method demonstrated improved speech quality and echo suppression performance compared with the simple ANN-based residual echo suppression [18].

Inspired by [56], in this chapter, we propose a novel approach in DNN framework for nonlinear acoustic echo suppression (NAES). Similar to residual gain estimation for nonlinear components, the proposed algorithm can directly enhance the input signal by applying the gain estimation based on DNNs which have the capability to automatically learn an arbitrary unknown mapping from the input to the target values. However, the architectures may be impossible to track the nonlinear echo signal in the various environments or room impulse responses (RIRs) because their networks are fixed on the contrary to adaptive filters. To compensate this weakness, additional inputs for echo information such as *a priori* and *a posteriori* signal-to-echo ratio (SER) are used in the DNN and this process is called echo aware training. Also, we introduce the multi-task learning (MTL) [57]–[60] to improve the gain estimates. In the MTL framework, related works are jointly trained with shared hidden layers to improve the generalization power of each task. In the proposed technique, the primary task of the gain estimation for NAES is jointly trained with an additional task of double-talk detection. Therefore, the proposed method makes the DNN be more robust in unseen conditions. Experimental results show that the proposed

method outperformed the conventional one, especially in double-talk situations.

5.2 DNN-based Nonlinear Acoustic Echo Suppression using Echo Aware Training

To remove nonlinear acoustic echo components in various environments, several echo cancellation methods have been researched over the last decades [19]–[27], [53], [54]. One of the well-known NAEC approaches is an adaptive Volterra filtering which can be viewed as a generalization of linear adaptive filters. In the past, various algorithms based on the Volterra filter were proposed to describe the characteristics of nonlinear echo and showed the better performance than the conventional linear filtering [19]–[26]. However, the performance might be not satisfied for hearing or using nonlinear echo-removed signal as features of other applications since these algorithms could not consider the full correlation for the nonlinear relationship between the input signal and far-end signal in time-frequency domain.

In speech recognition and enhancement areas, deep neural network (DNN) structures have been employed as a powerful tool to find the complicated mapping or functions and shown better performance than other conventional methods [47]–[49]. Recently, a DNN-based residual echo suppression (RES) was proposed to reduce the nonlinearity and the work has shown that its performance was more improved than the conventional ANN-based one [56]. Extending the scheme, we propose a novel DNN-based approach for NAES. Instead of estimating the residual gain for suppressing the nonlinear echo components, the proposed algorithm straightforwardly recovers the near-end speech signal through the direct gain estimation obtained from DNN frameworks on the input and far-end signal. However, DNN structures may be

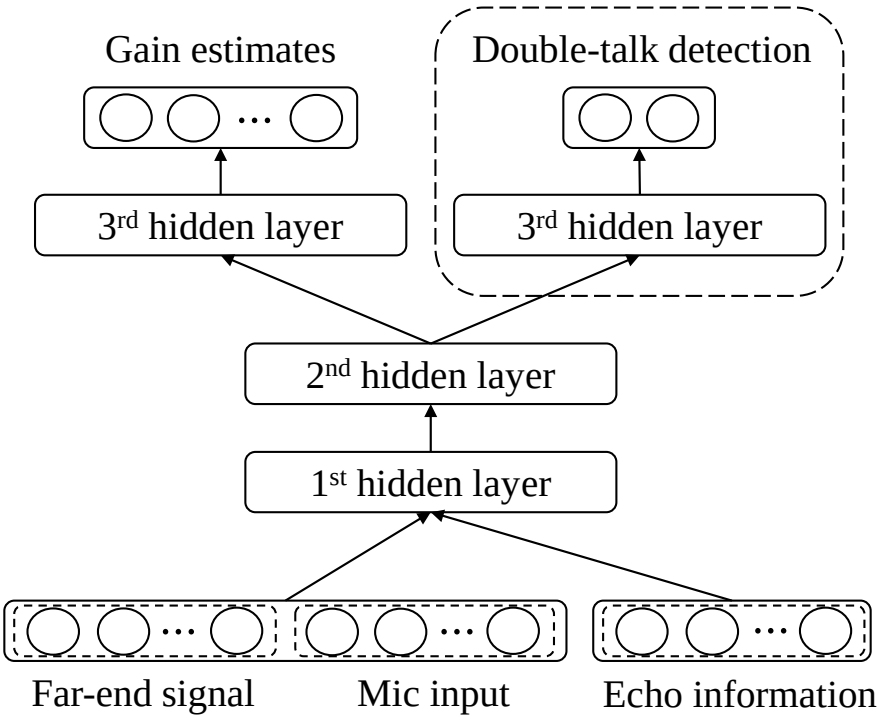

Figure 5.1: The proposed echo aware DNN structure with multi-task learning on double-talk detection for nonlinear acoustic echo suppression.

difficult to track the nonlinear echo signal in the various environments or room impulse responses (RIRs) because their networks cannot be changed in the test phase unlike the conventional NAEC or NAES based on adaptive filters.

Recent works have shown that the performance of DNN-based techniques can be improved by introducing the auxiliary information as other inputs which are extracted from the various environments or signals. Several DNN algorithms based on noise or room aware training have been widely used for speech recognition [57] or

speaker adaptation [61]. In noise aware training, the estimates of the additive noise corrupting the utterance were employed to improve the performance [57]. Room awareness information estimated in the speech signal was also used to work for dereverberation [61]. Similarly, in order to compensate the weakness of DNN-based NAES that the performance of the DNN system may be degraded in unseen and mismatched conditions, echo aware information is introduced as supplementary input in the proposed algorithm. In Fig. 5.1, it is found that input signals for DNN are not only microphone signal and far-end signal but also echo information. In this work, we focus on the signal-to-echo ratio (SER) information. The SER features such as *a priori* and *a posteriori* SERs are the well-known information for AES task and suitable things for tracking the change of echo signal or RIRs. Especially, the echo information can be used for implicit double-talk detectors in each frequency bins. *A priori* SER $\xi(n,k)$ and *a posteriori* SER $\gamma(n,k)$ can be defined as

$$\xi(n,k) \triangleq \frac{\lambda_S(n,k)}{\lambda_D(n,k)}, \tag{5.1}$$

$$\gamma(n,k) \triangleq \frac{|Y(n,k)|^2}{\lambda_D(n,k)} \tag{5.2}$$

where $\lambda_S(n,k)$ and $\lambda_D(n,k)$ denote the power spectral densities (PSDs) of the near-end signal and nonlinear echo, respectively. $|Y(n,k)|$ is the magnitude of the microphone input and n and k are the n-th frame index and k-th frequency bin in short-time Fourier transform (STFT) domain. The echo information can be obtained by applying the conventional AES approaches or training a DNN. In this work, one of the former methods was used for collecting the echo information. This approach [56] is based on decision-directed method for AES. Since the SER features are estimated per each frequency bin, we consider the a $2 \times (N/2+1)$-dimensional vector per frame as an additional input when taking the N-point STFT. Additionally, their

logarithms were finally taken for reducing the dynamic range of the SER values. The magnitude spectra of the far-end signal and the microphone input and the SER information in all frequency bins over T successive frames are fed to the input layer. These features are normalized to have zero mean and unit variance. The output is the gain vector in the current frame. Each hidden layer consists of binary units and the logistic sigmoid function is applied as nonlinear activation of the units.

Similar to the previous work [56], the optimal gain $G_{opt}(n,k)$ for NAES which is called phase-sensitive gain [62] is adopted and defined as follows

$$\begin{aligned} G_{opt}(n,k) &= \Re\left(\frac{S(n,k)}{Y(n,k)}\right) \\ &= \frac{|S(n,k)|}{|Y(n,k)|}\cos(\theta^S - \theta^Y) \end{aligned} \tag{5.3}$$

where θ^S and θ^Y denote the phase of the near-end speech and the phase of the microphone input, respectively. The gain is confined to the finite range (0,1) and the output of a sigmoid function is used as the activation function in the output layer. The phase-sensitive gain based on the error in the complex spectrum, which includes both amplitude and phase error may compensate for the use of the noisy phase.

5.3 Multi-Task Learning for NAES

In this section, we propose to combine a multi-task learning with the DNN-based NAES incorporating echo aware training for robust NAES. Generally, in the multi-task learning, the DNN model with shared hidden layers can be trained by performing several related tasks simultaneously. The advantage of the multi-task learning is that the information or representations obtained from one task may be

helpful for solving other tasks, and vice versa. Thus, multi-task learning allows to find internal information which cannot be discovered by training the model on each isolated task. Actually, through echo aware training, the DNN can learn the improved nonlinear mapping between the input features including the SER information and the gain estimates for NAES. However, the DNN may have difficulties in estimating the NAES gain estimates when there exists various mismatched conditions such as different devices or room environments.

In the proposed work, we introduce the multi-task learning technique which combines the gain estimation as the primary task with a double-talk detection as an additional task during the DNN training phase. The proposed approach is illustrated in Fig. 5.1. The main task in the left part of the DNN is to minimize the squared error between the optimal gain and the estimate gain which are corrupted by nonlinear echo. The objective function for this task is formulated by

$$J_{primary} = \sum_{n} \sum_{k} [G_{opt}(n,k) - G_{est}(n,k)]^2 \tag{5.4}$$

where G_{est} denotes the gain obtained from the NAES based on DNN. The right part of the network is treated as an additional DNN to classify as single-talk or double-talk period using cross entropy function. The second objective function for the binary classification can be defined as

$$J_{additional} = \sum_{n} \sum_{i} (-y_i(n) \log y_i'(n)) \tag{5.5}$$

where i denotes i-th class and y_i and y_i' is the true and the estimated value. Thus $y_1 = 1$ and $y_0 = 0$ when $i = 1$ and $i = 0$ are double-talk and single-talk, respectively. The softmax function is used for the output activation. Using (5.4) and (5.4), the

final objective function for multi-task learning is given by

$$
\begin{aligned}
J &= \lambda J_{primary} + (1-\lambda) J_{additional} \\
&= \lambda \sum_n \sum_k [G_{opt}(n,k) - G_{est}(n,k)]^2 + (1-\lambda) \sum_n \sum_i (-y_i(n) \log y_i'(n))
\end{aligned} \tag{5.6}
$$

where λ is the weight parameter between the gain estimation and double-talk detection tasks. In this architecture, a single hidden layer for each task is used and the two lower hidden layers are shared for multi-task learning. The activation function in the hidden layers is sigmoid. Similar to the conventional DNN approaches, the DNN is trained by passing through the pre-training and fine-tuning. To initialize the DNN, we pre-train a model built by stacking restricted Boltzmann machines (RBMs) [49] which is layer-wise unsupervised learning algorithm. Then, in the fine-tuning phase, the objective function (5.6) is applied on the DNN training.

Since the DNN-based NAES does not have any explicit double-talk detectors, the estimation of the NAES gain may be inaccurate in double-talk intervals. The information extracted from the subsidiary task may help to make the DNN be robust in this case. In other words, the network can learn the good representations which can suppress more in single-talk periods and improve the gain estimates in double-talk periods through the double-talk detection task. Another advantage of multi-task learning is that the DNN for the additional task is used only for training phase, so it can be discarded during the test phase. Thus, the complexity of the DNN does not increase during the test and the same processes without multi-task learning for recovering the near-end speech can be applied on the multi-task learning framework.

5.4 Experimental Results

To evaluate the performance of the proposed method, we conducted several experiments under real environments. From the TIMIT database, we created 150 files as the far-end signals of which 100 files were used for training while the other 50 files were used for the test. These files were sampled at 16 kHz. To record nonlinear echo signal in real conditions, we prepared 3 mobile devices (M1, M2, M3) and 2 room environments (R1, R2). The recording conditions are represented in Table. 5.1. The dimension of the first room is $5.30 \times 4.30 \times 2.35$ m^3 and the reverberation time T_{60} in the room is about 200 ms. The second room was designed with dimensions $8.09 \times 5.21 \times 2.70$ m^3 and its reverberation time $T_{60} \approx 382$ ms. For double-talk situation, the near-end speech is mixed with the recording echo data and the echo level measured at the microphone was on average 3.5 dB lower than that of the speech signal. For performance evaluation, the echo return loss enhancement (ERLE), the perceptual evaluation of speech quality (PESQ) [41], the ERLE in double-talk periods (DT) and the segmental speech-to-speech distortion ration (SSDR) [63] were used as objective measures. The ERLE is defined by

$$\mathrm{ERLE}(n) = 10 \log_{10} \left[\frac{E[y^2(n)]}{E[\hat{e}^2(n)]} \right] \text{(dB)}. \tag{5.7}$$

where $y(n)$ and $\hat{e}(n)$ denote the microphone input and the residual echo, respectively. The segmental SSDR is formulated as

$$SSDR_{seg} = \frac{1}{C(\Lambda)} \sum_{l \in \Lambda} \left[10 \log_{10} \frac{\Sigma_{k=1}^{K} S_l^2(k)}{\Sigma_{k=1}^{K} (\hat{S}_l(k) - S_l(k))^2} \right] \text{(dB)} \tag{5.8}$$

where the term $C(\Lambda)$ is the number of elements in set Λ representing a subset with speech being present and $\hat{S}_l(k)$ and $S_l(k)$ denote the enhanced and the clean speech, respectively. Since the amplifier and loudspeaker in the mobile devices were

Table 5.1: The recording conditions for training and test DB (mobile devices = {M1, M2, M3} and room environments = {R1, R2}).

Dataset	Conditions
Training	(M1,R1), (M1,R2), (M2,R2)
Test	(M1,R1), (M2,R1), (M3,R2)

so cheap, we expected the recorded echo signal to be nonlinear enough and found that the AES [56] yields the average ERLE of about 8 dB which is limited by the high level of nonlinear distortion.

First, the conventional NAES algorithm [56] was applied to the whole data set. The AES algorithm was slightly modified so that it fitted to single channel AES and the nonlinear RES was based on DNN in which each hidden and the output layer had 2048 and 129 units. The same parameters values and training processes were used as in [56] and the training dataset was used in Table. 5.1.

In the proposed algorithm, the frame length was set to 512 samples with 50% overlap, and a 512-point STFT was applied to each frame. Each hidden and the output layer had 2048 and 257 units, respectively. The input vector included the current frame, the previous two frames and the two estimated SERs in the current frame. The dimension of the vector becomes 1799. For multi-task learning, we fixed the weighted value λ to 0.9. In the pre-training for each task, the number of epochs for the RBM in each layer was 20 and the learning rate was 0.0005. In the fine-tuning for the main task, the learning rate was set to 0.1 for the first 10 epochs, then decreased by 2% after each epoch. In the case of the additional task, the learning

Table 5.2: ERLE and PESQ scores obtained with the matched and mismatched conditions.

Measure	Condition	AES + RES (DNN)	DNN_EAT	DNN_EAT_MTL
ERLE	(M1, R1)	37.10	43.94	44.59
	(M2, R1)	34.93	41.30	42.28
	(M3, R2)	30.03	39.20	39.25
PESQ	(M1, R1)	3.12	3.34	3.41
	(M2, R1)	3.11	3.28	3.35
	(M3, R2)	2.95	3.12	3.16

rate for 3rd hidden layer was set to 0.01 for the first 10 epochs, then decreased by 12.5% after each epoch. Total iteration number was 100 and the mini-batch size M was set to 128. For the tests, we used the test sets in Table. 5.1. We used the two sets in each test condition, which consist of the one set recorded for single-talk periods and the other copy made for double-talk periods.

In Table. 5.2, the overall results of the ERLEs for single-talk periods and PESQ scores for double-talk periods. The matched condition is (M1, R1) and the mismatched conditions are (M2, R1) and (M3, R2) for device mismatch. DNN_EAT is the DNN-based NAES using echo aware training (EAT) and DNN_EAT_MTL is the DNN_EAT with multi-task learning using double-talk task. The proposed method outperformed the conventional NAES. The proposed NAES on EAT improved both ERLE and PESQ scores compared with the conventional one and the DNN_EAT_MTL showed slightly better results in double-talk cases than DNN_EAT.

Table 5.3: ERLE in double-talk (DT) and segmental SSDR obtained with the matched and mismatched conditions.

Measure	Condition	AES + RES (DNN)	DNN_EAT	DNN_EAT_MTL
ERLE in DT	(M1, R1)	3.58	4.45	4.85
	(M2, R1)	3.92	4.43	4.51
	(M3, R2)	3.38	4.08	4.11
$SSDR_{seg}$	(M1, R1)	18.07	19.35	19.84
	(M2, R1)	16.17	17.45	17.96
	(M3, R2)	16.87	17.89	18.21

The proposed SER features for temporal dynamics and the additional task for double-talk detection may be helpful to improve the suppression of nonlinear echo. It is noted that through the results of the mismatched condition (M3, R2), the device mismatch which can make the nonlinearity in echo signal may be critical to improve the performance of NAES.

To evaluate the specific performance in double-talk periods, the results of ERLE in DT and segmental SSDR are represented in Table. 5.3. As expected, the proposed methods showed better performance than the combination of the AES and the DNN-based RES techniques. From the results, we concluded that DNN_EAT and DNN_EAT_MTL can remove the nonlinearity of echo and recover the near-end speech well in real conditions.

Fig. 5.2 shows that evolution of the ERLE over time in conjunction with the corresponding unprocessed echo waveform. The proposed method is DNN_EAT_MTL.

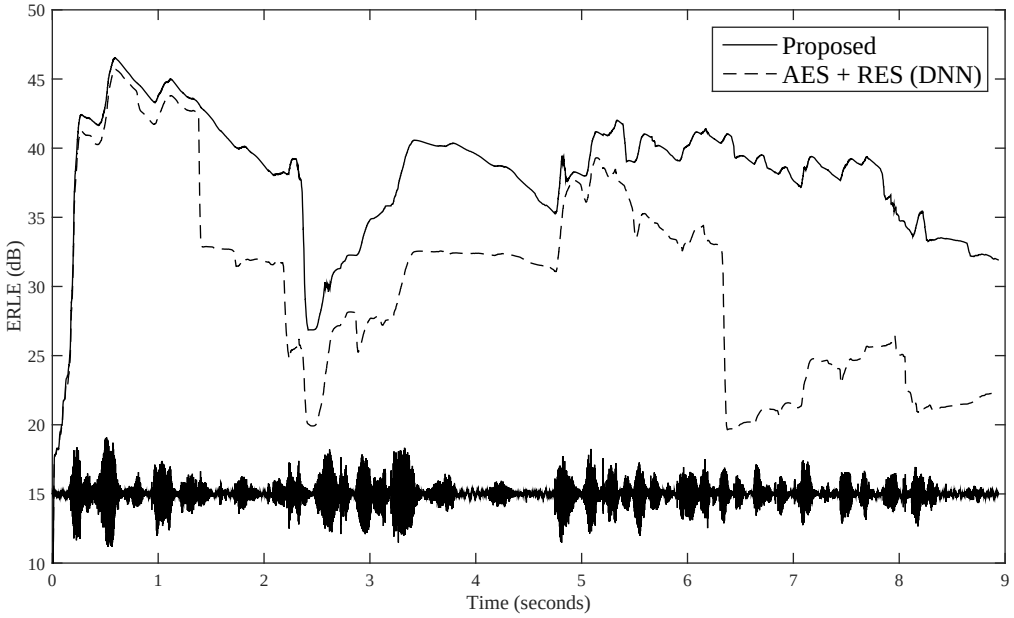

Figure 5.2: Comparison of ERLE in a single-talk situation of the mismatched condition (M3, R2).

The proposed algorithm attenuated the echo signal including nonlinear components more effectively than the conventional method. From Fig. 5.3 to Fig. 5.6, the waveforms and spectrograms are illustrated for checking echo suppression performance and the improvement of speech recovery. Consequently, DNN_EAT_MTL preserved the near-end speech better compared to the conventional method.

5.5 Summary

In this chapter, we proposed a novel DNN-based NAES approach using echo aware training and multi-task learning framework. The proposed algorithm can di-

rectly estimate the NAES optimal gain based on DNNs and the echo information such as *a priori* and *a posteriori* signal-to-echo ratio (SER) was used in the DNN. Also, we introduced the multi-task learning to improve the gain estimates in various conditions. In the framework, the main task of the gain estimation for NAES was jointly trained with an additional task of double-talk detection. Experimental results showed that the proposed method outperformed the conventional one, especially in double-talk situations.

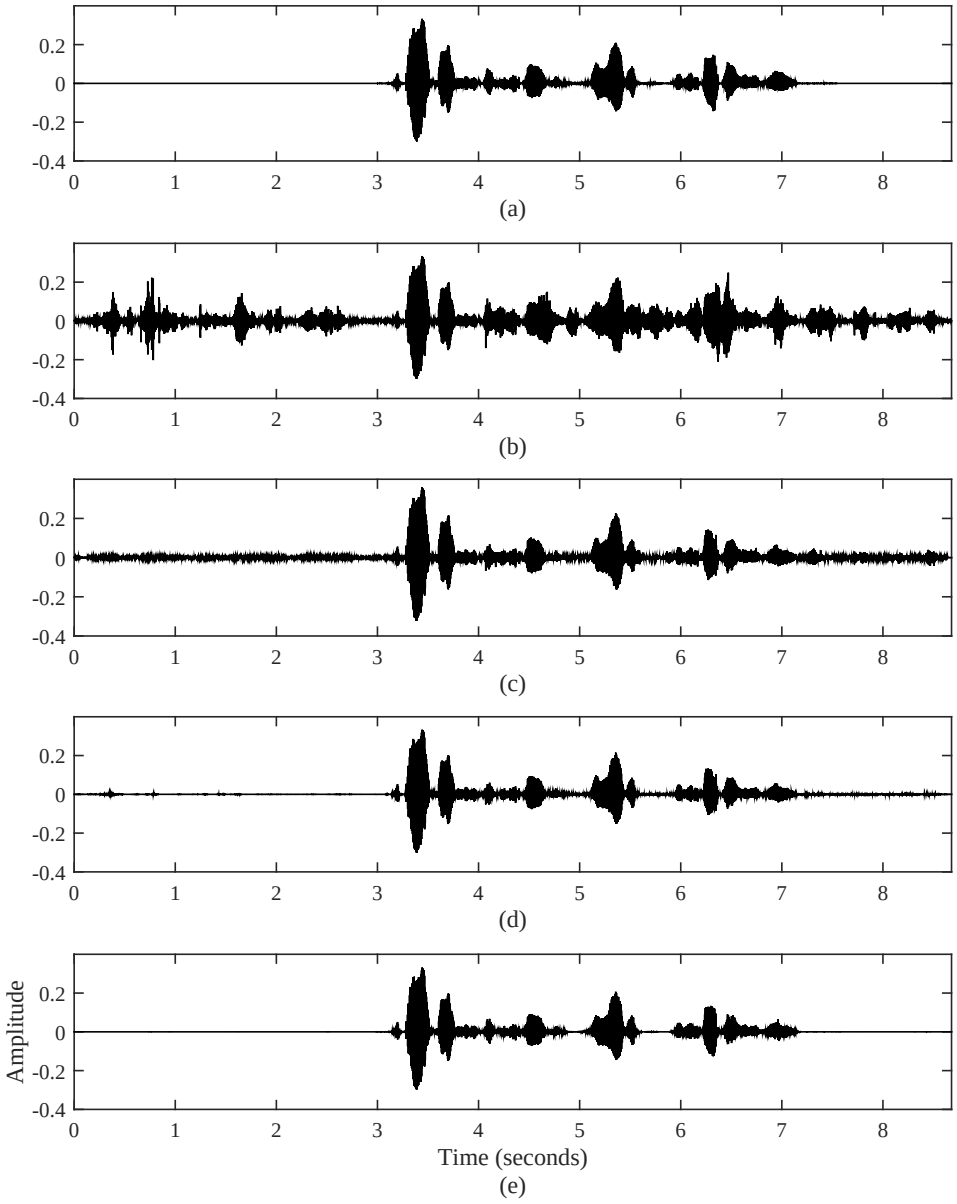

Figure 5.3: Waveforms for the double-talk case in the matched case (M1, R1). (a) clean near-end speech, (b) microphone signal, (c) output of AES (d) output of AES + RES (DNN) and (e) output of DNN_EAT_MTL.

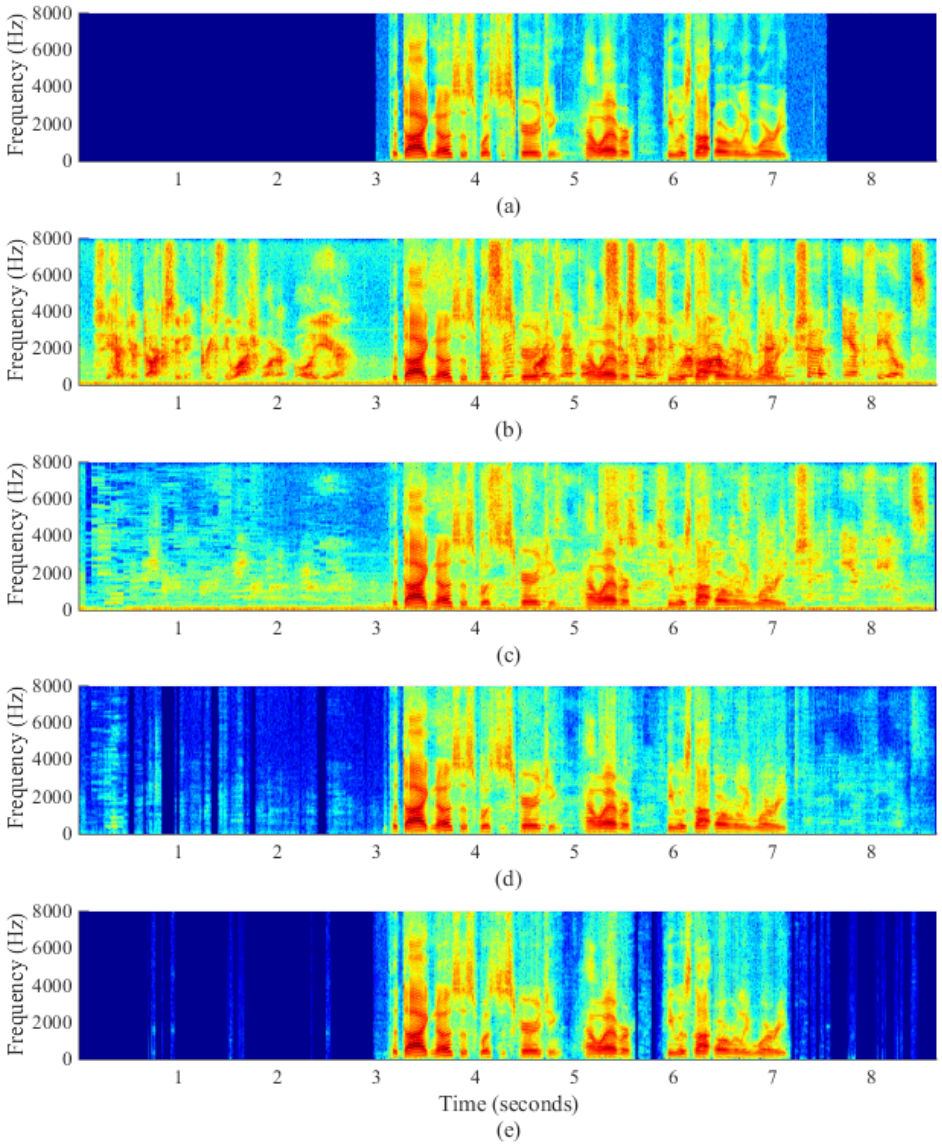

Figure 5.4: Spectrograms for the double-talk case in the matched case (M1, R1). (a) clean near-end speech, (b) microphone signal, (c) output of AES (d) output of AES + RES (DNN) and (e) output of DNN_EAT_MTL.

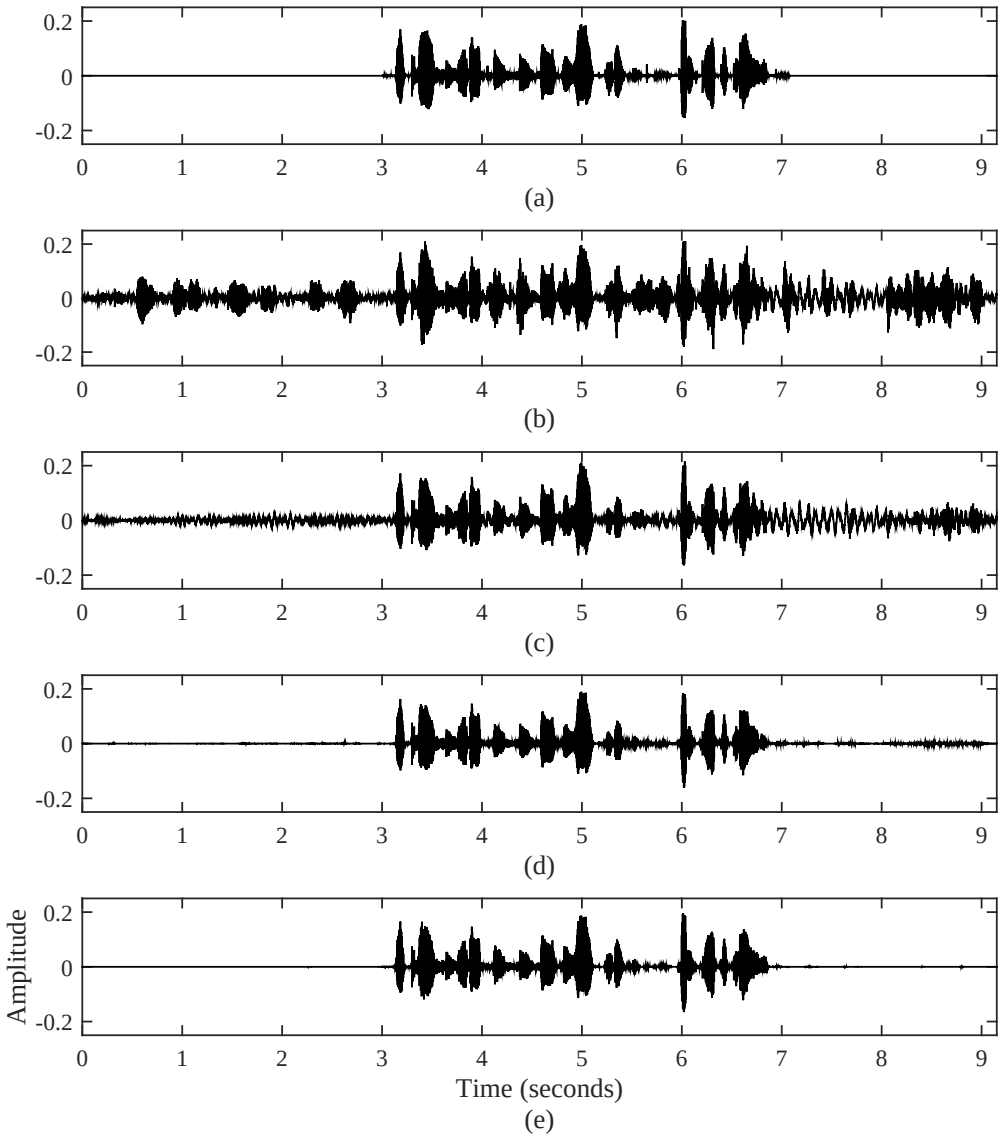

Figure 5.5: Waveforms for the double-talk case in the mismatched case (M3, R2). (a) clean near-end speech, (b) microphone signal, (c) output of AES (d) output of AES + RES (DNN) and (e) output of DNN_EAT_MTL.

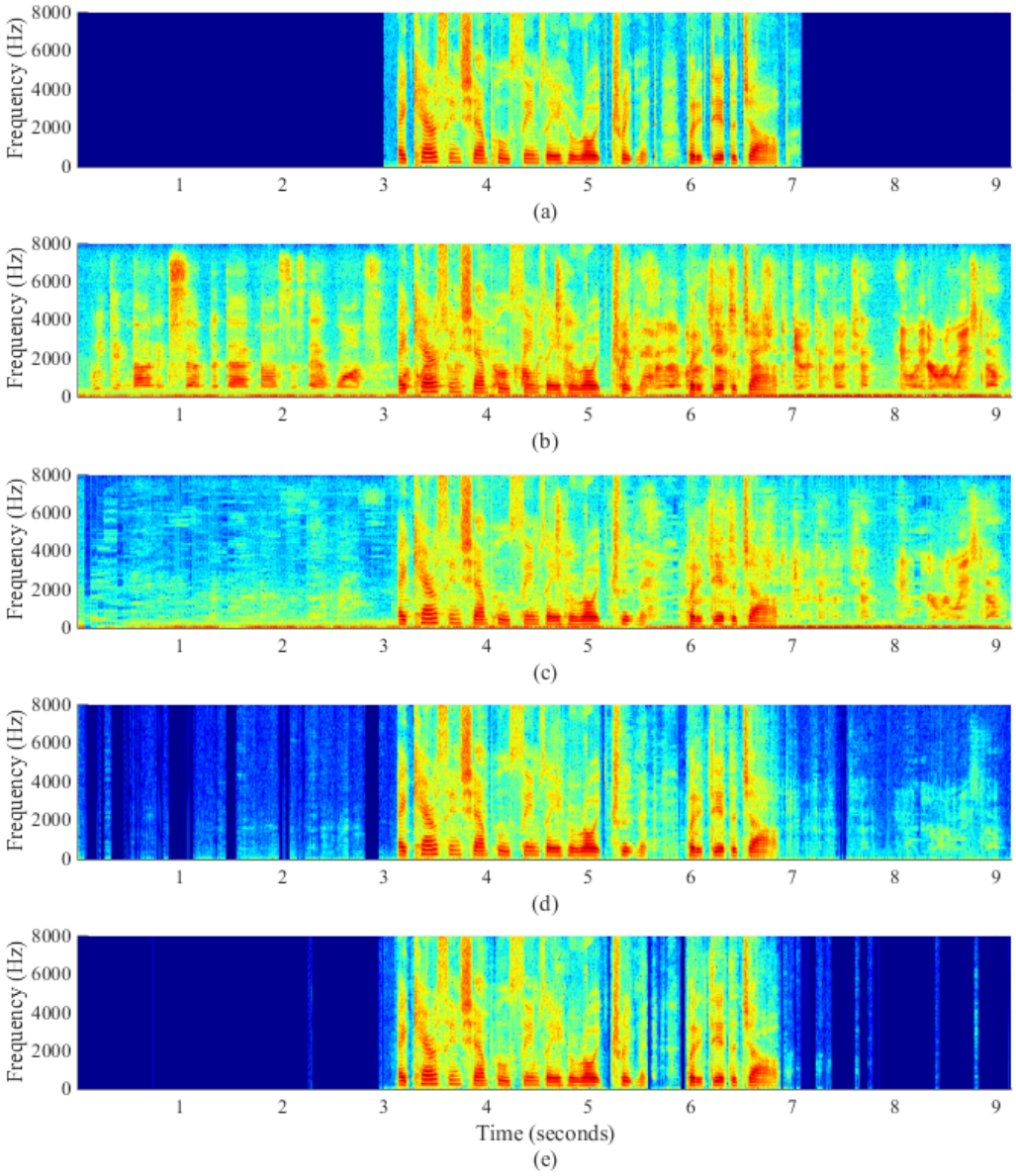

Figure 5.6: Spectrograms for the double-talk case in the mismatched case (M3, R2). (a) clean near-end speech, (b) microphone signal, (c) output of AES (d) output of AES + RES (DNN) and (e) output of DNN_EAT_MTL.

Chapter 6

Conclusions

In this thesis, three major approaches based on spectro-temporal correlations have been proposed for acoustic echo suppression. Even though many algorithms for acoustic echo cancellation and suppression have been proposed to solve the acoustic echo problem and produced some successful results in telecommunication systems during the last few decades, acoustic echo cancellation and suppression are still the major topics to transmit clear speech in full-duplex telecommunication systems. Especially, the non-uniqueness problem in stereophonic acoustic echo cancellation or the nonlinearity generated from cheap loudspeakers and amplifiers in echo signal should be overcome to both suppress the nonlinear echo and recover the near-end speech. Therefore, we have focused on the exact methods for echo representation based on cross-filtering and deep neural networks.

First, we have proposed the enhanced stereophonic AES algorithm using augmented vectors in order to incorporate spectral and temporal correlations. The approach takes advantage of the correlations among components in adjacent time frames and frequency bins in the STFT domain since a linear system can be mod-

eled more accurately through the correlations. To estimate the stereo echo signal, the extended PSD matrices and cross-PSD vectors are derived by using supervectors augmented with the adjacent components from the signal statistics. Experimental results demonstrated that the proposed method is superior to conventional SAES in terms of both ERLE and PESQ.

Second, an optimal residual echo suppression gain regression employing DNN has been proposed in single channel case. The DNN could represent the complicated mapping from the AES output and far-end signal in the whole frequency bins to RES gains. Furthermore, the proposed method does not need any explicit double-talk detectors as the DNN can accommodate the mapping for both single-talk and double-talk cases. The proposed RES algorithm outperformed the conventional one in terms of various objective measures in matched and mismatched conditions for various RIRs, SER, clipping type, and level of nonlinearity in loudspeaker.

Finally, we have proposed a novel DNN-based NAES approach based on echo aware training and multi-task learning framework. The proposed algorithm can directly attempt the NAES gain estimation based on DNNs and the echo information such as *a priori* and *a posteriori* signal-to-echo ratio (SER) has been introduced as additional features in the DNN. Also, we have combined the multi-task learning to improve the gain estimates in various conditions. In the framework, the main task of the gain estimation for NAES was jointly trained with an additional task of double-talk detection. Experimental results showed that the proposed method outperformed the conventional one, especially in double-talk situations.

Bibliography

[1] A. Gilloire, "Experiments with sub-band acoustic echo cancellers for teleconferencing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 12, Apr. 1987, pp. 2141–2144.

[2] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

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

[4] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation-an overview of the fundamental problem," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 148–151, Aug. 1995.

[5] J. H. Cho, D. R. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Transactions on Speech*

and Audio Processing, vol. 7, no. 6, pp. 718–724, Nov. 1999.

[6] S. Haykin, *Adaptive filter theory*. Pearson Education India, 2008.

[7] C. Avendano, "Acoustic echo suppression in the stft domain," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001, pp. 175–178.

[8] X. Lu and B. Champagne, "Acoustic echo cancellation with post-filtering in subband," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2003, pp. 29–32.

[9] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 1997, pp. 307–310.

[10] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1048–1062, Sep. 2005.

[11] C. Faller and C. Tournery, "Robust acoustic echo control using a simple echo path model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, May 2006, pp. 281–284.

[12] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, Mar. 1998.

[13] H. Buchner, J. Benesty, T. Gansler, and W. Kellermann, "Robust extended multidelay filter and double-talk detector for acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1633–1644, Sep. 2006.

[14] H. I. K. Rao and B. Farhang-Boroujeny, "Fast LMS/newton algorithms for stereophonic acoustic echo cancelation," *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 2919–2930, Aug. 2009.

[15] F. Yang, M. Wu, and J. Yang, "Stereophonic acoustic echo suppression based on wiener filter in the short-time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 19, no. 4, pp. 227–230, Apr. 2012.

[16] S. Y. Lee and N. S. Kim, "A statistical model-based residual echo suppression," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 758–761, Oct. 2007.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[18] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral feature-based nonlinear residual echo suppression," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2013, pp. 1–4.

[19] J. P. Costa, A. Lagrange, and A. Arliaud, "Acoustic echo cancellation using nonlinear cascade filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, Apr. 2003, pp. 389–392.

[20] A. Guerin, G. Faucon, and R. L. Bouquin-Jeannes, "Nonlinear acoustic echo cancellation based on Volterra filters," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 672–683, Nov. 2003.

[21] F. Kuech and W. Kellermann, "Partitioned block frequency-domain adaptive second-order Volterra filter," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 564–575, Feb. 2005.

[22] L. A. Azpicueta-Ruiz, M. Zeller, J. Arenas-Garcia, and W. Kellermann, "Novel schemes for nonlinear acoustic echo cancellation based on filter combinations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2009, pp. 193–196.

[23] L. A. Azpicueta-Ruiz, M. Zeller, A. R. Figueiras-Vidal, J. Arenas-Garcia, and W. Kellermann, "Adaptive combination of Volterra kernels and its application to nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 97–110, Jan. 2011.

[24] J. Park and J. H. Chang, "Frequency-domain Volterra filter based on data-driven soft decision for nonlinear acoustic echo suppression," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1088–1092, Sep. 2014.

[25] F. Kuech, A. Mitnacht, and W. Kellermann, "Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, Mar. 2005, pp. 105–108.

[26] A. N. Birkett and R. A. Goubran, "Nonlinear loudspeaker compensation for hands free acoustic echo cancellation," *Electronics Letters*, vol. 32, no. 12, pp. 1063–1064, Jun. 1996.

[27] J. M. Gil-Cacho, T. van Waterschoot, M. Moonen, and S. H. Jensen, "Nonlinear acoustic echo cancellation based on a parallel-cascade kernel affine projection algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2012, pp. 33–36.

[28] S. Gay and S. Tavathia, "The fast affine projection algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, May. 1995, pp. 3023–3026.

[29] T. Gansler, S. L. Gay, M. M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 656–663, Nov. 2000.

[30] M. A. Iqbal, J. W. Stokes, and S. L. Grant, "Normalized double-talk detection based on microphone and AEC error cross-correlation," in *IEEE International Conference on Multimedia and Expo*, Jul. 2007, pp. 360–363.

[31] P. Vary, "Noise suppression by spectral magnitude estimation —mechanism and theoretical limits—," *Signal Processing*, vol. 80, no. 4, pp. 387–400, 1985.

[32] H. Pobloth and W. B. Kleijn, "On phase perception in speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Mar. 1999, pp. 29–32.

[33] 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.

[34] C. M. Lee, J. W. Shin, and N. S. Kim, "Stereophonic acoustic echo suppression incorporating spectro-temporal correlations," *IEEE Signal Processing Letters*, vol. 21, no. 3, pp. 316–320, Mar. 2014.

[35] F. Kuech and W. Kellermann, "Nonlinear residual echo suppression using a power filter model of the acoustic echo path," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2007, pp. 73–76.

[36] J. Benesty and D. R. Morgan, "Frequency-domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2000, pp. 789–792.

[37] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 1997, pp. 219–222.

[38] A. W. H. Khong and P. A. Naylor, "Reducing inter-channel coherence in stereophonic acoustic echo cancellation using partial update adaptive filters," in *European Signal Processing Conference*, Sep. 2004, pp. 405–408.

[39] S. Emura, Y. Haneda, and A. Kataoka, "A solution to echo path imbalance problem in stereo echo cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, May. 2004, pp. 129–132.

[40] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May. 2007.

[41] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Rec.*, 2000.

[42] S. Eberli, D. Cescato, and W. Fichtner, "Divide-and-conquer matrix inversion for linear MMSE detection in sdr mimo receivers," in *NORCHIP*, Nov. 2008, pp. 162–167.

[43] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[44] K. Helwani, H. Buchner, J. Benesty, and J. Chen, "A single-channel MVDR filter for acoustic echo suppression," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 351–354, Apr. 2013.

[45] A. Chhetri, A. Surendran, J. W. Stokes, and J. Platt, "Regression-based residual acoustic echo suppression," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Sep. 2005, pp. 201–204.

[46] D. A. Bendersky, J. W. Stokes, and H. S. Malvar, "Nonlinear residual acoustic echo suppression for high levels of harmonic distortion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 261–264.

[47] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[48] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[49] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, Jul. 2006.

[50] R. Salakhutdinov and G. Hinton, "Using deep belief nets to learn covariance kernels for Gaussian processes," *in Proc. Advances in Neural Inform. Process. Syst.,*, vol. 20, pp. 1–8, 2007.

[51] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, Feb. 2015.

[52] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[53] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2065–2079, Sep. 2012.

[54] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-García, and A. Uncini, "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1502–1512, Jul. 2013.

[55] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[56] C. M. Lee, J. W. Shin, and N. S. Kim, "DNN-based residual echo suppression," in *Proc. Interspeech*, vol. 1, Sep. 2015, pp. 1775–1779.

[57] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning,"

in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 5014–5018.

[58] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May. 2013, pp. 6965–6969.

[59] A. Mohan and R. Rose, "Multi-lingual speech recognition with low-rank multi-task deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 4994–4998.

[60] T. G. Kang and N. S. Kim, "DNN-based voice activity detection with multi-task learning," *IEICE Trans. Inf. and Syst.*, vol. E99-D, no. 2, pp. 550–553, Feb. 2016.

[61] S. Xue, O. Abdel-Hamid, H. Jiang, and L. Dai, "Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May. 2014, pp. 6339–6343.

[62] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 708–712.

[63] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 825–834, May. 2008.

요 약

음향학적 에코로 인해 신호가 왜곡되는 것을 방지하기 위해 음향학적 에코 제거 및 억제 기법들의 적용은 필수적이다. 특히 양방향 전화통신 시스템이 급속히 확산되면서 에코 제거 기법들은 보다 빠르고 신뢰성 있는 알고리즘을 필요로 하였으며 이와 관련하여 오랫동안 연구가 이뤄졌다. 대부분의 음향학적 에코 제거 기법은 적응 필터를 기반으로 하지만 충분한 성능을 얻기 위해서는 길이가 긴 필터가 요구되며 선형 요소들만을 주로 고려할 수 있다는 점에서 한계를 지닌다.

본 논문에서는 음향학적 억제 기법을 활용하여 음향학적 에코로 인해 왜곡된 신호를 복원하고 에코를 효과적으로 감소시킬 수 있는 방법들을 제시한다. 첫 번째로, short-time Fourier transform (STFT) 영역에서 주파수 및 시간적 상관관계를 고려하여 스테레오 환경에서 발생하는 음향학적 에코를 제거하는 기법을 제안한다. 기존의 방법들과 달리, 신호를 비상관관계로 만드는 과정이 필요하지 않고 인접한 주파수 및 시간 요소들이 가지는 상호 관계를 고려하여 보다 정확한 에코 추정을 시도한다. 특히 기존 이중 통화 검출기가 없이 신호 대 에코비 (signal-to-echo ratio, SER) 정보만으로 에코 억제가 가능한 특징을 지닌다.

두 번째로, 비선형 환경에서 발생하는 음향학적 에코 성분 안의 비선형 요소 등을 억제하기 위한 방안으로 심층 신경망을 활용하는 잔여 에코 억제를 제안하였다. 일반적으로 스피커나 앰프가 가지는 비선형성은 음향학적 에코와 원본 신호 간의 관계를 복잡하게 만들어 간단한 선형 시스템 가정만으로는 충분한 에코 제거를 수행할 수 없

다. 따라서 주로 비선형 요소들로 이뤄진 잔여 에코를 억제하기 위해서 신경 심층망을 기반으로 하는 잔여 에코 억제 이득을 추정하는 기법을 제안했다. 심층 신경망은 선형 알고리즘으로는 모델링하기 어려운 복잡한 비선형 관계를 학습하기 용이하기 때문에 잔여 에코와 잔여 에코 억제 이득 간의 관계를 추정하기 쉽고 인접한 프레임과 한 프레임 내의 전체 주파수 요소를 모두 고려하는 입력을 사용함으로써 주파수 및 시간적 상관관계를 모두 고려할 수 있다. 단일 및 이중 통화 환경을 모두 고려하는 학습을 통해 동시 통화 검출 없이 동작하는 장점 또한 지닌다.

세 번째로, 비선형적 환경에서의 음향학적 에코 억제를 심층 신경망 학습으로 잔여 에코 억제 과정 없이 수행하기 위해 에코 억제 환경에 맞는 에코 어웨어 학습 (echo aware training)과 이중 통화 검출 정보를 활용한 멀티태스크 학습 (multi-task learning)을 제안하였다. 에코 어웨어 학습 과정에서는 심층 신경망으로 음향학적 에코 제거를 바로 시도할 경우 방이나 공간 환경, 에코 변화 등을 추정하기 어려울 수 있기 때문에 이를 보조할 특징 벡터를 이용하여 학습을 도운다. 이 특징 벡터는 사전 및 사후 SER 정보를 이용하며 기존의 에코 억제 알고리즘이나 심층 신경망을 통해 추정할 수 있다. 추가적으로 음향학적 에코 이득 추정을 개선하기 위해 이중 통화 검출 과정을 별도의 태스크로 만들어 기존 이득 추정 태스크와 함께 학습시키는 멀티태스크 학습을 제안한다. 이렇게 학습된 심층 신경망은 단일 통화 구간에서는 에코를 더 억제할 수 있고 이중 통화 구간에서는 음성 신호 향상에 도움을 주는 내부적인 은닉 신경망으로 구성된다. 또한 제안한 기법으로 학습된 심층 신경망은 다양한 환경에서 보다 강인할 가능성을 지닌다.

주요어: 음향학적 에코 억제, 주파수 및 시간적 상관관계, 잔여 에코 억제, 심층 신경망 (deep neural network), 에코 어웨어 학습 (echo aware training), 멀티태스크 학습 (multi-task learning)

학 번: 2009-20876