Beamforming-Based Multichannel Acoustic Echo Cancellation

Yuval Konforti

Beamforming-Based Multichannel Acoustic Echo Cancellation

Research Thesis

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

Yuval Konforti

Submitted to the Senate
of the Technion — Israel Institute of Technology
Kislev 5784 Haifa December 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 this thesis have been published by the author and research collaborators in a journal and conference papers during the course of the author's master's research period.

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 am extremely grateful to my research supervisors, Prof. Israel Cohen and Dr. Baruch Berdugo. This work is a result of their support and patience. They taught me how to approach research and think critically. They have been there through thick and thin, and for that, I am deeply indebted to them.

I would also like to thank my family, most notably my partner Galit. This work would not have been possible without their never-ending support. You make me who I am. Lastly, I would like to thank our late cat, Jasmine, for enabling the work on this thesis in good spirit, and for just being there.

This research was supported by the Pazy Research Foundation and the Israel Science Foundation (grant no. 1449/23).

List of Publications

The results of this thesis have been published by the author and research collaborators in the following publications:

1. Y. Konforti, I. Cohen, and B. Berdugo, "Array geometry optimization for region-of-interest broadband beamforming", in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2022.

2. Y. Konforti, I. Cohen, and B. Berdugo, "Multichannel acoustic echo cancellation with beamforming in dynamic environments", submitted to *IEEE Open Journal of Signal Processing*, 2023.

Contents

List of Figures

Abstract

In this thesis, we study the problem of beamformer design for acoustic echo cancellation (AEC) and region-of-interest (ROI) spatial filtering. Beamforming is a standard method to amplify signals from specific directions while attenuating signals from other directions. Beamformers are widely used in speech enhancement, direction-of-arrival (DOA) estimation, source localization, source separation, AEC, and others. Beamformers utilize an array of sensors, and according to the time difference of arrival (TDOA) between the sensors, can filter signals concerning their propagation direction.

This thesis considers two problems. The first considers multichannel AEC in dynamic environments. The objective is to eliminate the acoustic coupling between a loudspeaker and a microphone array in a dynamic environment where an additional speech signal may be present. The proposed method can operate with occurring double-talk, where we do not assume a static environment when transitioning to double-talk and do not rely on double-talk detection. The second problem considers array geometry optimization in a nonuniform linear structure for high directivity. The objective is to find the optimal locations of the microphones in the array to achieve a high directivity index. We do this by finding the array geometry and broadband beamformer coefficients, where the desired signal is in an ROI.

First, we introduce a multichannel echo canceller implemented by a microphone array beamformer that can adapt to a changing environment where the locations of both the far-end and near-end sources change during double-talk, with no double-talk detector. This is done by utilizing multiple recent frames in the short-time Fourier transform (STFT) domain. We show how can the acoustic paths be accurately estimated given the recent time frames of the far-end and microphone signals. Also, our beamformer aims to reduce background noise. Simulations are conducted in a reverberant room with nonlinear loudspeaker distortion and realistic low signal-to-echo ratio (SER) resembling a speakerphone. The experiments demonstrate the advantages of the proposed approach compared to normalized least-mean-squares (NLMS) based approaches.

Then, we present an efficient algorithm to find the optimal placements of microphones in a nonuniform linear array for broadband high-directivity beamforming. Optimization of a microphone array geometry has an important impact on the beamforming performance. Though the environment settings may change in time, microphone loca-

tions are typically preserved. Thus, the selected locations are of significant importance. Moreover, if the source of interest emits a broadband signal from a location that varies in time, finding the optimal geometry becomes challenging. The proposed method maintains high white noise gain (WNG) for sufficient robustness and considers several look directions in the ROI for a moving source. Our design achieves higher directivity toward any look direction in the ROI than standard designs.

Abbreviations

AEC	:	Acoustic Echo Cancellation
AES	:	Acoustic Echo Suppression
AOA	:	Angle of Arrival
CMTF	:	Cross-Multiplicative Transfer Function
DF	:	Directivity Factor
DI	:	Distortion Index
DMA	:	Differential Microphone Array
DNN	:	Deep Neural Network
DOA	:	Direction-of-Arrival
ERLE	:	Echo-Return Loss Enhancement
ISTFT	:	Inverse Short-Time Fourier Transform
LCMV	:	Linear-Constraint-Minimum-Variance
LMS	:	Least-Mean-Squares
MTF	:	Multiplicative Transfer Function
NLMS	:	Normalized Least-Mean-Squares
PNLMS	:	Proportionate Normalized Least-Mean-Squares
RIR	:	Room Impulse Response
ROI	:	Region-of-Interest
RTF	:	Relative Transfer Function
SER	:	Signal-to-Echo Ratio
SNR	:	Signal-to-Noise Ratio
STFT	:	Short-Time Fourier Transform
TDOA	:	Time Difference of Arrival
TF	:	Transfer Function
UCA	:	Uniform Circular Array
UCCA	:	Uniform Concentric Circular Array
ULA	:	Uniform Linear Array
VAD	:	Voice Activity Detector
WNG	:	White Noise Gain

Notations

A	:	Maximal aperture of the array.
$\mathcal{B}\left[\mathbf{h}\left(\mathbf{x},\omega,\tilde{\theta}\right),\theta\right]$	:	Beampattern.
c	:	Speed of sound.
$\mathcal{D}\left[\mathbf{h}\left(\mathbf{x},\omega,\theta\right)\right]$	:	DF.
$\mathcal{DI}_{[\omega_L,\omega_H]}\left[\mathbf{h}\left(\mathbf{x},\omega,\theta\right)\right]$	:	Directivity index.
$\mathbf{d}\left(k,n\right)$	:	Signal received by the array.
$\mathbf{d}\left(\mathbf{x},\omega,\theta\right)$	:	Steering vector toward the desired source.
$\mathbf{d}_{\text{tot}}\left(\omega,\theta\right)$	:	$\mathbf{d}\left(\mathbf{x},\omega,\theta\right)$ utilizing all potential sensor placements.
$D_m\left(k,n\right)$	:	STFT domain signal received by the m-th microphone.
d_c	:	Minimal distance between two adjacent microphones.
$d_m\left(t\right)$	:	Time domain signal received by the m-th microphone.
$E\left[\cdot\right]$	:	Expectation operator.
f	:	Temporal frequency.
f_H	:	High temporal frequency for broadband directivity.
f_L	:	Low temporal frequency for broadband directivity.
$\mathbf{g}\left(k\right)$	:	Steering vector toward the far-end source.
G	:	Number of restricted areas in the aperture.
$G_m\left(k,n\right)$	:	TF from the loudspeaker to the m-th microphone.
$g_m\left(t\right)$	:	RIR from the loudspeaker to the m-th microphone.
$\mathbf{h}\left(k,n\right)$	:	STFT domain beamforming filter.
$\mathbf{h}\left(\mathbf{x},\omega,\theta\right)$	:	Frequency domain beamforming filter.
$\mathbf{h}^*\left(\mathbf{x},\omega,\theta\right)$	:	Optimal frequency domain beamforming filter.
$\mathbf{h}_{\text{opt}}\left(k,n\right)$	:	Optimal STFT domain beamforming filter.
$\mathbf{h}_{\text{tot}}\left(\omega,\theta\right)$	:	Beamformer with all sensor placements.
$\mathbf{h}^*_{\text{tot}}\left(\omega,\theta\right)$	:	Optimal beamformer with all sensor placements.
$\mathbf{h}_\epsilon\left(\mathbf{x}^*,\omega,\theta\right)$	:	Robust superdirective beamformer.
$H_m\left(k,n\right)$	:	STFT domain beamforming coefficients.
$H_m\left(\mathbf{x},\omega,\theta\right)$	:	Frequency domain beamforming coefficients.
$H_{\text{tot},m}\left(\omega,\theta\right)$	:	Beamforming coefficients with all sensor placements.
$\mathbf{I}_M$	:	Identity matrix of dimensions $M \times M$.
$\mathbf{i}_N$	:	Column vector of length N consisting of ones.

L	:	Number of frames used.
M	:	Number of microphones in the array.
$\mathbf{q}(k)$	:	Steering vector toward the near-end source.
$Q_m(t)$	:	TF from the talker to the m-th microphone.
$q_m(t)$	:	RIR from the talker to the m-th microphone.
$\mathbf{s}$	:	Microphone position vector binary optimization variable.
$\mathbf{s}^*$	:	Optimal binary microphone position vector.
S_i	:	i-th element of $\mathbf{s}$.
$S(\omega)$	:	Frequency domain signal emitted by the desired source.
$\hat{S}(\omega)$	:	Estimate of $S(\omega)$.
$S(k,n)$	:	Near-end STFT domain signal.
$s(t)$	:	Near-end time domain signal.
$\mathbf{u}(k,n)$	:	STFT domain received near-end speech.
$U_f(k,n)$	:	STFT domain filtered near-end signal.
$U_m(k,n)$	:	STFT domain m-th microphone near-end speech .
$\hat{U}(k,n)$	:	STFT domain estimate of the desired signal.
$u_f(t)$	:	Time domain filtered near-end signal.
$u_m(t)$	:	Time domain m-th microphone near-end speech.
$\hat{u}(t)$	:	Time domain estimate of the desired signal.
$\mathbf{v}(\omega)$	:	Frequency domain noise received by the array.
$\mathbf{v}(k,n)$	:	STFT domain noise received by the array.
$V_m(k,n)$	:	STFT domain m-th microphone noise.
$V_{\mathrm{rn}}(k,n)$	:	STFT domain residual noise component.
$v_m(t)$	:	Time domain m-th microphone noise.
$\mathcal{W}[\mathbf{h}(\mathbf{x},\omega,\theta)]$	:	WNG.
$\mathbf{x}$	:	Microphone position vector.
$\mathbf{x}^*$	:	Optimal microphone position vector.
$X(k,n)$	:	STFT domain far-end signal.
$X_{\mathrm{NL}}(k,n)$	:	STFT domain loudspeaker output.
$x(t)$	:	Time domain far-end signal.
x_m	:	m-th microphone location.
$x_{\mathrm{NL}}(t)$	:	Time domain loudspeaker output.
$\mathbf{y}(\omega)$	:	Frequency domain received signal by the array.
$\mathbf{y}(k,n)$	:	STFT domain received far-end speech.
$Y_m(\omega)$	:	Frequency domain m-th microphone signal.
$Y_m(k,n)$	:	STFT domain m-th microphone far-end speech.
$Y_{\mathrm{re}}(k,n)$	:	STFT domain residual echo signal.
$y_m(t)$	:	Time domain m-th microphone far-end speech.
$y_{\mathrm{re}}(t)$	:	Time domain residual echo signal.
$\mathbf{\Gamma}(\mathbf{x},\omega)$	:	Pseudo-coherence matrix.
$\mathbf{\Gamma}_{\mathrm{tot}}(\omega)$	:	Pseudo-coherence matrix with all sensor placements.

Δx	:	Spacing between possible microphone locations.
$\Delta\omega$	:	Spacing between sampled angular frequencies.
$\Delta\theta$	:	Spacing between sampled AOAs.
δ	:	Minimal WNG.
ϵ	:	Tradeoff parameter between WNG and DF.
Θ	:	Range of angles marking the ROI.
θ	:	AOA.
θ_H	:	Largest AOA in the ROI.
θ_L	:	Lowest AOA in the ROI.
$\nu\left(t\right)$	:	DI.
$\xi\left(t\right)$	:	ERLE.
$\sigma_v^2\left(k\right)$	:	Residual noise component variance.
Ω	:	Angular frequency range.
ω	:	Angular frequency.
ω_H	:	High angular frequency for broadband directivity.
ω_L	:	Low angular frequency for broadband directivity.
$\mathbf{0}_i$	:	Column vector of length i consisting of zeros.
$\{\cdot\}^T$	:	Transpose operator.
$\{\cdot\}^H$	:	Transpose conjugate operator.
$\{\cdot\}^*$	:	Complex conjugate operator.
$\{\cdot\}^\dagger$	:	Pseudo-inverse operator.

Chapter 1

Introduction

1.1 Background and Motivation

Acoustic echo control is a substantial part of any hands-free teleconferencing system [1, 2], and sensor array beamforming is a widely-used method for spatial filtering [3–5]. Such systems have gained popularity in recent years. Hence, beamformer design, acoustic echo cancellation (AEC), and acoustic echo suppression (AES) emerge as topics of great importance.

The echo cancellation problem is eliminating the undesired echo from the acoustic coupling between a loudspeaker and a microphone. This echo component should be removed so that talkers, on each side of the conversation, will not hear themselves as feedback. Numerous AEC methods were proposed following the work of Sondhi [6]. Some methods also consider the effects of background noise received by the microphone, and loudspeaker nonlinear distortion induced by physical properties of the electrodynamic loudspeaker model [7].

Typically, echo cancellers are implemented by an adaptive filter operating on the reference loudspeaker signal that aims to portray the acoustic echo path. Then, the transmitted signal can be found by subtracting this echo component from the received signal by the microphone. The undesired echo (far-end) signal and the desired (near-end) signal may be active at the same time, making the problem a challenging one in the low signal-to-echo ratio (SER) case. Such periods are referred to as double-talk and can be found by double-talk detectors and voice activity detectors (VADs). The adaptive filter must be adapted during periods when there is no double-talk, with techniques such as least-mean-squares (LMS), normalized LMS (NLMS), proportionate NLMS (PNLMS), and others [5]. While advanced multimodal detectors are available [8], some AEC applications do not contain additional modalities or lack the required data for network training. Therefore, it is inevitable that during some undetected double-talk periods, the filter adapts inaccurately. Some works have also explored implementing the echo canceller with a deep neural network (DNN) [9–12].

One method to improve echo cancellation performance is to utilize a microphone

array. Adding more microphones enables the use of spatial filtering. Furthermore, the adaptive filter coefficients can be adapted using information obtained by all microphone signals. Such a multi-sensor spatial filter is called a beamformer [4]. Beamformers can be used to amplify speech from some directions while attenuating speech from other directions in noisy environments [5, 13–16]. As illustrated by Figure 1.1, the time difference of arrival (TDOA) of the impinging source signal between the sensors is dependent on the propagation direction. Beamformers exploit this physical property to create a direction-dependent gain. Strategies for combining echo cancellers and beamformers were presented by Kellermann [17], and later on, more methods were explored by [18–22]. Typically, the whole process is divided into two stages, echo canceling and beamforming, and careful attention should be paid to what stage comes first. If beamforming is used, one must also consider the location of the desired source for it to be preserved. In the case of a dynamic system, this requires dealing with moving sources, which can be challenging [23–25]. If significant reverberation takes place on the way to the reference microphone, a dereverberation stage might be necessary subsequent to echo cancellation [26, 27].

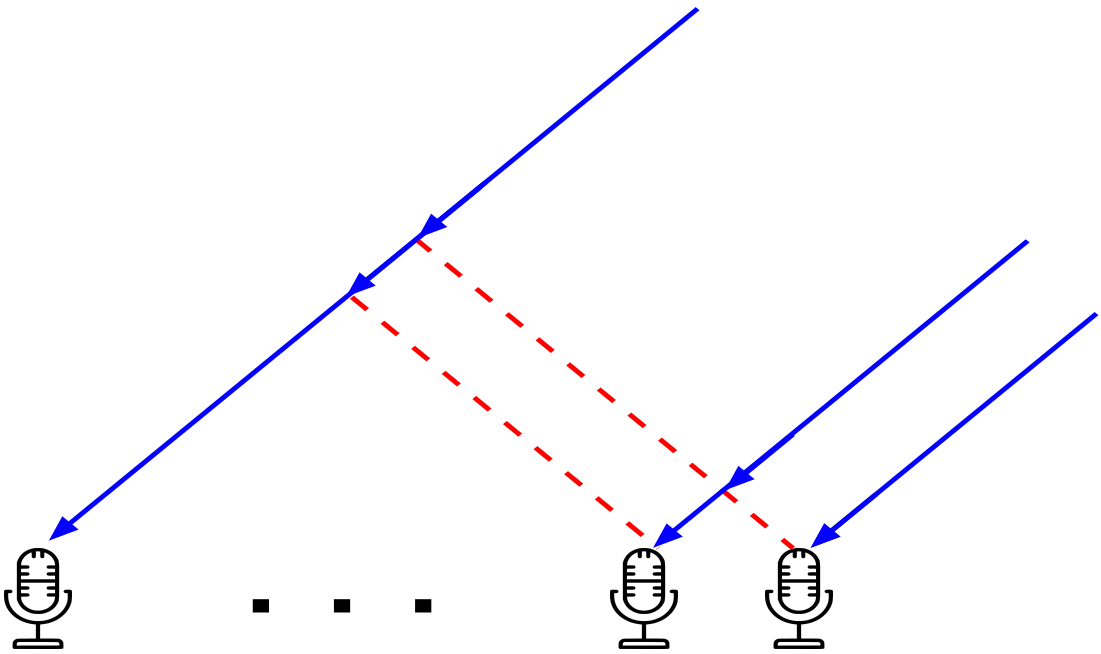

Figure 1.1: Illustration of an impinging source on an array of sensors. The TDOA of the signal is dependent on the source propagation direction, hence, it is dependent on the source location. The red dotted lines mark the time when the source hits the first two microphones.

Another method to better echo control is to utilize multiple frames in the short-time Fourier transform (STFT) domain. This way, an adaptive filter operates on multiple frames of a single microphone. The idea of utilizing multiple frames was introduced by Benesty and Huang [28, 29] and was later used in [30–32], in the context of noise reduction. Naturally, these works were extended to the problem of echo cancellation

in [12, 33–37]. The common property of these approaches is that a filter is adapted according to the inter-frame correlations. The difference lies in how the inter-frame correlations are estimated. In [34, 35], an initial guess of the desired signal is provided by finding an LMS solution, [33] use the statistics of the far-end signal, [36, 37] use an estimate obtained by NLMS, and in [12] a neural network is used. It should be noted that utilizing the inter-frame correlations can, at best, preserve only the expected component of the near-end signal, which is not necessarily the near-end signal itself. Furthermore, these methods assume that the far-end and near-end talkers are statistically uncorrelated, which is not valid in real scenarios.

The works in [38–40] consider the multi-microphone speech separation and noise reduction problems, which are inherently different from the AEC problem. In these scenarios, no available far-end signal produces an echo. The availability of the echo-producing loudspeaker signal is critical to diminishing the received echo component. Thus, for the problem of AEC, combining the two methods is beneficial.

Most importantly, typical echo cancellers rely on the assumption that during double-talk periods, the adapted filter from previous time segments can still provide a good estimate of the echo path. This assumption may be problematic in the case of a dynamic environment where the acoustic paths change during double-talk. In such environments, the experimental results show that performance is severely degraded once the acoustic path is changed for up to several seconds [20, 37]. A study has been conducted in [27] that can manage such a dynamic environment, but still requires a double-talk detector. In the following thesis, we propose a combined multi-sensor multi-frame approach for the problem of AEC in dynamic environments.

Two factors impact the performance of any beamformer design: the array geometry and the filter coefficients. Concerning the array geometry, typical microphone arrays use simple symmetric geometries such as uniform linear arrays (ULAs), uniform circular arrays (UCAs), and uniform concentric circular arrays (UCCAs). Recently, more efforts were made to find optimal geometries for several tasks [41–49]. Such methods optimize the sensor locations, usually with a genetic algorithm [41–46] or with a greedy-based approach [47–49]. These methods may converge to an undesired local optimum, and some works consider narrowband signals only [41–43]. There is a particular interest in finding geometries that enable high directivity to obtain a design that attenuates all directions that are not of interest. It has been shown that, when a high directivity factor (DF) is desired, a vanishingly small linear array is best compared to all geometries with vanishingly small separations [50]. However, when such a geometry is utilized white noise is severely amplified. In [51] practical methods were proposed to alleviate this effect. If the array is large, or the source of interest is close to the array, a near-field model for directivity should be used [52]. Interestingly, the average DF over all look directions is constant regardless of the geometry [53].

Several works have investigated region-based [54–62] and constant-beamwidth [63–65] beamformers directed toward a region of interest (ROI). These designs are practical

when several angles of arrival are considered. Such a scenario is encountered when the source is distributed, moving, or there is some uncertainty in the source direction. Inspired by these works, we revisit the problem of array geometry optimization for region-based beamforming. A similar study was conducted in [41] but did not consider broadband signals.

1.2 Main Contributions

The research in this thesis focuses on filling in the knowledge gaps discussed in the previous section. We list our main contributions:

- A novel multichannel AEC method is introduced, utilizing both multiple sensors and multiple frames in the STFT domain. Inter-sensor and inter-frame relations are both taken into account this way, yielding better echo cancellation compared to an NLMS-based approach. As opposed to existing multi-frame AEC methods, the far-end and near-end signals are not assumed to be statistically uncorrelated. Furthermore, we aim to preserve the near-end signal itself, rather than just the near-end expected signal component.

- The proposed AEC scheme does not include a double-talk detector and can operate in a dynamic environment where the acoustic paths change. The array, near-end source, and far-end source may all move during double-talk. This is demonstrated with simulations in a reverberant room with nonlinear loudspeaker distortion and a dynamic environment.

- A convex framework is presented, which enables us to find the best nonuniform linear geometry for ROI beamforming. This framework guarantees that the obtained solution is the global optimum to the problem, under reasonable assumptions.

- Array geometry optimization is done alongside beamformer coefficient optimization. While typically, the beamformer coefficients are obtained per given geometry, we find the best geometry considering all possible broadband beamforming coefficients towards several look directions in the ROI. The derived coefficients also adhere to a constraint on white noise gain (WNG), for sufficient noise robustness.

1.3 Research Overview

The first part of this research focuses on multichannel AEC systems in dynamic environments. We develop an AEC scheme for any arbitrary array geometry and focus on finding the beamformer coefficients rather than the optimal array geometry. We adopt the linear-constraint-minimum-variance (LCMV) beamforming technique to formulate a beamformer that theoretically eliminates the undesired echo component, preserves

the desired near-end component, and reduces background noise impact. The use of this beamformer is only justified once the acoustic paths are estimated accurately, to this end, we develop an acoustic path estimation mechanism. Using the multiplicative transfer function (MTF) approximation [66], mathematical expressions of the received signals by the array in recent time frames are developed. These expressions can be written as one matrix-form equation. Then, assuming that the environment is static during these recent frames, we show how can the acoustic paths be estimated utilizing the LMS solution of the structured matrix equation. With the proposed method, the acoustic paths can be estimated accurately even with background noise and nonlinear loudspeaker distortion present. A simulative study is conducted in a noisy, dynamic environment with nonlinear loudspeaker distortion during double-talk. In this study, we investigate the impact of the number of utilized sensors, as well as the number of utilized recent frames. Also, we show that the echo cancellation performance achieved by the proposed method is superior to the methods in [36, 37].

The second part of the research considers finding the optimal geometry of the array where the source is located in an ROI. We formulate a problem aimed at optimizing performance in the worst-case look direction in the ROI. The array geometry must remain static before exactly locating the source, hence, this principle ensures the best performance possible in the worst-case scenario where the source is located in the worst possible location in the ROI. The solution to this problem is the optimal array geometry. Since speech and audio signals are broadband, we extend the narrowband DF performance measure to a broadband directivity index performance measure. Our problem is formulated to find the geometry that maximizes the directivity index in the worst-case look direction, under a restriction on WNG. When considering a discrete set of frequencies, look directions, and candidate sensor locations, the formulated problem is shown to be a convex one. This means that a solution can be provided with any industrial mixed-integer convex optimization solver. Since the problem targets only the worst-case look direction, the resulting coefficients toward any other look directions in the ROI may not be optimal. This is circumvented by a post-processing scheme that finds the optimal beamformer coefficients for all other look directions as well. This scheme produces the robust superdirective beamformer, maximizing the DF for each frequency and each look direction while adhering to the WNG constraint. We do this with a bisection search on a directivity-robustness tradeoff parameter. Experiments show that the resulting geometry achieves higher directivity in all the ROI compared to standard uniform linear array (ULA) and differential microphone array (DMA) geometries.

The research presented in this thesis takes into account practical considerations in both aforementioned parts. For the AEC problem, we study the performance in a realistic low SER environment, as is typical in applications utilizing a speakerphone with significant acoustic coupling between the loudspeaker and microphone array. For the array optimization problem, minimal spacing between microphones is guaranteed so that a real scenario with non-zero microphone volume can use the proposed approach.

1.4 Organization

This thesis is organized as follows. The scientific background related to this work, as well as the problem formulation and performance measures, are presented in Chapter 2. Chapter 3 introduces the first original contribution of this research, where a multichannel AEC utilizing a beamformer is suggested for dynamic environments. A novel design is suggested, taking into account multiple microphones and multiple STFT frames. Then, in Chapter 4, the second original contribution of this research is introduced, where the microphone locations are optimized for broadband ROI beamforming. The optimization scheme finds the best geometry considering all possible coefficients. Finally, Chapter 5 summarizes the main findings of this research, concludes the thesis, and presents possible directions for future work.

Chapter 2

Preliminaries

This chapter provides the required scientific background to the problems tackled in this thesis and an overview of the mathematical models used. In Section 2.1, the basics of AEC are laid out. Section 2.2 provides the required background on array processing for ROI beamforming. Then, performance measures are defined in Section 2.3. Lastly, using the different performance measures, we formulate our problem in Section 2.4.

2.1 Acoustic Echo Cancellation

The AEC problem considers a scenario where two distinct speakers, located in different rooms, converse using a conferencing system. The room where the discussed echo canceller is being implemented is termed the 'near-end' room, and the other room, where the transmitted signal is sent to, is termed the 'far-end' room. The far-end room sends a speech signal $x(t)$, that is played by the loudspeaker in the near-end room. In addition, the person located in the near-end room also emits speech, this signal is termed the near-end signal and is marked by $s(t)$. Both $x(t)$ and $s(t)$ are captured by a microphone in the near-end room, as illustrated in Figure 2.1. Echo cancellers use the available received signal from the far-end room, and the received microphone signal in the near-end room, to cancel the echo component that originates from the acoustic coupling between the loudspeaker and microphone.

This setup may be over-simplistic in a real-life environment. Two important effects should be taken into account as well: nonlinear loudspeaker distortion and background noise. The first effect is due to the physical properties of the loudspeaker, the actual signal emitted is $x_{\mathrm{NL}}(t)$, which is different than $x(t)$. The term $x_{\mathrm{NL}}(t)$ is used since the distortion induced by the loudspeaker is nonlinear, i.e. it cannot be modeled by a linear system operating on $x(t)$. The second effect is due to thermal noise in the microphone itself. To alleviate the impact of these effects on performance, an array of microphones can be employed.

Consider a system with M microphones, as depicted in Figure 2.2. In addition to the far-end and near-end sources, background noise $v_m(t)$ assumed to be independent

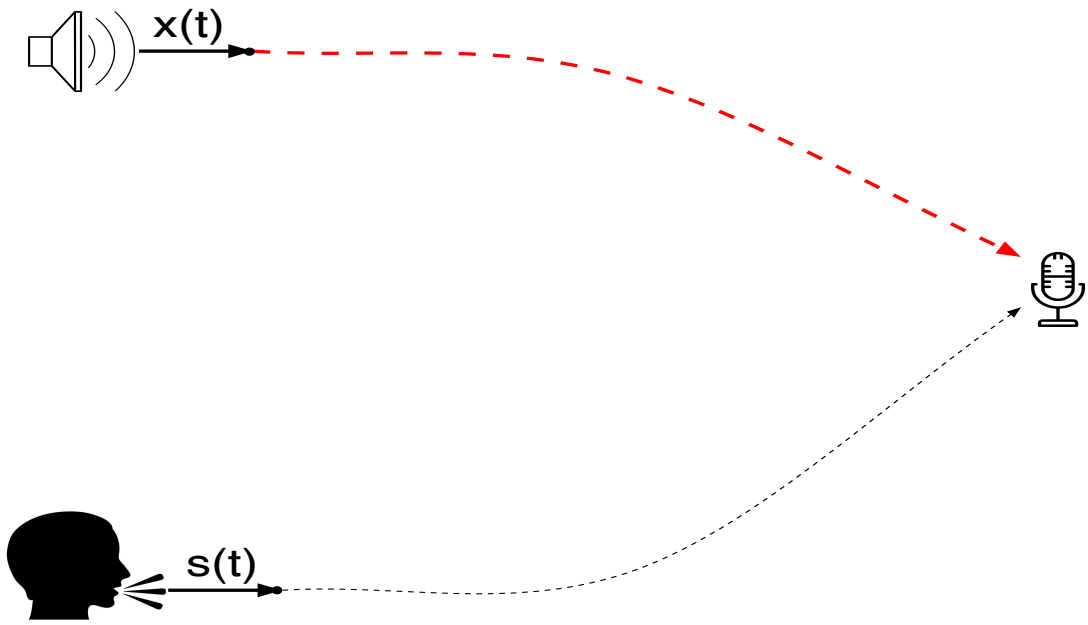

Figure 2.1: Illustration of acoustic echo cancellation. A microphone captures the far-end speech signal and the near-end speech signal. The far-end source is depicted by the loudspeaker and the near-end source is depicted by the speaker. The red dotted line marks the echo component received by the microphone, which the echo canceller should eliminate before transmitting the signal to the far-end room.

and uncorrelated between all microphones, is received. Using the signal model for acoustic echo, the m-th microphone signal $d_m(t)$ is given by

$$d_m(t) = g_m(t) * x_{\mathrm{NL}}(t) + q_m(t) * s(t) + v_m(t). \tag{2.1}$$

Here $g_m(t)$ is the impulse response from the loudspeaker to the m-th microphone, $q_m(t)$ is the impulse response from the talker to the m-th microphone, and $*$ represents the convolutional operator. This can also be written as

$$d_m(t) = y_m(t) + u_m(t) + v_m(t) \tag{2.2}$$

where

$$y_m(t) = g_m(t) * x_{\mathrm{NL}}(t) \tag{2.3}$$

is the received far-end speech by the m-th microphone, and

$$u_m(t) = q_m(t) * s(t) \tag{2.4}$$

is the received near-end speech by the m-th microphone. We arbitrarily define the reference microphone as the first, i.e., by $m = 1$.

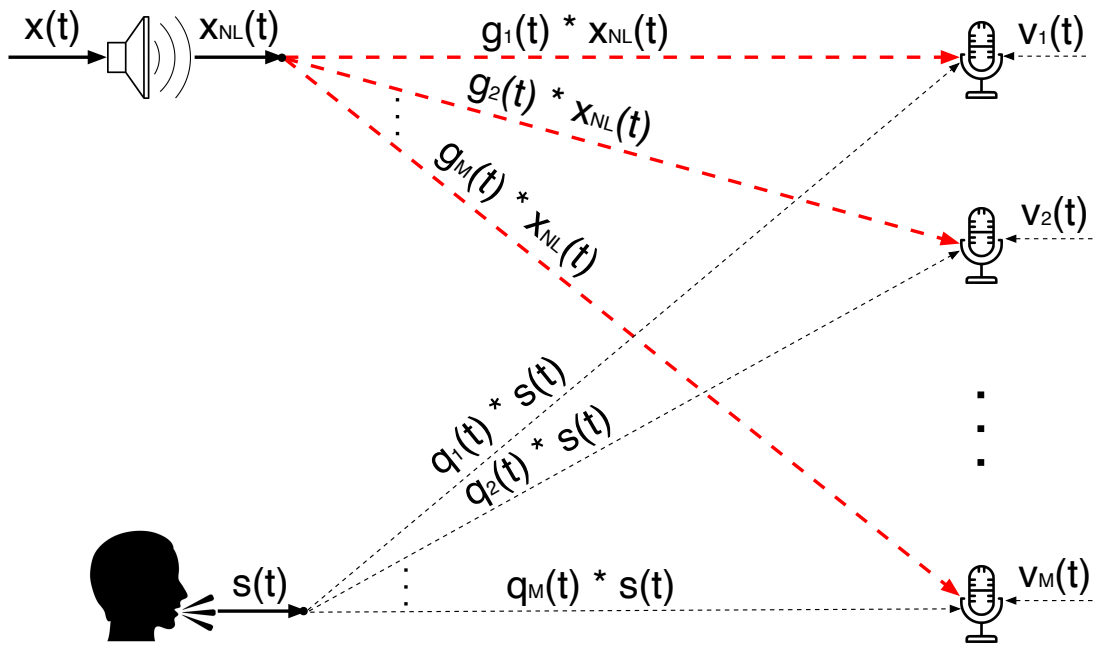

Figure 2.2: Illustration of multichannel acoustic echo cancellation. M microphones capture the far-end speech signal, the near-end speech signal, and background noise. The far-end source is depicted by the loudspeaker and the near-end source is depicted by the speaker. The red dotted lines mark the echo component received by the microphones, which the echo canceller should eliminate before transmitting the signal to the far-end room.

Applying the STFT on 2.2, we get

$$D_m(k,n) = Y_m(k,n) + U_m(k,n) + V_m(k,n) \tag{2.5}$$

where $D_m(k,n)$, $Y_m(k,n)$, $U_m(k,n)$, and $V_m(k,n)$ are the STFTs of $d_m(t)$, $y_m(t)$, $u_m(t)$, and $v_m(t)$, respectively, at frequency bin k and time frame n. We will also use the approximations

$$Y_m(k,n) \approx G_m(k) X_{\mathrm{NL}}(k,n) \tag{2.6}$$

$$U_m(k,n) \approx Q_m(k) S(k,n) \tag{2.7}$$

where $G_m(k)$ is the transfer function (TF) from the loudspeaker to the m-th microphone, $Q_m(k)$ is the TF from the talker to the m-th microphone, and $X_{\mathrm{NL}}(k,n)$ and $S(k,n)$ are the STFTs of $x_{\mathrm{NL}}(t)$ and $s(t)$, respectively. These approximations hold when the lengths of the filters $g_m(t)$ and $q_m(t)$ are significantly shorter than the STFT window length, i.e. the MTF approximation is used [66]. The TFs may include any element in the acoustic paths such as reverberation, attenuation, and time of arrival. Therefore, the far-field model is not assumed.

We define the array steering vectors with the relative TFs (RTFs):

$$\mathbf{g}(k) = \left[1, \frac{G_2(k)}{G_1(k)}, ..., \frac{G_M(k)}{G_1(k)}\right]^T \tag{2.8}$$

$$\mathbf{q}(k) = \left[1, \frac{Q_2(k)}{Q_1(k)}, ..., \frac{Q_M(k)}{Q_1(k)}\right]^T \tag{2.9}$$

where $\mathbf{g}(k)$ is the steering vector toward the far-end source, $\mathbf{q}(k)$ is the steering vector toward the near-end source, and the superscript ^{T} represents the transpose operator.

Finally, we can apply a beamformer:

$$\hat{U}(k,n) = \sum_{m=1}^{M} H_m^*(k,n) D_m(k,n) \tag{2.10}$$

where $\hat{U}(k,n)$ is an estimate of the desired signal $U_1(k,n)$, $H_m(k,n)$ are the beamformer coefficients at bin k and frame n on the m-th sensor, and the superscript * marks the complex conjugate operator. This can also be written in vector form:

$$\hat{U}(k,n) = \mathbf{h}^H(k,n)\,\mathbf{d}(k,n) \tag{2.11}$$

where

$$\mathbf{d}(k,n) = \mathbf{y}(k,n) + \mathbf{u}(k,n) + \mathbf{v}(k,n), \tag{2.12}$$

$$\mathbf{y}(k,n) = [Y_1(k,n), Y_2(k,n), ..., Y_M(k,n)]^T, \tag{2.13}$$

$$\mathbf{u}(k,n) = [U_1(k,n), U_2(k,n), ..., U_M(k,n)]^T, \tag{2.14}$$

$$\mathbf{v}(k,n) = [V_1(k,n), V_2(k,n), ..., V_M(k,n)]^T, \tag{2.15}$$

$$\mathbf{h}(k,n) = [H_1(k,n), H_2(k,n), ..., H_M(k,n)]^T, \tag{2.16}$$

and the superscript ^{H} marks the transpose conjugate operator.

2.2 Region-of-Interest Beamforming

In many applications, a spatial filter is needed but the exact location of the desired source is unknown. However, there is usually an ROI where the desired source is assumed to be. This is illustrated by Figure 2.3 in the far-field case with a linear array. In such applications, one can utilize region-based beamformers or constant-beamwidth beamformers. This way, the desired source can be preserved, but other undesired sources in the ROI may deteriorate performance. Another way to solve this is to employ direction-of-arrival (DOA) estimation on θ, and subsequently apply a beamformer directed toward the estimated direction. In the following, we consider the latter method, where a beamformer should be designed and directed towards a specific estimated DOA in the ROI.

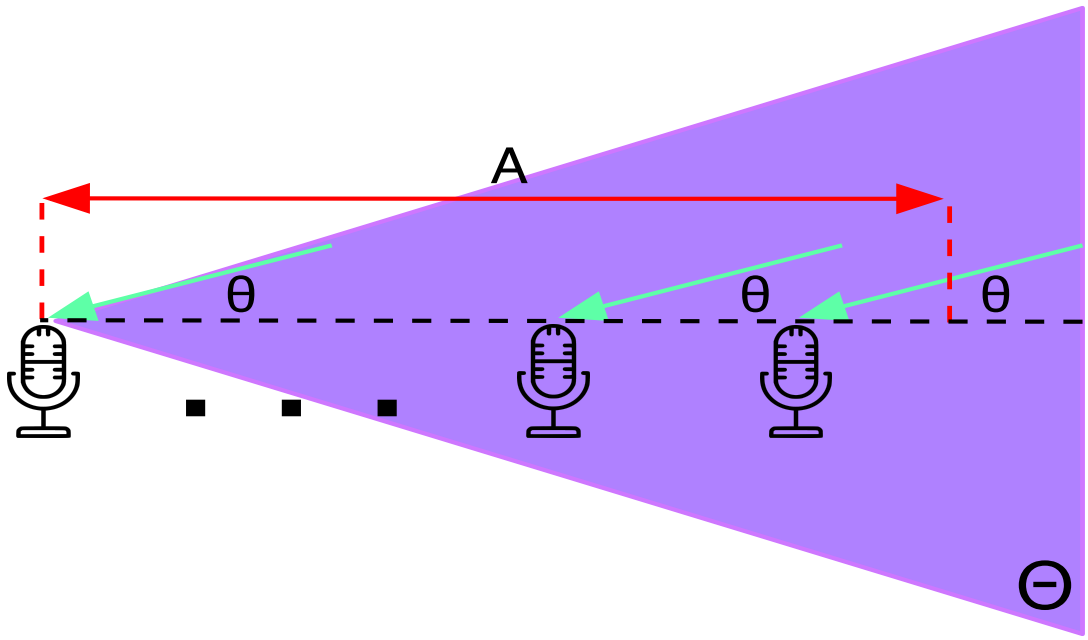

Figure 2.3: Illustration of a signal impinging a microphone array from within an ROI Θ. The angle of arrival (AOA) θ is within the ROI. The array is within an aperture A. The purple area marks the ROI.

Consider a system with M omnidirectional microphones placed nonuniformly across the aperture. The observed signal in the frequency domain is given by the vector

$$\mathbf{y}(\omega) \triangleq [Y_1(\omega), Y_2(\omega), ..., Y_M(\omega)]^T = \mathbf{d}(\mathbf{x}, \omega, \theta) S(\omega) + \mathbf{v}(\omega) \tag{2.17}$$

where $Y_m(\omega)$ is the signal captured by the m-th microphone, $S(\omega)$ is the propagated signal, $\mathbf{v}(\omega)$ is the additive noise vector, and

$$\mathbf{d}(\mathbf{x}, \omega, \theta) = \left[e^{-\jmath \frac{\omega}{c} x_1 \cos\theta}, e^{-\jmath \frac{\omega}{c} x_2 \cos\theta}, ..., e^{-\jmath \frac{\omega}{c} x_M \cos\theta}\right]^T \tag{2.18}$$

is the array steering vector, where $\mathbf{x} \triangleq [x_1, x_2, ..., x_M]^T$ is the microphone position vector, $\omega = 2\pi f$ is the angular frequency, f is the temporal frequency, $\jmath$ is the imaginary unit, and c is the speed of sound, i.e., 340[m/s].

Utilizing all sensors in the array, we may design a beamforming filter

$$\mathbf{h}(\mathbf{x}, \omega, \theta) \triangleq [H_1(\mathbf{x}, \omega, \theta), H_2(\mathbf{x}, \omega, \theta), ..., H_M(\mathbf{x}, \omega, \theta)]^T \tag{2.19}$$

where $H_m(\mathbf{x}, \omega, \theta)$ can be used to estimate the source signal of interest at angle θ by

$$\hat{S}(\omega) = \mathbf{h}^H(\mathbf{x}, \omega, \theta) \mathbf{y}(\omega), \tag{2.20}$$

The beampattern is then defined by

$$\mathcal{B}\left[\mathbf{h}\left(\mathbf{x},\omega,\tilde{\theta}\right),\theta\right]=\mathbf{d}^{H}\left(\mathbf{x},\omega,\theta\right)\mathbf{h}\left(\mathbf{x},\omega,\tilde{\theta}\right)=\sum_{m=1}^{M}H_{m}\left(\mathbf{x},\omega,\tilde{\theta}\right)e^{\jmath\frac{\omega}{c}x_{m}\cos\theta} \tag{2.21}$$

which measures the response of the beamformer directed toward $\tilde{\theta}$ at an AOA θ.

2.3 Performance Measures

To understand how the beamformer in (2.11) impacts the echo component, we define the residual echo signal by

$$Y_{\mathrm{re}}\left(k,n\right)=\mathbf{h}^{H}\left(k,n\right)\mathbf{y}\left(k,n\right). \tag{2.22}$$

A good measure of echo attenuation is the echo-return loss enhancement (ERLE):

$$\xi\left(t\right)=\frac{\mathrm{LPF}\left\{y_{1}^{2}\left(t\right)\right\}}{\mathrm{LPF}\left\{y_{\mathrm{re}}^{2}\left(t\right)\right\}} \tag{2.23}$$

where $y_{\mathrm{re}}\left(t\right)$ is the inverse STFT (ISTFT) of $Y_{\mathrm{re}}\left(k,n\right)$, and LPF$\{\cdot\}$ describes a low pass filter. As the residual echo is diminished, the ERLE grows. Thus, the ERLE should be as large as possible.

To examine how the filter in (2.11) impacts the desired signal component, we define the filtered near-end signal by

$$U_{\mathrm{f}}\left(k,n\right)=\mathbf{h}^{H}\left(k,n\right)\mathbf{u}\left(k,n\right). \tag{2.24}$$

Then, the near-end signal distortion can be assessed by the distortion index (DI):

$$\nu\left(t\right)=\frac{\mathrm{LPF}\left\{\left[u_{1}\left(t\right)-u_{\mathrm{f}}\left(t\right)\right]^{2}\right\}}{\mathrm{LPF}\left\{u_{1}^{2}\left(t\right)\right\}} \tag{2.25}$$

where $u_{f}\left(t\right)$ is the ISTFT of $U_{f}\left(k,n\right)$. As the filter distorts the signal, the DI grows. Thus, a small DI is desired.

Additionally, the Perceptual Evaluation of Speech Quality (PESQ) measure [67] can be used to evaluate performance when comparing $\hat{u}\left(t\right)$ to $u_{1}\left(t\right)$, where $\hat{u}\left(t\right)$ is the ISTFT of $\hat{U}\left(k,n\right)$. With PESQ, the residual noise at the filter output

$$V_{\mathrm{rn}}\left(k,n\right)=\mathbf{h}^{H}\left(k,n\right)\mathbf{v}\left(k,n\right) \tag{2.26}$$

is also taken into account.

A good measure of beamformer robustness in the presence of white noise is WNG:

$$\mathcal{W}\left[\mathbf{h}\left(\mathbf{x},\omega,\theta\right)\right] \triangleq \frac{\left|\mathbf{d}^{H}\left(\mathbf{x},\omega,\theta\right)\mathbf{h}\left(\mathbf{x},\omega,\theta\right)\right|^{2}}{\mathbf{h}^{H}\left(\mathbf{x},\omega,\theta\right)\mathbf{h}\left(\mathbf{x},\omega,\theta\right)}. \tag{2.27}$$

As the look direction amplification increases with respect to noise amplification, the WNG increases.

Another important measure is the DF, which measures beamformer performance in the presence of a diffuse noise field:

$$\mathcal{D}\left[\mathbf{h}\left(\mathbf{x},\omega,\theta\right)\right] \triangleq \frac{\left|\mathbf{d}^{H}\left(\mathbf{x},\omega,\theta\right)\mathbf{h}\left(\mathbf{x},\omega,\theta\right)\right|^{2}}{\mathbf{h}^{H}\left(\mathbf{x},\omega,\theta\right)\mathbf{\Gamma}\left(\mathbf{x},\omega\right)\mathbf{h}\left(\mathbf{x},\omega,\theta\right)} \tag{2.28}$$

where

$$\mathbf{\Gamma}_{i,j}\left(\mathbf{x},\omega\right) = \frac{\sin\left(\omega\left(x_i - x_j\right)/c\right)}{\omega\left(x_i - x_j\right)/c}, \quad 1 \leq i,j \leq M. \tag{2.29}$$

This narrowband measurement may also be extended to the broadband case. We define the broadband directivity index over frequencies $\omega_L \leq \omega \leq \omega_H$ by

$$\mathcal{DI}_{[\omega_L,\omega_H]}\left[\mathbf{h}\left(\mathbf{x},\omega,\theta\right)\right] \triangleq \frac{\int_{\omega_L}^{\omega_H}\left|\mathbf{d}^{H}\left(\mathbf{x},\omega,\theta\right)\mathbf{h}\left(\mathbf{x},\omega,\theta\right)\right|^{2}d\omega}{\int_{\omega_L}^{\omega_H}\mathbf{h}^{H}\left(\mathbf{x},\omega,\theta\right)\mathbf{\Gamma}\left(\mathbf{x},\omega\right)\mathbf{h}\left(\mathbf{x},\omega,\theta\right)d\omega}. \tag{2.30}$$

We also denote $\omega_L = 2\pi f_L$ and $\omega_H = 2\pi f_H$.

2.4 Problem Formulation

The first objective, in the AEC scenario, is to find the near-end signal received by the reference microphone $U_1\left(k,n\right)$, given the far-end signal $X\left(k,n\right)$ and all microphone signals $D_m\left(k,n\right)$. This signal can then be sent to the far-end room. Our aim is to maximize $\xi\left(t\right)$ for maximum echo cancellation, minimize $\nu\left(t\right)$ for minimum distortion, and maximize the PESQ score.

The second objective is to find the optimal array geometry $\mathbf{x}$, that maximizes the worst-case directivity index, as in (2.30), in a ROI around the endfire direction $|\theta| \leq \theta_H$. Each beamformer, directed toward θ, must admit to the distortionless constraint, have sufficient WNG, and maintain a minimal distance between two microphones. This problem can be expressed mathematically as

$$\begin{aligned}
\mathbf{x}^{*} = \arg\max_{\mathbf{x}} \quad & \min_{\theta\in\Theta}\mathcal{DI}_{[\omega_L,\omega_H]}\left[\mathbf{h}\left(\mathbf{x},\omega,\theta\right)\right] \\
\text{s.t.} \quad & \mathcal{B}\left[\mathbf{h}\left(\mathbf{x},\omega,\theta\right),\theta\right] = 1 && \forall\theta\in\Theta,\forall\omega\in\Omega \\
& \mathcal{W}\left[\mathbf{h}\left(\mathbf{x},\omega,\theta\right)\right] \geq \delta && \forall\theta\in\Theta,\forall\omega\in\Omega \\
& \left|x_i - x_j\right| \geq d_c && \forall i,j\in\left[1,M\right], i\neq j \\
& 0 \leq x_m \leq A && \forall m\in\left[1,M\right]
\end{aligned} \tag{2.31}$$

where δ is the minimal WNG, d_c is the minimal distance between two adjacent microphones (half of microphone physical space), $\Omega = \{\omega : \omega_L \leq \omega \leq \omega_H\}$ marks the frequency range, $\Theta = \{\theta : |\theta| \leq \theta_H\}$ marks the ROI, and $\mathbf{x}^*$ is the optimal array geometry.

Chapter 3

Multichannel Acoustic Echo Cancellation with Beamforming in Dynamic Environments

This chapter presents an echo cancellation system that combines multiple sensors and multiple frames. The system utilizes the adaptive LCMV beamforming technique. To achieve good ERLE and DI performance, the acoustic paths are estimated using past STFT frames. In Section 3.1, we present a scheme for the proposed echo canceller. We show how can the acoustic paths be estimated with past STFT frames in Section 3.2. A simulative study with a comparison to NLMS-based approaches is provided in Section 3.3. The chapter is summarized in Section 3.4.

3.1 Beamformer Design

Considering the aforementioned performance measures, the beamformer coefficients can be determined by various methods. This section presents a method that eliminates the echo component, maintains a distortionless response for the desired component, and reduces noise.

Substituting (2.6) into (2.13), we get

$$\mathbf{y}(k,n) = X_{\mathrm{NL}}(k,n)\left[G_1(k), G_2(k), ..., G_M(k)\right]^T, \tag{3.1}$$

then, substituting (3.1) into (2.22)

$$\begin{aligned} &Y_{\mathrm{re}}(k,n) = \\ &X_{\mathrm{NL}}(k,n)\,\mathbf{h}^H(k,n)\left[G_1(k), G_2(k), ..., G_M(k)\right]^T. \end{aligned} \tag{3.2}$$

Finally, substituting (2.8) into (3.2), we get:

$$Y_{\mathrm{re}}(k,n) = G_1(k)\, X_{\mathrm{NL}}(k,n)\, \mathbf{h}^H(k,n)\, \mathbf{g}(k). \tag{3.3}$$

Therefore, to eliminate the echo component, we impose the constraint:

$$\mathcal{C}_1[\mathbf{h}(k,n)] : \mathbf{h}^H(k,n)\, \mathbf{g}(k) = 0. \tag{3.4}$$

Notice that this constraint eliminates the overall echo component, including the nonlinear distortion induced by the loudspeaker. Similarly, by substituting (2.7) into (2.14), we get

$$\mathbf{u}(k,n) = S(k,n)\, [Q_1(k), Q_2(k), ..., Q_M(k)]^T, \tag{3.5}$$

then substituting (3.5) into (2.24)

$$\begin{aligned} &U_{\mathrm{f}}(k,n) = \\ &S(k,n)\, \mathbf{h}^H(k,n)\, [Q_1(k), Q_2(k), ..., Q_M(k)]^T. \end{aligned} \tag{3.6}$$

Finally, substituting (2.9) into (3.6), we get:

$$U_{\mathrm{f}}(k,n) = Q_1(k)\, S(k,n)\, \mathbf{h}^H(k,n)\, \mathbf{q}(k). \tag{3.7}$$

Therefore, to preserve the desired component, we impose the constraint:

$$\mathcal{C}_2[\mathbf{h}(k,n)] : \mathbf{h}^H(k,n)\, \mathbf{q}(k) = 1. \tag{3.8}$$

In this way, the near-end signal received by the reference microphone is preserved.

Now, for noise reduction, we consider the residual noise component

$$V_{\mathrm{rn}}(k,n) = \mathbf{h}^H(k,n)\, \mathbf{v}(k,n) \tag{3.9}$$

and minimize

$$E\left[|V_{\mathrm{rn}}(k,n)|^2\right] = \mathbf{h}^H(k,n)\, E\left[\mathbf{v}(k,n)\, \mathbf{v}^H(k,n)\right] \mathbf{h}(k,n) = \sigma_v^2(k)\, ||\mathbf{h}(k,n)||^2 \tag{3.10}$$

where E is the expectation operator, and $\sigma_v^2(k) = E\left[|V_m(k,n)|^2\right]$ is the noise variance. Overall considering (3.4), (3.8), and (3.10), we can formulate the following problem:

$$\begin{aligned} \mathbf{h}_{\mathrm{opt}}(k,n) = \underset{\mathbf{h}(k,n)}{\arg\min}\, ||\mathbf{h}(k,n)||^2 \\ \text{s.t.} \quad \mathcal{C}_1[\mathbf{h}(k,n)] \\ \mathcal{C}_2[\mathbf{h}(k,n)] \end{aligned} \tag{3.11}$$

where $\mathbf{h}_{\mathrm{opt}}(k,n)$ are the optimal coefficients that eliminate the echo component, pre-

serve the desired component, and minimize the noise component. This is the LCMV beamformer, which is given by

$$\mathbf{h}^{*}(k,n) = \mathbf{C}(k)\left[\mathbf{C}^{H}(k)\,\mathbf{C}(k)\right]^{-1}\mathbf{i}_{c} \tag{3.12}$$

where

$$\mathbf{C}(k) = [\mathbf{q}(k), \mathbf{g}(k)] \tag{3.13}$$

and

$$\mathbf{i}_{c} = [1, 0]^{T}. \tag{3.14}$$

Theoretically, this beamformer eliminates the echo component and preserves the desired component, producing ideal ERLE and DI. However, it can be constructed only when the steering vectors $\mathbf{g}(k)$ and $\mathbf{q}(k)$ are known. In practice, these may change when the loudspeaker or talker moves or when there are changes in the acoustic paths. Furthermore, they must be estimated accurately. Performance may severely degrade due to steering vector inaccuracies. In the following section, we present a method to estimate both $\mathbf{g}(k)$ and $\mathbf{q}(k)$. These estimates are then used to construct the beamformer in (3.12), as indicated by the scheme given in Figure 3.1.

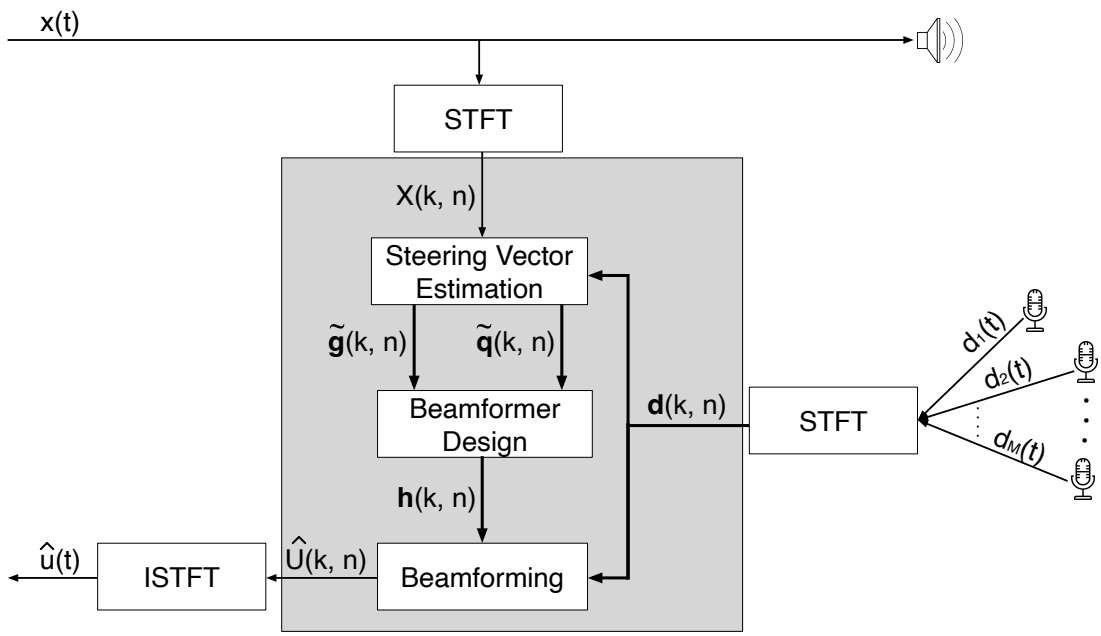

Figure 3.1: Proposed scheme for acoustic echo cancellation. The beamformer in (3.12) is applied on the M microphones with (2.11). The steering vector estimates determine the beamformer coefficients.

In many echo cancellation algorithms, an additional NLMS-based echo canceller is utilized after the beamformer to reduce residual echo further. While this may improve performance when a fixed beamformer is considered, it can degrade performance when an adaptive beamformer is considered [3]. In the adaptive case, the target filter of the

NLMS may change in time, making the convergence of the NLMS filter unstable. In the case of a dynamic environment where the steering vectors change in time, an adaptive beamformer should be utilized. Therefore, such a stage is not utilized in the proposed design.

3.2 Steering Vector Estimation

In this section, we provide an estimate for the steering vectors $\mathbf{g}(k)$ and $\mathbf{q}(k)$ used for the beamformer design in Section 3.1. To this end, we assume no movements in the room in the last L frames, i.e., the loudspeaker, talker, and microphones did not move in the last L frames.

Neglecting nonlinear echo and noise, we can utilize (2.5), (2.6), and (2.7) on all the $l \in [1, ..., L]$ last frames and get

$$D_m(k, n-l+1) = G_m(k) X(k, n-l+1) + Q_m(k) S(k, n-l+1) \tag{3.15}$$

Then, by taking $m = m_1$ and $m = m_2$ for any two sensors m_1 and m_2, we find the ratio

$$\frac{Q_{m_1}(k)}{Q_{m_2}(k)} = \frac{D_{m_1}(k, n-l+1) - G_{m_1}(k) X(k, n-l+1)}{D_{m_2}(k, n-l+1) - G_{m_2}(k) X(k, n-l+1)}. \tag{3.16}$$

Notice that the left-hand side of (3.16) is independent of l. Thus, we can state that the right-hand side for any $l = l_1$ and $l = l_2$ are equal

$$\begin{aligned} &\frac{D_{m_1}(k, n-l_1+1) - G_{m_1}(k) X(k, n-l_1+1)}{D_{m_2}(k, n-l_1+1) - G_{m_2}(k) X(k, n-l_1+1)} = \\ &\frac{D_{m_1}(k, n-l_2+1) - G_{m_1}(k) X(k, n-l_2+1)}{D_{m_2}(k, n-l_2+1) - G_{m_2}(k) X(k, n-l_2+1)} \end{aligned} \tag{3.17}$$

Multiplying (3.17) with the common denominator and simplifying both sides of the equation, the quadratic components of $G_{m_1}(k) G_{m_2}(k)$ are reduced. This yields a linear equation with respect to $G_{m_1}(k)$ and $G_{m_2}(k)$:

$$\begin{aligned} &G_{m_1}(k) \left[X(k, n-l_1+1) D_{m_2}(k, n-l_2+1) - X(k, n-l_2+1) D_{m_2}(k, n-l_1+1) \right] + \\ &G_{m_2}(k) \left[X(k, n-l_2+1) D_{m_1}(k, n-l_1+1) - X(k, n-l_1+1) D_{m_1}(k, n-l_2+1) \right] = \\ &D_{m_1}(k, n-l_1+1) D_{m_2}(k, n-l_2+1) - D_{m_1}(k, n-l_2+1) D_{m_2}(k, n-l_1+1) \end{aligned} \tag{3.18}$$

Overall, for any pick of $1 \leq m_1$, $m_2 \leq M$ and $1 \leq l_1$, $l_2 \leq L$, we get an equation as in (3.18). This set of equations can be used to find $G_m(k)$ for any $1 \leq m \leq M$, since $X(k, n-l+1)$ and $D_m(k, n-l+1)$ are given. Some of these equations are trivial. For any case where $l_1 = l_2$ or $m_1 = m_2$, the equation is reduced to a degenerate one. Also, notice that when substituting l_1 by l_2 and vice versa, or m_1 by m_2 and vice versa,

we get essentially the same equation. Thus, the informative equations are the ones taking (3.18) for any $l_1 \neq l_2$ and $m_1 \neq m_2$, where the pairs (l_1, l_2) and (m_1, m_2) are not reiterated. In the special case where the far-end speech is absolutely silent, we also arrive at a trivial expression. To circumvent this case, it is possible to utilize a VAD on the loudspeaker so that when there is no speech, no echo cancellation is needed.

Since every equation corresponds to a different pick of (m_1, m_2) and (l_1, l_2), we arrive at $\binom{M}{2}\binom{L}{2}$ linearly independent equations. The solution to this system can be found only if the number of variables is lower than the number of equations, i.e.,

$$M \leq \binom{M}{2}\binom{L}{2} \tag{3.19}$$

which, when simplified, can be written as:

$$L\left(L-1\right)\left(M-1\right) \geq 4. \tag{3.20}$$

It is clear from (3.20) that only a small L and M are needed, i.e., $L \geq 3$ or $M \geq 3$, for us to solve the system. In general, there may be more equations than variables, meaning not all equations are needed. However, due to noise and loudspeaker non-linearity, more equations may help us find a better estimate for $G_m\left(k\right)$.

Now, we express this system in matrix form and solve it, as described in Algorithm 3.1. The matrix $\mathbf{A}\left(k,n\right)$ and vector $\mathbf{b}\left(k,n\right)$ are constructed so that they define our system. Notice that in the general case, where there are more equations than variables, the system is unsolvable due to conflicting equations. Nevertheless, these conflicts stem from the appearance of noise and loudspeaker non-linearity, which introduce relatively small perturbations. Thus, we can find the least-squares estimator for the system to mitigate the effect of these perturbations. The least-squares estimator can be found by $\mathbf{A}^{\dagger}\left(k,n\right)\mathbf{b}\left(k,n\right)$, where the superscript † marks the pseudo-inverse operator and

$$\mathbf{A}^{\dagger}\left(k,n\right) \triangleq \left[\mathbf{A}^{H}\left(k,n\right)\mathbf{A}\left(k,n\right)\right]^{-1}\mathbf{A}^{H}\left(k,n\right). \tag{3.21}$$

Once we have found all $\tilde{G}_m\left(k,n\right)$ the steering vector towards the loudspeaker can be found by utilizing (2.8)

$$\tilde{\mathbf{g}}\left(k,n\right) = \left[1, \frac{\tilde{G}_2\left(k\right)}{\tilde{G}_1\left(k\right)}, ..., \frac{\tilde{G}_M\left(k\right)}{\tilde{G}_1\left(k\right)}\right]^T \tag{3.22}$$

and the steering vector toward the talker can be found by substituting (2.9) and (3.16)

$$\begin{aligned}\tilde{\mathbf{q}}\left(k,n\right) = \Bigg[1, &\frac{D_2\left(k,n\right) - \tilde{G}_2\left(k,n\right)X\left(k,n\right)}{D_1\left(k,n\right) - \tilde{G}_1\left(k,n\right)X\left(k,n\right)}, ... \\ &, \frac{D_M\left(k,n\right) - \tilde{G}_M\left(k,n\right)X\left(k,n\right)}{D_1\left(k,n\right) - \tilde{G}_1\left(k,n\right)X\left(k,n\right)}\Bigg]^T.\end{aligned} \tag{3.23}$$

Algorithm 3.1

Inputs: M, L,
$X(k, n-l+1)$, $D_m(k, n-l+1)$ $\quad 1 \leq l \leq L \quad 1 \leq m \leq M$
Outputs: $\tilde{G}_m(k, n)$ $\quad 1 \leq m \leq M$
Create list M_{list} of $\binom{M}{2}$ non-repetitive pairs (m_1, m_2)
Create list L_{list} of $\binom{L}{2}$ non-repetitive pairs (l_1, l_2)
$\mathbf{A}(k, n) \leftarrow \mathbf{0}_{\binom{M}{2}\binom{L}{2} \times M}$
for $i = 1, 2, ..., \binom{M}{2}$ **do**
  $(m_1, m_2) \leftarrow M_{\text{list}}(i)$
  for $j = 1, 2, ..., \binom{L}{2}$ **do**
    $(l_1, l_2) \leftarrow L_{\text{list}}(j)$
    $\mathbf{A}_{\left[(i-1)\binom{L}{2}+j, m_1\right]}(k, n) \leftarrow X(k, n-l_1+1) D_{m_2}(k, n-l_2+1) - X(k, n-l_2+1) D_{m_2}(k, n-l_1+1)$
    $\mathbf{A}_{\left[(i-1)\binom{L}{2}+j, m_2\right]}(k, n) \leftarrow X(k, n-l_2+1) D_{m_1}(k, n-l_1+1) - X(k, n-l_1+1) D_{m_1}(k, n-l_2+1)$
    $\mathbf{b}_{\left[(i-1)\binom{L}{2}+j\right]}(k, n) \leftarrow D_{m_1}(k, n-l_1+1) D_{m_2}(k, n-l_2+1) - D_{m_1}(k, n-l_2+1) D_{m_2}(k, n-l_1+1)$
  end for
end for
$\left[\tilde{G}_1(k, n), \tilde{G}_2(k, n), ..., \tilde{G}_M(k, n)\right]^T \leftarrow \mathbf{A}^{\dagger}(k, n)\, \mathbf{b}(k, n)$

The larger M and L are, the more information is used when estimating the steering vectors. Specifically, as L grows, the system obtains more equations while the number of variables remains, producing more accurate results. In this case, however, the acoustic paths are assumed to be static for long periods, which may be problematic in real scenarios. The larger M is, we have more equations and variables in the system, so while spatial sampling is increased by adding microphones, the steering vector estimation task is more challenging.

Analyzing the computational complexity of Algorithm 3.1, there are 2 nonzero elements in a row of $\mathbf{A}(k, n)$ that require 2 multiplications each, and all elements of $\mathbf{b}(k, n)$ require 2 multiplications as well. Therefore, the construction of $\mathbf{A}(k, n)$ and $\mathbf{b}(k, n)$ is of complexity

$$\mathcal{O}\left\{\binom{M}{2}\binom{L}{2}\right\} = \mathcal{O}\left\{M^2 L^2\right\}. \tag{3.24}$$

Then, we must find $\mathbf{A}^{\dagger}(k, n)$. From (3.21), this contains a matrix multiplication, an inverse operation, and another matrix multiplication with overall complexity

$$\mathcal{O}\left\{M\left[\binom{M}{2}\binom{L}{2}\right]^2 + M^3 + M^2\binom{M}{2}\binom{L}{2}\right\} = \mathcal{O}\left\{M^5 L^4\right\} \tag{3.25}$$

Finally, solving the system requires

$$\mathcal{O}\left\{M\binom{M}{2}\binom{L}{2}\right\} = \mathcal{O}\left\{M^3L^2\right\} \tag{3.26}$$

multiplications. Thus, the overall complexity of Algorithm 3.1 is $\mathcal{O}\left\{M^5L^4\right\}$. Due to this, M and L must be carefully selected. While performance should be maximized, utilizing large values of M and L may affect runtime. In section 3.3, we show that relatively small values are sufficient for high performance. Also, notice that $\mathbf{A}\left(k,n\right)$ is a sparse matrix, thereby reducing runtime and hardware resources.

Note that no double-talk detector is used throughout the proposed scheme in Figure 3.1. Our proposed algorithm cancels acoustic echo regardless of the scenario, be it double-talk or single-talk of any of the speakers. This is because the RTFs can be estimated during double-talk, as opposed to RTF estimation methods [5, 68–74] that assume the existence of only the corresponding source. The method in [72] is capable of tracking multiple sources, but assumes that sources do not become simultaneously active. Specifically, cross-relation system identification methods [73, 74] also exploit the relations between sensor pairs but assume that only one source exists. Furthermore, in a dynamic environment where the talker or loudspeaker moves or a different talker speaks in the near-end room, our algorithm adjusts to the new paths after L time frames. During this time, double-talk may also take place.

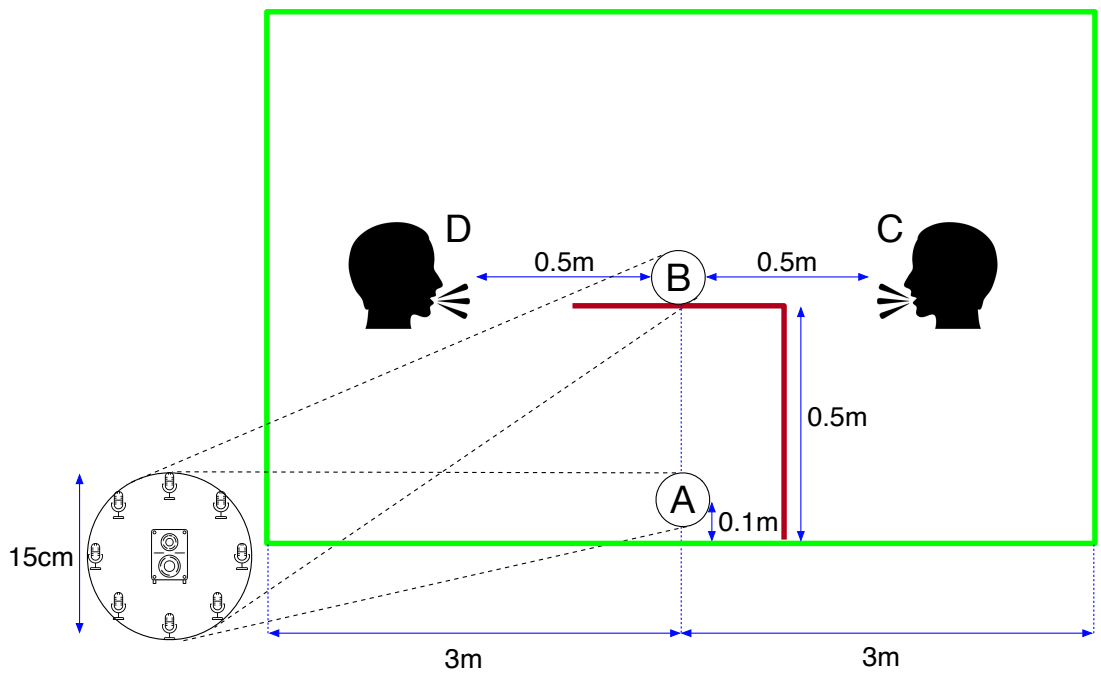

Figure 3.2: Simulated room. The speakerphone, consisting of a microphone array in a UCA geometry and a loudspeaker, can be at $\mathcal{A}$ or $\mathcal{B}$. The active talker can be at $\mathcal{C}$ or $\mathcal{D}$.

3.3 Simulations

In this section, we evaluate the proposed beamformer. The ERLE (2.23), DI (2.25), and PESQ measurements are studied. Also, the signals received by the microphones and the beamformer output are presented to give better visual perception.

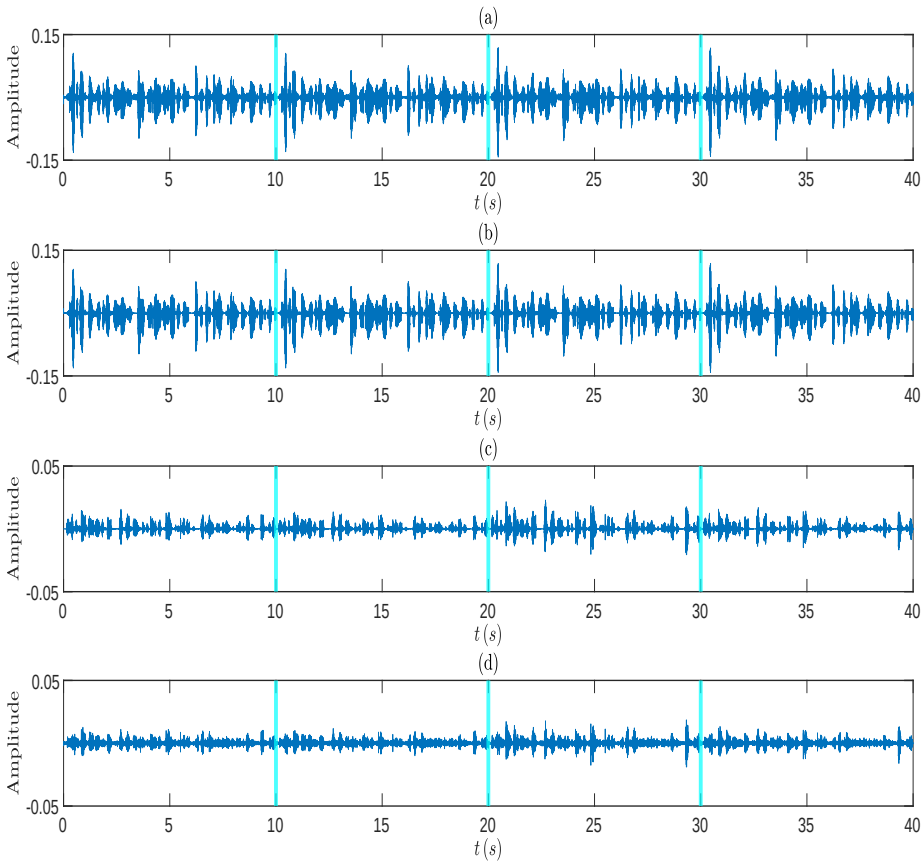

Figure 3.3: Received signals by the first microphone and the beamformer output as a function of time. (a) the total received signal in the reference microphone $d_1(t)$, (b) the echo component signal in the reference microphone $y_1(t)$, (c) the desired component signal in the reference microphone $u_1(t)$, and (d) the beamformer output signal $\hat{u}(t)$. The cyan lines mark the different time segments. $M = L = 4$. Note the difference in scale between (a), (b) and (c), (d).

We evaluate the proposed method in a simulated room of dimensions 6 m $\times$6 m $\times$4.5 m. To simulate a realistic low SER, we use a speakerphone consisting of a microphone array structured in a UCA geometry of radius 7.5 cm, and a loudspeaker at the center of the array. Significant acoustic coupling takes place in this configuration. This array geometry was chosen due to its ability to produce high-directivity spatial filters toward any location [75]. Overall, four location configurations depicted by Figure 3.2 are considered in our experiment, where each configuration defines a time segment of

length 10 s. The speakerphone may be located at $\mathcal{A}$ - the room center on the floor in coordinates $[3, 3, 0.1]$, or at $\mathcal{B}$ - the room center installed on a surface in coordinates $[3, 3, 0.5]$. The active talker may be located at $\mathcal{C}$ - 0.5 m to the right of $\mathcal{B}$, or at $\mathcal{D}$ - 0.5 m to the left of $\mathcal{B}$. This can fit a scenario where two distinct speakers converse, and the loudspeaker is moved between the floor and the surface. To understand how both loudspeaker and near-end talker locations affect performance, the following four configurations are simulated in this working order:

1. Speakerphone at $\mathcal{A}$, talker at $\mathcal{C}$.
2. Speakerphone at $\mathcal{A}$, talker at $\mathcal{D}$.
3. Speakerphone at $\mathcal{B}$, talker at $\mathcal{C}$.
4. Speakerphone at $\mathcal{B}$, talker at $\mathcal{D}$.

The transition between the configurations is done instantly to maintain an environment that is not constantly changing and to reduce any uncertainties in the device locations.

The speech signals $x(t)$ and $s(t)$ were taken from the TIMIT database [76] and $x_{\mathrm{NL}}(t)$ was generated with the model by Thompson [77, 78]. The signals $y_m(t)$ and $u_m(t)$ were found by convolving $x_{\mathrm{NL}}(t)$ and $s(t)$ with the room impulse responses (RIRs) corresponding to the loudspeaker and talker locations, respectively, and the m-th microphone location. The RIRs were found with the RIR generator by Habets [79]. Speech is sampled at 16kHz. Unless stated otherwise, a reverberation time of $T_{60} = 0.3$s is simulated, and white Gaussian noise is added subsequently to all microphones with SNR= 30dB. Here, the SNR is found with respect to the overall received signal, i.e. SNR= $10 \log_{10} E\{\frac{d_m^2(t)}{v_m^2(t)}\}$, this is done to describe noise which is proportionate to the overall obtained signal. The SERs measured in the reference microphone during the four configurations are -17.89 dB, -18.7 dB, -15.27 dB, and -16.89 dB, respectively. For the STFT, a Kaiser window with $\beta = 5$ is used with a length of 512 samples (32 ms) and 75% overlap.

The rest of this section is organized as follows. First, we demonstrate our echo-canceling ability visually with the observed and resulting signals. Then, we investigate how environmental reverberation and noise impact performance. Later, we evaluate how the algorithm parameters M and L impact performance. Finally, we compare the proposed beamformer with the NLMS-based methods in [36, 37] that utilize multiple frames.

3.3.1 Echo Cancellation

Figure 3.3 shows the received signals by the reference microphone and the beamformer output as a function of time. Notice the scale difference in Figure 3.3 between (a), (b) and (c), (d). The echo component is the main contributor to the received signal due to the small distance between the loudspeaker and the microphone, as is typical

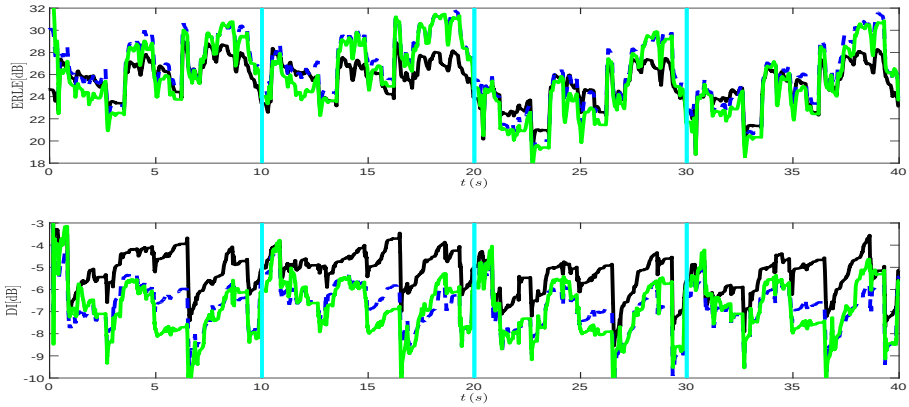

Figure 3.4: ERLE and DI as a function of time for various SNRs. The black, blue-dotted, and green lines mark SNR= 10dB, 20dB, and 30dB, respectively. The cyan lines mark the different time segments. $M = L = 4$.

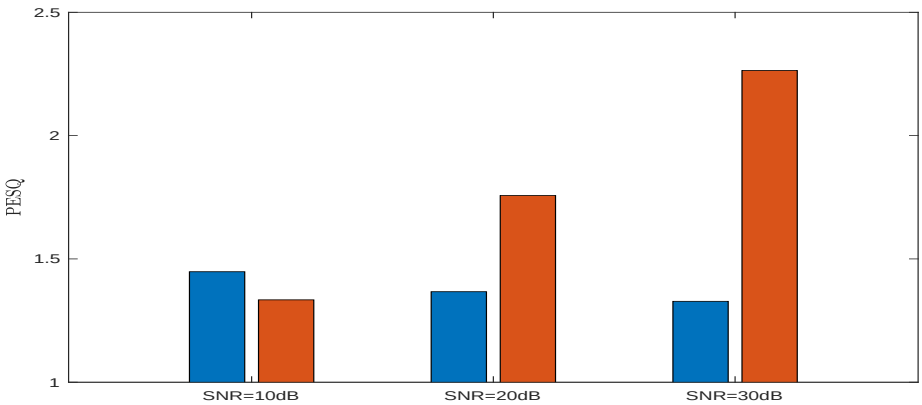

Figure 3.5: PESQ over the entire simulation for various SNRs. The blue and red bars mark the PESQ before and after filtering, respectively. $M = L = 4$.

in speakerphones. Also, one can see how the near-end component is preserved, despite the significant echo component.

3.3.2 Performance as Function of Noise and Reverberation

Let us start by studying how environmental noise impacts performance. It has been shown that design immunity to white noise increases robustness to microphone mismatch errors [80]. Therefore, analyzing noise robustness can also be viewed from the perspective of microphone mismatch robustness. The ERLE and DI as a function of time are presented in Figure 3.4 for various SNRs. It appears there is some improvement in both ERLE and DI when comparing SNR=20dB to SNR=10dB, and that the performance of SNR=30dB and SNR=20dB is comparable. The PESQ for various SNRs over the entire simulation is in Figure 3.5. The PESQ before filtering (using

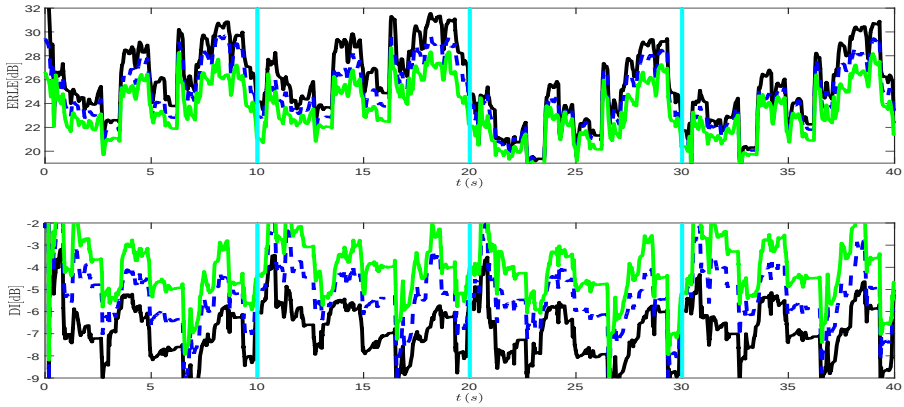

Figure 3.6: ERLE and DI as a function of time for various values of T_{60}. The black, blue-dotted, and green lines mark $T_{60} = 0.3$s, 0.4s, and 0.5s, respectively. The cyan lines mark the different time segments. $M = L = 4$.

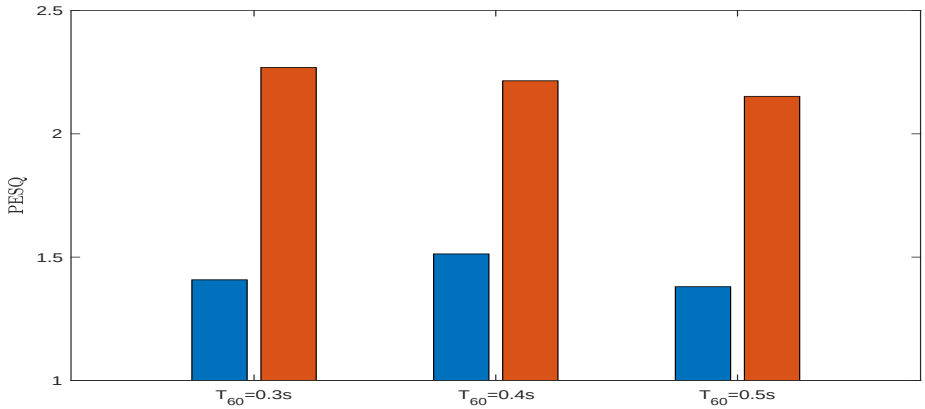

Figure 3.7: PESQ over the entire simulation for various values of T_{60}. The blue and red bars mark the PESQ before and after filtering, respectively. $M = L = 4$.

$u_1(r)$ and $d_1(t)$) is also presented for reference. Notice that since the SNR varies, the PESQ achieved before filtering varies as well. In contrast to the ERLE and DI, a clear improvement in the PESQ can be observed comparing SNR=30dB and SNR=20dB. This is because PESQ also takes into account the residual noise at the filter output, which is diminished as SNR increases.

Now, we continue by analyzing how reverberation impacts performance. The ERLE and DI as a function of time are presented in Figure 3.6 for various values of T_{60}. A steady decline in both ERLE and DI can be observed as reverberation time increases. The PESQ for various values of T_{60} over the entire simulation is in Figure 3.7. Overall, as more reverberations occur, performance is degraded. This can be attributed to the longer impulse responses in the room. In this case, the approximations in (2.6) and (2.7) are less accurate, thereby degrading performance.

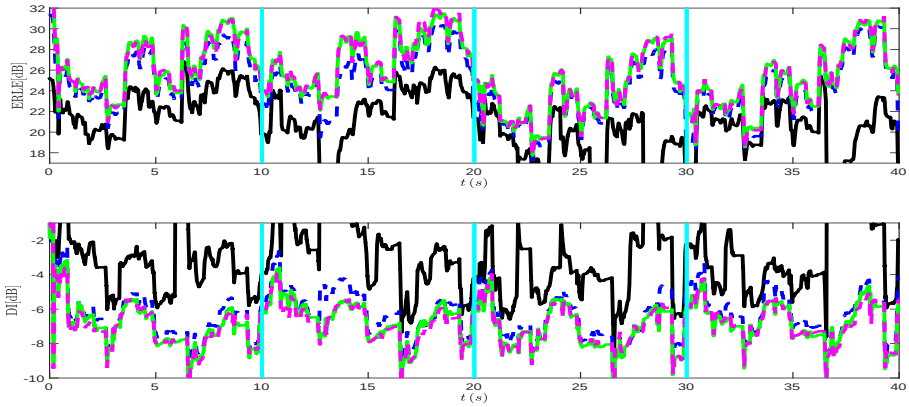

Figure 3.8: ERLE and DI as a function of time for various M. The black, blue-dotted, green, and magenta-dotted lines mark $M = 2$, 3, 4, and 5, respectively. The cyan lines mark the different time segments. $L = 4$.

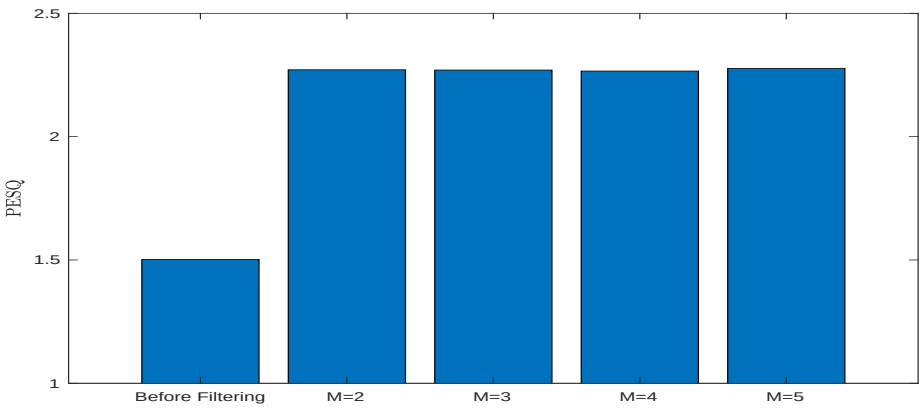

Figure 3.9: PESQ over the entire simulation for various M. $L = 4$.

3.3.3 Performance as a Function of Microphones and Frames

The ERLE and DI as a function of time are presented in Figure 3.8 for various M. As M grows, a slight improvement can be seen in the ERLE, which is most significant when M grows from 2 to 3. The DI barely changes, except from $M = 2$ to $M = 3$. The PESQ for various values of M over the entire simulation is in Figure 3.9. Here, all values of M are comparable. The only slight change in performance as a function of M can be attributed to the fact that increasing M has a dual impact on steering vector estimation. On the one hand, more equations are added to the system; on the other hand, more variables are added. This is translated to a performance limit.

Now, we examine how L impacts performance. The ERLE and DI as a function of time are presented in Figure 3.10 for various L. We can clearly state that $L = 2$ frames are insufficient for the proposed approach. Utilizing just the two recent frames does not give an accurate steering vector estimation. This may be because insufficient

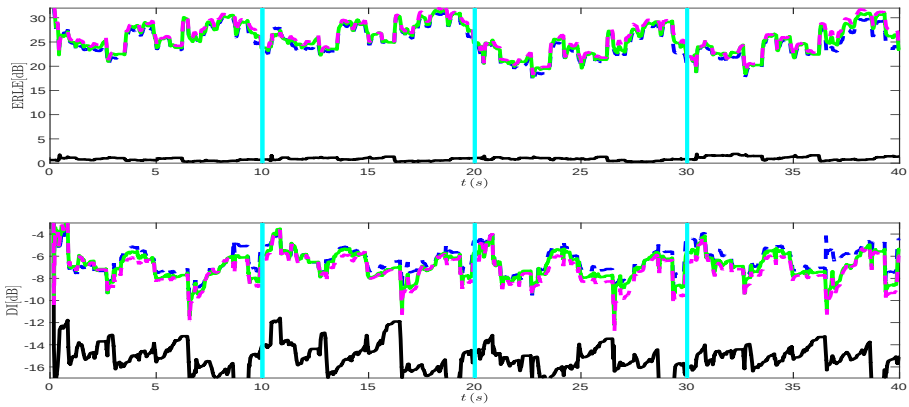

Figure 3.10: ERLE and DI as a function of time for various L. The black, blue-dotted, green, and magenta-dotted lines mark $L = 2$, 3, 4, and 5, respectively. The cyan lines mark the different time segments. $M = 4$.

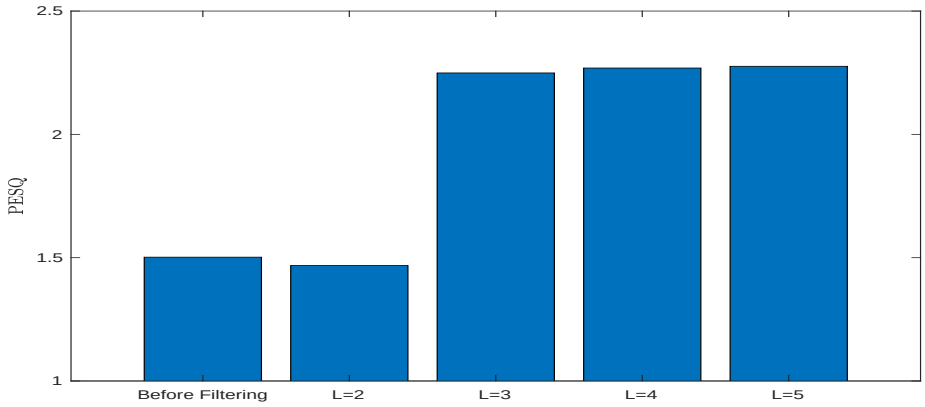

Figure 3.11: PESQ over the entire simulation for various values of L. $M = 4$.

equations are utilized in the system. Only 6 equations are used to estimate 4 variables. Therefore the appearance of noise and nonlinear echo significantly impacts performance. Furthermore, since the impulse responses are longer than the STFT window length, the TFs may vary between frames; therefore, utilizing only 2 frames degrades performance. Utilizing a large number of frames can help us successfully portray the TFs. This can also be observed when examining the PESQ for various L in Figure 3.11. For $L \geq 3$ the ERLE and DI performance slightly improve as L increases.

To sum up, to achieve good echo cancellation with acceptable distortion, both M and L should be larger than 3, although a solution can be produced for lower values that guarantee (3.19).

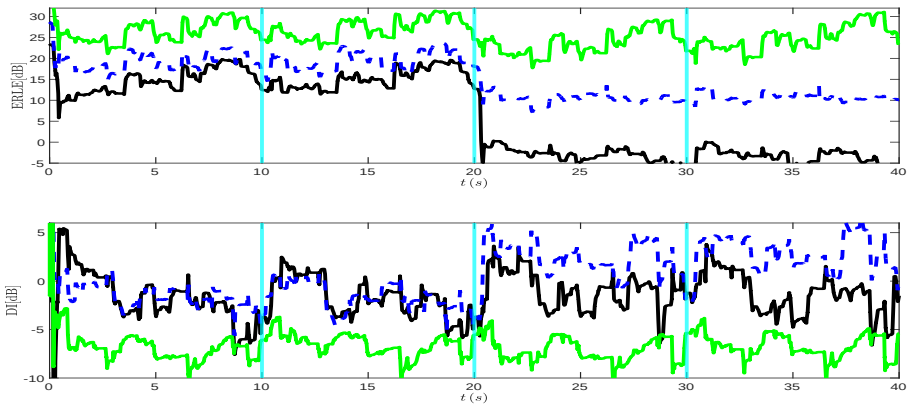

Figure 3.12: ERLE and DI as a function of time for all methods. The black, blue-dotted, and green lines mark the methods [36], [37], and the proposed method, respectively. The cyan lines mark the different time segments. $M = L = 4$.

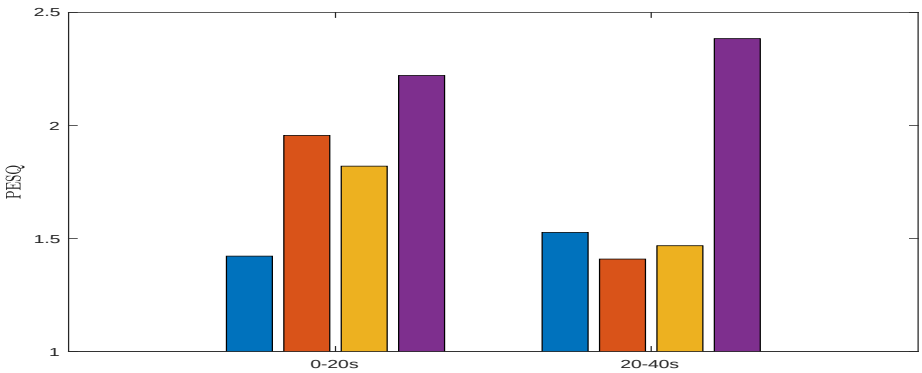

Figure 3.13: PESQ before echo reflections change (0-20s) and after echo reflections change (20-40s), for all competing methods. The blue, red, yellow, and purple bars mark the PESQ before filtering, using [36], using [37], and using the proposed method, respectively. $M = L = 4$.

3.3.4 Method Comparison

We now compare the proposed method with [36] and [37]. In both competing methods the NLMS filter was adapted during a previous segment that is identical to our first segment, only that the near-end talker is silent. The ERLE and DI as a function of time are presented for all methods in Figure 3.12. Higher ERLE and lower DI are obtained with the proposed method, even more so after the speakerphone moves.

Notice that, neglecting reflections, the echo paths do not change throughout the experiment. Therefore one can expect NLMS-based methods to work well even when the speakerphone moves during double-talk. However, this is not what happens in practice. Both ERLE and DI worsen for the competing methods once the speakerphone moves. This means that varying reflections significantly impact the echo path, degrading the

accuracy produced by the NLMS algorithm. The PESQ for all methods, before and after the change in reflections, is in Figure 3.13. The PESQ is degraded due to changing reflections as well in the competing methods. Indeed, the adapted NLMS filter is irrelevant once the echo path has changed. This explains the clear advantage of the proposed method in the last two segments. Considering the first two segments, the advantage may be explained by the ability of our method to contain background noise and nonlinear echo during double-talk, as the least-squares estimator in Algorithm 3.1 is designed to reduce the impact of these on the estimate.

3.4 Summary

An adaptive beamformer for AEC was developed, where the adaptation process considers recent frames of the reference loudspeaker signal and the received microphone signals. This enabled the beamformer to adapt appropriately in a dynamic environment during double-talk, with no double-talk detection. Furthermore, in theory, if there is no background noise and no nonlinear distortion from the loudspeaker, our method completely cancels the echo component while preserving the desired component, as the steering vectors can be accurately estimated from the recent frames. The steering vectors were estimated using a least-squares estimator approach designed to reduce the impact of those two factors specifically. Finally, experiments in a simulated room were conducted. Our experiments indicate that far-end component attenuation, near-end component distortion, and PESQ, all achieve higher performance when compared to NLMS-based methods. This improvement is mainly attributed to our method's ability to adjust to a changing echo path during double-talk, as it responds even to secondary echo path variations that stem from reflections. Future research may focus on more advanced strategies for steering vector estimation utilizing multiple frames. For example, incorporating a nonlinear loudspeaker distortion model in the steering vector estimation process may improve the steering vector estimates for beamformer design.

Chapter 4

Array Geometry Optimization for Region-of-Interest Broadband Beamforming

This chapter presents a method to optimize the array geometry for ROI beamforming. The optimal array is found by maximizing the worst-case directivity index in the ROI considering a broadband frequency range. A constraint on WNG is also employed for noise robustness. In Section 4.1, we find the optimal geometry. In Section 4.2, a post-processing scheme is developed, that finds the best beamformer coefficients in other directions than the worst-case direction in the ROI, given the obtained optimal geometry. A simulative study is conducted in Section 4.3. Section 4.4 summarizes this chapter.

4.1 Optimal Array Design

To find $\mathbf{x}^*$, we formulate the problem as a convex one. First, we present the constraints, then the target function.

4.1.1 Constraints

We start by sampling our search space. Consider a grid of N possible microphone locations $[0 : \Delta x : A]$, where

$$\Delta x = \frac{A}{N-1}. \tag{4.1}$$

We define a selection vector optimization variable

$$\mathbf{s} = [S_1, ..., S_{N-1}, 1]^T \tag{4.2}$$

which consists of binary values. Each element S_i is 1 if a microphone is placed at distance $(i-1)\,\Delta x$ with respect to the rightmost placement, and is 0 otherwise. Note

that, without loss of generality, the leftmost coordinate is always occupied, i.e., $S_N = 1$. This is done to increase the search grid of the array effectively. To guarantee the existence of only M microphones, we should constrain $\mathbf{s}$ to $\mathcal{C}_1$:

$$\mathcal{C}_1 [\mathbf{s}] : \mathbf{s}^H \mathbf{i}_N = M \tag{4.3}$$

where $\mathbf{i}_N$ is a column vector of length N consisting of ones.

To guarantee minimal distances, we must ensure that all adjacent selected placements will be separated by at least d_c. This means that there are restricted areas where no more than a single microphone can be present. All elements of S that correspond to such an area are summed and constrained to be no more than 1. Mathematically, this is described by $\mathcal{C}_2$:

$$\mathcal{C}_2 [\mathbf{s}] : \mathbf{s}^H U \leq \mathbf{i}_G^T \tag{4.4}$$

where $G = N - \lfloor \frac{d_c}{\Delta x} \rfloor$ is the number of restricted areas, U is a matrix of dimensions $N \times G$, whose i-th column is of the form

$$u_i = \left[\mathbf{0}_{i-1}^T, \mathbf{i}_{N+1-G}^T, \mathbf{0}_{G-i}^T \right]^T , \tag{4.5}$$

and $\mathbf{0}_i$ is a column vector consisting zeros of length i.

Now, in addition to the variable $\mathbf{s}$, we must take into account the coefficient variables. To this end, we denote by

$$\mathbf{h}_{\text{tot}} (\omega, \theta) = [H_{\text{tot},1} (\omega, \theta) , H_{\text{tot},2} (\omega, \theta) , ..., H_{\text{tot},N} (\omega, \theta)]^T \tag{4.6}$$

a vector that corresponds to a beamformer directed toward θ that utilizes all N potential sensor placements. Sampling in frequency space and angle space, we consider several frequencies $[\omega_L : \Delta\omega : \omega_H]$ where

$$\Delta\omega = \frac{\omega_H - \omega_L}{Q - 1}, \quad \omega_q = \omega_L + (q - 1) \Delta\omega, \quad q \in [1, Q] \tag{4.7}$$

and several look directions $[0 : \Delta\theta : \theta_H]$ where

$$\Delta\theta - \frac{\theta_H}{P - 1}, \quad \theta_p = (p - 1) \Delta\theta, \quad p \in [1, P] . \tag{4.8}$$

Thus, we can sample the coefficients $\mathbf{h}_{\text{tot}} (\omega, \theta)$ on several values of ω and θ, overall involving $N \times Q \times P$ coefficient variables in our optimization. Note that only positive values of θ are taken into account due to the performance symmetry of linear arrays with respect to the endfire direction.

To admit to the distortionless constraint, the beampattern of any beamformer at any frequency toward the look direction should be 1. This is constrained by $\mathcal{C}_3$:

$$\mathcal{C}_3 [\mathbf{h}_{\text{tot}} (\omega, \theta)] : \mathbf{d}_{\text{tot}}^H (\omega_q, \theta_p) \mathbf{h}_{\text{tot}} (\omega_q, \theta_p) = 1 \qquad \forall p \in [1, P] , \quad \forall q \in [1, Q] , \tag{4.9}$$

where

$$\mathbf{d}_{\text{tot}}(\omega,\theta) = \left[1, e^{-\jmath\frac{\omega}{c}\Delta x\cos\theta}, ..., e^{-\jmath\frac{\omega}{c}A\cos\theta}\right]^T . \tag{4.10}$$

When the distortionless constraint is satisfied, it is sufficient to use $\mathcal{C}_4$ to maintain the desired WNG:

$$\mathcal{C}_4\left[\mathbf{h}_{\text{tot}}(\omega,\theta)\right] : \mathbf{h}_{\text{tot}}^H(\omega_q,\theta_p)\,\mathbf{h}_{\text{tot}}(\omega_q,\theta_p) \leq \frac{1}{\delta} \qquad \forall p\in[1,P], \quad \forall q\in[1,Q]. \tag{4.11}$$

Finally, utilizing only M microphones in practice, the following must hold:

$$\mathcal{C}_5\left[\mathbf{s},\mathbf{h}_{\text{tot}}(\omega,\theta)\right] : \left|H_{\text{tot},i}(\omega_q,\theta_p)\right|^2 \leq \frac{S_i}{\delta} \qquad \forall i\in[1,N], \quad \forall p\in[1,P], \quad \forall q\in[1,Q]. \tag{4.12}$$

Essentially, if $S_i = 0$, then all beamformers will not utilize the i-th microphone placement. In practice, this means that no microphone is placed at $(i-1)\,\Delta x$. However, if $S_i = 1$, then there is indeed a microphone placed, and all beamformers may utilize that position. The factor $1/\delta$ provides an upper bound so that $\mathcal{C}_5$ is convex. As long as $\mathcal{C}_4$ is also maintained, this factor does not restrict the coefficients further.

4.1.2 Target Function

Notice that when the distortionless constraint is met, the numerator in (2.30) is constant:

$$\int_{\omega_L}^{\omega_H} \left|\mathbf{d}^H(\mathbf{x},\omega,\theta)\,\mathbf{h}(\mathbf{x},\omega,\theta)\right|^2 d\omega = \int_{\omega_L}^{\omega_H} \left|\mathcal{B}\left[\mathbf{h}(\mathbf{x},\omega,\theta),\theta\right]\right|^2 d\omega = \int_{\omega_L}^{\omega_H} d\omega = \omega_H - \omega_L. \tag{4.13}$$

Therefore, when maximizing the directivity index, we may focus on minimizing the denominator of (2.30). When approximating the integral to a discrete sum, using our optimization variables, we get

$$\int_{\omega_L}^{\omega_H} \mathbf{h}^H(\mathbf{x},\omega,\theta)\,\mathbf{\Gamma}(\mathbf{x},\omega)\,\mathbf{h}(\mathbf{x},\omega,\theta)\,d\omega \propto \sum_{q=1}^{Q} \mathbf{h}_{\text{tot}}^H(\omega_q,\theta)\,\mathbf{\Gamma}_{\text{tot}}(\omega_q)\,\mathbf{h}_{\text{tot}}(\omega_q,\theta), \tag{4.14}$$

where $\mathbf{\Gamma}_{\text{tot}}(\omega)$ is of dimensions $N\times N$ with elements

$$\mathbf{\Gamma}_{\text{tot},i,j}(\omega) = \frac{\sin\left[\omega\,(i-j)\,\Delta x/c\right]}{\omega\,(i-j)\,\Delta x/c}, \quad 1\leq i,j\leq N. \tag{4.15}$$

To maximize the worst-case directivity index as in (2.31), the maximal value of

(4.14) over $\theta \in \Theta$ should be minimized. Thus, we should find the minimum of

$$R\left[\mathbf{h}_{\text{tot}}\left(\omega,\theta\right)\right] = \max_{p\in[1,P]} \sum_{q=1}^{Q} \mathbf{h}_{\text{tot}}^{H}\left(\omega_q,\theta_p\right) \mathbf{\Gamma}_{\text{tot}}\left(\omega_q\right) \mathbf{h}_{\text{tot}}\left(\omega_q,\theta_p\right). \tag{4.16}$$

Notice that $R\left[\mathbf{h}_{\text{tot}}\left(\omega,\theta\right)\right]$ is a convex function, since it is the maximum of convex functions [81].

Since the target function and constraints are all convex, we can solve the mixed-integer convex optimization problem

$$\begin{aligned} \min_{\mathbf{s},\mathbf{h}_{\text{tot}}(\omega,\theta)} \quad & R\left[\mathbf{h}_{\text{tot}}\left(\omega,\theta\right)\right] \\ \text{s.t.} \quad & \mathcal{C}_1\left[\mathbf{s}\right], \mathcal{C}_2\left[\mathbf{s}\right], \mathcal{C}_3\left[\mathbf{h}_{\text{tot}}\left(\omega,\theta\right)\right], \\ & \mathcal{C}_4\left[\mathbf{h}_{\text{tot}}\left(\omega,\theta\right)\right], \mathcal{C}_5\left[\mathbf{s},\mathbf{h}_{\text{tot}}\left(\omega,\theta\right)\right]. \end{aligned} \tag{4.17}$$

The non-zero elements of the optimal binary vector $\mathbf{s}^*$ yield the optimal microphone locations $\mathbf{x}^*$. The non-zero elements of the optimal coefficients $\mathbf{h}_{\text{tot}}^*\left(\omega,\theta\right)$ yield the optimal coefficients $\mathbf{h}^*\left(\mathbf{x}^*,\omega,\theta\right)$.

4.2 Coefficient Post Processing

Once $\mathbf{x}^*$ is found, the coefficients $\mathbf{h}^*\left(\mathbf{x}^*,\omega,\theta\right)$ are chosen so that the worst-case directivity index in the ROI is maximized. Thus, beamformers directed toward other directions in the ROI may not yield the best possible directivity. To circumvent this, given $\mathbf{x}^*$, a post-processing scheme is introduced.

The post-processed coefficients must have sufficient WNG and maximize the DF. Given the geometry $\mathbf{x}^*$, this can be done by finding the robust superdirective beamformer

$$\mathbf{h}_{\epsilon}\left(\mathbf{x}^*,\omega,\theta\right) = \frac{\mathbf{\Gamma}_{\epsilon}^{-1}\left(\mathbf{x}^*,\omega\right)\mathbf{d}\left(\mathbf{x}^*,\omega,\theta\right)}{\mathbf{d}^H\left(\mathbf{x}^*,\omega,\theta\right)\mathbf{\Gamma}_{\epsilon}^{-1}\left(\mathbf{x}^*,\omega\right)\mathbf{d}\left(\mathbf{x}^*,\omega,\theta\right)} \tag{4.18}$$

where

$$\mathbf{\Gamma}_{\epsilon}\left(\mathbf{x}^*,\omega\right) = \mathbf{\Gamma}\left(\mathbf{x}^*,\omega\right) + \epsilon\mathbf{I}_M, \tag{4.19}$$

$\mathbf{I}_M$ is the identity matrix of dimensions $M \times M$, and ϵ is a tradeoff parameter between WNG and DF.

We can decompose $\mathbf{\Gamma}\left(\mathbf{x}^*,\omega\right)$ as

$$\mathbf{\Gamma}\left(\mathbf{x}^*,\omega\right) = \mathbf{Q}\left(\mathbf{x}^*,\omega\right)\mathbf{\Lambda}\left(\mathbf{x}^*,\omega\right)\mathbf{Q}^T\left(\mathbf{x}^*,\omega\right) \tag{4.20}$$

where $\mathbf{\Lambda} = \text{diag}\left[\lambda_1,\lambda_2,...,\lambda_M\right]$ is the eigenvalue matrix such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_M$, and $\mathbf{Q}$ is the eigenvector matrix. As in [41], a robust superdirective beamformer that

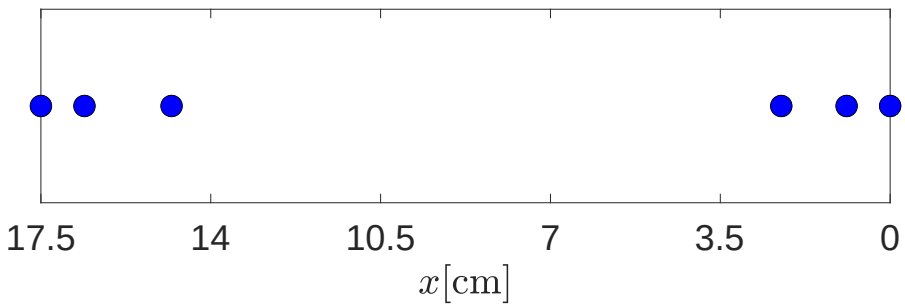

Figure 4.1: Optimal array geometry for $M = 6$, $d_c = 0.5$ cm, $A = 17.5$ cm, $\theta_H = 30°$, $f_L = 2$ kHz, $f_H = 6$ kHz, and $\delta = -10$ dB.

maintains sufficient WNG can be found for some ϵ in

$$0 \leq \epsilon \leq \frac{\lambda_1 - \sqrt{M/\delta}\lambda_M}{\sqrt{M/\delta} - 1}. \tag{4.21}$$

Thus, for any ω and θ, we can run a bisection search on ϵ in this range to find the robust superdirective beamformer that yields the highest directivity yet has sufficient WNG. Subsequently, the beamformer is normalized so that the distortionless constraint is met.

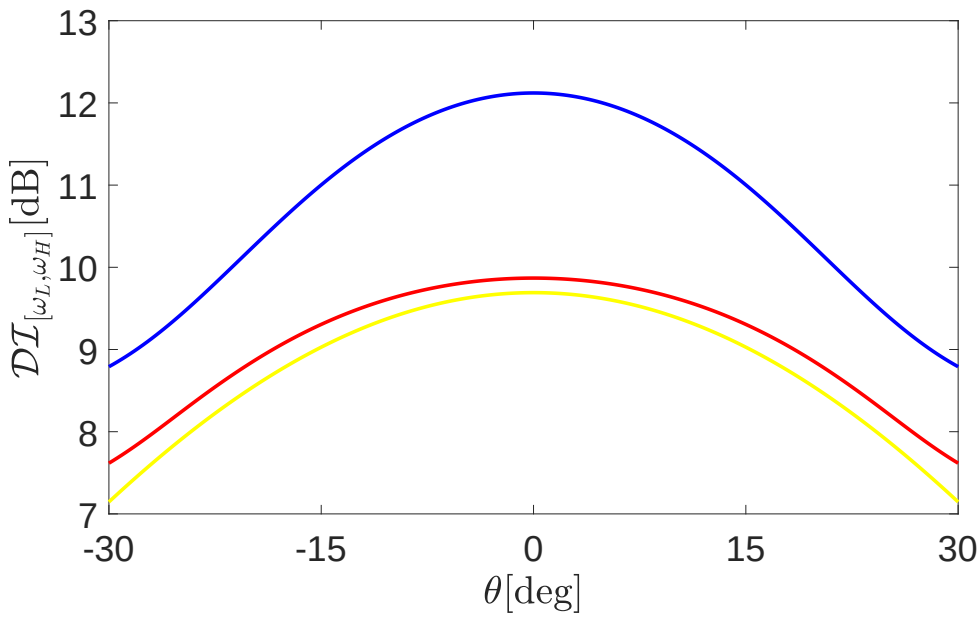

Figure 4.2: Directivity index as function of θ for the competing methods. The blue, red, and yellow lines mark the proposed, ULA, and dense geometries, respectively. $M = 6$, $d_c = 0.5$ cm, $A = 17.5$ cm, $\theta_H = 30°$, $f_L = 2$ kHz, $f_H = 6$ kHz, and $\delta = -10$ dB.

4.3 Simulations

To solve the mixed-integer convex problem in (4.17), the MATLAB CVX toolbox [82] is used with the MOSEK [83] solver. We search for the optimal placement of $M = 6$ microphones on an aperture of length $A = 17.5$ cm where microphones are separated by at least $d_c = 0.5$ cm. Frequencies from $f_L = 2$ kHz to $f_H = 6$ kHz and look directions up to $\theta_H = 30°$ are considered. Minimum WNG is set to $\delta = -10$ dB. Placements, frequencies, and look directions are sampled by $N = 40$, $Q = 15$, and $P = 15$, respectively. Our results are compared with a ULA geometry spread on all A, and the most dense feasible geometry (i.e., ULA with spacing d_c). The compared geometries also use M microphones. Theoretically, the dense geometry has an advantage in the endfire direction [4]. Coefficient post-processing, as described in Section 4.2 was applied to all geometries.

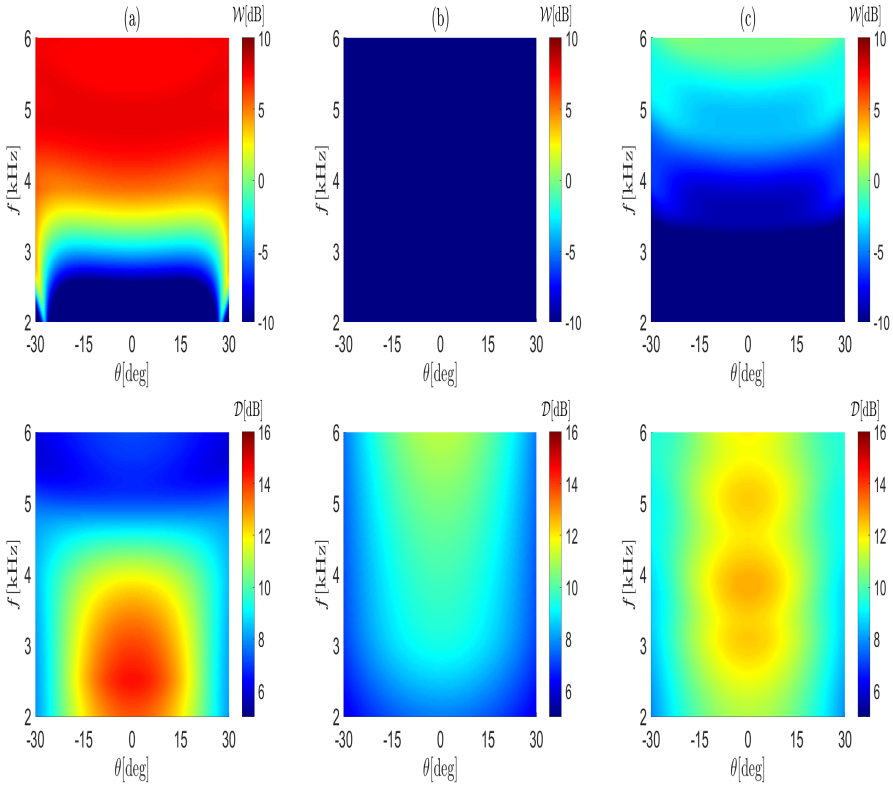

Figure 4.3: WNG and DF as function of f and θ for the competing methods: (a) ULA geometry, (b) dense geometry, and (c) proposed geometry. $M = 6$, $d_c = 0.5$ cm, $A = 17.5$ cm, $\theta_H = 30°$, $f_L = 2$ kHz, $f_H = 6$ kHz, and $\delta = -10$ dB.

Figure 4.1 shows the optimal geometry $\mathbf{x}^*$. The resulting positions of the microphones are dense near the edges, and a large gap is present in the middle. The reasoning is that some microphones are placed close to each other to avoid low directivity due to spatial aliasing in high frequencies. On the other hand, some microphones are placed far from each other to maintain high spatial resolution for lower frequencies.

In Figure 4.2, the broadband directivity index as a function of look direction θ is shown for the competing methods. The proposed method has the highest minimum directivity index across $\theta \in \Theta$. Most importantly, superior performance is also achieved in any look direction by itself.

Figure 4.3 illustrates the WNG and DF as a function of frequency f and look direction θ for the competing methods. Using the ULA, a high directivity is achievable for low frequencies, yet the directivity deteriorates for high frequencies. This is due to spatial aliasing; the high WNG can no longer be exchanged for the sake of directivity. Considering the dense geometry, high directivity can be obtained for high frequencies, yet it is still comparatively low. This happens because the WNG is at its lowest possible value, a well-known problem associated with DMAs. The optimal geometry achieves high directivity across all frequencies, thereby maximizing the broadband directivity index.

4.4 Summary

We have presented an algorithm that finds the optimal microphone locations for broadband directivity in a ROI. Our method places some microphones closely to avoid spatial aliasing in high frequencies, and sets others further apart for spatial resolution in lower frequencies. We have shown that our design outperforms standard designs considering the worst-case look direction. Furthermore, excellent performance is achieved over all possible look directions as well. For simplicity, we have demonstrated our approach for a nonuniform linear geometry, but it can be extended to other geometries, including two and three-dimensional arrays.

Chapter 5

Conclusions

5.1 Summary

In this research, we designed beamformers for AEC and ROI beamforming. In both designs, the acoustic paths from the speech sources to the array are unknown. For AEC, we estimated the steering vectors of the far-end and near-end talkers by utilizing multiple frames. Then, we adopted the LCMV framework to eliminate the echo component, preserve the desired component, and reduce background noise. For ROI beamforming, rather than estimating the direction of the desired source, we maximized the broadband directivity index in the worst-case look direction in the ROI. The optimal array geometry was found regardless of the look direction, and beamforming coefficients were found for each possible look direction. Furthermore, a broadband range of frequencies was considered.

In Chapter 3 we presented an AEC scheme that combines both multiple sensors and multiple frames, and can operate in a dynamic environment under challenging SER conditions. A traditional AEC is implemented by subtracting the microphone signal by a filter output, where the filter input is the reference loudspeaker and the filter aims to portray the far-end acoustic path. While this approach may work in some simulated environments it may not work well in a realistic environment. A double-talk detector must be utilized, and if a double-talk period is missed only for a small period of time, a mismatch between the filter and the acoustic path can be present. The proposed approach does not utilize a double-talk detector. It is capable of beamformer adaptation in periods where both far-end and near-end talkers are active. The acoustic paths were estimated by finding the least-squares solution to a linear system of equations. The number of equations and variables in this system is dependent on the number of sensors used and the number of frames used. To overcome the impact of nonlinear loudspeaker distortion and background noise, more equations may be utilized. Our design considers an arbitrary array geometry. To imitate a real-life scenario with challenging SER, a speakerphone was simulated, so that significant acoustic coupling is present. Simulations were run in a dynamic, reverberant, and noisy environment during double-talk.

The proposed design achieved higher ERLE and PESQ, as well as lower DI, compared to existing NLMS-based designs.

In Chapter 4 we searched for the optimal geometry for ROI beamforming. Existing geometry optimization methods are based on genetic, or greedy-based algorithms. Instead, we proposed to find the best geometry by formulating a convex optimization problem. This allowed us to find the global optimum for our problem. To guarantee good performance in any scenario, and since the source may be located anywhere within the ROI, the worst-case look direction was considered. The physical volume of microphones was taken into account so that microphones have sufficient space when placed adjacently in a nonuniform linear array geometry. For sufficient noise robustness, WNG adheres to a constraint too. A grid of candidate microphone locations was considered for microphone placements. Also, a discrete set of frequencies in a broadband frequency range, and a discrete set of look directions in the ROI, were considered. The array geometry and beamformer coefficients were found simultaneously in the formulated problem. Later, a post-processing scheme was introduced to improve the directivity further. The advantage of the proposed method compared to existing geometries was demonstrated. ULAs may experience low directivity in higher frequencies due to spatial aliasing, and DMAs suffer from low WNG since their high directivity comes at the expense of noise robustness. Utilizing the proposed method, we showed that a good compromise is achieved between directivity and noise robustness. Some of the sensors are placed close together, and some are placed further apart. This enabled the obtained geometry to trade WNG, whenever possible, for higher broadband directivity.

5.2 Future Research

In this thesis, we proposed two novel beamformer design methods. One may be integrated into multi-microphone AEC systems, and the other may be integrated into any multi-microphone system with an ROI. While the proposed methods achieve higher performance compared to existing ones, they may be improved further. Some ideas left for future research are:

1. When deriving the AEC beamformer, the acoustic paths from the far-end source to the array were estimated with a least-squares solution to a set of linear equations. This approach works well under the assumption that background noise and nonlinear loudspeaker distortion components are relatively small. In the case they are not, however, these components should be modeled. Therefore, an extended set of equations, that takes into account an explicit mathematical model of these components, may improve performance.

2. The MTF approximation [66] was used, i.e. we assumed that the received signal can be modeled by the source signal multiplied by a TF. This approximation

holds only for the case where the RIR length is significantly shorter than the STFT window length. A more accurate model, such as the cross-multiplicative transfer function (CMTF) approximation [69], may be used. This in turn will yield different inter-frame and inter-sensor relations. In this way, better echo cancellation may be possible.

3. When deriving the ROI beamformer, we limited the search to nonuniform linear arrays. Exploring different types of geometries, such as rectangular or circular arrays, might better performance. When such geometries are considered, an ROI defined by both azimuth and elevation angles can be taken into account. Therefore, this should enable us to find the optimal array for more complex ROIs.

4. To obtain a solvable convex problem, we sampled the search space of sensor locations and also sampled specific frequencies and look directions. Such discretization is justified only when the search space is sampled well. In scenarios where a large aperture is considered, the propagated signal contains a large frequency band, or the ROI is large, more samples should be taken into account. In this scenario, the formulated problem significantly grows in terms of complexity and runtime. In this case, a different approach may be necessary, such as learning-based methods or deep neural networks (DNNs).

Bibliography

[1] B. Widrow and S. D. Stearns, *Adaptive Signal Processing.* Prentice-Hall, 1985.

[2] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach.* Wiley, 2004.

[3] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications.* Springer-Verlag, 2001.

[4] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing.* Wiley, 2018.

[5] J. Benesty, M. M. Sondhi, and Y. A. Huang, *Springer Handbook of Speech Processing.* Springer, 2008.

[6] M. M. Sondhi, "An adaptive echo canceller," *The Bell System Technical Journal*, vol. 46, pp. 497–511, 1967.

[7] W. Klippel, "Loudspeaker nonlinearities – causes, parameters, symptoms," *Audio Engineering Society Convention*, vol. 54, pp. 907–939, 2006.

[8] I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using wavenet encoder and residual networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 265–274, 2019.

[9] H. Zhang and D. Wang, "A deep learning approach to multi-channel and multi-microphone acoustic echo cancellation," in *Proc. Interspeech*, pp. 1139–1143, 2021.

[10] T. Haubner and W. Kellermann, "Deep learning-based joint control of acoustic echo cancellation, beamforming and postfiltering," in *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 752–756, 2022.

[11] H. Zhang, S. Kandadai, H. Rao, M. Kim, T. Pruthi, and T. Kristjansson, "Deep adaptive AEC: hybrid of deep learning and adaptive acoustic echo cancellation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 756–760, 2022.

[12] Y. Tsai, Y. Hsu, and M. R. Bai, "Acoustic echo suppression using a learning-based multi-frame minimum variance distortionless response (MFMVDR) filter," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022.

[13] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1071–1086, 2009.

[14] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1420–1434, 2009.

[15] R. Berkun, I. Cohen, and J. Benesty, "Combined beamformers for robust broadband regularized superdirective beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 877–886, 2015.

[16] E. A. P. Habets, J. Benesty, S. Gannot, and I. Cohen, "The MVDR Beamformer for Speech Enhancement," in *Speech Processing in Modern Communication: Challenges and Perspectives*. Springer, 2010.

[17] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 219–222, 1997.

[18] G. Reuven, S. Gannot, and I. Cohen, "Joint acoustic echo cancellation and transfer function GSC in the frequency domain," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, pp. 412–415, 2004.

[19] G. Reuven, S. Gannot, and I. Cohen, "Multichannel acoustic echo cancellation and noise reduction in reverberant environments using the transfer-function GSC," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 81–84, 2007.

[20] T. Burton and R. Goubran, "A new structure for combining echo cancellation and beamforming in changing acoustical environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 77–80, 2007.

[21] G. Reuven, S. Gannot, and I. Cohen, "Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller," *Speech Communication*, vol. 49, pp. 623–635, 2007.

[22] A. Cohen, A. Barnov, S. Markovich-Golan, and P. Kroon, "Joint beamforming and echo cancellation combining QRD based multichannel AEC and MVDR for reducing noise and non-linear echo," in *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 6–10, 2018.

[23] Z. Yermeche, N. Grbic, and I. Claesson, "Blind subband beamforming with time-delay constraints for moving source speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2360–2372, 2007.

[24] S. Markovich Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 201–204, 2010.

[25] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Online localization and tracking of multiple moving speakers in reverberant environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 88–103, 2019.

[26] E. A. P. Habets, S. Gannot, I. Cohen, and P. C. W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 1433–1451, 2008.

[27] N. Cohen, G. Hazan, B. Schwartz, and S. Gannot, "An online algorithm for echo cancellation, dereverberation and noise reduction based on a kalman-EM method," *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 1–17, 2021.

[28] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 273–276, 2011.

[29] Y. A. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 1256–1269, 2012.

[30] D. Fischer and T. Gerkmann, "Single-microphone speech enhancement using MVDR filtering and wiener post-filtering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 201–205, 2016.

[31] D. Fischer and S. Doclo, "Robust constrained MFMVDR filtering for single-microphone speech enhancement," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 41–45, 2018.

[32] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for single-microphone speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8443–8447, 2021.

[33] H. Huang, J. Benesty, J. Chen, K. Helwani, and H. Buchner, "A study of the MVDR filter for acoustic echo suppression," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 615–619, 2013.

[34] K. Helwani, H. Buchner, J. Benesty, and J. Chen, "A single-channel MVDR filter for acoustic echo suppression," *IEEE Signal Processing Letters*, vol. 20, pp. 351–354, 2013.

[35] K. Helwani, *Adaptive Identification of Acoustic Multichannel Systems Using Sparse Representations*. Springer, 2015.

[36] H. Huang, C. Hofmann, W. Kellermann, J. Chen, and J. Benesty, "A multiframe parametric wiener filter for acoustic echo suppression," in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[37] H. Huang, C. Hofmann, W. Kellermann, J. Chen, and J. Benesty, "Multiframe echo suppression based on orthogonal signal decompositions," in *Proc. Speech Communication; ITG Symposium*, pp. 287–291, 2016.

[38] E. A. P. Habets, J. Benesty, and J. Chen, "Multi-microphone noise reduction using interchannel and interframe correlations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 305–308, 2012.

[39] Z. Zhang, Y. Xu, M. Yu, S. X. Zhang, L. Chen, D. S. Williamson, and D. Yu, "Multi-channel multi-frame ADL-MVDR for target speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3526–3540, 2021.

[40] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for binaural noise reduction," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022.

[41] X. Chen, C. Pan, J. Chen, and J. Benesty, "Planar array geometry optimization for region sound acquisition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 756–760, 2021.

[42] K. Chen, X. Yun, Z. He, and C. Han, "Synthesis of sparse planar arrays using modified real genetic algorithm," *IEEE Transactions on Antennas and Propagation*, vol. 55, pp. 1067–1073, 2007.

[43] R. Haupt, "Thinned arrays using genetic algorithms," *IEEE Transactions on Antennas and Propagation*, vol. 42, pp. 993–999, 1994.

[44] F. L. Courtois, J. H. Thomas, F. Poisson, and J. C. Pascal, "Genetic optimisation of a plane array geometry for beamforming. application to source localization in a high speed train," *Journal of Sound and Vibration*, vol. 371, pp. 78–93, 2016.

[45] J. Yu, F. Yu, and Y. Li, "Optimization of microphone array geometry with evolutionary algorithm," *Journal of Computers*, vol. 8, pp. 200–207, 2013.

[46] Z. Li, K. F. C. Yiu, and Z. Feng, "A hybrid descent method with genetic algorithm for microphone array placement design," *Applied Soft Computing*, vol. 13, pp. 1486–1490, 2013.

[47] A. Frank and I. Cohen, "Constant-beamwidth kronecker product beamforming with nonuniform planar arrays," *Frontiers in Signal Processing*, vol. 2, pp. 1–17, 2022.

[48] Y. Gershon, Y. Buchris, and I. Cohen, "Greedy sparse array design for optimal localization under spatially prioritized source distribution," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4607–4611, 2020.

[49] Y. Buchris, I. Cohen, J. Benesty, and A. Amar, "Joint sparse concentric array design for frequency and rotationally invariant beampattern," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1143–1158, 2020.

[50] A. T. Parsons, "Maximum directivity proof for three-dimensional arrays," *The Journal of the Acoustical Society of America*, vol. 82, pp. 179–182, 1987.

[51] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 393–398, 1986.

[52] D. Y. Levin, S. Markovich-Golan, and S. Gannot, "Near-field superdirectivity: an analytical perspective," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1661–1674, 2021.

[53] D. Y. Levin, E. A. P. Habets, and S. Gannot, "On the average directivity factor attainable with a beamformer incorporating null constraints," *IEEE Signal Processing Letters*, vol. 22, pp. 2122–2126, 2015.

[54] Z. Zhong, M. Shakeel, K. Itoyama, K. Nishida, and K. Nakadai, "Assessment of a beamforming implementation developed for surface sound source separation," in *Proc. IEEE/SICE International Symposium on System Integration*, pp. 369–374, 2021.

[55] Z. Zhong, K. Itoyama, K. Nishida, and K. Nakadai, "Design and assessment of a scan-and-sum beamformer for surface sound source separation," in *Proc. IEEE/SICE International Symposium on System Integration*, pp. 808–813, 2020.

[56] M. Taseska and E. A. P. Habets, "Spotforming: spatial filtering with distributed arrays for position-selective sound acquisition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1291–1304, 2016.

[57] J. Martinez, N. Gaubitch, and W. B. Kleijn, "A robust region-based near-field beamformer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2494–2498, 2015.

[58] M. Taseska and E. A. P. Habets, "Spotforming using distributed microphone arrays," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–4, 2013.

[59] H. Jo, Y. Park, and Y. S. Park, "A sound telescope: a control of zone of interest," in *Proc. International Congress on Acoustics (ICA)*, pp. 1–5, 2010.

[60] A. Davis, S. Y. Low, S. Nordholm, and N. Grbic, "A subband space constrained beamformer incorporating voice activity detection [speech enhancement applications]," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 65–68, 2005.

[61] J. Chen, L. Shue, H. Sun, and K. Phua, "An adaptive microphone array with local acoustic sensitivity," in *Proc. IEEE International Conference on Multimedia and Expo*, pp. 1–4, 2005.

[62] N. Grbić and S. Nordholm, "Soft constrained subband beamforming for hands-free speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 885–888, 2002.

[63] A. Kleiman, I. Cohen, and B. Berdugo, "Constant-beamwidth beamforming with concentric ring arrays," *Sensors*, vol. 21, pp. 1–19, 2021.

[64] T. Long, I. Cohen, B. Berdugo, Y. Yang, and J. Chen, "Window-based constant beamwidth beamformer," *Special Issue of Sensors on Speech, Acoustics, Audio, Signal Processing and Applications in Sensors*, vol. 19, pp. 1–20, 2019.

[65] O. Rosen, I. Cohen, and D. Malah, "FIR-based symmetrical acoustic beamformer with a constant beamwidth," *Signal Processing*, vol. 130, pp. 365–376, 2017.

[66] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, pp. 337–340, 2007.

[67] I.-T. P.862, "Perceptual evaluation of speech quality (pesq): an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, itu-t recommendation p.862." Available: https://www.itu.int/rec/T-REC-P.862, 2001.

[68] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 451–459, 2004.

[69] Y. Avargel and I. Cohen, "Adaptive system identification in the short-time fourier transform domain using cross-multiplicative transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 162–173, 2008.

[70] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 546–555, 2009.

[71] Y. Avargel and I. Cohen, "Modeling and identification of nonlinear systems in the short-time fourier transform domain," *IEEE Transactions on Signal Processing*, vol. 58, pp. 291–304, 2010.

[72] D. Cherkassky and S. Gannot, "Successive relative transfer function identification using blind oblique projection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 474–486, 2019.

[73] H. Liu, G. Xu, and L. Tong, "A deterministic approach to blind identification of multi-channel FIR systems," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 581–584, 1994.

[74] A. Aissa-El-Bey, M. Grebici, K. Abed-Meraim, and A. Belouchrani, "Blind system identification using cross-relation methods: further results and developments," in *Proc. Seventh International Symposium on Signal Processing and Its Applications (ISSPA)*, pp. 649–652, 2003.

[75] J. Chen, J. Benesty, and I. Cohen, *Design of Circular Differential Microphone Arrays.* Springer, 2015.

[76] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus." Available: https://catalog.ldc.upenn.edu/LDC93s1, 1993.

[77] S. Thompson, "Nonlinear modeling of a moving coil loudspeaker." Available: https://www.mathworks.com/matlabcentral/fileexchange/121263-nonlinear-modeling-of-a-moving-coil-loudspeaker, 2022.

[78] S. Thompson, "Acoustical domain for simscape." Available: https://www.mathworks.com/matlabcentral/fileexchange/109029-acoustical-domain-for-simscape, 2023.

[79] E. A. P. Habets, "Room impulse response generator." Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, 2010.

[80] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 617–631, 2007.

[81] S. P. Boyd and L. Vandenbergh, *Convex Optimization.* Cambridge university press, 2004.

[82] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming." Available: http://cvxr.com/cvx/, 2014.

[83] M. ApS, "The MOSEK optimization toolbox for Matlab." Available: https://docs.mosek.com/10.0/toolbox/index.html, 2019.

העושות גם הן שימוש בכמה מסגרות במישור התמרת הפורייה בזמן קצר. הגישה שלנו עדיפה גם מבחינת ביטול ההד וגם מבחינת שימור הדובר הרצוי.

לאחר מכן, נציג שיטה למציאת הגאומטריה האופטימאלית של מערכי מיקרופונים בשביל תכנון אלומה לכיוון איזור עניין. לעתים קרובות וביישומים רבים, על אף שאין בידינו את המיקום המדויק של המקור המבוקש אנו יודעים לשייך את המיקום לאיזור כל שהוא. כלומר אנחנו יודעים שהמקור תחום להיות באיזור ידוע מראש. איזור זה נקרא איזור העניין. במקרה זה, כדי להעביר את האות המבוקש, ניתן להשתמש בשתי טכניקות אפשריות. טכניקה אחת היא יצירת אלומה עם רוחב אונה קבוע על תחום תדרים. טכניקה זו פשוטה למימוש, אך אם יש עוד מקורות שברצוננו לסנן באיזור העניין, היא לא תסנן מקורות נוספים אלו. טכניקה שנייה היא קודם לשערך את כיוון ההגעה של המקור המבוקש במדויק, ולאחר מכן ליצור אלומה שתעביר אך ורק את המקור הזה. טכניקה זו מורכבת יותר למימוש אך דואגת להעביר אך ורק את המקור המבוקש. בעבודה זו, אנו מחפשים את הגאומטריה האופטימאלית לשימוש בטכניקה השנייה. לכן, מציאת הגאומטריה חייבת לקחת בחשבון מספר אלומות כאשר כל אלומה מכוונת לכיוון אחר באיזור העניין, מה שהופך את הבעיה למאתגרת מאוד. אנו מתחילים מלהציג את הבעיה כבעיה קמורה, ולאחר מכן פותרים את הבעיה, דבר המאפשר לנו להגיע לגאומטריה האופטימלית. למרות שהסביבה ומיקומי המקורות עלולים להשתנות עם הזמן, בדרך כלל הגאומטריה, שבעלת השפעה מכרעת על ביצועי המערכת, נשארת קבועה. על כן, למציאת המיקומים האופטימאליים של המיקרופונים חשיבות רבה. בניגוד לעבודות קודמות, השיטה המוצעת לוקחת בחשבון מקורות רחבי סרט כמו למשל מקורות דיבור. הגישה שומרת על חסינות גבוהה מרעש לבן בעזרת הצבת סף למדד הגבר הרעש הלבן, ולוקחת בחשבון את גודלם הפיזי של המיקרופונים כך שלא יושמו קרוב מדי אחד לשני. ניתן לפרש ולהסביר בצורה מעניינת את הגאומטריה המתקבלת מהרצת הסימולציות ולראות איך המרחקים בין המיקרופונים משפיעים על ביצועי הכיווניות והגבר הרעש של האלומות כתלות בתדר וכיוון ההגעה. בהשוואה לגאומטריות סימטריות המקובלות לשימוש, אנחנו מקבלים כיווניות גבוהה יותר כלפי כל כיוון באיזור העניין.

תקציר

תזה זו עוסקת בשני נושאים. הנושא הראשון הוא ביטול הד אקוסטי רב ערוצי בסביבה דינאמית בעזרת תכנון אלומה, והנושא השני הוא מציאת הגאומטרייה האופטימאלית של מערך מיקרופונים בשביל תכנון אלומה לאיזור עניין. המשותף לשני הנושאים המוצגים בחיבור זה הוא השימוש באלומה. כשברצוננו לבטל הד אקוסטי, בדרך כלל נעשה שימוש במיקרופון אחד אך אפשר לשפר ביצועים עם שימוש במספר מיקרופונים ותכנון אלומה העושה שימוש בהם. בנוסף, אם ברצוננו להגביר אות שנמצא באיזור עניין, ניתן לעשות זאת בעזרת תכנון אלומה גם כן. יש שימוש נרחב באלומות ביישומים טכנולוגיים רבים. למשל, בתחום של עבוד אותות אקוסטי, אלומות משמשות לסינון מרחבי, ביטול הד אקוסטי, לוקליזציה של מקורות, הערכת כיווני הגעה של מקורות, הפרדת מקורות ועוד. הרעיון הבסיסי העומד במרכז השימוש באלומה מרחבית הוא דגימת המרחב. דגימה זו מתבצעת בעזרת מספר מיקרופונים הפרושים במרחב בצורה גאומטרית כל שהיא. כיוון שגלים אקוסטיים מתפשטים במרחב החל ממקור הפליטה שלהם והלאה, הגל מגיע לכל מיקרופון בזמן אחר, כך שהפרש זמני ההגעה של האות בין המיקרופונים תלוי בכיוון התפשטות הגל. התופעה הפיסיקלית הזו מאפשרת לנו לעשות שימוש בכלים מתוך תחום עבוד האותות כדי לסנן את האותות שמתפשטים במרחב כתלות בכיוון התפשטותם. שימוש בטכניקה זו מאפשרת הגברת אותות מכיוונים מסוימים והנחתת אותות מכיוונים אחרים.

ראשית, נציג שיטה לביטול הד אקוסטי רב ערוצי בסביבה דינאמית. ביטול הד אקוסטי נדרש במגוון יישומים כגון שיחות טלפוניות, מערכות שיחות ועידה, יישומי רכב, מוצרים מסחריים שעליהם להגיב לפקודות שמע ועוד. השיטה המוצגת בחיבור זה מבטלת הד אקוסטי בעזרת אלומה הפועלת על מספר מיקרופונים, היכולה לפעול גם כאשר הסביבה משתנה. כלומר, יתכן שמיקום הרמקול, הדובר, או המיקרופונים משתנים עם הזמן. בנוסף, השיטה לא עושה שימוש בגלאי דיבור-כפול. האלומה פועלת במישור התמרת הפורייה לזמן קצר, ומקדמי האלומה מסתגלים מחדש עם כל מסגרת זמן חדשה כך שתוקפם נשמר למסגרת הזמן הבאה. בכל מסגרת זמן נעשה שימוש במספר מסגרות הזמן האחרונות כדי לשערך את המסלולים האקוסטיים, כאשר ההנחה היא שבמסגרות אלו הסביבה סטטית. שיטת השערוך מבוססת על מזעור מיצוע של שגיאת ריבועים כאשר השגיאה הריבועית מחושבת כסטייה ממשוואות שאמורות להתקיים באופן תאורטי. התהליך עושה שימוש הן באות הרמקול והן באותות המיקרופונים במערך. בנוסף לביטול ההד, האלומה מסוגלת למזער את השפעת רעשי הרקע במיקרופונים על מוצא המערכת. כדי לבחון את השיטה המוצעת, הורצו סימולציות בסביבה אתגרית הכוללת דיבור כפול, רעש רקע, עיוות רמקול, ויחס אות להד נמוך האופייני להתקן משולב רמקול - מיקרופונים. הסימולציות מדגימות מצד אחד ביטול מוצלח של רכיב ההד, ומצד שני שימור מוצלח של רכיב הדובר הרצוי. השפעת פרמטרי האלגוריתם על ביצועי המערכת גם נבחנת. כדי להראות את עדיפות השימוש של הגישה המוצעת, נערכת השוואה בין השיטה ושיטות מתחרות

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

התוצאות של חיבור זה פורסמו כמאמרים מאת המחבר ושותפיו למחקר בכתב-עת ובכנס במהלך תקופת מחקר המגיסטר של המחבר.

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 תודה רבה למנחי המחקר שלי, פרופ' ישראל כהן ודר' ברוך ברדוגו. עבודה זו היא תוצאה של תמיכתם וסובלנותם. הם לימדו אותי כיצד לגשת למחקר ולחשוב בצורה ביקורתית. הם היו איתי בכל מצב באש ובמים, ועל כך, אני מודה להם מעומק לבי.

ברצוני גם להודות למשפחתי, ובמיוחד לבת זוגי גלית. עבודה זו לא הייתה מתאפשרת ללא תמיכתם האין סופית. בזכותכם, אני מי שאני. לבסוף, ארצה לומר תודה לחתולתנו המנוחה, ג'זמין, על שאפשרה לעבוד על תזה זו ברוח טובה, ועל שפשוט הייתה שם.

ביטול הד אקוסטי רב ערוצי על בסיס תכן אלומה

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים בהנדסת חשמל

יובל קונפורטי

הוגש לסנט הטכניון — מכון טכנולוגי לישראל
כסלו התשפ"ד חיפה דצמבר 2023

ביטול הד אקוסטי רב ערוצי על בסיס תכן אלומה

יובל קונפורטי