Acoustic Echo Cancellation Combined with Deep-Learning-Based Residual Echo Suppression

Eran Shachar

Acoustic Echo Cancellation Combined with Deep-Learning-Based Residual Echo Suppression

Research Thesis

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical Engineering

Eran Shachar

Submitted to the Senate
of the Technion — Israel Institute of Technology
Adar 5783 Haifa February 2023

This research was carried out under the supervision of Prof. Israel Cohen and Dr. Baruch Berdugo, in the Faculty of Electrical and Computer Engineering.

The results of Chapter 4 of this thesis have been published as an article by the author and research collaborators in a journal during the course of the author's masters research period, the most up-to-date version of which is:

> Eran Shachar, Israel Cohen, and Baruch Berdugo. Double-talk detection-aided residual echo suppression via spectrogram masking and refinement. *Acoustics*, 4(3):637–655, 2022.

The results of Chapter 3 were organized as another article, entitled "Acoustic Echo Cancellation with the Normalized Sign-Error Least Mean Squares Algorithm and Deep Residual Echo Suppression", which was accepted for publication in the Algorithms journal of the MDPI publisher, but is yet unpublished at the time of submitting the thesis.

The author of this thesis states that the research, including the collection, processing, and presentation of data, addressing and comparing to previous research, etc., was done entirely in an honest way, as expected from scientific research that is conducted according to the ethical standards of the academic world. Also, reporting the research and its results in this thesis was done in an honest and complete manner, according to the same standards.

Acknowledgements

I would like to express my deepest gratitude and appreciation to my research supervisors, Prof. Israel Cohen and Dr. Baruch Berdugo. This endeavor would not have been possible without their guidance and support. Throughout this journey, they have taught me how to become a better researcher, allowed me to develop meaningful skills, and helped me overcome many difficulties, and for that I am sincerely grateful.

I would also like to thank my partner Batel and my family, for accompanying and supporting me through the downs and sharing my joy through the ups of this process. This accomplishment would not have been possible without them.

Contents

List of Figures

Abstract

In this thesis, we study the problems of acoustic echo cancellation and residual echo suppression. Acoustic echo is a common problem in full-duplex telecommunication systems. An acoustic echo is generated when the signal produced by a loudspeaker is captured by a microphone along with the desired signal. This echo can cause conversation quality degradation, which poses a problem in many real-life situations, such as a meeting in which the remote participants' speech is played in the meeting room by loudspeakers. Abundant research was conducted to mitigate the acoustic echo problem. In recent years, acoustic echo cancellers (AECs) have achieved outstanding performance thanks to deep-learning technology. Nevertheless, several aspects were not studied in previous research. This thesis aims to fill this gap by studying three aspects: A proper choice of a linear AEC in a deep-learning-based residual echo suppression system, a proper integration of a double-talk detector (DTD) with a deep-learning residual echo suppression model, and residual echo suppression (RES) in the low signal-to-echo ratio (SER) scenario.

First, we present an echo suppression system that combines a linear AEC with a deep-complex convolutional recurrent network (DCCRN) for residual echo suppression. The filter taps of the AEC are adjusted in subbands by using the normalized sign-error least mean squares (NSLMS). We compare the NSLMS with the normalized least mean squares (NLMS) and study the combination of each with a deep RES model. We also study the utilization of a pre-trained deep-learning speech denoising model as an alternative to a RES model. Results show that the performance of the NSLMS is superior to that of the NLMS in all settings. With the NSLMS output, the proposed RES model achieves better performance than the larger, pre-trained speech denoiser model. Furthermore, the denoiser performs better on the NSLMS output than the NLMS output, indicating that the residual echo in the NSLMS output is more akin to noise than speech.

The acoustic echo cancellation problem is especially challenging in low SER scenarios, such as hands-free conversations over mobile phones when the loudspeaker volume is high. In this thesis, we propose a two-stage deep-learning approach to residual echo suppression focused on the low SER scenario. The first stage consists of a speech spectrogram masking model integrated with a DTD. The second stage consists of a spectrogram refinement model optimized for speech quality by minimizing a perceptual

evaluation of speech quality (PESQ) related loss function. The proposed integration of DTD with the masking model outperforms several other configurations based on previous studies. We conduct an ablation study that shows the contribution of each part of the proposed system. We evaluate the proposed system in several SERs and demonstrate its efficiency in the challenging setting of a very low SER. Finally, the proposed approach outperforms competing methods in several residual echo suppression metrics. We conclude that the proposed system is well-suited for the task of low SER residual echo suppression.

Abbreviations

AEC	:	Acoustic Echo Canceller
AECMOS	:	Acoustic Echo Cancellation Mean Opinion Score
BCE	:	Binary Cross-Entropy
BLSTM	:	Bidirectional Long Short-Term Memory
CAD-AEC	:	Context-Aware Deep Acoustic Echo Cancellation
CRM	:	Complex Ratio Mask
CRN	:	Convolutional Recurrent Network
DCCRN	:	Deep Complex Convolutional-Recurrent Network
DFT	:	Discrete Fourier Transform
DNN	:	Deep Neural Network
DNS	:	Deep Noise Suppression
DNSMOS	:	Deep Noise Suppression Mean Opinion Score
DTD	:	Double-Talk Detector
DTLN	:	Dual-Signal Transformation LSTM Network
ELU	:	Exponential Linear Unit
ENR	:	Echo-to-Noise Ratio
ERLE	:	Echo Return Loss Enhancement
FC	:	Fully-Connected
FCRN	:	Fully-Convolutional Recurrent Network
GRU	:	Gated Recurrent Unit
IRM	:	Ideal Ratio Mask
iSTFT	:	inverse Short-Time Fourier Transform
LMS	:	Least Mean Squares
MRI	:	Magnetic Resonance Imaging
MSE	:	Mean Squared Error
NLMS	:	Normalized Least Mean Squares
NSLMS	:	Normalized Sign-Error Least Mean Squares
PESQ	:	Perceptual Evaluation of Speech Quality
PReLU	:	Parametric Rectified Linear Unit
PSF	:	Phase-Sensitive Filter
ReLU	:	Rectified Linear Unit
RES	:	Residual Echo Suppressor

RT_{60}	:	Reverberation Time
RNN	:	Recurrent Neural Network
RTF	:	Real-Time Factor
SER	:	Signal-to-Echo Ratio
SLMS	:	Sign-Error Least Mean Squares
SNR	:	Signal-to-Noise Ratio
STFT	:	Short-Time Fourier Transform
T-F	:	Time-Frequency
VAD	:	Voice Activity Detector

Notations

$a(n)$	:	Output signal of the AEC at time-point n
$A(f,k)$	:	Spectrogram magnitudes of the AEC's output signal at frequency bin f and time bin k
B	:	Batch size
$\mathbf{c}(n)$	:	Linear AEC's filter tap weights vector at time-point n
$d(n)$	:	Near-end signal at time-point n
$\tilde{d}(n)$	:	Estimated near-end signal at time-point n
$D(f,k)$	:	Spectrogram magnitudes of the near-end signal at frequency bin f and time bin k
$\tilde{D}(f,k)$	:	Spectrogram magnitudes of the estimated near-end signal at frequency bin f and time bin k
$e(n)$	:	Error signal at time-point n
$E(f,k)$	:	Spectrogram magnitudes of the error signal at frequency bin f and time bin k
F_c	:	Output of a complex LSTM layer
$h(t)$	:	Impulse response at time t
$H(f,k)$	:	Log of the ratio between the spectrogram magnitudes of the clean near-end speech and that of the error signal at frequency bin f and time bin k
$\tilde{H}(f,k)$	:	Masking model's output at frequency bin f and time bin k
$H_{\text{in/out}}$	:	Height of a convolution layer's input/output feature maps
$j(t)$	:	Continuous-time input to a linear filter at time t
l	:	Masking model's loss function
l_{DTD}	:	DTD's loss function
$l_{\text{DTD-farend}}$	:	The DTD's far-end speech loss term
$l_{\text{DTD-nearend}}$	:	The DTD's near-end speech loss term
l_{mask}	:	Masking loss function
l_{MSE}	:	Refinement model's MSE loss term
l_{PESQ}	:	Refinement model's PESQ loss term
LSTM_r	:	Real LSTM layer
LSTM_i	:	Imaginary LSTM layer
M	:	Number of STFT resolutions in the STFT loss

$m(n)$	:	Microphone signal at time-point n
$M(f,k)$	:	Spectrogram magnitudes of the microphone signal at frequency bin f and time bin k
N	:	Number of linear AEC's filter taps
O_c	:	Output of a complex 2-D convolution layer
$P(f,k)$	:	DTD's output at frequency bin f and time bin k
T	:	Number of time frames
$t_{\text{inference}}$	:	The time it takes a model to make an inference
t_{signal}	:	Duration of a model's input signal
$v(n)$	:	Noise signal at time-point n
v_k	:	DTD's ground-truth label at time-point k
$\hat{v}_k$	:	DTD's prediction at time-point k
W_r, W_i	:	Real and imaginary convolution kernels
$x(n)$	:	Far-end reference signal at time-point n
$X(f,k)$	:	Spectrogram magnitudes of the far-end reference signal at frequency bin f and time bin k
$\mathbf{x}_N(n)$	:	Far-end reference signal vector of length N at time-point n
X_r, X_i	:	Real and imaginary parts of a complex features map
$y(n)$	:	Echoic loudspeaker signal at time-point n
$z(t)$	:	Continuous-time output of a linear filter at time t
$\alpha(n)$	:	Step size of the NSLMS algorithm at time-point n
$\epsilon, \epsilon_1, \epsilon_2$	:	Small constants for numerical stability
λ_{DTD}	:	DTD's loss weight parameter
λ_{MSE}	:	Refinement model's MSE loss weight parameter
μ	:	Step size of the LMS algorithm

Chapter 1

Introduction

1.1 Background and Motivation

Modern telecommunication systems often suffer speech intelligibility degradation caused by an acoustic echo. A typical scenario includes two speakers communicating between a far-end and a near-end point. At the near-end point, a microphone captures both the near-end speaker's signal and the acoustic echo of a loudspeaker playing the far-end signal [1]. When the far-end speaker speaks, he hears the echo of his voice, thus reducing the quality of the conversation. Therefore, canceling the acoustic echo while preserving near-end speech quality is desired in any full-duplex communication system. An illustration of an acoustic echo scenario is depicted in Figure 1.1. Typically, the acoustic paths include reflections from the walls or other objects. Only direct paths between the speakers, microphones, and loudspeakers are illustrated for simplification.

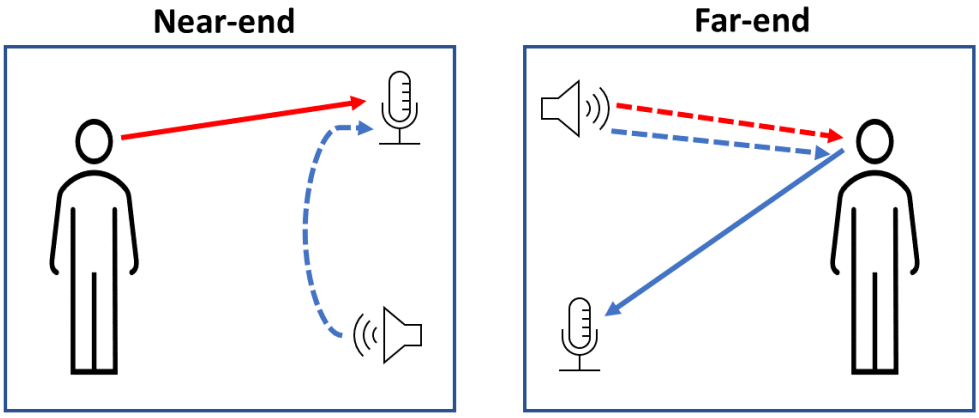


Figure 1.1: Illustration of an acoustic echo scenario. Red lines represent the near-end speaker's speech and blue lines represent the far-end speaker's speech.

Acoustic echo cancellers (AECs) are commonly employed to cancel the echo component of the microphone signal. Traditionally, AECs are based on linear adaptive filters [2]. Linear AECs estimate the acoustic path from the loudspeaker to the microphone. The estimated filters are applied to the far-end reference signal (i.e. the near-end loudspeaker's signal before propagating through the room), resulting in an estimate of the echo signal as received by the microphone. Then, the estimated echo-free near-end signal is obtained by subtracting the estimated echo from the microphone signal.

Linear AECs commonly use the least mean squares (LMS) algorithm [3] and its normalized version, the normalized LMS (NLMS) [4]. The normalization allows the step size to be independent of the input signal's power. Variants of the LMS and NLMS algorithms are the sign-error LMS (SLMS), and normalized SLMS (NSLMS) algorithms [5]. In contrast to the NLMS, the NSLMS adjusts the weight for each filter tap based on the polarity (sign) of the error signal. Several studies have shown the advantage of the NSLMS over the NLMS. For example, Freire and Douglas [6] used the NSLMS adaptive filter to cancel geomagnetic background noise in magnetic anomaly detection systems and demonstrated its benefit over the NLMS. Pathak et al. [7] utilized the NSLMS adaptive filter to perform speech enhancement in noisy magnetic resonance imaging (MRI) environments. According to their experiments, the NSLMS achieves faster convergence than the NLMS, and residual noise produced by the NSLMS has characteristics of white noise. In contrast, residual noise produced by the NLMS is more structured. However, due to their linear nature, residual non-linear components of the echo remain at the output of the linear AECs. In most cases, the residual echo still interferes and degrades the near-end speech's quality.

In recent years, deep-learning neural networks (DNNs) achieved unprecedented performance in many fields, e.g., computer vision, natural language processing, audio and speech processing, and more. Possessing high nonlinear modeling capabilities, DNNs became a natural choice for nonlinear acoustic echo cancellation. Zhang and Wang [8] employ a bi-directional long short-term memory (BLSTM) [9] recurrent neural network (RNN) operating in the time-frequency (T-F) domain to capture dependency between time frames. The model predicts an ideal ratio mask (IRM) [10] applied to the microphone signal's spectrogram magnitudes to estimate the near-end signal's spectrogram magnitudes. Kim and Chang [11] propose a time-domain U-Net [12] architecture with an additional encoder that learns features from the far-end reference signal. An attention mechanism [13] accentuates the meaningful far-end features for the U-Net's encoder. Westhausen and Meyer [14] combine T-F and time-domain processing by adapting the dual-signal transformation LSTM network (DTLN) [15] to the task of acoustic echo cancellation. Although DNN AECs achieve performance superior to linear AECs and allow for end-to-end training and inference, they are prone to introducing distortion to the estimated near-end signal, especially when the signal-to-echo ratio (SER) is low.

An alternative for end-to-end DNN acoustic echo cancellation is residual echo suppression. While traditional residual echo suppression relies on filter-based techniques [16,17], recent advances in deep learning have shifted the focus towards neural network-based approaches. In a typical residual echo suppression setting, a linear AEC is followed by a DNN aimed at suppressing the residual echo at the output of the linear AEC. Linear AECs introduce little distortion to the near-end signal. Their estimation of the echo and near-end signals provides the residual echo suppressor (RES) with better features, allowing for better near-end estimation with smaller model sizes. Carbajal et al. [18] propose a simple fully-connected architecture that receives the spectrogram magnitudes of the far-end reference signal and the linear AEC's outputs and predicts a phase-sensitive filter (PSF) [19] to recover the near-end signal from the linear AEC's error signal. Pfeifenberger and Pernkopf [20] suggest utilizing an LSTM to predict a T-F gain mask from the log differences between the power of the microphone signal and the AEC's echo estimate. Chen et al. [21] propose a time-domain RES based on the well-known Conv-TasNet architecture [22]. They employ a multi-stream modification of the original architecture, where the outputs of the linear AEC are separately encoded before being fed to the main Conv-TasNet. Fazel et al. [23] propose context-aware deep acoustic echo cancellation (CAD-AEC), which incorporates a contextual attention module to predict the near-end signal's spectrogram magnitudes from the microphone and linear AEC output signals. Halimeh et al. [24] employ a complex-valued convolutional recurrent network (CRN) to estimate a complex T-F mask which is applied to the complex spectrogram of the AEC's error signal to recover the near-end signal's spectrogram. Ivry et al. [25] employ a 2-D U-Net operating on the spectrogram magnitudes of the linear AEC's outputs. A custom loss function with a tunable parameter allows a dynamic tradeoff between the levels of echo suppression and estimated signal distortion. Franzen and Fingscheidt [26] propose a 1-D fully convolutional recurrent network (FCRN) operating on discrete Fourier transform (DFT) inputs. An ablation study is performed to study the effect of different combinations of input signals on the joint task of residual echo suppression and noise reduction. Although achieving state-of-the-art residual echo suppression performance, none of the above studies focus on the challenging scenario of extremely low SER. Low SER may occur in typical real-life situations, such as a conversation over a mobile phone where the loudspeaker plays the echo at a high volume.

Under challenging real-life conditions, for example, low SER and changing acoustic echo paths, the performance of the linear AEC preceding the RES model has a significant impact on the overall performance. Hence, investigating the AECs in conjunction with deep-learning models for RES may be beneficial. Furthermore, the output of a linear AEC is expected to contain a distorted weaker version of the echo signal while keeping the near-end signal almost distortionless. Therefore, denoising the linear AEC's estimated near-end signal with a designated speech denoiser might suppress the residual echo while also eliminating other noises, i.e., the speech denoiser may act

as a RES. Research on deep-learning-based speech enhancement algorithms has seen significant progress over the last few years, with many models exhibiting excellent performances [27–29]. Speech denoisers are commonly trained on speech data containing various types of noise where the clean speech utterances are the labels. In the residual echo suppression setting, the residual echo may sometimes resemble speech, thus the denoiser could possibly preserve it. Therefore, it is assumed that for a speech denoiser to achieve good performance as a RES, the AEC must produce residual echo that closely resembles noise rather than human speech.

In a typical residual echo suppression scenario, one of four situations may occur at each time point: both speakers are silent, only the far-end speaker speaks, only the near-end speaker speaks, and double-talk, where both speakers speak at the same time. When only the near-end speaker speaks, the microphone signal should remain unchanged to keep the near-end speech distortionless. Ideally, the microphone signal should be completely canceled when only far-end speech is present to remove any echo component. The challenging situation is double-talk, where it is desired to cancel the echo of the far-end speech while keeping the near-end speech distortion to a minimum. Therefore, it is natural to integrate a double-talk detector (DTD) into the system. Linear AECs typically employ a DTD to prevent the cancellation of the near-end speech in double-talk situations [30, 31]. Several studies also integrate double-talk detection in deep-learning acoustic echo cancellation or residual echo suppression models. Zhang et al. [32] employ an LSTM, which operates on the spectrogram magnitudes of the microphone and far-end reference signals, and predicts near-end speech presence via a binary mask that is applied to the output of the DNN AEC. Zhou and Leng [33] formulate the problem as a multi-task learning problem where a single DNN learns to perform residual echo suppression and double-talk detection in tandem. The model consists of two output branches, the first branch predicts a PSF and acts as a RES, and the second branch detects double-talk. The RES is conditioned on the DTD's predictions by supplying it with features before the classification. Ma et al. [34] propose to perform double-talk detection with two voice activity detectors (VADs), one for detecting near-end speech and the other for detecting far-end speech. Features from several layers of the VADs are fed to a gated recurrent unit (GRU) [35] RNN that performs residual echo suppression. Ma et al. [36] propose a multi-class classifier that receives the encoded features of the time-domain microphone and far-end signals and classifies each time frame independently of the AEC's predictions. Zhang et al. [37] also incorporate a VAD as an independent output branch in a residual echo suppression model. While exhibiting high residual echo performance, their results show that adding the VAD does not lead to improved objective metrics. The rest of the works mentioned above do not study the effect of the DTD/VAD on the RES's performance. Therefore, it is worth studying the effect of DTD and RES integration configurations on the system's performance, especially in the low SER setting where the echo may screen the near-end speech completely.

1.2 Main Contributions

This research aims to fill the gaps and address the drawbacks discussed in the previous section. The research yielded several main contributions:

- An echo suppression system employing NSLMS to perform linear acoustic echo cancellation and DCCRN [28] to perform residual echo suppression is suggested. The performance of the system with the NSLMS is superior to that of the same system employing the commonly-used NLMS. Furthermore, the DCCRN RES achieves superior performance compared to a speech-denoiser RES, which was pre-trained on a large corpus with diverse conditions and despite the denoiser comprising substantially more model parameters.

- The utilization of a speech denoiser to the output of the linear AEC to perform residual echo suppression and denoising in tandem is evaluated. The combination of the denoiser with the NSLMS results in a notable performance improvement compared to using NLMS. The results indicate that the NSLMS output contains residual echo that resembles noise more closely than speech.

- A novel two-stage residual echo suppression deep-learning system focused on the challenging low SER scenario is proposed. By integrating a DTD in the first stage and employing a perceptual speech-quality loss function in the second stage, the proposed system achieves the highest performance gain in the extremely-low SER setting. Furthermore, the proposed system achieves the best performance compared to other RES systems in this challenging setting.

- The integration of the DTD with the DNN is studied and the proposed configuration is compared to several others based on previous research. Results show that while all other configurations result in minor or no performance improvement, the proposed configuration achieves notable performance gain.

1.3 Research Overview

This research is focused on residual echo suppression systems based on linear adaptive AECs and DNN RES. Different aspects of the systems are studied, including the choice of the AEC, utilization of a pre-trained speech denoiser as an alternative to a designated RES, integration of a DTD, and an extremely-low SER setting. The first part of the research focuses on the choice of the linear AEC and its effect on the DNN RES. Two different linear AECs are studied and compared: one based on the NLMS algorithm and the other based on the NSLMS algorithm. Although the NLMS is a common choice in many DNN-based residual echo suppression systems, several studies showed the superiority of NSLMS over NLMS in other fields [6, 7]. The presented results show that the NSLMS is superior to NLMS in linear acoustic echo cancellation as well. We

propose a RES based on the DCCRN architecture [28], initially proposed for speech enhancement and adapted to the residual echo suppression task. The performance of the RES with the NSLMS is compared to that of the RES with the NLMS. The performance gap between the two settings is larger than the performance gap between the NSLMS AEC and NLMS AEC, which indicates that NSLMS is a better choice than NLMS for RES as well. Furthermore, a pre-trained deep-learning speech denoiser [27] is utilized as an alternative to a RES. Although the speech denoiser was pre-trained on a larger corpus with diverse conditions, and despite its model comprising a greater number of parameters by an order of magnitude, the proposed RES model achieves better performance. Nevertheless, the performance gap between the denoiser with the NSLMS and the denoiser with the NLMS is greater than the respective gap in the RES setting. These results indicate that the NSLMS produces output more akin to noise than speech. This observation further strengthens the claim that the proper choice of linear AEC is crucial for the RES's performance.

The second part of the research focuses on the challenging and little-studied scenario of extremely-low SER residual echo suppression. We propose a two-stage DNN RES inspired by [38], where a two-stage approach was taken to tackle the low signal-to-noise ratio (SNR) speech enhancement task. In the proposed system, the first stage consists of spectrogram masking. A different architecture than the masking stage of [38] is employed, consisting of fewer model parameters and exhibiting a shorter algorithmic delay. Furthermore, a DTD is integrated with the model. An ablation study shows that the proposed DTD configuration contributes to the performance, contrary to configurations proposed in previous studies, which showed little performance gain. The second stage of the proposed system consists of spectrogram refinement. In [38], the second stage is spectrogram inpainting. The mask produced by the first stage is used to create holes in spectrogram bins that contained speech but were dominated by noise. The inpainting operation aims to reconstruct the speech components while eliminating noise. In our experiments, we found that this approach is less suitable for residual echo suppression since it is more challenging to reconstruct the near-end speech and completely discard the echoic far-end speech. Instead, the proposed refinement stage aims to refine the first stage's outputs rather than create holes and perform inpainting. The refinement is achieved by minimizing a speech-quality-related loss function. The proposed system outperforms compared RES systems in the low-SER setting. Furthermore, the system is evaluated in different SERs. Results show that the performance gain is increased when the SER is decreased, further showing the proposed system's efficacy in this challenging scenario.

This research proposes solutions to the common, real-life problems of acoustic echo cancellation and residual echo suppression. Therefore, all proposed systems are implemented with real-life considerations, such as small model sizes, low memory consumption, and short algorithmic latency. Furthermore, the systems were tested on real-life, independently-recorded data rather than synthetic data, which is often used to evaluate

AEC and RES systems.

1.4 Organization

This thesis is organized as follows. Chapter 2 presents the problem formulation, including the different notations used throughout the thesis and the related scientific background. Chapter 3 presents the first set of contributions: the proposed DCCRN RES with the NSLMS linear AEC, the comparison to NLMS, and the utilization of the speech denoiser as an alternative to RES. The second set of contributions is presented in Chapter 4, where the two-stage residual echo suppression system is proposed to tackle the extremely-low SER problem. A novel DTD integrated with the masking stage is also proposed and evaluated. Chapter 5 concludes the thesis, summarizes the main contributions, and proposes future research directions.

Chapter 2

Preliminaries

This chapter provides background to the different aspects and methods described in this thesis. In Section 2.1, we formulate the problem of residual echo suppression and denote the different signals. Section 2.2 provides background to linear adaptive filters and their application to acoustic echo cancellation. Lastly, we describe the different performance measures used to evaluate the methods described in this thesis in Section 2.3.

2.1 Problem Formulation

To formulate the problem of residual echo suppression, we denote the different signals as follows. $x(n)$ denotes the far-end reference signal at time point n. We denote the echoic loudspeaker signal received by the microphone by $y(n)$ and the near-end speaker's signal by $d(n)$. The noise signal is denoted by $v(n)$. The microphone signal is denoted by $m(n)$ and is given by

$$m(n) = y(n) + d(n) + v(n)\,. \tag{2.1}$$

The inputs to the linear AEC are $m(n)$ and $x(n)$, and its outputs are $a(n)$ and $e(n) = m(n) - a(n)$, the estimated echo signal $y(n)$ and the error signal, respectively. The filter tap weights vector is denoted by $\mathbf{c}(n) = [c_1(n), ..., c_N(n)]^T$, where N is the number of filter taps, and $(\cdot)^T$ is the transpose operation. We also denote the far-end reference signal vector of length N at time n by $\mathbf{x}_N(n) = [x(n), x(n-1), ..., x(n-N+1)]^T$.

The error signal $e(n)$ contains noise and residual echo components. The goal is to enhance $e(n)$ to obtain a better estimate of $d(n)$ by further suppressing the residual echo and possibly removing noise. This is done either by a speech denoising model, in which case it receives $e(n)$ as a single input to be denoised, or by an RES model, in which case it also receives as inputs $x(n)$, $m(n)$, and $a(n)$. $\tilde{d}(n)$ denotes the estimated near-end speaker's signal at the entire system's output. Figure 2.1 depicts the residual echo suppression setup and the different signals. The following chapters will refer to the

spectrogram magnitudes of the different signals' short-time Fourier transform (STFT). These will be denoted by capital letters of their respective time-domain signal notation, e.g., $X(f, k)$ is the STFT spectrogram magnitude of $x(n)$, where f and k denote the frequency-bin and time-bin indices, respectively.

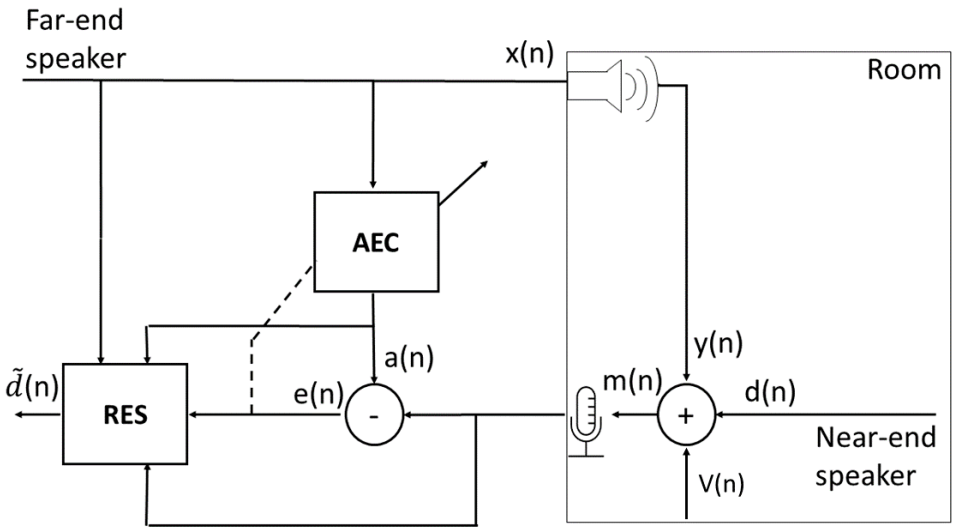


Figure 2.1: Residual echo suppression setup.

2.2 Linear Adaptive AEC

Digital filters are a fundamental part of digital signal processing (DSP). Among the common tasks of digital filters is signal separation, where a superposition of two signals is decomposed into two different signals with the help of a reference signal. Linear filters are filters whose outputs are a linear function of their inputs. In the continuous time domain, the output of the linear filter $z(t)$ can be mathematically expressed as the convolution of the input signal $j(t)$ with the filter's impulse response $h(t)$:

$$z(t) = \int_0^T j(t-\tau)h(\tau)\partial\tau \tag{2.2}$$

Contrary to filters with fixed coefficients, where the coefficients are set in advance and do not vary over time, adaptive filters allow flexibility when the filter coefficients that provide the best performance cannot be determined in advance. Figure 2.2 shows the basic schema of a linear adaptive filter. In the figure, digital signals are considered, where $x(n)$ is the filter's input, $\mathbf{c}(n)$ is the filter's coefficients vector of length N, $a(n)$ is the filter's output, $m(n)$ is the superposition of the signals to be separated (or the desired signal in the case of signal reconstruction), and $e(n)$ is the error signal. The filter's coefficients are adapted using the error signal according to some optimization

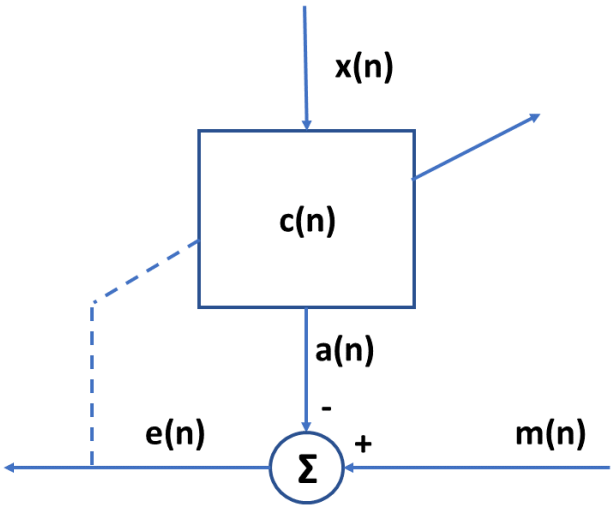


Figure 2.2: Linear adaptive filter schema.

algorithm. One of the most basic and widely used algorithms is the least-mean squares (LMS) algorithm, defined in Algorithm 2.1.

Algorithm 2.1 The LMS algorithm

Parameters: μ - step size, N - number of filter coefficients
for $n = 0, 1, 2, ...$ **do**
  $\mathbf{c}(n) = [c_1(n), ..., c_N(n)]^T$
  $\mathbf{x}_N(n) = [x(n), x(n-1), ..., x(n-N+1)]^T$
  $e(n) = m(n) - a(n) = m(n) - \mathbf{c}^T(n)\mathbf{x}_N(n)$
  $\mathbf{c}(n+1) = \mathbf{c}(n) + 2\mu e(n)\mathbf{x}_N(n)$
end for

For the purpose of acoustic echo cancellation, linear adaptive filters are used to estimate the echo signal $y(n)$ from the microphone signal $m(n)$ using the far-end reference signal $x(n)$. The output of the linear AEC, $a(n)$, is the estimate of the echo signal, and the error signal $e(n) = m(n) - a(n)$ is the estimate of the (noisy) near-end signal.

2.3 Performance Measures

To evaluate the performance of the proposed and compared systems, two scenarios are considered: far-end only and double-talk. Except for some results presented in Chapter 3, near-end-only periods are not considered for performance evaluation and comparison since all systems introduce little distortion to the input signal when no echo is present. Furthermore, since it is a trivial task to determine that the far-end speaker is silent, during these periods, the microphone signal can be directly passed to the system's output. Thus, no distortion will be applied to it.

During far-end-only periods, we expect the enhanced signal to have as low energy as possible (ideally, it is completely silent). Therefore, performance is evaluated during these periods using the echo return loss enhancement (ERLE), which measures the echo reduction between the microphone signal and the enhanced signal. ERLE is measured in dB and is defined as

$$\text{ERLE} = 10\log_{10}\frac{||m(n)||^2}{||\tilde{d}(n)||^2}. \tag{2.3}$$

ERLE may not always correlate well with human subjective ratings [39]. AEC mean opinion score (AECMOS) [40] provides a speech quality assessment metric for evaluating echo impairment that overcomes the drawbacks of conventional methods. AECMOS is a DNN trained to directly predict subjective ratings for echo impairment using ground-truth human ratings of more than 148 hours of data. The model predicts two scores in the range $[1, 5]$, one for echo impairment (AECMOS-echo) and the other for other degradations (AECMOS-degradations). The model distinguishes between three scenarios: near-end single-talk, far-end single-talk, and double-talk. In the far-end single-talk case, only AECMOS-echo is considered. AECMOS is used to evaluate performance in Chapter 4.

We aim to suppress the residual echo during double-talk periods while maintaining near-end speech quality. During these periods, performance is evaluated using three different measures. The first measure is perceptual evaluation of speech quality (PESQ) [41]. PESQ is an intrusive speech quality metric based on an algorithm designed to approximate a subjective evaluation of a degraded audio sample. PESQ score range is $[-0.5, 4, 5]$, where a higher score indicates better speech quality. However, like ERLE, PESQ does not always correlate well with subjective human ratings. Therefore, the second performance measure, used to evaluate performance in Chapter 3, is deep noise-suppression mean opinion score (DNSMOS) [42]. DNSMOS is a perceptual objective speech quality metric that was initially developed to evaluate noise suppressors and does not require a clean reference signal. Similarly to AECMOS, DNSMOS is a DNN trained to predict subjective ratings of noise suppressors. The third performance measure, used to evaluate performance in Chapter 4, is AECMOS-echo which measures the echo reduction during double-talk periods. We do not use AECMOS-degradations for performance evaluation for two reasons: (i) we focus on the low SER scenario without including intense noise or distortions which may cause additional degradations, and (ii) as we show in the results, AECMOS-degradations fails to capture the true residual echo suppression performance in the low SER case.

Finally, although not performance measures, we formally define SER and echo-to-noise ratio (ENR). SER is measured in double-talk periods and used to measure the

near-end signal's energy relative to the echo signal's energy. SER is expressed in dB as

$$\text{SER} = 10\log_{10}\frac{||d||^2}{||y||^2}\,. \tag{2.4}$$

ENR is also expressed in dB and is measured during far-end-only periods. ENR is defined as

$$\text{ENR} = 10\log_{10}\frac{||y||^2}{||v||^2}. \tag{2.5}$$

Chapter 3

Acoustic Echo Cancellation with the Normalized Sign-Error Least Mean Squares Algorithm and Deep Residual Echo Suppression

This chapter presents an echo suppression system that combines a linear AEC with the NSLMS algorithm with a deep-complex convolutional recurrent network (DCCRN) for residual echo suppression. The main focus of the research in this chapter is the utilization of the NSLMS algorithm for acoustic echo cancellation and its effect on the performance of RES. Two alternatives are considered for RES: the proposed deep-learning model and a pre-trained speech enhancement model. In Section 3.1, we present the different components of the systems: the weights update equations of the NSLMS and NLMS algorithms, the proposed RES, and the pre-trained speech enhancement model. Details regarding the implementation, training procedures, and data, are given in Section 3.2. We provide experimental results and a comparison between the different systems and conditions in Section 3.3. The chapter is summarized in Section 3.4.

3.1 System Components

An RES system comprises a linear AEC and an RES model. Two linear AECs are being compared: NSLMS and NLMS. For RES, two alternatives are considered: the proposed RES model and a pre-trained speech-denoising model.

3.1.1 Linear Acoustic Echo Cancellers

For linear acoustic echo cancellation, we employ an AEC with the NSLMS algorithm. The algorithm operates in the subband domain by transforming the signals using uniform single-sideband filter banks [43, Section 7.6]. The filters' tap weights update

equation for each subband is given by

$$\mathbf{c}(n+1) = \mathbf{c}(n) + \frac{\alpha(n)\text{sgn}(e(n))\mathbf{x}_N(n)}{||\mathbf{x}_N(n)||^2} \tag{3.1}$$

where $\alpha(n)$ is the step size, and $\text{sgn}(\cdot)$ is the signum function. The performance of NSLMS is compared to that of NLMS, for which the tap weights update equation is given by

$$\mathbf{c}(n+1) = \mathbf{c}(n) + \frac{\alpha(n)e(n)\mathbf{x}_N(n)}{||\mathbf{x}_N(n)||^2}. \tag{3.2}$$

The normalization factor allows the steady-state error of the AEC to be independent of the far-end signal power [44].

3.1.2 Residual Echo Suppression Model

For residual echo suppression, we adopt the DCCRN [28] architecture, which employs a complex convolutional encoder-decoder structure and a complex LSTM. The model was originally developed for speech enhancement in the time-frequency domain. It estimates a complex ratio mask (CRM) applied to the input signal's short-time Fourier transform (STFT). For the purpose of residual echo suppression, we adapt the model to have 4 input channels instead of one and feed it with all available signals: $e(n)$, $a(n)$, $x(n)$, and $m(n)$. The estimated CRM is applied to the STFT of the error signal, $E(f,k)$. Fig. 3.1 depicts the model architecture. The encoder and decoder branches

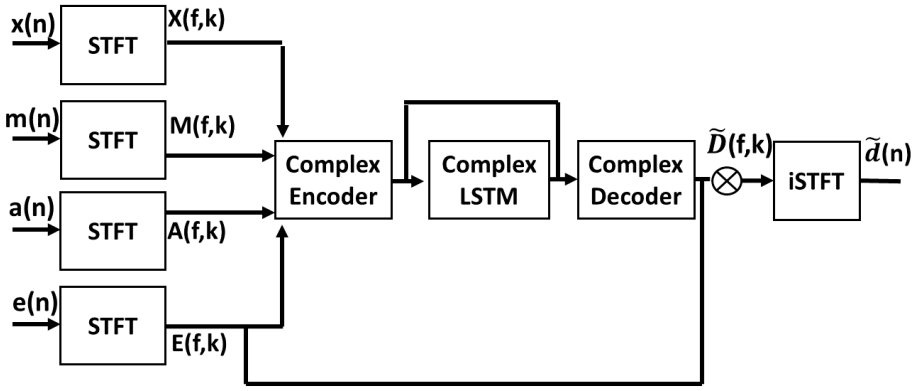


Figure 3.1: Residual echo suppression model architecture.

of the network are symmetrical, where the outputs of each encoder block are used as the inputs of the next encoder block as well as additional inputs to the decoder block of the same level. These connections between the different encoder and decoder blocks

are termed skip connections. Skip connections have two advantages: they provide an alternative path for the gradient during back-propagation, which is beneficial for model convergence, and they allow re-using of features from the encoder in the decoder. Each encoder/decoder block is comprised of a complex 2-D convolution layer, a complex batch-normalization layer, and a real PReLU activation function, as depicted in Figure 3.2. A complex 2-D convolution layer is comprised of two real 2-D convolution layers,

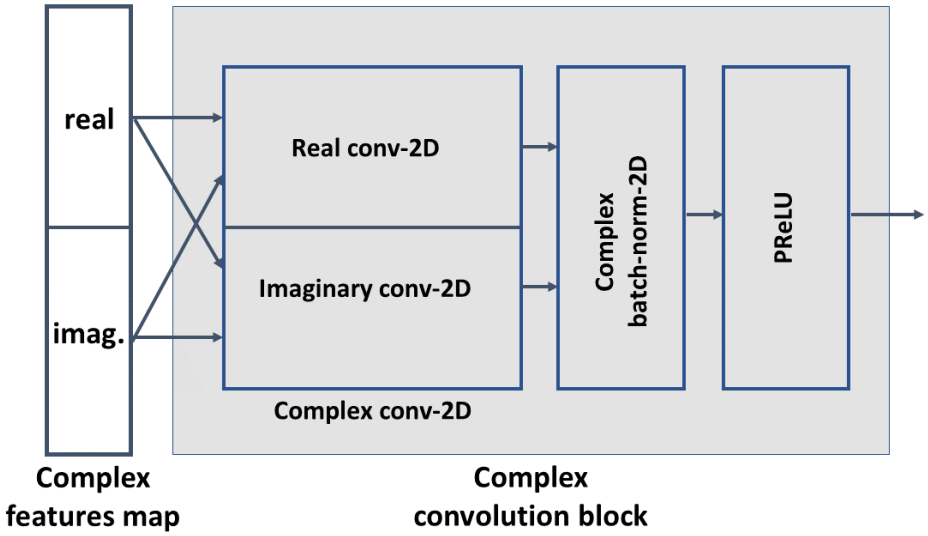


Figure 3.2: Structure of a complex convolution block. The input features map, consisting of real and imaginary parts, is fed to a complex 2-D convolution layer, the outputs of which are fed to a complex 2-D batch normalization layer. A PReLU activation function provides the block's output.

each operating on both the real and imaginary parts of its input. The output of a complex 2-D convolution layer, denoted by O_c, is formulated as

$$O_c = (X_r * W_r - X_i * W_i) + j(X_r * W_i + X_i * W_r)\,, \tag{3.3}$$

where X_r, X_i are the real and imaginary parts of the input, respectively, W_r, W_i are the real and imaginary convolution kernels, respectively, and $*$ is the convolution operation. We note that the complex convolution was implemented this way since, at the time of developing this model, the framework used (PyTorch) did not support convolution with complex numbers. Support was added since then, although we expect that the performance of a single convolution layer with double the number of channels is comparable to the performance of the implementation used in this thesis.

Similar to the complex 2-D convolution layer, the complex LSTM layer is comprised of two real LSTM layers, denoted by LSTM_r and LSTM_i. The output of the complex

LSTM, denoted by F_c, is formulated as

$$F_c = (\text{LSTM}_r(X_r) - \text{LSTM}_i(X_i)) + j(\text{LSTM}_i(X_r) + \text{LSTM}_r(X_i))\,. \tag{3.4}$$

Since a clean near-end signal is unavailable when training with real, recorded data, the training target is the noisy near-end signal $d(n) + v(n)$. As a training objective, we use the waveform ℓ_1 loss, combined with the multi-resolution STFT magnitude loss, adopted from [27]. For an estimated signal $\tilde{\mathbf{y}}$ and its ground-truth $\mathbf{y}$, the loss is defined as

$$\text{Loss} = \frac{1}{T}[||\mathbf{y} - \tilde{\mathbf{y}}||_1 + \sum_{i=1}^{M} L_{\text{mag}}^{(i)}(\mathbf{y}, \tilde{\mathbf{y}})] \tag{3.5}$$

$$L_{\text{mag}}^{(i)}(\mathbf{y}, \tilde{\mathbf{y}}) = \frac{1}{T}||\log|\text{STFT}(\mathbf{y})| - \log|\text{STFT}(\tilde{\mathbf{y}})|||_1 \tag{3.6}$$

where T is the number of time steps, $||\cdot||_1$ is the ℓ_1 norm, M is the number of STFT resolutions, and i is the resolution index.

3.1.3 Speech Denoising Model

As an alternative to the RES model, we utilize an off-the-shelf, pre-trained speech-denoising deep-learning model [27] which accepts a single input $e(n)$ and outputs $\tilde{d}(n)$. The model operates in the time domain, and similarly to DCCRN, it employs a convolutional encoder-decoder structure and an LSTM between the encoder and the decoder. The model is pre-trained on the Valentini dataset [45] and the INTERSPEECH 2020 deep noise suppression (DNS) challenge dataset [46]. The model is subsequently fine-tuned with the same training data used for training the RES models, once with the NSLMS outputs and once with the NLMS outputs. The loss function that is minimized is given in (3.5).

3.2 Experimental Setup

3.2.1 Datasets

Two datasets were employed for training the different models: the ICASSP 2021 AEC challenge synthetic dataset [47] and an independently recorded dataset. The independent recordings were taken to train and test the systems in real-life conditions with low SERs. Some variations in recording conditions include different near-end source positions and distances from the microphone, echo-path changes, and different room sizes with varying reverberation times. The dataset was created as follows. 5.5 hours of speech from the TIMIT [48] corpus and 5.5 hours of speech from the LibriSpeech [49]

corpus were used in the recordings. Double-talk utterances were generated with an average overlap of 90% and contained two different speakers. The generated dataset contains an equal amount of female and male speakers. To simulate a low SER scenario, such as a conversation over a mobile phone where the loudspeaker plays the far-end signal with high volume, Spider MT503$^{\text{TM}}$ or Quattro MT301$^{\text{TM}}$ speakerphones were employed, in which the microphone and loudspeaker are enclosed within a distance of 5 cm. In order to introduce echo path changes, in some of the recordings, the echo was played by a Logitech type Z120$^{\text{TM}}$ loudspeaker. The loudspeaker was moved 1, 1.5, or 2 meters away from the microphone during recordings. In order to simulate near-end speech, mouth simulator type 4227-A$^{\text{TM}}$ of Bruel&Kjaer was employed to generate the near-end signal. Three different positions were used for the mouth simulator, either at 1, 1.5, or 2 meters from the microphone. Additional variations in recording conditions include 4 different room sizes (between $3 \times 3 \times 2.5\, m^3$ and $5 \times 5 \times 4\, m^3$) and different reverberation times (RT_{60}), which vary between 0.3 and 0.6 s. Further details concerning the recordings can be found in [25]. The training data SER is distributed on $[-24, 18]$ dB, and the test data SER is distributed on $[-18, 5]$ dB. Test data speakers are unique and not used in the training set.

The ICASSP 2021 AEC challenge synthetic dataset was used to augment the training data. About 27.7 hours of data were generated, with different scenarios including near-end only, far-end only, double-talk, with/without near-end noise, and likewise for far-end. In addition, several nonlinear distortions were applied, with different SERs and signal-to-noise ratios. Further details regarding the dataset can be found in [47].

3.2.2 Implementation details

Before being fed to the linear AECs, the input signals, with a sampling rate of 16 kHz, are transformed using uniform 32-band single-sideband filter banks [43, Section 7.6]. Both AECs comprise filters of 150 taps in each subband (equivalent to time-domain filters of 150 ms length with 2400 taps).

For the RES model, all input signals are transformed to the time-frequency domain with a 512-point STFT, resulting in 257 frequency bins. The STFT window length is 25 ms, and the hop length is 6.25 ms. The number of convolution kernels for the different encoder layers is $[16, 32, 64, 128, 256, 256]$. The LSTM has 2 layers with a 128 hidden size. The model comprises 2.07 M parameters. Training optimization is done with the Adam optimizer [50] and an initial learning rate of $5e^{-4}$. The learning rate is decreased by a factor of 2 if there was no validation loss improvement for 3 consecutive epochs. Mini-batch size is 16, and the training continues for a maximum number of 100 epochs, where early-stopping is applied if no validation loss improvement occurs for 10 consecutive epochs.

The denoiser was pre-trained using the Valentini dataset [45] and the INTERSPEECH 2020 DNS challenge dataset [46]. The model comprises 18.87 M parameters.

Table 3.1: Performance comparison of the different systems. FE stands for far-end only, NE stands for near-end only, and DT stands for double-talk.

	ERLE	**DNSMOS**		**PESQ**	
	FE	**DT+NE**	**DT**	**DT+NE**	**DT**
NLMS	16.60	2.81	2.62	3.33	2.42
NSLMS	**21.17**	**2.86**	**2.71**	**3.66**	**2.98**
NLMS+ Denoiser	32.63	2.72	2.44	3.23	2.32
NSLMS+ Denoiser	**39.44**	**2.84**	**2.65**	**3.63**	**3.13**
NLMS+ RES	38.55	2.76	2.46	3.34	2.53
NSLMS+ RES	**40.34**	**2.84**	**2.64**	**3.70**	**3.11**

For a fair comparison with the RES model, we employ the causal version of the denoiser. For both linear AECs, the model is fine-tuned using the same data used to train the RES model. Training continues for 20 epochs with a learning rate of $3e^{-4}$ using the Adam optimizer [50]. Further details regarding the model architecture can be found in [27].

3.3 Experimental results

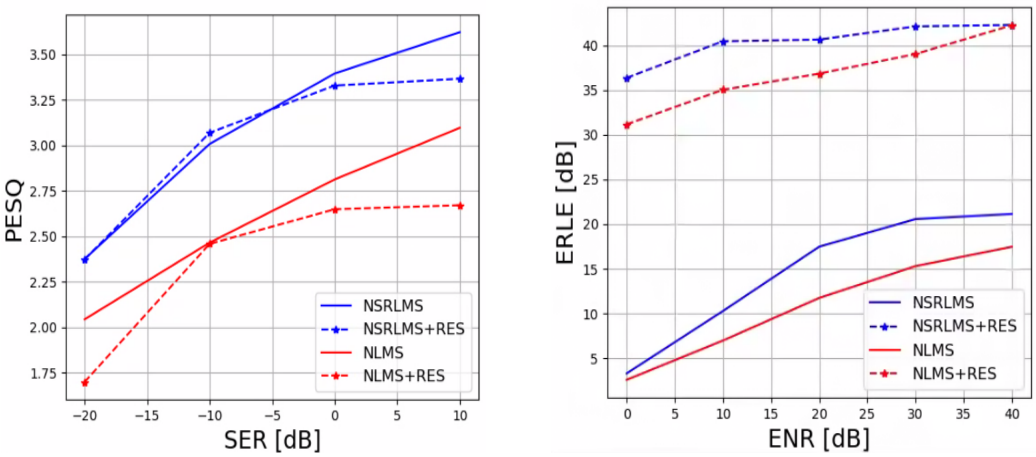
Table 3.1 shows the different methods' performance on the test set: the linear AECs (NLMS and NSLMS), the denoiser [27] operating on the outputs of each of the linear AECs (NLMS+Denoiser and NSLMS+Denoiser), and the RES model combined with each of the linear AECs (NLMS+RES and NSLMS+RES). As seen from the table, NSLMS achieves superior results over NLMS both in the cancellation of far-end echo when only the far-end signal is present (as indicated by ERLE) and in preserving near-end speech quality when the near-end signal is present (as indicated by DNSMOS and PESQ). We differentiate PESQ and DNSMOS for the double-talk-only scenario from the respective results when also including the near-end-only scenario. As expected, there is a degradation in results in the double-talk scenario for both linear AECs. The NLMS PESQ degrades by 0.91 while the NSLMS PESQ degrades by a smaller amount of 0.68 - further showing the superiority of NSLMS AEC over NLMS AEC in double-talk scenarios. NSLMS achieves superior results over NLMS in the denoiser setting and the RES setting as well.

The NSLMS+RES system achieves better residual echo suppression capabilities than the NSLMS+Denoiser system, as seen from the 0.9 decibel gap in ERLE. The near-end speech quality of both systems is on-par, as seen from the DNSMOS and PESQ scores. When taking these measures in double-talk scenarios, the denoiser system has a

negligible advantage over RES. When including near-end-only scenarios, both systems achieve identical DNSMOS, and the RES system achieves a higher PESQ score. When comparing the two systems to the baseline NSLMS linear AEC, it can be seen that while both systems achieve improved PESQ over the baseline in double-talk-only scenarios, the denoiser sees degradation in PESQ in near-end-only scenarios. In contrast, the RES system improves PESQ compared to the baseline. These results show the efficacy of the proposed RES model, as it achieves better echo suppression and on-par near-end speech quality with the denoiser while requiring 10 times fewer parameters than the denoiser model, which was also pre-trained on a large corpus with diverse speakers and noises.

When comparing the performance of both RES systems, it can be seen from the table that NSLMS+RES is favorable over NLMS+RES. The gap in ERLE between the two systems is smaller than the gap in ERLE between the respective linear AECs outputs. Both systems see degradation in DNSMOS compared to their respective baseline DNSMOS, but the NSLMS+RES degradation is smaller than that of NLMS+RES. Both RES systems achieve improved near-end speech quality compared to the baseline linear AECs as measured by PESQ, and the improvement in the NSLMS setting is more significant than that in the NLMS setting by a small margin. Overall, both NLMS and NSLMS perform well when combined with the proposed RES model, where NSLMS+RES shows superior results over NLMS+RES - both in all reported measures and the near-end speech quality gap compared to the linear AEC.

Denoising the output error signal $e(n)$ of the linear AECs results in a significant gap in performance between NSLMS and NLMS compared to the gap in performance in the RES setting. The NSLMS+Denoiser system achieves ERLE that is higher by 6.81 decibels than the NLMS+Denoiser system. This ERLE gap is more significant than the respective ERLE gap in the linear AEC and the RES settings. It is due to the NLMS output residual echo, which is more structured and less akin to noise than the residual echo in the output of NSLMS, as was suggested in [7]. This results in the denoiser being unable to cancel some of the residual echoes that resemble human speech more closely than noise. Significant differences are also observed in PESQ scores. While NSLMS+Denoiser achieves improved double-talk scenario PESQ compared to the baseline linear AEC, NLMS+Denoiser sees degradation in PESQ compared to the baseline. This result further strengthens the claim that NSLMS outputs residual echo that more resembles noise than the residual echo in the NLMS output - the denoiser is better able to suppress the residual echo and preserve the near-end speech in the NSLMS setting. In contrast, in the NLMS setting, it identifies some of the residual echoes as speech and cannot distinguish them from the near-end speech. When measuring PESQ in the near-end only scenario as well, the gap between NSLMS+Denoiser PESQ and NLMS+Denoiser PESQ is smaller. This further shows that the gap in performance between the two is mainly due to the denoiser being better able to suppress the far-end echo in the NSLMS setting. The above results and observations show that when



(a) PESQ in double-talk only scenario

(b) ERLE in far-end only scenario

Figure 3.3: Comparison of the linear AECs with or without RES.

employing an off-the-shelf speech denoising model to the task of RES, NSLMS is better suited to the preceding task of linear acoustic echo cancellation than NLMS.

Next, we compare the performance of NSLMS and NLMS as a baseline AEC as well as combined with the proposed RES model for different SERs and ENRs. Figure 3.3(a) shows the PESQ scores for different values of SER in the double-talk scenario. NSLMS achieves superior PESQ over NLMS for all SERs, both for AEC and RES. Furthermore, when the SER is low, NSLMS+RES achieves improved PESQ over the baseline NSLMS AEC, while NLMS+RES never surpasses the baseline PESQ. Figure 3.3(b) shows ERLE for different values of ENR when only far-end speech is present. NSLMS achieves superior ERLE over NLMS for all ENRs, both in the AEC and the RES settings. For the RES systems, it can be observed that the performance gap is more significant for lower ENRs.

3.4 Summary

In this chapter, we have presented an echo suppression system based on the NSLMS-AEC and the DCCRN speech enhancement model. We conducted experiments in challenging real-life conditions with low SER. We compared the performance of the proposed system to the performance of a pre-trained speech-denoising model operating on the error signal at the output of the linear AEC and fine-tuned with the same training data. Results show that although the speech denoising model was pre-trained on a large corpus with diverse speakers and conditions and is 10 times larger concerning the number of parameters, the proposed RES model achieves better residual echo suppression capabilities and on-par near-end speech quality. We also compared the performances of all the systems using NSLMS-AEC and NLMS-AEC. Results show that NSLMS achieves superior results over NLMS in all settings and for a wide range of SER and ENR val-

ues. Notably, the results support the claim that the NSLMS produces a residual echo that is less structured than the output produced by the NLMS, as observed from the denoiser performance gap between the two. Therefore, when the complexity of the model is not an important consideration, fine-tuning a readily available denoiser could be a reasonable alternative to creating a new RES model. However, the choice of linear AEC becomes more critical, and NSLMS should be preferred.

Chapter 4

Double-talk Detection-aided Residual Echo Suppression via Spectrogram Masking and Refinement

This chapter presents a two-stage residual echo suppression system that focuses on the low SER scenario. The system employs the NSLMS linear AEC discussed in Chapter 3. The first stage of the RES includes spectrogram masking and double-talk detection. The second stage performs spectrogram refinement. In Section 4.1, we present the spectrogram masking and double-talk detection model. In Section 4.2, we present the spectrogram refinement model. Section 4.3 discusses the data used for training and evaluation and the models' training procedures. Results are presented and discussed in Section 4.4. Section 4.5 summarizes this chapter.

4.1 Masking and Double-Talk Detection

The spectrogram masking model aims to perform a significant portion of the residual echo suppression. The greatest challenge in residual echo suppression is suppressing the echo during double-talk periods while reducing the near-end speech distortion. Therefore, the model may benefit from optimization for double-talk detection in tandem with residual echo suppression.

In the masking stage, we employ the U-Net architecture [12]. This architecture differs from the one in [38], where the architecture comprises convolutional blocks consisting of residual connections, requiring more model parameters and resulting in a longer inference time than the U-Net architecture while achieving similar performance. U-Net has a fully-convolutional encoder-decoder structure with skip connections between levels of the encoder and the decoder. The proposed model is a concatenation of

two U-Nets. The first U-Net performs double-talk detection and is also used to learn a feature representation from the double-talk predictions and the input signals used for the masking task. The second U-Net receives the outputs of the first U-Net and all input signals and predicts a spectrogram ratio mask.

The first U-Net's input is the log of the input signals $X(f,k)$, $A(f,k)$, $M(f,k)$, and $E(f,k)$, concatenated along the channel dimension. The encoder comprises down-sampling convolution blocks (referred to as "down-blocks" from here on), where each block consists of a 2-D convolution layer, instance normalization layer [51], and leaky rectified linear unit (leaky ReLU) [52] activation function. The convolution window stride is 2 along the frequency dimension and 1 along the time dimension - effectively down-sampling the inputs along the frequency dimension while preserving the time dimension. The output of the encoder is fed to a uni-directional GRU, which learns time dependency between the different frames. The GRU's output has two purposes - it is used both as features utilized by a classifier that predicts double-talk for each time frame and as inputs to the decoder, which learns a representation from the DTD's features. We frame the double-talk detection task as a binary multi-label classification task, where each time frame is labeled as either containing near-end speech or not, as well as either containing far-end speech or not. We empirically found that this approach leads to better classification performance than the more common approach of multi-class classification, where each time frame is assigned a single label (most commonly, the labels are: silence, near-end speech only, far-end speech only, or double-talk). To classify each time frame, the outputs of the GRU are fed to a fully-connected layer responsible for reducing the feature dimension (while preserving the time dimension as we want to classify each time frame) to 2, which corresponds to the two possible labels.

The features learned by the encoder for double-talk detection are employed to assist the task of learning a spectrogram mask. Instead of directly feeding the masking U-Net with the encoder's features, the decoder learns a feature representation. The decoder comprises up-blocks similar to down-blocks, except that the inputs are first up-sampled via nearest-neighbor up-sampling with a factor of 2 along the frequency dimension and 1 along the time dimension. The up-sampled inputs are concatenated along the channel dimension with the outputs of the matching level of the encoder. The output of the decoder, $P(f,k)$, has a single channel and is of the same frequency and time dimensions as the input signals. To learn a spectrogram mask, an additional U-Net is concatenated to the first U-Net. This U-Net accepts as inputs the log of all input signals $X(f,k)$, $A(f,k)$, $M(f,k)$, and $E(f,k)$, as well as the output of the first U-Net $P(f,k)$, resulting in 5 input channels. The second U-Net's structure is similar to that of the first U-Net with a few exceptions - the down-sampling (as well as the up-sampling) factor is 2 for both frequency and time dimensions, and the last decoder block contains neither an activation function nor a normalization layer. The model's output, denoted by $\hat{H}(f,k)$, consists of one channel and has the same frequency and time dimensions as the input signals. The entire DTD and masking model's architecture is depicted in Figure 4.1.

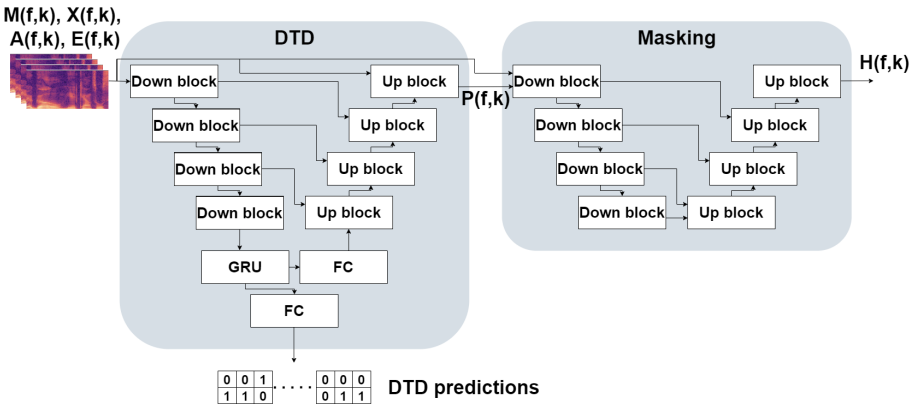


Figure 4.1: Structure of the double-talk detector (DTD) and masking model architecture. FC stands for fully connected.

As previously mentioned, we empirically found that the DTD performs better when trained to detect near-end and far-end speech separately for each time frame. Therefore, the DTD's training target for each utterance is a tensor of shape $(B, 2, T)$, where B is the batch size, and T is the number of time frames. The first and second rows of the second dimension represent the presence of near-end speech and far-end speech, respectively, where 1 indicates the presence of speech and 0 represents its absence. The training target of the masking model is the log of the ratio between the spectrogram magnitudes of the clean near-end speech and that of the error signal, denoted by $H(f, k)$ and given by

$$H(f,k) = \log_{10}\left(\frac{D(f,k)}{E(f,k) + \epsilon_1} + \epsilon_2\right) \ , \tag{4.1}$$

where ϵ_1 and ϵ_2 are small constants for numerical stability. The loss function used for the double-talk detection task is denoted by l_{DTD} and given by

$$l_{\text{DTD}} = \frac{1}{2}(l_{\text{DTD-nearend}} + l_{\text{DTD-farend}}) \ , \tag{4.2}$$

where $l_{\text{DTD-nearend}}$ and $l_{\text{DTD-farend}}$ are binary cross entropy (BCE) loss terms for near-end and far-end speech detection, respectively. $l_{\text{DTD-nearend}}$ is given by

$$l_{\text{DTD-nearend}} = -\frac{1}{T}\sum_{k=1}^{T}[v_k \cdot \log\, \sigma(\hat{v}_k) + (1 - v_k) \cdot \log(1 - \sigma(\hat{v}_k))] \ , \tag{4.3}$$

where v_k is the ground-truth label for time frame k, $\hat{v}_k$ is the predicted label for time frame k, and $\sigma(\cdot)$ is the sigmoid function. $l_{\text{DTD-farend}}$ is similarly defined. For the masking task, we use the mean squared error (MSE) loss between the labels and the outputs, denoted by l_{mask} and given by

$$l_{\text{mask}} = \frac{1}{n}\sum_{f}\sum_{k}(H(f,k) - \hat{H}(f,k))^2 \ , \tag{4.4}$$

where n is the total number of spectrogram bins. The overall loss function used to optimize the model is a weighted sum of the two loss functions with a weight parameter λ_{DTD} applied to l_{DTD}:

$$l = \lambda_{\text{DTD}}\, l_{\text{DTD}} + l_{\text{mask}}\,. \tag{4.5}$$

4.2 Spectrogram Refinement

The spectrogram masking approach alone may not be sufficient to both suppress the residual echo and preserve the near-end speech's quality. It is especially true in the low SER scenario, where the echo signal's energy is considerably higher than the near-end signal's. In this case, spectrogram masking can suppress the residual echo to a large degree, at the cost of degrading the near-end speech quality. In the most severe cases, the masking operation completely cancels parts of the near-end speech during double-talk. In [38], following the masking stage, the speech is further enhanced by a spectrogram inpainting stage. The inpainting operation aims to reconstruct spectrogram bins containing speech canceled in the masking stage. In the residual echo suppression case, near-end speech is screened by far-end echoic speech rather than noise. The screening renders the reconstruction operation more challenging as it may be difficult to distinguish the speech components of the near-end signal from those of the far-end signal. Instead, we frame this stage as spectrogram refinement, where the mask learned by the masking model is used as an additional feature along with the input signals rather than to mask the signal from which we want to obtain the desired near-end speech.

For spectrogram refinement, we adopt the architecture used in [38]. In our experiments, we found that when using the U-Net architecture for this stage, the model's performance was almost identical to the performance of the masking model. Since the masking model achieves good performance on its own, and due to the skip connection between the inputs and the decoder outputs, the refinement model with the U-Net architecture achieved negligible performance gain compared to the masking model. Instead, we employ a fully-convolutional architecture consisting of residual connection blocks, as proposed in [38].

The input to the model is the log of the input signals $X(f,k)$, $A(f,k)$, $M(f,k)$, and $E(f,k)$, the output of the masking model $\hat{H}(f,k)$, and the double-talk features $P(f,k)$, concatenated along the channel dimension. The input is first fed to two consecutive down-blocks, similar to the encoder blocks in the masking stage. The inputs are downsampled by a factor of 2 along both time and frequency dimensions. Instead of leaky ReLU, we employ an exponential linear unit (ELU) activation function [53] as proposed in [38]. Following the down-blocks is a series of identical residual blocks. A residual block comprises two consecutive down-blocks with a convolution kernel stride $(1,1)$. The output of the second convolution block is summed element-wise with the input to the residual block. Following the last residual block are two up-blocks with an up-

sampling factor of 2 along both time and frequency dimensions. The output layer is a 2-D convolution layer with one output channel. Figure 4.2 depicts the refinement model's architecture and the residual blocks.

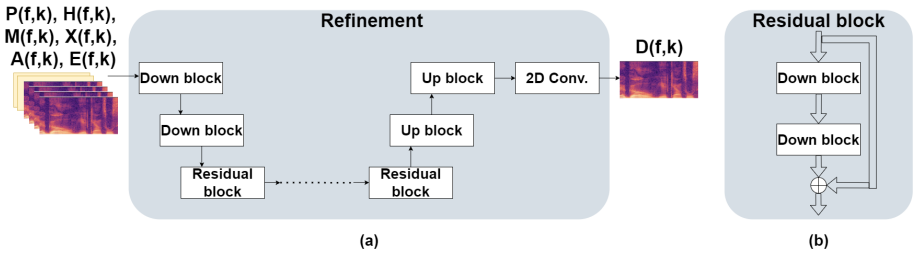


Figure 4.2: Structure of refinement model architecture and residual blocks. (**a**) Refinement model architecture. (**b**) Structure of the residual blocks.

We frame the refinement stage as a regression task, where the model learns to predict the near-end spectrogram magnitudes directly. Therefore, the training target is the log of the near-end signal's spectrogram magnitudes $\log_{10}(D(f,k)+\epsilon)$, where ϵ is a small constant for numerical stability. Inverting the log operation and applying inverse STFT (iSTFT) using the error signal's phase, we obtain the time-domain near-end signal $d(n)$. Since a significant portion of the residual echo was suppressed in the masking stage, the main goal of the refinement stage is to improve the estimated near-end speech quality. We achieve this goal by optimizing the model for speech quality measured by PESQ. Since the PESQ function is non-differentiable, it cannot be used as a loss function in gradient-descent-based algorithms. The PMSQE loss function [54] aims to approximate PESQ with a differentiable function. PMSQE, unlike MSE, takes into account perceptual-related features of the predicted signal by incorporating two disturbance terms inspired by the PESQ algorithm. We denote the PMSQE loss term by l_{PESQ} (for brevity, we do not formulate the loss function and its different components here - the reader is referred to [54] for additional details). We empirically found that minimizing the PMSQE loss function alone does not achieve the desired results since, although the loss value converges, the different evaluation metrics diverge. Therefore, we add a regularizing MSE loss term defined as

$$l_{\text{MSE}} = \frac{1}{n}\sum_{f}\sum_{k}(\log_{10}(\tilde{D}(f,k)+\epsilon) - \log_{10}(D(f,k)+\epsilon))^2\,. \tag{4.6}$$

The complete loss function minimized during the refinement model training is given by

$$l = l_{\text{PESQ}} + \lambda_{\text{MSE}} l_{\text{MSE}}\,, \tag{4.7}$$

where λ_{MSE} is a weight parameter.

The following tables detail the specifications of the different layers of the two stages' models.

Table 4.1: Double-talk detection and spectrogram masking model specifications. Module names with an asterisk are model outputs. For down-blocks and up-blocks, the numbers in the Details column represent input channels, output channels, kernel size, and stride of the convolution window or up-sampling factor, respectively. For the GRU layer, the Details column's numbers represent hidden-layer size and the number of layers, respectively. For fully-connected (FC) layers, the number represents the number of neurons. More than one module in the Input column means concatenation of the modules in parentheses.

Module	**Details**	**Inst. norm**	**Activation**	**Input**
Down-block 1	$(4, 32, 3, (2, 1))$	$\surd$	Leaky ReLU	Model's input
Down-block 2	$(32, 64, 3, (2, 1))$	$\surd$	Leaky ReLU	Down-block 1
Down-block 3	$(64, 128, 3, (2, 1))$	$\surd$	Leaky ReLU	Down-block 2
Down-block 4	$(128, 256, 3, (2, 1))$	$\surd$	Leaky ReLU	Down-block 3
GRU	$(128, 1)$	-	-	Down-block 4
FC 1*	2	-	Sigmoid	GRU
FC 2	2816	-	Leaky ReLU	GRU
Up-block 1	$(384, 128, 3, (2, 1))$	$\surd$	Leaky ReLU	(FC2, Down-block 3)
Up-block 2	$(192, 64, 3, (2, 1))$	$\surd$	Leaky ReLU	(Up-block 1, Down-block 2)
Up-block 3	$(96, 32, 3, (2, 1))$	$\surd$	Leaky ReLU	(Up-block 2, Down-block 1)
Up-block 4*	$(36, 1, 3, (2, 1))$	$\surd$	Leaky ReLU	(Up-block 3, Model's input)
Down-block 5	$(5, 32, 3, (2, 2))$	$\surd$	Leaky ReLU	(Up-block 4, Model's input)
Down-block 6	$(32, 64, 3, (2, 2))$	$\surd$	Leaky ReLU	Down-block 5
Down-block 7	$(64, 128, 3, (2, 2))$	$\surd$	Leaky ReLU	Down-block 6
Down-block 8	$(128, 256, 3, (2, 2))$	$\surd$	Leaky ReLU	Down-block 7
Up-block 5	$(384, 128, 3, (2, 2))$	$\surd$	Leaky ReLU	(Down-block 8, Down-block 7)
Up-block 6	$(192, 64, 3, (2, 2))$	$\surd$	Leaky ReLU	(Up-block 5, Down-block 6)
Up-block 7	$(96, 32, 3, (2, 2))$	$\surd$	Leaky ReLU	(Up-block 6, Down-block 5)
Up-block 8*	$(37, 1, 3, (2, 2))$	-	-	(Up-block 7, Up-block 4, Model's input)

Table 4.2: Refinement model specifications. Module names with an asterisk are model outputs. For down-blocks, up-blocks, and residual blocks (Res. blocks), the numbers in the Details column represent input channels, output channels, kernel size, and stride of the convolution window or up-sampling factor, respectively.

Module	**Details**	**Inst. norm**	**Activation**	**Input**
Down-block 1	$(6, 64, 3, (2, 2))$	$\surd$	ELU	Model's input
Down-block 2	$(64, 128, 3, (2, 2))$	$\surd$	ELU	Down-block 1
Res. block 1	$(128, 128, 3, (1, 1))$	$\surd$	ELU	Down-block 2
Res. block 2	$(128, 128, 3, (1, 1))$	$\surd$	ELU	Res. block 1
Res. block 3	$(128, 128, 3, (1, 1))$	$\surd$	ELU	Res. block 2
Res. block 4	$(128, 128, 3, (1, 1))$	$\surd$	ELU	Res. block 3
Res. block 5	$(128, 128, 3, (1, 1))$	$\surd$	ELU	Res. block 4
Up-block 1	$(128, 64, 3, (2, 2))$	$\surd$	ELU	Res. block 5
Up-block 2	$(64, 32, 3, (2, 2))$	$\surd$	ELU	Up-block 1
Up-block 3*	$(32, 1, 3, (1, 1))$	-	-	Up-block 2

4.3 Data and Training Procedures

We employ the independently recorded dataset, discussed in Chapter 3, to train and evaluate the proposed system in real-life conditions. Recorded data were split between the training and test sets, such that the test set contains unique speakers not shared by the training set and unique conditions and setups not seen during training. To augment the training dataset, synthetic data from the ICASSP 2021 AEC challenge dataset [47] were also used during training. In 80% of the cases, the far-end signal in the synthetic dataset was processed with a nonlinear function. Some examples of nonlinear functions are clipping of the maximum value, a sigmoidal function, or a learned nonlinear distortion function. More details regarding the synthetic data can be found in [47]. Since this part of the research focuses on the low SER scenario, The SER in both datasets (synthetic and independent recordings) was set to -20 ± 3 dB. For analysis in different SERs, the same data were used in every experiment, where the SER was set to -15 ± 3 dB, -10 ± 3 dB, or -5 ± 3 dB. The combined dataset consists of 34.1 hours of data with a 16 kHz sampling rate.

As mentioned in Chapter 3, the NSLMS linear AEC operates in the subband domain. Therefore before being fed to the AEC, the input signals are transformed using uniform 32-band single-sideband filter banks [43, Section 7.6]. The linear AEC comprises filters of 150 taps in each subband, equivalent to time-domain filters of length 150 ms with 2400 taps.

All inputs to the RES system are transformed to the time-frequency domain using a 320-point STFT with a window length of 20 ms and a hop length of 10 ms. For utterances of 2 s, this results in an input tensor of size $(B, 4, 161, 201)$ where B is the

batch size, 4 corresponds to the four input signals, and 161 and 201 are the frequency and time bins, respectively. Both stages' models are optimized with the Adam optimizer [50]. The initial learning rate of the masking model is $6e^{-4}$, and the initial learning rate of the refinement model is $1e^{-4}$. For both models, learning-rate scheduling is applied such that it is multiplied by a factor of 0.5 each time there was no validation loss improvement for 4 consecutive epochs. Early stopping is applied if there was no validation loss improvement for 8 consecutive epochs. We set $\lambda_{\text{DTD}} = 0.5$ to balance the size of the two loss terms of the masking and DTD model. λ_{MSE} is set to 1 since the regularizing loss term l_{MSE} is a magnitude-of-order smaller than l_{PESQ}. We set $\epsilon_1 = \epsilon_2 = \epsilon = 1e^{-8}$. For both models, the mini-batch size is 32, and the maximum number of epochs is 100. All models are implemented with Pytorch, and a single Nvidia GeForce GTX 1080 is used for training.

4.4 Experimental Results

In this Section, we present the experimental results for this chapter. First, we present the results of the ablation study, where we show the contribution of each part of the system. We also show the efficacy of the proposed DTD and compare it to other DTD configurations, based on previous studies. The DTD's classification results are also presented and discussed. Next, we present the comparison of the proposed system to other systems from previous studies.

4.4.1 Ablation study

First, we present the ablation study's results, showing how each part of the proposed system contributes to the performance. Table 4.3 shows the performance of the AEC, the performance of the AEC followed by the masking stage with and without double-talk detection (AEC+M+D and AEC+M, respectively), the performance of the AEC followed by the refinement stage without the masking stage's outputs (AEC+R, using only the input signals), and the entire system's performance - AEC followed by masking and double-talk detection followed by the refinement model (AEC+M+D+R).

Table 4.3: Ablation study results. M stands for masking, D for DTD, and R for refinement.

	Far-end only		**Double-talk**	
	ERLE	**AECMOS**	**PESQ**	**AECMOS**
AEC	18.80	4.67	2.25	4.15
AEC+M	40.39	4.67	2.74	4.66
AEC+M+D	42.28	4.67	2.84	4.69
AEC+R	40.69	4.66	2.75	4.57
AEC+M+D+R	**44.32**	**4.68**	**2.94**	**4.71**

From the table, combining the DTD with the masking model improves ERLE by almost 2 dB while achieving on-par far-end only AECMOS, which indicates better echo suppression performance when there is no near-end speech. During double-talk, there is a notable increase of 0.1 in the PESQ score and a minor increase of 0.03 in AECMOS. These results indicate that combining the DTD with the masking model improves performance compared to not combining a DTD during double-talk periods. When adding the refinement stage to the masking+DTD stage, there is an additional improvement in all measures. Most notably, ERLE is increased by an additional 2.04 dB, and PESQ is increased by 0.1. Far-end AECMOS and double-talk AECMOS are also improved, albeit by a negligible amount. It can also be observed how without first employing the masking stage, the refinement stage on its own achieves on-par performance with the masking model without the DTD. This further asserts the efficacy of the proposed system; the masking stage, aided by the DTD, performs the initial residual echo suppression, and the refinement stage, which relies on the features provided by the masking stage, further improves performance. It can be concluded from the ablation study that the proposed configuration of the DTD aids the masking model's performance and that the refinement stage indeed performs refinement to the outputs of the first stage since its stand-alone performance is inferior. Figure 4.3 shows examples of spectrograms from different stages of the system.

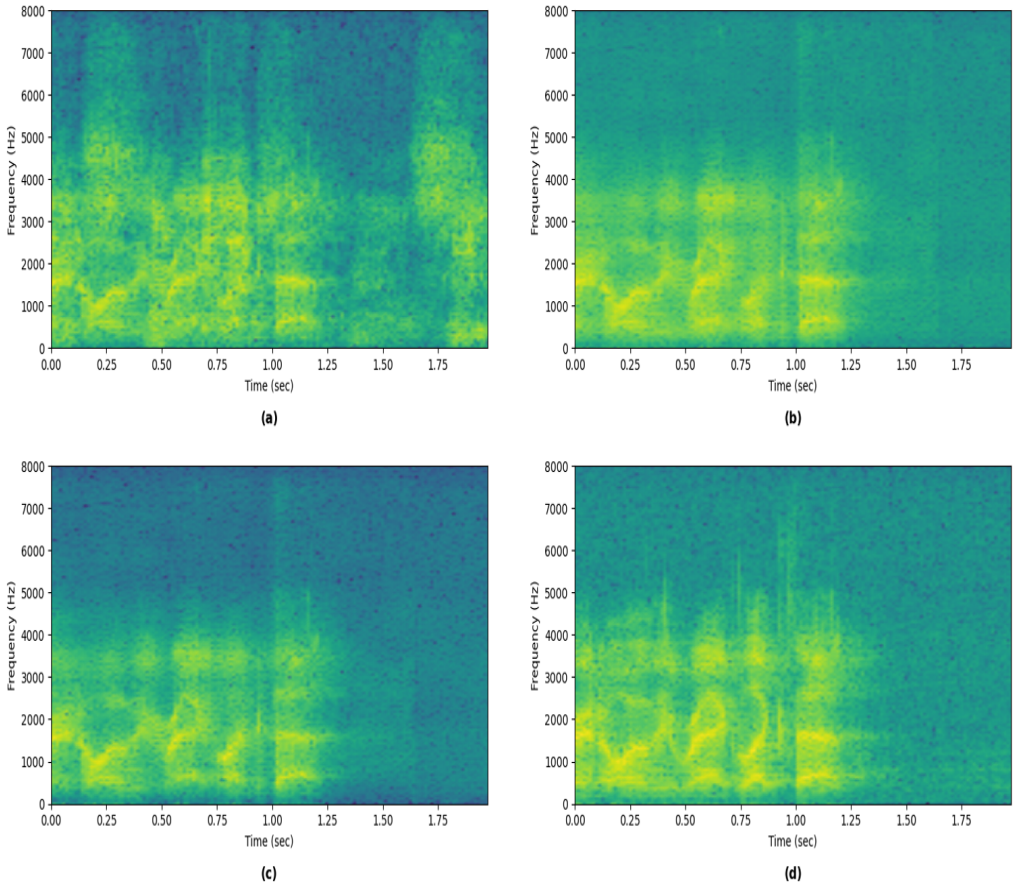


Figure 4.3: Visualization of spectrograms of the different stages' outputs. **(a)** Error signal spectrogram. **(b)** Spectrogram of the signal reconstructed from the masking stage's output. **(c)** Spectrogram of the refinement stage's output. **(d)** Near-end signal's spectrogram.

It can be observed from the figure that the masking model suppresses the majority of the residual echo, notably evident after 1.25 seconds and above 4000 Hz. The finer details of the near-end speech are blurred compared to the near-end spectrogram. The refinement model refines the output of the masking model, resulting in a finer-detailed spectrogram that closely resembles the near-end spectrogram.

Next, we study different ways to combine the DTD with the masking model. We compare five different configurations:

- **No double-talk detection** - A single U-Net is utilized to perform spectrogram masking (similar to the proposed system, without the first U-Net).
- **Configuration 1: Shared encoder** - A single U-Net, where the outputs of the encoder are used both by a double-talk classifier and by the decoder that outputs the spectrogram mask. This is similar to the configuration proposed in [36].
- **Configuration 2: Separate encoders, shared features** - Two identical encoders are employed. The first encoder learns features used for double-talk detec-

tion. The second encoder receives all input signals, and each level's features are concatenated with features from the matching level of the DTD's encoder. This is similar to the configuration proposed in [24].

- **Configuration 3: Separate decoders, conditioning** - The features learned by a single encoder are fed into two separate decoders. The first decoder performs double-talk detection. The second decoder learns a spectrogram mask, its outputs conditioned on the DTD's predictions by sharing the decoders' features in each matching level. This is similar to the configuration proposed in [33].

- **Proposed** - the configuration proposed in this study, as detailed in Section 4.1.

Table 4.4 shows the performance of the masking model combined with the DTD in each of the above configurations. The proposed configuration achieves the best residual echo suppression performance during far-end-only periods, as indicated by ERLE and AECMOS. The proposed configuration's ERLE is more than 1 dB greater than the second-best ERLE (Conf. 3), and the AECMOS equals the no-DTD baseline AECMOS. In contrast, all other configurations see a minor degradation. In the double-talk scenario, the proposed configuration's PESQ score is nearly 0.1 greater than the second-best PESQ (Conf. 2), which is only 0.01 greater than the no-DTD baseline PESQ. The AECMOS is also the highest among all compared configurations' AECMOS. Overall, results show that the proposed configuration of DTD combined with the masking model achieves a notable performance improvement compared to not combining a DTD, where all other configurations have little to no effect on performance. We conclude that combining a DTD with the masking model is beneficial when the double-talk detection is performed before the masking and that it is necessary to learn a feature representation from the DTD's predictions to enable the masking model to use these predictions effectively.

Table 4.4: Study of different configurations of the masking model with a DTD. Conf. stands for configuration.

	Far-end only		**Double-talk**	
	ERLE	**AECMOS**	**PESQ**	**AECMOS**
No DTD	40.39	**4.67**	2.74	4.66
Conf. 1	41.07	4.61	2.69	4.56
Conf. 2	39.88	4.66	2.75	4.60
Conf. 3	41.17	4.66	2.72	4.65
Proposed	**42.28**	**4.67**	**2.84**	**4.69**

For completion, we provide the DTD's performance in Table 4.5. Since the proposed DTD operates as a multi-label classifier where the labels are the presence of near-end speech and far-end speech, double-talk is not an actual class for the classifier. Instead,

it is determined for time-frames containing both near-end and far-end speech. The provided results for near-end and far-end include time frames where both are present (double-talk). Multi-class classification results are also provided for comparison. We can observe from the table that both near-end and far-end performance is high and that precision and recall are balanced. The far-end performance is slightly better than that of the near-end. This small performance gap is expected in the low SER setting since, during double-talk periods, the near-end speech may be almost indistinguishable. This observation is also evident in the double-talk results, notably degraded. During these periods, the DTD may predict a time frame as containing far-end speech and not containing near-end speech. When using the DTD's prediction directly as inputs to the subsequent masking model, it may cancel these time frames, as it learns to do so from the actual far-end only time frames. Learning a representation from the DTD's predictions helps overcome this issue. It can also be observed from the table that the proposed multi-label classifier outperforms the multi-class classifier. While near-end performance is on-par, the far-end performance and overall accuracy of the multi-label classifier are superior to that of the multi-class classifier. In the double-talk scenario, the multi-label classifier achieves superior precision and inferior recall, and its overall accuracy is notably superior to that of the multi-class classifier.

Table 4.5: Performance of the DTD. Numbers in parentheses represent the respective results of the multi-class classifier.

	Precision	**Recall**	**Accuracy**
Near-end	0.96 (0.95)	0.95 (0.96)	0.97 (0.97)
Far-end	0.98 (0.94)	0.97 (0.89)	0.98 (0.97)
Double-talk	0.90 (0.88)	0.91 (0.93)	0.86 (0.80)
Overall	-	-	0.98 (0.95)

Finally, we address an issue with double-talk AECMOS-degradations in the low SER scenario. Figure 4.4 shows double-talk AECMOS-degradations at different SERs, where the 'degraded' signals used to obtain the scores are $m(n)$, $e(n)$, $d(n)$, and $\tilde{d}(n)$. The graphs show how the microphone signal's AECMOS is substantially higher than the clean near-end speech's AECMOS. Furthermore, the gap between the two is more significant when the SER is lower. When the SER is low, the far-end speech is loud (and its quality is high since we do not consider noise or additional distortions in our data), while the near-end speech is nearly indistinguishable. Thus, the microphone signal's AECMOS-degradations are high, despite mainly containing undesired echo. On the other hand, the clean near-end speech signal's AECMOS-degradations are considerably lower, degrading further when the SER is lowered. This may indicate that the AECMOS model was not trained on such extreme cases since we expect this score to be high regardless of the SER as it contains no noise or distortions. Nevertheless, we can see that at all SERs, the enhanced signal $\tilde{d}(n)$ obtains slightly better AECMOS-

degradations than the error signal $e(n)$, indicating that the proposed model improves AECMOS-degradations compared to its input.

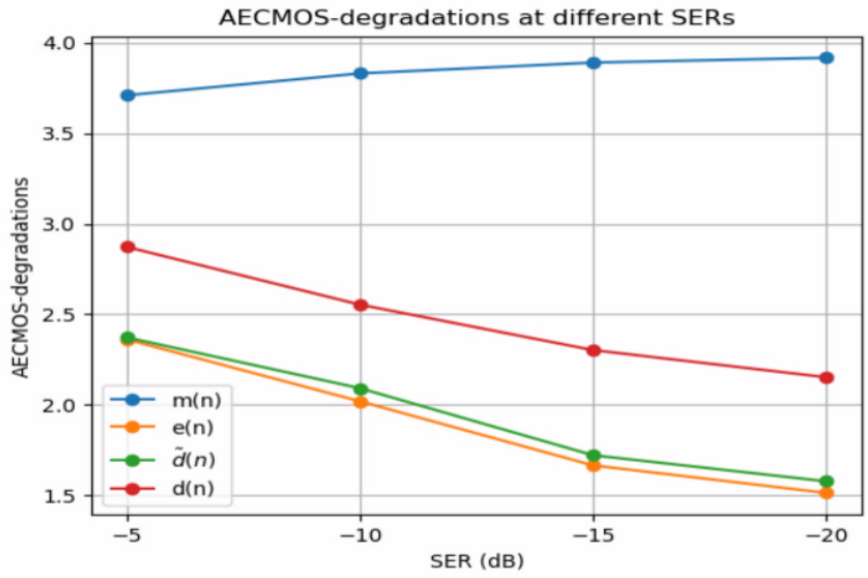


Figure 4.4: AECMOS-degradations of the different signals at various signal-to-echo ratios (SERs).

4.4.2 Comparative results

We compare the proposed system to two recent RES systems: Regression-U-Net [25], and Complex-Masking [24]. Both systems operate in the T-F domain. Regression-U-Net's inputs are the spectrogram magnitudes of $e(n)$ and $a(n)$. The model predicts the spectrogram magnitudes of $\tilde{d}(n)$. Since we optimize our refinement model to increase the PESQ score, we choose $\alpha = 0$ in the Regression-U-Net's implementation, as it yields the best PESQ [25]. Complex-Masking's model consists of a convolutional encoder and decoder and a GRU between them. All layers in the model are complex, which allows the model to learn a phase-aware mask while utilizing the complete information from the input signals. The model's inputs are the complex spectrograms of $e(n)$ and $x(n)$, and its output is a complex mask applied to the spectrogram of $e(n)$. We note the differences between the two systems: Regression-U-Net is real-valued and performs regression (outputs the desired signal directly). At the same time, Complex-Masking is complex-valued and performs masking rather than regression. Both systems were trained using the original code provided by the authors and the same training data used to train the proposed system, and they were evaluated using the same test data. Since our work focuses on the RES part, all systems used the same preceding linear AEC. Table 4.6 shows the performance of the different systems, their number of parameters

and memory consumption, and their real-time factor (RTF), defined as

$$\text{RTF} = \frac{t_{\text{inference}}}{t_{\text{signal}}}, \tag{4.8}$$

where $t_{\text{inference}}$ is the time it takes the model to infer an output for an input of duration t_{signal}. All systems' RTF is measured on the standard Intel Core i7-11700K CPU @ 3.60 GHz.

Table 4.6: Comparison of the proposed, the Regression-U-Net (U-Net), and the Complex-Masking (Masking) systems. Param. stands for parameters and Mem. for memory.

	Far-end only		**Double-talk**		**# Param.**	**Mem. (Bytes)**	**RTF**
	ERLE	**AECMOS**	**PESQ**	**AECMOS**			
U-Net	39.39	4.62	2.56	4.04	0.14 M	0.5 M	0.03
Masking	**44.54**	4.67	2.73	4.55	1.86 M	7.0 M	0.32
Proposed	44.32	**4.68**	**2.94**	**4.71**	5.1 M	21.3 M	0.04

Results show that the proposed and Complex-Masking systems achieve on-par performance during far-end-only periods. Complex-Masking achieves negligibly better ERLE, and the proposed system achieves negligibly better AECMOS-echo. Regression-U-Net's performance is inferior to the other two systems - most notably, its ERLE is 4.93 dB less than that of the proposed system. Regression-U-Net's performance is also inferior to the other systems during double-talk periods. This performance gap may be due to the model's low complexity; it has only 0.14 M parameters, which is 1.72 M fewer than Complex-Masking. Therefore, it may be hard for the model to learn the input-output relations in such extreme conditions properly. Contrary to far-end-only periods, during double-talk, the proposed system's performance is notably superior to that of Complex-Masking. The proposed system's PESQ is higher by more than 0.2, and AECMOS is higher by 0.16 dB. Although the proposed system's number of parameters is about three times greater than that of Complex-Masking, its RTF is significantly lower. Thus, when inference time is a more critical constraint than memory consumption, the proposed system is favorable over Complex-Masking. It is worth noting how the proposed system's RTF is only slightly larger than Regression-U-Net's RTF, despite having significantly more parameters and higher memory consumption. It is due to the difference in the systems' input sizes; the proposed model was trained on 2 seconds-long segments while Regression-U-Net was trained on 0.3 seconds-long segments. Although the proposed system's architecture allows for variable-size input, it provides the best performance for 2 seconds-long inputs. Thus, in cases where low memory consumption and short algorithmic delay are high priorities while performance

is not, Regression-U-Net might be favorable. We also note Complex-Masking's high RTF despite the relatively small parameter number. This is due to the complex operations, which are more time-consuming.

Next, we study the different systems' performance in different SERs. We focus on far-end only ERLE and double-talk PESQ. Figure 4.5 (a) shows the ERLE difference between the systems' output signal $\tilde{d}(n)$ and the error signal $e(n)$. Similarly, Figure 4.5 (b) shows the PESQ difference.

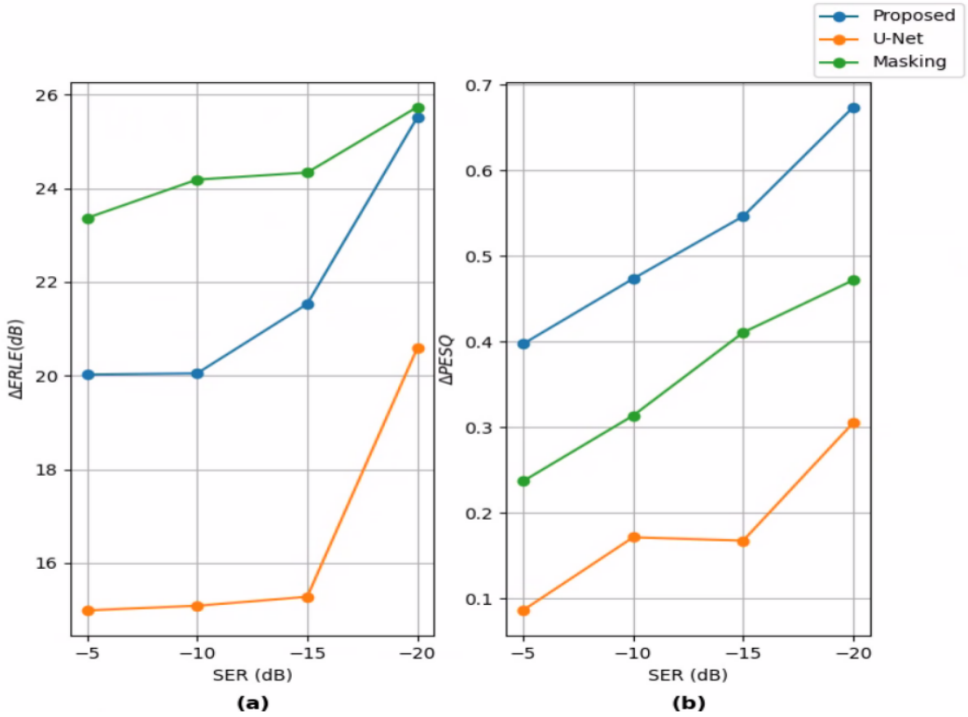


Figure 4.5: Systems' performance in different SERs. **(a)** Echo return loss enhancement (ERLE) difference between the systems' outputs and the error signal. **(b)** Perceptual evaluation of speech quality (PESQ) difference between the systems' outputs and the error signal.

The proposed system's graphs show its efficiency in lower SERs - it can be seen that both ΔERLE and ΔPESQ are increased when the SER is lowered, and the increase rate is also increasing (the graphs' slopes are higher in lower SERs). In other words, the proposed system is more effective in lower SERs. A similar trend can be seen in Regression-U-Net's performance, although the ΔPESQ increase rate is lower. Regarding Complex-Masking, which is more comparable to the proposed system, it can be seen that although its ERLE is consistently higher than the proposed system's ERLE, the rate at which ΔERLE increases is lower. At -20 dB SER, the gap between the two graphs is negligible. The increase rate of ΔPESQ is lower at lower SERs, while for the proposed system, it grows larger, i.e., the proposed system is more effective at lower SERs than Complex-Masking.

Finally, we compare the performance of the proposed masking architecture (AEC+M,

without the DTD) with the performance of the masking architecture proposed in [38] (Masking-inpainting). Table 4.7 shows the different performance measures, the number of parameters, the memory consumption, and the RTF of the models.

Table 4.7: Comparison of the proposed masking architecture without the DTD (AEC+M) and the masking architecture in Masking-inpainting.

	Far-end only		**Double-talk**		**#**	**Mem.**	**RTF**
	ERLE	**AECMOS**	**PESQ**	**AECMOS**	**Param.**	**(Bytes)**	
Masking-inpainting	40.21	**4.67**	2.72	**4.68**	2.56 M	9.76 M	0.031
AEC+M	**40.39**	**4.67**	**2.74**	4.66	1.01 M	3.85 M	0.007

Results show that the performance measures of the two models are on-par with negligible differences. On the contrary, the proposed model is preferable to Masking-inpainting's model concerning memory and running-time performance. Masking-inpainting's parameter number and memory consumption are roughly 2.5 times that of AEC+M, and its RTF is an order of magnitude greater than AEC+M's RTF. Hence the choice of the proposed masking architecture over the one proposed in [38].

4.5 Summary

We have presented a two-stage deep-learning residual echo suppression and double-talk detection system focused on the low SER scenario. The first stage combines the DTD with a spectrogram masking model based on the U-Net architecture. We conducted experiments with different configurations (based on previous studies) of the DTD with the masking model. The results show that the proposed configuration outperforms all other configurations. To the best of our knowledge, this is the first study of different ways to combine a DTD with a residual echo suppression model and the first study to report improved results due to the DTD. The second stage performs spectrogram refinement. The architecture is based on convolution blocks consisting of residual connections. The model is optimized to maximize the desired speech quality by minimizing the PMSQE loss function, which approximates PESQ. We performed an ablation study which shows the contribution of each stage of the system. Furthermore, we conducted experiments at different levels of SER. We showed that the proposed algorithm achieves the best performance gain in the low SER setting, approving its effectiveness in this challenging scenario. Lastly, we compared the proposed system to several other systems. The proposed system outperforms all others in near-end speech quality during double-talk periods, as measured by PESQ and AECMOS. During far-end-only periods, the system's performance is on par with one of the compared systems and outperforms the other system.

Chapter 5

Conclusions

5.1 Summary

In this research, we have focused on the field of acoustic echo cancellation and residual echo suppression. Traditionally, AECs and residual echo suppressors are based on adaptive filters. In recent years, with the rapid improvement in deep-learning technology, DNNs have become a popular choice for acoustic echo cancellation or residual echo suppression based on the outputs of a linear AEC. While exhibiting high performance, DNN-based residual echo suppression studies mainly focus on the RES model. This research showed that the proper choice of the preceding linear AEC is crucial for the RES's performance. We proposed to use an AEC based on the NSLMS algorithm rather than the common NLMS and showed that it results in increased performance, especially with the proposed RES model. Since the most significant challenge for AECs and residual echo suppressors is during double-talk periods, it is natural to integrate a DTD into the system. While some previous studies utilized DTDs in their work, none focused on their proper integration or effect on results. In this research, we proposed a novel DTD integration and showed that it improves performance, while other integrations, based on previous studies, do not improve results. Furthermore, none of the previous studies focus on the challenging scenario of extremely-low SER, an example of which is the common real-life situation of a conversation over a mobile phone when the loudspeaker volume is high. We proposed a deep-learning-based RES comprising two stages - double-talk detection and spectrogram masking, and spectrogram refinement. The proposed system outperforms competing systems while exhibiting significant performance gain, particularly in the low-SER setting.

In Chapter 3, we proposed a complex-valued deep-learning RES model. The preceding linear AEC is based on the NSLMS algorithm. Commonly, linear AECs are based on the NLMS algorithm. We showed that the NSLMS outperforms the NLMS as a baseline AEC and when combined with a deep-learning RES. Previous studies in other fields show that the residual signal in the output of the NSLMS is more akin to noise than speech. Therefore, in addition to the proposed RES, we utilized an off-the-

shelf, pre-trained speech denoiser trained on hundreds of hours of speech with varying noises and conditions to perform residual echo suppression. While the proposed RES outperforms the speech denoiser, the performance gap between the NLMS and NSLMS is greater for the speech denoiser. This affirms that, indeed, NSLMS produces a residual echo that is less structured than the residual echo produced by the NLMS. We concluded that the NSLMS is a better choice for residual echo suppression than the commonly-used NLMS. Furthermore, an off-the-shelf pre-trained speech denoiser can be employed for the task of residual echo suppression. In this case, the proper choice of the preceding linear AEC is even more crucial, and the NSLMS, which produces a residual echo that is more akin to noise than speech, is preferable over NLMS.

Chapter 4 proposed a two-stage deep-learning RES designed explicitly for the low-SER scenario. The first stage consists of spectrogram masking and double-talk detection. Previous studies that combined DTD in their residual echo suppression system did not study its effect on performance. In addition to the proposed DTD integration, we studied other integrations based on previous studies. We showed that while all other integrations bring little or no improvement to performance, the proposed integration does improve results. The second stage of the system is spectrogram refinement. Although a significant portion of the residual echo is eliminated in the first stage, the near-end speech's quality is degraded, especially during double-talk periods. Therefore, this stage is focused on improving speech quality. This is done by employing a network architecture that has shown good speech synthesis performance in previous studies and by minimizing a PESQ-related loss function. While improving performance in all measures compared to the first stage, the proposed system also outperforms compared systems. Specifically, the proposed system exhibits the highest performance gain in lower SERs. We concluded that the proposed system is effective in the challenging and little-explored scenario of low SER.

5.2 Future Research

In this thesis, we have proposed several acoustic echo cancellation systems, focusing on the deep-learning residual echo suppressors while emphasizing the importance of the preceding linear AEC. While obtaining high performance compared to existing systems and providing novel insights into little-explored aspects of this field, some questions remain that can provide a basis for future research. These include:

1. In Chapter 3, we showed that the NSLMS is superior to the NLMS, especially when combined with a pre-trained speech denoiser utilized as a RES. However, there are many more types of linear AECs, some of which may be better suited for deep-learning residual echo suppressors. Furthermore, it may be that some other off-the-shelf deep-learning model rather than a speech denoiser is available to the user. Thus, studying other linear AEC algorithms for deep-learning-based

residual echo suppression may be worthwhile, especially when little training data is available and an off-the-shelf solution is desirable.

2. Our research focused on the single-channel scenario where a single microphone captures the related signals. In many real-life applications, numerous microphones are available. The information provided by the different microphones can be utilized to improve performance with multi-channel methods such as beamforming. While several previous works study multi-channel residual echo suppression, none focus on the aspects studied in this thesis - proper choice of linear AEC, proper integration of DTDs, and low SER.

3. While all systems proposed in this thesis were designed with real-time performance considerations, no real-time evaluation was performed. While the analysis windows in the different systems' components comply with real-time standards used in various acoustic echo cancellation challenges, the systems are not causal. Furthermore, the system proposed in Chapter 4 comprises a relatively large parameter number, which may render it unusable in small portable devices. Thus, it may be essential to design systems that provide solutions to the challenges addressed in this thesis while also being utterly compatible with real-time and memory consumption requirements.

Bibliography

[1] M. Sondhi, D. Morgan, and J. Hall, "Stereophonic Acoustic Echo Cancellation-an Overview of the Fundamental Problem," *IEEE Signal Process. Lett.*, vol. 2, no. 8, pp. 148–151, 1995.

[2] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. Heidelberg: Springer Berlin, 1st ed., 2001.

[3] O. Macchi, *Adaptive Processing: the Least Mean Squares Approach*. USA: John Wiley and Sons, Inc., 1995.

[4] N. Bershad, "Analysis of the Normalized LMS Algorithm with Gaussian Inputs," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 793–806, 1986.

[5] B. Farhang-Boroujeny, *Adaptive Filters: Theory and Applications*. USA: John Wiley and Sons, Inc., 1998.

[6] N. Freire and S. Douglas, "Adaptive Cancellation of Geomagnetic Background Noise Using a Sign-error Normalized LMS Algorithm," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 3, pp. 523–526, 1993.

[7] N. Pathak, I. Panahi, P. Devineni, and R. Briggs, "Real Time Speech Enhancement for the Noisy MRI Environment," in *Proc. Annu. Int. Conf. IEEE Engineering in Medicine and Biology Soc.*, pp. 6950–6953, 2009.

[8] H. Zhang and D. Wang, "Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios," in *Interspeech*, 2018.

[9] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," *Neural Comput.*, vol. 9, pp. 1735–80, 1997.

[10] Y. Wang, A. Narayanan, and D. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[11] J.-H. Kim and J.-H. Chang, "Attention Wave-U-Net for Acoustic Echo Cancellation.," in *Interspeech*, pp. 3969–3973, 2020.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Med. Image Comput. Comput. Assist. Interv.* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[13] R. Giri, U. Isik, and A. Krishnaswamy, "Attention Wave-U-Net for Speech Enhancement," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 249–253, 2019.

[14] N. L. Westhausen and B. T. Meyer, "Acoustic Echo Cancellation with the Dual-Signal Transformation LSTM Network," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 7138–7142, 2021.

[15] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," *arXiv e-prints*, 2020.

[16] A. Guerin, G. Faucon, and R. Le Bouquin-Jeannes, "Nonlinear Acoustic Echo Cancellation Based on Volterra Filters," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 672–683, 2003.

[17] S. Malik and G. Enzner, "State-Space Frequency-Domain Adaptive Filtering for Nonlinear Acoustic Echo Cancellation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 2065–2079, 2012.

[18] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Multiple-Input Neural Network-Based Residual Echo Suppression," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 231–235, 2018.

[19] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and Recognition-boosted Speech Separation Using Deep Recurrent Neural Networks," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 708–712, 2015.

[20] L. Pfeifenberger and F. Pernkopf, "Nonlinear Residual Echo Suppression Using a Recurrent Neural Network," in *Interspeech*, pp. 3950–3954, ISCA, 2020.

[21] H. Chen, T. Xiang, K. Chen, and J. Lu, "Nonlinear Residual Echo Suppression Based on Multi-stream Conv-TasNet," *arXiv e-prints*, 2020.

[22] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[23] A. Fazel, M. El-Khamy, and J. Lee, "CAD-AEC: Context-Aware Deep Acoustic Echo Cancellation," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 6919–6923, 2020.

[24] M. M. Halimeh, T. Haubner, A. Briegleb, A. Schmidt, and W. Kellermann, "Combining Adaptive Filtering and Complex-valued Deep Postfiltering for Acoustic Echo Cancellation," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 121–125, 2021.

[25] A. Ivry, I. Cohen, and B. Berdugo, "Deep Residual Echo Suppression with A Tunable Tradeoff Between Signal Distortion and Echo Suppression," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 126–130, 2021.

[26] J. Franzen and T. Fingscheidt, "Deep Residual Echo Suppression and Noise Reduction: A Multi-Input FCRN Approach in a Hybrid Speech Enhancement System," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 666–670, 2022.

[27] A. Defossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," *preprint arXiv:2006.12847*, 2020.

[28] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," *preprint arXiv:2008.00264*, 2020.

[29] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech Enhancement Using Self-Adaptation and Multi-Head Self-Attention," in *Proc. ICASSP. IEEE*, pp. 181–185, 2020.

[30] H. Buchner, J. Benesty, T. Gansler, and W. Kellermann, "Robust Extended Multidelay Filter and Double-talk Detector for Acoustic Echo Cancellation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 5, pp. 1633–1644, 2006.

[31] M. Hamidia and A. Amrouche, "A New Robust Double-talk Detector Based on the Stockwell Transform for Acoustic Echo Cancellation," *Digit. Signal Process.*, vol. 60, pp. 99–112, 2017.

[32] H. Zhang, K. Tan, and D. Wang, "Deep Learning for Joint Acoustic Echo and Noise Cancellation with Nonlinear Distortions," in *Interspeech*, pp. 4255–4259, ISCA, 2019.

[33] X. Zhou and Y. Leng, "Residual Acoustic Echo Suppression Based on Efficient Multi-task Convolutional Neural Network," *arXiv e-prints*, 2020.

[34] L. Ma, H. Huang, P. Zhao, and T. Su, "Acoustic Echo Cancellation by Combining Adaptive Digital Filter and Recurrent Neural Network," *arXiv e-prints*, 2020.

[35] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv e-prints*, 2014.

[36] L. Ma, S. Yang, Y. Gong, X. Wang, and Z. Wu, "EchoFilter: End-to-End Neural Network for Acoustic Echo Cancellation," *arXiv e-prints*, 2021.

[37] S. Zhang, Z. Wang, J. Sun, Y. Fu, B. Tian, Q. Fu, and L. Xie, "Multi-Task Deep Residual Echo Suppression with Echo-Aware Loss," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 9127–9131, 2022.

[38] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, "Masking and Inpainting: A Two-Stage Speech Enhancement Approach for Low SNR and Non-Stationary Noise," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 6959–6963, 2020.

[39] R. Cutler, B. Naderi, M. Loide, S. Sootla, and A. Saabas, "Crowdsourcing Approach for Subjective Evaluation of Echo Impairment," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 406–410, 2021.

[40] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "AECMOS: A Speech Quality Assessment Metric for Echo Impairment," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 901–905, 2022.

[41] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 2, pp. 749–752, 2001.

[42] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in *Proc. ICASSP. IEEE*, pp. 6493–6497, 2021.

[43] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. USA: Prentice Hall PTR, 1983.

[44] S. Koike, "Analysis of Adaptive Filters Using Normalized Signed Regressor LMS Algorithm," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2710–2723, 1999.

[45] C. Valentini-Botinhao, "Noisy Speech Database for Training Speech Enhancement Algorithms and TTS Models." University of Edinburgh, School of Informatics, Centre for Speech Technology Research (CSTR), 2017, doi: https://doi.org/10.7488/ds/2117.

[46] C. K. A. Reddy, E. Beyrami, H. Dube, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework," *preprint arXiv:2001.08662*, 2020.

[47] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "ICASSP 2021 Acoustic Echo Cancellation Challenge: Datasets, Testing Framework, and Results," in *Proc. ICASSP. IEEE*, pp. 151–155, 2021.

[48] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM. NIST speech disc 1-1.1." Tech. Rep. LDC93S1, Nat. Inst. Standards Technol., Gaithersburg, MD, USA, 1993.

[49] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 5206–5210, 2015.

[50] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *preprint arXiv:1412.6980*, 2014.

[51] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," *arXiv e-prints*, 2016.

[52] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical Evaluation of Rectified Activations in Convolutional Network," *arXiv e-prints*, 2015.

[53] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," *arXiv e-prints*, 2016.

[54] J. M. Martin-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A Deep Learning Loss Function Based on the Perceptual Evaluation of the Speech Quality," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1680–1684, 2018.

אות-הד נמוך במיוחד. השלב הראשון כולל מודל למיסוך של ספקטרוגרמת דיבור אשר משלב זיהוי דיבור-כפול. השלב השני כולל מודל מיטוב ספקטרוגרמה. מודל זה מאומן למיטוב איכות הדיבור. האימון נעשה על ידי מזעור של פונקציית הפסד שקשורה למדד נפוץ שמשמש להערכת איכות אותות דיבור. השילוב המוצע של מזהה דיבור-כפול עם מודל המיסוך מביא לביצועים טובים יותר בהשוואה לשילובים אחרים שנבחנו, המבוססים על מחקרים קודמים. בנוסף, אנו מבצעים מחקר השוואתי על מנת להראות את החשיבות של כל אחד מהשלבים והמרכיבים של המערכת המוצעת, ושל השילוב הנכון ביניהם. כמו כן אנו בוחנים את ביצועי המערכת המוצעת ביחסי אות-הד שונים, ומראים שהיא יעילה במיוחד בתנאים מאתגרים של יחסי אות-הד נמוכים מאוד. לבסוף, אנו מראים כי הביצועים של המערכת המוצעת טובים יותר מהביצועים של מערכות אחרות, מבוססות מחקרים קודמים, במספר מדדים שקשורים לביטול הד שיורי. אנו מסיקים כי המערכת המוצעת מתאימה היטב לבעיה של ביטול הד שיורי ביחס אות-הד נמוך.

תקציר

בחיבור זה אנו חוקרים את התחומים של ביטול הד אקוסטי וביטול הד שיורי. הד אקוסטי הינו בעיה נפוצה במערכות תקשורת דו-כיווניות. הד אקוסטי נוצר כאשר האות המופק על ידי הרמקול נקלט על ידי המיקרופון, יחד עם האות הרצוי. צימוד זה בין האות הרצוי להד עלול לגרום לירידה באיכות השיחה, מה שמהווה בעיה בסיטואציות יום-יומיות רבות, כגון פגישה אשר מתרחשת בחדר ישיבות בו הדיבור של משתתפים מקוונים מושמע באמצעות רמקול. מחקר רב נעשה לצורך פתרון בעיית ההד האקוסטי. בשנים האחרונות, מבטלי הד אקוסטי הגיעו לביצועים מרשימים הודות לטכנולוגיית למידה עמוקה. למרות זאת, מספר היבטים של התחום לא נחקרו לעומק במחקרים קודמים. חיבור זה שואף לסגור את הפער הזה על ידי מחקר של שלושה היבטים שונים: בחירה נכונה של מבטל הד אקוסטי לינארי במערכות ביטול הד שיורי מבוססות למידה עמוקה, השילוב הנכון של מזהה דיבור-כפול עם מערכת ביטול הד שיורי מבוססת למידה עמוקה, וביטול הד שיורי במצבים של יחס אות-הד נמוך במיוחד.

תחילה, אנו מציגים מערכת ביטול הד אשר משלבת מבטל הד אקוסטי לינארי עם רשת קונבולוציה מרוכבת סדרתית עמוקה אשר מבצעת ביטול הד שיורי. מקדמי המסנן של מבטל ההד האקוסטי הלינארי נקבעים בעזרת אלגוריתם סימן-שגיאה-ריבועים-פחותים מנורמל (אלגוריתם סימן-שגיאה). אנו משווים את האלגוריתם הזה עם אלגוריתם ריבועים-פחותים מנורמל (אלגוריתם ריבועים-פחותים) וחוקרים את השילוב של כל אחד מהם עם מודל הלמידה העמוקה לביטול ההד השיורי. כמו כן, אנו חוקרים את השימוש של מודל למידה עמוקה שאומן מראש לצורך סינון רעשים מדיבור כחלופה למודל ביטול הד שיורי ייעודי. התוצאות מראות שהביצועים של אלגוריתם סימן-שגיאה עדיפים על פני הביצועים של אלגוריתם ריבועים-פחותים בכל המצבים השונים – כאשר הם עומדים בפני עצמם, וכאשר משתמשים בפלט שלהם כקלט למערכות ביטול ההד השיורי השונות. הביצועים של מודל ביטול ההד השיורי המוצע עדיפים על הביצועים של מודל סינון הרעשים. זאת, למרות שמודל סינון הרעשים גדול ומורכב יותר, ובנוסף הוא אומן על שעות רבות של אותות דיבור שונים במצבים שונים ומגוונים. בנוסף, אנו מראים שהביצועים של מודל סינון הרעשים טובים יותר כאשר הקלט שלו הוא הפלט של אלגוריתם סימן-שגיאה מאשר הפלט של אלגוריתם ריבועים-פחותים. ההבדלים בין ביצועי המודל עם שני האלגוריתמים הללו הם הגדולים ביותר בהשוואה לשאר המערכות, מה שמהווה אינדיקציה לכך שההד השיורי במוצא של אלגוריתם סימן-שגיאה דומה יותר לרעש מאשר לדיבור.

בעיית ביטול ההד האקוסטי מאתגרת במיוחד במצבים שבהם יחס האות-הד הוא נמוך במיוחד, למשל שיחה בטלפון סלולרי כאשר הרמקול מופעל בעוצמה גבוהה, כך שהאדם איתו משוחחים שומע את ההד של הדיבור שלו עצמו יחד עם הדיבור של האדם שמשתמש בטלפון הסלולרי. בחיבור זה, אנו מציעים מודל למידה עמוקה בעל שני שלבים לביטול הד שיורי אשר מתמקד במצבים של יחס

המחקר בוצע בהנחייתם של פרופסור ישראל כהן וד"ר ברוך ברדוגו בפקולטה להנדסת חשמל ומחשבים.

התוצאות של פרק 4 של חיבור זה פורסמו כמאמר מאת המחבר ושותפיו למחקר בכתב-עת במהלך תקופת מחקר המגיסטר של המחבר, אשר גרסתו העדכנית ביותר הינה:

> Eran Shachar, Israel Cohen, and Baruch Berdugo. Double-talk detection-aided residual echo suppression via spectrogram masking and refinement. *Acoustics*, 4(3):637–655, 2022.

מאמר נוסף, שכותרתו "ביטול הד אקוסטי עם אלגוריתם סימן-שגיאה-מינימום-ריבועים וביטול הד שיורי בעזרת למידה עמוקה" התקבל לפירסום בכתב-העת Algorithms MDPI אך עדיין לא פורסם בזמן הגשת החיבור.

מחבר חיבור זה מצהיר כי המחקר, כולל איסוף הנתונים, עיבודם והצגתם, התייחסות והשוואה למחקרים קודמים וכו', נעשה כולו בצורה ישרה, כמצופה ממחקר מדעי המבוצע לפי אמות המידה האתיות של העולם האקדמי. כמו כן, הדיווח על המחקר ותוצאותיו בחיבור זה נעשה בצורה ישרה ומלאה, לפי אותן אמות מידה.

תודות

ברצוני להביע את הכרת התודה והערכתי הרבה למנחי המחקר שלי, פרופסור ישראל כהן ודר' ברוך ברדוגו. המחקר הזה לא היה מתאפשר ללא התמיכה וההכוונה שלהם. במהלך המסע הזה, הם לימדו אותי איך להיות חוקר טוב יותר, איפשרו לי ללמוד כישורים משמעותיים, ועזרו לי להתגבר על הקשיים הרבים שבדרך, ועל זה אני מוקיר תודה.

ברצוני גם להודות לבת זוגתי בתאל ולמשפחה שלי, שליוו אותי ותמכו בי ברגעים הקשים וחלקו איתי את הרגעים הטובים של התהליך הזה. ההישג הזה לא היה מתאפשר בלעדיהם.

ביטול הד אקוסטי בשילוב ביטול הד שיורי מבוסס למידה עמוקה

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים בהנדסת חשמל

ערן שחר

הוגש לסנט הטכניון — מכון טכנולוגי לישראל
אדר התשפ"ג חיפה פברואר 2023

ביטול הד אקוסטי בשילוב ביטול הד שיורי מבוסס למידה עמוקה

ערן שחר