Speech Enhancement Based on Adaptive Line Enhancer

Aviva Atkins

Speech Enhancement Based on Adaptive Line Enhancer

Research Thesis

In Partial Fulfillment of the Requirements
for the Degree of Master of Science in Electrical Engineering

Aviva Atkins

Submitted to the Senate of the Technion - Israel Institute of Technology

Shvat 5780 Haifa February 2020

The research thesis was done under the supervision of Professor Israel Cohen from the Faculty of Electrical Engineering at the Technion.

Acknowledgments

First and foremost, I would like to express my deepest gratitude and appreciation to Professor Israel Cohen for his unwavering guidance, support, and patience throughout every stage of this research.
I'm also extremely grateful to Professor Jacob Benesty for his professional support and valuable contribution throughout my research.
I would like to thank Professor Ronen Talmon and Professor Amir Averbuch for their review of my research.
I also had the pleasure of working with Yuval Ben-Hur on Chapter 5, and would like to extend my thanks for the collaboration and stimulating discussions.
Special thanks to my managers Max Perri and Benjamin Pilatovsky as well as my colleagues at MVS HW Intel for their support and understanding throughout the period of my research.
Last but not least, I would like to thank my family, my beloved husband, and my friends for their constant support, love and encouragement. This accomplishment would not have been possible without them.

The generous financial support of the Technion, the Israel Science Foundation (grant no. 576/16), and the ISF-NSFC joint research program (grant No. 2514/17) is gratefully acknowledged.

Publications

The research thesis is based on the following publications:

1. A. Atkins, I. Cohen and J. Benesty, Adaptive Line Enhancer for Non-Stationary Harmonic Noise Reduction, submitted to Computer Speech and Language, 2019.

2. A. Atkins and I. Cohen, Speech Enhancement Using ARCH model, in Proceedings of IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), 2016.

3. A. Atkins, Y. Ben-Hur, I. Cohen and J. Benesty, Robust superdirective beamformer with optimal regularization, in Proceedings of IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), 2016.

Contents

Contents (Continued)

List of Figures

List of Figures (Cont.)

List of Tables

Abstract

In a real-life acoustic system noise is unavoidable. From the microphone self noise, to additive noise from other sources, to reverberations which are reflections of your own speech from walls and other obstacles arriving at different delays to the microphone, to echos caused by the coupling of the microphone and loudspeaker, the noise degrades the quality and the intelligibility of the speech signal. The process of suppressing the additive noise, to "clean" the noisy speech signal and improve its quality and intelligibility, is known as noise reduction or also as speech enhancement. The speech enhancement problem has been extensively studied, and developed methods are widely used in numerous applications such as telecommunications, teleconferencing, hearing aids, human-machine interfaces, and more; however, it remains a challenging problem to this day, as in many of the methods it is possible to improve the quality of the signal and reduce the noise but at the expense of some distortion to the signal. The most difficult scenario is when only a single microphone is available, known as the single-channel case, and though multiple microphones and microphone arrays are becoming more popular as it is getting easier to miniaturize the structures, the single microphone is still used.

One of the most difficult types of noise to reduce is nonstationary noise, i.e., noise that changes quickly over time, though if it has an underlying structure, such as being composed of harmonics, it is possible to exploit this in the noise reduction. In this thesis we develop a method to reduce harmonic nonstationary noise for the single channel case. We propose the use of a frequency-domain adaptive line enhancer (ALE), which consists of a delay element to create a reference signal from the input signal and an adaptive filter, where we use a combination of a forward adaptive filter and a backward non-causal filter with a harmonic noise indicator. We apply our proposed combined method to synthetic and real noise signals and demonstrate that our proposed method yields improved performance compared to other methods.

As noise signals typically contain wide-band noise as well, our method to remove harmonic noise can be followed by a conventional spectral domain noise reduction method to remove the remaining wide-band noise. In this thesis we investigate the use of the autoregressive conditional heteroscedasticity (ARCH) model, whose generalized form is used extensively in financial applications, as part of the spectral domain noise reduction algorithm, and compare it to one of the most commonly used methods.

In the final part of the thesis, we consider the multi-microphone case. We use a combined noise field approach and develop a robust superdirective beamformer that enables control of the trade-off between white noise amplification and the directivity factor. We propose a one dimensional search algorithm to find the optimal regularization factor employed in the beamformer and demonstrate by simulations improved performance compared to a recent method.

Glossary

Abbreviations

ALE	Adaptive Line Enhancer
ANC	Adaptive Noise canceling
ARCH	Autoregressive Conditional Heteroscedasticity
BLP	Backward Linear Predictor
BMLNLMS	Backward Mapped L Normalized Least Mean Squares
CASA	Computational Auditory Scene Analysis
CLP	Combined Linear Predictor
CMLNLMS	Combined Mapped L Normalized Least Mean Squares
CNN	Convolutional Neural Network
DD	Decision Directed
DF	Directivity Factor
DNN	Deep Neural Network
DS	Delay and Sum
FLP	Forward Linear Predictor
FMLNLMS	Forward Mapped L Normalized Least Mean Squares
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
iid	independent and identically distributed
iSNR	input Signal to Noise Ratio
KNN	k-Nearest Neighbor
LKR	Log of the Kurtosis Ratio
LMS	Least Mean Squares
LS	Least Squares
LSA	Log Spectral Amplitude
MAS	Minimize and Search
MI	Mutual Information
ML	Maximum Likelihood
MLP	Multilayer Perceptron
MMSE	Minimum Mean Square Error

MS	Minimum Statistics
MSE	Mean Square Error
MVDR	Minimum Variance Distortionless Response
NLMS	Normalized Least Mean Squares
NMF	Non-negative Matrix Factorization
NRR	Noise Reduction Ratio
OD	Orthogonal Decomposition
OM-LSA	Optimally Modified Log Spectral Amplitude
oSNR	output Signal to Noise Ratio
PDF	Probability Distribution Function
PESQ	Perceptual Evaluation of Speech Quality
RLS	Recursive Least Squares
RNN	Recurrent Neural Networks
SA	Spectral Amplitude
SD	Superdirective
SNR	Signal to Noise Ratio
SP	Spectral Power
STFT	Short Time Fourier Transform
STOI	Short Time Objective Intelligibility
STSA	Short Time Spectral Amplitude
ULA	Uniform Linear Array
VAD	Voice Activity Detector
WGN	White Gaussian Noise
WNG	White Noise Gain

Operators and known functions

$(\cdot)^T$	transpose operator
$(\cdot)^*$	conjugate operator
$(\cdot)^H$	conjugate-transpose operator
$(\cdot)_{\|}$	magnitude of complex $(\cdot)$
$(\cdot)_{\angle}$	phase of complex $(\cdot)$
$\mathbf{A}^{-1}$	matrix inverse
$[\mathbf{A}]_{ij}$	the (i,j) element of matrix $\mathbf{A}$
$\mathrm{tr}(\cdot)$	trace
$\mathrm{medianfilt}(\cdot)$	median filter
$E[\cdot]$	mathematical expectation
$H(\cdot)$	entropy
$H(U,W)$	joint entropy of random variables U and W
$H(U\|W)$	conditional entropy of random variable U given random variable W
$I_0(\cdot)$	modified Bessel function of order zero
$I_1(\cdot)$	modified Bessel function of first order
$I(U;W)$	mutual information of random variables U and W
$p(\cdot)$	probability function
$P(\cdot)$	rectifying function
$\Gamma(\cdot)$	Gamma function
$\Phi(a,b;c)$	Confluent hypergeometric function
$\psi(\cdot)$	Digamma function
$\mathcal{O}$	order of complexity
$\mathcal{N}(\mu,\sigma^2)$	Normal distribution with mean μ and variance σ^2

Notations

$\mathbf{0}_{\tau_2}$	vector of zeros with length τ_2
$A_\ell(k)$	spectral magnitude of $X_\ell(k)$
$\hat{A}_\ell(k)$	estimate of the spectral magnitude of $X_\ell(k)$
c	speed of sound
$d(n)$	noise signal
$d_{i,j}$	distance measure between point i and point j
$\mathbf{d}(\omega)$	steering vector at the source direction (θ=0)
$\mathbf{d}(\omega,\theta)$	steering vector
$D_\ell(k)$	STFT of signal $d(n)$
$e(n)$	error signal
$e(\cdot,\cdot)$	error function
$E(k,m)$	error signal of the ALE in the STFT domain
$E_c(k,m)$	ALE combined predictor
$E_b(k,m)$	ALE backward predictor
$E_b^\ell(k,m)$	ALE backward predictor per filter length
$E_f(k,m)$	ALE forward predictor
$E_f^\ell(k,m)$	ALE forward predictor per filter length
$E_V(k,m)$	error signal for noise estimation
f	temporal frequency
$g(U)$	random variable
$\mathbf{g}(k,m)$	filter in the STFT domain for the ALE backward predictor
$G(\cdot,\cdot)$	gain function
$G_{\min}$	lower bound threshold for spectral gain function
$G_i(k,m)$	filter coefficient i in the STFT domain for the ALE backward predictor
$G_p(\cdot,\cdot)$	pth order spectral gain function
$G_{LSA}(\cdot,\cdot)$	LSA spectral gain function
$G_{MMLSA}(\cdot,\cdot)$	multiplicative modified LSA gain function
$G_{OMLSA}(\cdot,\cdot)$	optimally modified LSA gain function
$G_{SP}(\cdot,\cdot)$	spectral power gain function
$G_{STSA}(\cdot,\cdot)$	STSA spectral gain function
$G_W(\cdot,\cdot)$	Wiener spectral gain function
$\mathbf{h}(k,m)$	vector of narrow-band filter coefficients in the STFT domain
$\mathbf{h}(m)$	vector of full-band filter coefficients in the STFT domain
$\mathbf{h}(\omega)$	complex valued linear filter vector in the frequency domain
$\mathbf{h}_v$	filter for the noise estimate
$\mathbf{h}_{v,W}$	Wiener filter for the noise estimate

$\mathbf{h}_W$	Wiener filter
$\mathbf{h}_{\mathrm{DS}}(\omega)$	Delay and Sum beamformer
$\mathbf{h}_{\mathrm{mDF}}(\omega)$	maximum DF beamformer
$\mathbf{h}_{\mathrm{R},\epsilon}(\omega)$	robust Superdirective beamformer
$\mathbf{h}_{\mathrm{SD}}(\omega)$	Superdirective beamformer
$\mathbf{h}_{\alpha}(\omega)$	proposed robust Superdirective beamformer
$\mathbf{h}_{\alpha,\epsilon}(\omega)$	combined beamformer
$H_i(k,m)$	filter coefficient i in the STFT domain
$H_i(\omega)$	filter coefficient i in the frequency domain
$\mathbf{i}_1$	identity filter, with 1 at location 1
I_{thr}	threshold value for MI decision coefficient calculation
$I^A(k)$	MI for the log-magnitude spectrum at frequency k
$I^P(k)$	MI for the phase spectrum at frequency k
$\bar{I}^A(k)$	normalized MI for the log-magnitude spectrum at frequency k
$\bar{I}^P(k)$	normalized MI for the phase spectrum at frequency k
$\bar{I}^{total}(k)$	total MI at frequency k
$\hat{I}^P(k)$	filtered MI for the phase spectrum at frequency k
$\hat{I}^{total}(k)$	filtered total MI at frequency k
$\mathbf{I}_M$	$M \times M$ identity matrix
j	imaginary unit
J	MSE criterion
J_d	MSE criterion of desired speech
J_r	MSE criterion of residuals noise and interference
J_X	speech distortion
k	frequency index in STFT domain
k	amount of neighbors in Section 2.1.2 and Section 2.1.3
k^*	frequency boundary between bands for MI algorithm
K	amount of frequency indexes in STFT domain
ℓ	frame index in STFT domain
ℓ	filter length index in Chapter 3
ℓ'	frame index in STFT domain
ℓ'	filter length index in Chapter 3
L	filter length
m	frame index in STFT domain
m	microphone index in Section 2.3 and Chapter 5
M	amount of frame indexes in STFT domain
M	number of microphones in Section 2.3 and Chapter 5
n	discrete time index
n_1	filter order applied on the normalized phase MI
n_2	filter order applied on the total MI

$n_u(i)$	number of points u_j whose distance to u_i is less than $\epsilon(i)/2$
$n_w(i)$	number of points w_j whose distance to w_i is less than $\epsilon(i)/2$
N	number of samples
$\bar{N}$	number of STFT frames in 3 s block of the signal
p	GARCH model order parameter for standard deviation lag
p	filter length index in Chapter 3
q	GARCH model order parameter for error lag
q_ℓ	probability for speech absence
Q_μ	MI decision coefficient
t	propagation time from the source to microphone 1
u	observation of random variable U
u_i	sample of observation of random variable U
U	discrete random variable
$v(n)$	noise signal
v_{sd}	distortion index
$v_0(n)$	noise signal at reference input
$v_m(n)$	noise signal at microphone m
$\hat{v}(n)$	noise signal estimate
$\mathbf{v}(n)$	additive noise signal vector in the time domain
$\mathbf{v}(\omega)$	additive noise signal vector in the frequency domain
$\mathbf{v}(k,m)$	vector of length L of STFT coefficients of $V(k,m)$
$\mathbf{v}'(k,m,\tau)$	interference vector to $\mathbf{v}(k,m)$
$V(k,m)$	STFT of signal $v(n)$
$V_m(\omega)$	frequency domain representation of the noise signal at microphone m
$V_{\mathrm{rn}}(k,m)$	residual noise
$V'(k,m,\tau)$	interference to $V(k,m)$ using delay τ
w	observation of random variable W
w_i	sample of observation of random variable W
$\mathbf{w}$	filter in the time domain
W	discrete random variable
$x(n)$	speech signal
$x_m(n)$	desired signal at microphone m
$\hat{x}(n)$	speech signal estimate
$\mathbf{x}(n)$	desired signal vector in the time domain
$\mathbf{x}(\omega)$	desired signal vector in the frequency domain
$\mathbf{x}(k,m)$	vector of length L of STFT coefficients of $X(k,m)$
$\mathbf{x}_1(k,m)$	vector of length $L+\tau$ of STFT coefficients of $X(k,m)$
$\mathbf{x}'(k,m,\tau)$	interference vector to $\mathbf{x}(k,m)$
$X(\omega)$	desired signal

$X_m(\omega)$	frequency domain representation of the desired signal at microphone m
$X(k,m)$	STFT of signal $x(n)$
$X_{\text{est}}(k,m)$	speech estimate
$X_{\text{fd}}(k,m)$	filtered desired signal
$X_\ell(k)$	STFT of signal $x(n)$
$\hat{X}_\ell(k)$	STFT estimate of the clean speech signal
$X'(k,m,\tau)$	interference to $X(k,m)$ using delay τ
$X'_{\text{ri}}(k,m)$	residual interference
$y(n)$	observed noisy speech signal
$y_m(n)$	observed signal at microphone m
$\mathbf{y}(n)$	observed signal vector in the time domain
$\mathbf{y}(\omega)$	observed signal vector in the frequency domain
$\mathbf{y}(k,m)$	vector of length L of STFT coefficients of $Y(k,m)$
$\mathbf{y}_A(k)$	vector of log-magnitude of the noisy speech coefficients at frequency k
$\mathbf{y}_P(k)$	vector of phase of the noisy speech coefficients at frequency k
$\tilde{\mathbf{y}}_A(k)$	delayed vector of log-magnitude of the noisy speech coefficients at frequency k
$\tilde{\mathbf{y}}_P(k)$	delayed vector of phase of the noisy speech coefficients at frequency k
$Y(k,m)$	STFT of signal $y(n)$
$Y_m(\omega)$	frequency domain representation of the observed signal at microphone m
$Y_\ell(k)$	STFT of signal $y(n)$
z	observation of random variable Z
$z(n)$	filter output
z_i	sample of observation of random variable Z
Z	joint random variable
$Z(\omega)$	beamformer output in the frequency domain
$Z(k,m)$	filter output in the STFT domain
α	DD smoothing parameter
$\alpha(k)$	MI coefficient for step size control
$\alpha(\omega)$	frequency dependent regularization parameter
$\alpha_\epsilon(\omega)$	ϵ dependent combined beamformer regularization parameter
α_ℓ	weighting factor for the proposed two step estimator
α_{min}	regularization parameter that minimizes the gain
γ_ℓ	a-posteriori SNR
$\boldsymbol{\gamma}_V(k,m,\tau)$	correlation vector of signal $V(k,m)$

$\boldsymbol{\gamma}_X(k,m,\tau)$	correlation vector of signal $X(k,m)$
$\Gamma_V(k,m,\tau)$	inter-frame correlation coefficient of signal $V(k,m)$
$\Gamma_X(k,m,\tau)$	inter-frame correlation coefficient of signal $X(k,m)$
$\boldsymbol{\Gamma}_{\mathbf{v}}(\omega)$	pseudo-coherence matrix of $\mathbf{v}(\omega)$
$\boldsymbol{\Gamma}_{\mathrm{d}}(\omega)$	pseudo-coherence matrix for diffuse noise
$\boldsymbol{\Gamma}_{\mathrm{d},\alpha}$	pseudo-coherence matrix for the combined noise field
$\boldsymbol{\Gamma}_{\epsilon}(\omega)$	regularized form of the pseudo-coherence matrix of the diffuse noise
δ	adaptive algorithm regularization
δ	distance between two successive sensors in Section 2.3 and Chapter 5
δ	GARCH model parameter in Chapter 4
ϵ	regularization parameter
ϵ	threshold for confining the dynamic range of the log spectrum in Chapter 4
ϵ_0	tolerance parameter
$\epsilon(i)$	twice the distance from z_i to its kth neighbor
$\epsilon_u(i)$	twice the distance from z_i to its kth neighbor projected into the U subspace
$\epsilon_w(i)$	twice the distance from z_i to its kth neighbor projected into the W subspace
$\epsilon_{\mathrm{inherent}}(k,m)$	inherent estimation error
θ	direction of arrival azimuth angle
θ_s	source signal direction of arrival
κ	GARCH/ARCH model parameter
λ_ℓ	short term spectrum of the speech signal
$\lambda_{\ell\|\ell'}$	conditional short term spectrum of the speech signal
$\Lambda(Y_\ell, q_\ell)$	ratio between probabilities
μ	adaptive algorithm step size
μ	ARCH model parameters in Chapter 4
$\mu(k)$	frequency dependent adaptive algorithm step size
μ_0	constant for step size control based on MI
μ_m	mth order moment of a signal
ν_ℓ	ratio between the a-priori SNR multiplied by the a-posteriori SNR and the a-priori SNR plus 1
ν'_ℓ	ratio between the hypothesis conditional a-priori SNR multiplied by the a-posteriori SNR and the hypothesis conditional a-priori SNR plus 1
ξ_ℓ	a-priori SNR
ξ'_ℓ	hypothesis conditional a-priori SNR

$\xi_{\ell\|\ell'}$	conditional a-priori SNR
$\hat{\xi}_{\ell\|\ell}$	a-priori SNR estimate
$\hat{\xi}_{\ell\|\ell-1}$	one frame ahead a-priori SNR estimate
$\xi_{\min}$	noise floor parameter
ξ_{nr}	noise reduction factor
ξ_{sr}	speech reduction factor
σ	tolerance parameter
σ_ℓ^2	short term spectrum of the noise signal
τ	decorrelation delay
τ_0	delay between successive microphones
τ_1	decorrelation delay for MI algorithm
τ_2	delay for MI calculation
τ_m	relative delay between the mth microphone to microphone 1
$\varphi_\epsilon(\omega)$	angle between vectors $\mathbf{d}(\omega)$ and $\mathbf{\Gamma}_\epsilon^{-1}(\omega)\mathbf{d}(\omega)$
$\phi_\cdot(\omega)$	variance of $\cdot(\omega)$
$\phi_\cdot(k,m)$	variance of $\cdot(k,m)$
$\mathbf{\Phi}_\cdot(k,m)$	covariance matrix of $\cdot(k,m)$
$\mathbf{\Phi}_\cdot(k,m,\tau)$	covariance matrix of interference $\cdot(k,m,\tau)$
ω	angular frequency
$\mathcal{B}[\mathbf{h}(\omega),\theta_s]$	Beam pattern
$\mathcal{D}[\mathbf{h}(\omega)]$	directivity factor
$\mathcal{D}_{\max}[\mathbf{h}(\omega),\theta]$	maximum DF
$\mathcal{G}[\mathbf{h}(\omega)]$	SNR gain
$\mathcal{G}_0$	fixed value of SNR gain
$\mathcal{G}_k$	calculated SNR gain in mid-section as defined per algorithm 1
$\mathcal{H}_0$	speech absence hypothesis
$\mathcal{H}_0$	noise presence hypothesis in Chapter 3
$H_0^{(\ell,k)}$	speech absence hypothesis in frame ℓ and frequency bin k
$\mathcal{H}_1$	speech presence hypothesis
$\mathcal{H}_1$	noise absence hypothesis in Chapter 3
$H_1^{(\ell,k)}$	speech presence hypothesis in frame ℓ and frequency bin k
$\mathcal{I}(k,m)$	noise indicator
$\mathcal{W}[\mathbf{h}(\omega)]$	white noise gain
$\mathcal{W}_0$	fixed value of WNG
$\mathcal{W}_{\max}$	maximum WNG

Chapter 1

Introduction

1.1 Background and Motivation

In real environments it is generally unavoidable to capture an acoustic signal without contaminating noise which degrades the signal quality and intelligibility. The noise can be caused both by the surrounding environment and by the very system recording the signal. It can be divided into four categories; *additive noise*, additive sound from additional sources, *echo*, a reflection of the sound arriving to the listener after a delay from the direct sound signal, occurs when there is coupling between loudspeakers and a microphone, *reverberation*, in an enclosed environment there are multiple propagation paths due to reflections that arrive at the microphone after different delays from the direct sound signal, and *interference*, when there are multiple competing sound sources. For each category of noise, there will be different approaches to suppress the noise and extract a clean signal. Our focus will be on the additive noise. The process of suppressing the detrimental additive noise before the signal is stored, transmitted, or played is known as noise reduction or alternatively as speech enhancement.

The observed signal is generally modeled as a superposition of the clean desired speech signal and the noise signal, when the additive noise and the desired speech signal are statistically independent. As the observed signal is a mixture of signals, its characteristics can be very different from those of the speech signal or the noise. As the noise characteristics can change and as there are numerous applications with different goals, the problem of noise reduction is very challenging and has been extensively researched over the past few decades. To name a few applications where noise reduction processing is required:

- telecommunications - voice communication suffers from the background

noise present at the transmitting end and can be processed to reduce the noise and improve speech quality prior to being played at the receiver end.

- hearing aids - processing the noisy signal to reduce the noise before the signal gets amplified for the hearing impaired.

- speech recognition - to improve the accuracy of the recognition the noisy signal can be processed and noise reduced prior to being analyzed for recognition.

When a single channel is available for the processing of the acoustic signal, this is known as single channel speech enhancement. This could be imposed by the system used (single microphone applications) or by the availability of the desired signal (prerecorded application). This case, where only the noisy signal is available and there is no access to additional reference signals, is one of the most challenging cases and it has been studied extensively. Many methods exist for combating this problem, and they can be divided into the following main categories:

- Spectral subtraction methods:
  First developed using analog circuits by Schroeder [2], and later in the digital Fourier domain by Boll [3], they operate on the principle of subtracting an estimated amplitude spectrum of the noise (which is estimated when the speech is not present or by other means) from the amplitude spectrum of the observed signal to get the amplitude spectrum of the estimated desired signal, while using the phase of the observed signal. A rectification step is then applied to set any negative components to zero. Though popular and simple to implement, these methods suffer from the musical tones artifacts. Throughout the years modifications have been proposed, such as using over-subtraction factor in combination with a spectral floor [4], adaptive gain averaging [5], frequency dependent over-subtraction [6] and more, to reduce the unpleasant musical tones at the cost of added heuristics for the control of the algorithms.

- Subspace methods:
  These methods are based on a decomposition of the observed noisy signal subspace into a clean signal subspace and a noise subspace and using the components from the signal subspace to estimate the desired signal. The decomposition can be done using orthogonal matrix factorizations such as the singular value decomposition, initiated by

the work of Dendrinos et al. [7] for white noise, or the eigenvalue decomposition initiated by the work of Ephraim and Van Trees [8].

- Binary Mask methods:
  Originating from the computational auditory scene analysis (CASA), see [9] for information on CASA, these methods estimate a binary mask which is applied on the observed signal, meaning that specific frequency bins are retained for the estimate of the desired signal per a specific criterion while the rest are discarded. The noise reduction problem becomes a binary classification problem, an elementary form of supervised learning. The binary masks methods have been shown to improve speech intelligibility in single channel condition [10], though it remains challenging as it still relies on the statistical properties of the speech and/or noise signal.

- Deep Learning methods:
  These methods are a class of machine learning algorithms that exploit multiple layers of information processing for supervised or unsupervised feature extraction, pattern analysis, and classification. They have gained popularity in the past decade due to the increase in processing capabilities and the availability of large data for training. Deep learning was first introduced to speech separation/enhancement by Wang and Wang in 2012 [11, 12] where they used deep neural networks (DNNs) for subband classification to estimate a binary mask. In supervised learning there are three main components: the learning machines, the training targets, and the acoustic features. Different types of DNNs can be employed such as feedforward multilayer perceptrons (MLPs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). Though DNNs are widely-used, there are alternative deep learning machines, for example, deep non-negative matrix factorization (NMF) [13].

- Statistical model based methods:
  In these methods the noise reduction problem is formulated as a statistical estimation problem, where a cost function is defined and the desire is to find an estimator (either linear or non-linear) of the parameter of interest to optimize the cost function. These methods were initiated by the work of McAulay and Malpass [14] where they proposed to use a maximum-likelihood (ML) estimator for the speech spectral amplitude with a soft decision for the speech presence probability, followed by the work of Ephraim and Malah [15, 16] where

they proposed the minimum mean square error (MMSE) estimator for spectral amplitude and log spectral amplitude. Lim and Oppenheim [17] initiated the use of the Wiener filter, which minimizes the MMSE between the filter output and the desired speech signal.

The vast majority of speech enhancement algorithms require an estimate of the noise spectrum which can have a large impact on the enhanced signal. If the estimate is too low then residuals of the noise will remain, if it is too high then the speech will be distorted. This led to extensive research resulting in many methods. If the noise is stationary, i.e., it does not change over time, e.g. white noise and fan noise, the simplest approach is to estimate and update the noise spectrum during segments when the signal is absent. This can be done using a voice activity detector (VAD) which determines per segment whether the speech signal is present or absent by extracting features from the observed signal (such as energy level and zero-crossings [18], cepstral features [19], and more), which are then compared against a threshold value. However, in practice, the noise is rarely stationary, for example multiple people speaking in the background, medical equipment beeps, and alarms. This can lead to erroneous estimates of the VAD, and even if the VAD is accurate it might not be sufficient, as an accurate noise estimate is required also during speech segments. When the noise is nonstationary, it is not enough to update the noise only when speech is absent; it needs to be tracked continuously. Clearly, it is more difficult to estimate and hence to suppress nonstationary noise.

There are three main classes for continuous noise estimation, though other methods exist as well. The first is the minimum tracking algorithms which are based on the observation that the power of the noise speech signal in each frequency bin decays to the power level of the noise even in segments where the speech is present. So by tracking the minimum of the noisy speech in each frequency bin, it is possible to estimate the noise level. This was originally considered in the minimum statistics (MS) method proposed by Martin [20] and later refined to include a bias compensation and better smoothing factor [21]. A disadvantage of the MS methods is that they don't respond quickly to rapid changes in the noise level: there is a lag to the noise true behavior due to the window analysis used. A different approach was proposed by Doblinger [22] where the noise is continuously updated by using a non-linear smoothing function per frequency bin instead of a fixed window; this however results in overestimation of the noise during speech activity.

The second class of methods is the recursive averaging algorithms. These

are based on the observation that different frequency bins will have different SNR and speech presence probability. Consequently, it is possible to update the noise estimate per frequency bin when the speech presence probability is low or when the SNR is low. The noise is estimated using a recursive weighted average of the past estimates and the present spectrum of the noisy speech, where the weights are updated according to the SNR, e.g. [23], or the speech presence probability, e.g. [24]. These methods can achieve better tracking of fast changing noise than the MS methods as the recursive averaging is carried out immediately.

The third class of methods is the histogram based algorithms which are derived from the observation that the most frequent energy value per frequency bin, i.e. the histogram maximum, corresponds to the noise level per the frequency bin. The basic formulation of these methods, e.g. [25], is per frame to calculate the histogram of a window of past spectrum values, and the noise estimate would be the maximum of that histogram, when this is done per frequency bin. A smoothing function on the noise estimate can be used to avoid spikes. These methods are computationally expensive. In addition, the noise can be overestimated if the window length is too short; however, taking a long window length will inhibit the tracking ability.

When the noise has an underlying structure characteristic, this can be exploited in the noise estimation method. Harmonic noise is an example of this. Harmonic noise, also known as periodic noise, refers to deterministic tones present in real-life noises such as the sound of a vacuum cleaner, vehicle engine noise, medical equipment beeping sounds, alarm sounds, and more. The Adaptive Line Enhancer (ALE) first introduced by Widrow [26, 27] is a method developed to cancel periodic interference without an external reference source, i.e. for the single channel case. It consists of a single microphone and a delay element to produce a delayed version of the noisy signal to be used as the reference signal. The delay enables separation of the periodic and stochastic components in a signal; it de-correlates the stochastic components between the input and the delayed input, while leaving the periodic noise components correlated, which are then extracted using an adaptive filter. The enhanced speech signal is then the result of the subtraction of the noise estimate from the noisy input. The ALE reduces the periodic or harmonic noise but does not remove any wide-band background noise; hence it can be used in conjunction with a traditional noise estimator or even with another ALE to remove any additional wide-band background noise present [28]. It can be implemented in the Frequency domain [29], where adaptive filters have less computational cost and can be calculated separately per frequency bin. The de-correlation delay parameter and the

step size of the adaptive filter have a large impact of the performance of the algorithm. In the conventional methods they are fixed or else heuristically optimized, hence noise residuals remain, especially for nonstationary noise. In a recent study [30] it was proposed to implement the ALE with a frequency dependent step size based on mutual information (MI) to detect the presence of harmonic noise and reduce it. The algorithm is implemented in a block-wise manner to deal with nonstationary noise, where it is assumed that the span of the stationarity of the noise is at least as large as the block length. This approach is not an effective solution for highly nonstationary noise signals such as medical equipment beeps or alarm sounds.

Ideally, we would like to improve both quality and intelligibility of the observed signal; however, when reducing noise we introduce speech distortion which could degrade the intelligibility. There is a compromise between noise reduction and speech distortion, between the quality and intelligibility of the enhanced signal. The challenge is to suppress the noise without introducing perceptible distortion to the speech signal.

To better address this drawback of single channel noise reduction, the use of microphone arrays has been proposed. Microphone arrays consist of multiple sensors organized in a specific structure, spatially sampling the sound field. Beamforming is the process of combining the inputs using spatial filters to get a single output, where different methods have been developed to determine the filter weights per the desired application, such as signal presence detection, estimation of the direction of arrival, sounds source separation, noise reduction and more. The shape and size of the array and the number of microphones used as well as their position impact the performance of the array. Also, the array behavior can change under different noise environments. Beamformers can be divided into two categories, fixed and adaptive. Fixed beamformers are designed taking into account the array geometry, assumed look direction, and noise statistics. The filter weights are computed and then fixed regardless of the actual environment. Some examples of well-known fixed beamformers are the Delay and Sum beamformer [31], which is designed to maximize the SNR gain under white noise conditions, and the Superdirective beamformer [32, 33], which is designed to maximize the SNR gain under diffuse noise conditions. As it turns out, there is a trade-off in beamformer performance under white noise and diffuse noise. A beamformer designed to be optimal for white noise preforms poorly for diffuse noise conditions and vice versa. As realistic environments contain both, extensive work has been done to find a superdirective beamformer with increased robustness to white noise. In comparison, adaptive beamformers are designed taking into account the noise statistics, hence their performance

can be more optimal as long as the statistics are estimated correctly.

As the noise is everywhere and is here to stay, can we improve on what we have for the single channel case, particularly for nonstationary harmonic noise? For the multichannel fixed beamformer, can we find an optimal solution under different noise conditions? In this thesis, we try to address these questions.

1.2 Research Overview

The major part of the presented research deals with single-channel noise reduction. Specifically, the main objective is to develop a method for removing nonstationary harmonic noise from a speech signal recorded with a single microphone. We propose a frequency-domain adaptive line enhancer (ALE) to reduce nonstationary harmonic noise, such as medical equipment beeps, from a noisy speech signal captured by a single microphone. The reduction of nonstationary noise is very challenging with the tradeoff between noise reduction and speech distortion, often resulting with much noise residuals. The proposed ALE is a combination of the commonly-used forward adaptive linear filter and a non-causal backward adaptive linear filter used together with an indicator for the presence of transient noise. The proposed combined filter results in less noise residuals while preserving the speech components. We compare the proposed approach to conventional and recent methods and show that it can outperform these methods, achieving lower distortion, more noise reduction, and overall better speech quality and intelligibility.

The adaptive line enhancer, which reduces the harmonic noise (nonstationary and stationary), can be followed by a second stage of processing to reduce remaining wide-band background noise. This second stage can be implemented by traditional spectral speech enhancement methods which require an estimator for the a-priori SNR. In the second part of this research we investigate the use of the autoregressive conditional heteroscedasticity (ARCH) model as a replacement for the well-known decision-directed method in the log-spectral amplitude estimator for speech enhancement. We employ three sound quality measures: speech distortion, noise reduction and musical noise, and explain the effect the ARCH model parameters have on these measures. We demonstrate and compare the use of the decision-directed and ARCH estimators and show that the ARCH model achieves better results than the decision-directed for some of these measures, while compromising between the speech distortion and noise reduction.

In the third and final part of this research, we set aside the single-channel case and touch lightly on microphone array beamforming. We introduce an

optimal beamformer design that facilitates a compromise between high directivity and low white noise amplification. The proposed beamformer involves a regularization factor, whose optimal value is determined using a simple and efficient one dimensional search algorithm. Simulation results demonstrate controlled tuning of various gain properties of the desired beamformer and improved performance compared to a competing method.

1.3 Organization

This thesis is organized as follows. In Chapter 2, we provide the necessary background for the following chapters. In Chapter 3, we discuss the problem of reducing nonstationary harmonic noise and propose the use of a combined backward and forward adaptive line enhancer. We analyze the performance of the proposed solution and compare it to existing methods. In Chapter 4, we analyze and compare the use of an ARCH estimator to the well-known decision-directed estimator for log-spectral amplitude. In Chapter 5, we diverge from the single-channel analysis and introduce microphone array beamforming, where we discuss the design of an optimal beamformer that facilitates a compromise between high directivity and low white noise amplification. Finally, in Chapter 6, conclusions of the research are presented along with possible directions for future research.

Chapter 2

Background

In this chapter we provide background for reading this thesis, where each section is relevant to the specified chapter. Section 2.1 contains the required background for Chapter 3. Section 2.2 contains the required background for Chapter 4, and finally, Section 2.3 contains the required background for Chapter 5.

2.1 Adaptive Noise Cancellation and Adaptive Line Enhancer

The work on adaptive noise cancellation started in the late 1950s and proliferated in the 1960s. In 1975 Widrow et al. [26] laid out the concept of adaptive noise cancelling, and since then the technique has been applied to various communication and industrial appliances. The concept for the adaptive noise canceller (ANC) introduced is a variation of optimal filtering that derives an estimate of the noise by filtering a reference input signal and then subtracting this noise estimate from the primary input containing both signal and noise. As a result, the noise is reduced or even eliminated. The reference signal is obtained from one or more sensors placed in the noise field in such a way that they detect the noise but the signal is weak or undetectable. By using an adaptive process as opposed to a fixed filter, there is little risk of distorting the signal or amplifying the noise from the noise estimate subtraction.

Figure 2.1 illustrates the ANC solution, where $x(n)$ is the signal, $v(n)$ is the noise at the primary input, $v_0(n)$ is the noise at the reference input, and n is the discrete time index. The signal $x(n)$ is uncorrelated with the noise $v(n)$ and $v_0(n)$, yet $v(n)$ is correlated with $v_0(n)$. $v_0(n)$ is then filtered to produce the estimate $\hat{v}(n)$ that we desire to be as close as possible to $v(n)$.

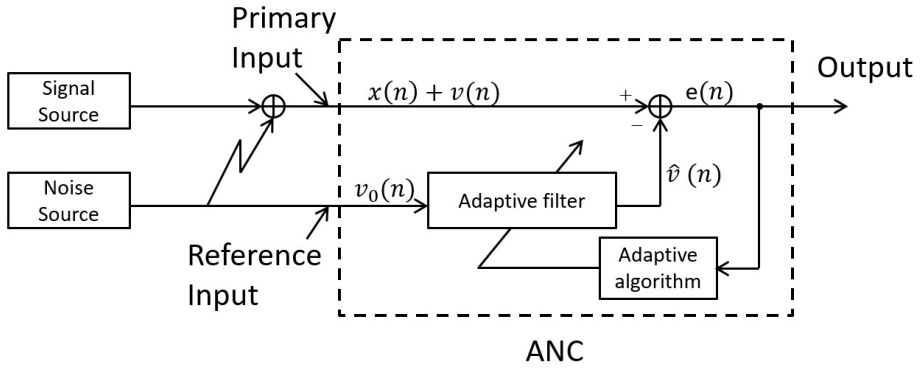

Figure 2.1: Adaptive noise cancelling concept

The estimate is subtracted from the primary input to obtain the output error signal $e(n) = x(n) + v(n) - \hat{v}(n)$. The error signal is used to continuously adjust the filter and minimize itself. For the ANC the objective is to produce an output which is the signal estimate $\hat{x}(n) = e(n)$ and is optimal in the least squares sense.

$$\hat{x}(n) = x(n) + v(n) - \hat{v}(n) \tag{2.1}$$

Squaring this equation

$$\hat{x}^2(n) = x^2(n) + (v(n) - \hat{v}(n))^2 + 2x(n)(v(n) - \hat{v}(n)) \tag{2.2}$$

We assume that $x(n)$, $v(n)$, and $v_0(n)$ are statistically stationary and with zero mean, $x(n)$ is uncorrelated with $v(n)$ and $v_0(n)$, and $v(n)$ is correlated with $v_0(n)$. Taking the expectation of Equation (2.2), we get

$$\begin{aligned} E\left[\hat{x}^2(n)\right] &= E\left[x^2(n)\right] + E\left[(v(n) - \hat{v}(n))^2\right] + 2E\left[x(n)(v(n) - \hat{v}(n))\right] \\ &= E\left[x^2(n)\right] + E\left[(v(n) - \hat{v}(n))^2\right]. \end{aligned} \tag{2.3}$$

As the filter is adjusted to minimize $E\left[\hat{x}^2(n)\right]$, the signal power is unaffected, and the minimum output power is

$$\min E\left[\hat{x}^2(n)\right] = E\left[x^2(n)\right] + \min E\left[(v(n) - \hat{v}(n))^2\right], \tag{2.4}$$

meaning that $E\left[(v(n) - \hat{v}(n))^2\right]$ is also minimized, and the filter output is the best least squares estimate of the noise $v(n)$. Moreover, if we rearrange Equation (2.1) to

$$\hat{x}(n) - x(n) = v(n) - \hat{v}(n), \tag{2.5}$$

we can clearly see that when we minimize $E\left[\left(v\left(n\right)-\hat{v}\left(n\right)\right)^2\right]$ the estimate $\hat{x}\left(n\right)$ is the best least square estimate of the signal as we also minimize $E\left[\left(x\left(n\right)-\hat{x}\left(n\right)\right)^2\right]$. Minimizing the output power minimizes the output noise and hence maximizes the output signal to noise ratio. The minimum feasible for the output power, i.e. the ideal case, is $E\left[\hat{x}^2\left(n\right)\right]=E\left[x^2\left(n\right)\right]$ which occurs when $E\left[\left(v\left(n\right)-\hat{v}\left(n\right)\right)^2\right]=0$, thus $\hat{x}\left(n\right)=x\left(n\right)$ and $\hat{v}\left(n\right)=v\left(n\right)$.

As seen, the ANC works on the principle of correlation cancellation, meaning that the correlated components are removed. If additional noise is present in the primary input, which is uncorrelated to the signal and to the noise in the reference input, this primary noise remains uncancelled and is passed through to the output. And if some signal components are present in the reference signal, some of the signal will be cancelled, causing signal distortion and degradation of the ANC performance.

In many applications a reference signal is difficult to obtain or is not available. For this case Widrow et al., in the same study [26], introduced the Adaptive line Enhancer (ALE), which uses a delayed version of the input signal to be used as the reference. The delay, denoted by τ, must be long enough to decorrelate the random components in the signal, while the periodic components of the input remain correlated to the delayed signal and therefore will be cancelled out.

We define the observed noisy signal by $y\left(n\right)$, where

$$y\left(n\right)=x\left(n\right)+v\left(n\right). \tag{2.6}$$

We then apply a delay τ to the observed signal, and pass this delayed signal through a filter $\mathbf{w}$. For a filter of length L we can define the vector

$$\begin{aligned}\mathbf{y}\left(n-\tau\right)&=\left[y\left(n-\tau\right),y\left(n-\tau-1\right),\ldots,y\left(n-\tau-L+1\right)\right]^T\\&=\mathbf{x}\left(n-\tau\right)+\mathbf{v}\left(n-\tau\right),\end{aligned} \tag{2.7}$$

where the superscript $\left(\cdot\right)^T$ is the transpose operator, and $\mathbf{x}\left(n-\tau\right)$ and $\mathbf{v}\left(n-\tau\right)$ are defined in a similar fashion to $\mathbf{y}\left(n-\tau\right)$. The output of the filter is then

$$z\left(n\right)=\mathbf{w}^T\mathbf{y}\left(n-\tau\right), \tag{2.8}$$

and the error signal is

$$e\left(n\right)=y\left(n\right)-z\left(n\right) \tag{2.9}$$

$$=y\left(n\right)-\mathbf{w}^T\mathbf{y}\left(n-\tau\right) \tag{2.10}$$

Two general cases can be considered. The first is when the signal is broadband and the noise is periodic. In this case, the delay decorrelates the signal while leaving the noise correlated, hence the filter estimates the noise, meaning that $z\left(n\right) = \hat{v}\left(n\right)$, and the signal estimate is the error signal $e\left(n\right) = \hat{x}\left(n\right)$, which is marked output in Figure 2.2. The second case is the opposite case, where the signal is narrowband and the noise is broadband. Here, the delay decorrelates the noise and the filter estimates the signal. Hence, in this case, $z\left(n\right) = \hat{x}\left(n\right)$, which is marked output' in Figure 2.2.

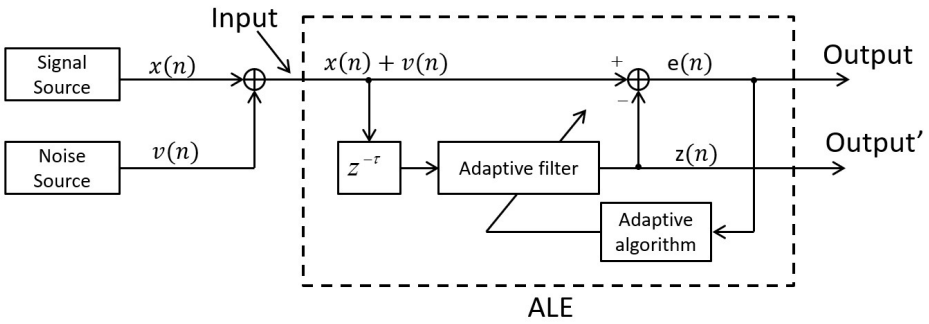

Figure 2.2: Adaptive Line Enhancer concept

Both cases can be used for speech signal estimation, depending on the noise characteristics and the decorrelation delay used. If the noise is broadband background noise, it is possible to estimate the speech signal using the pitch period for the delay. However, estimating the pitch period can be challenging. If the noise is periodic such as ventilation fan noise, it is possible to first extract the noise and then estimate the signal using the subtraction. It is even possible to cascade ALEs, where the first ALE removes the harmonic components of the noise and the second the remaining broadband components. However, some correlation of the speech signal to the delayed speech signal will remain, and signal distortion will occur.

The ALE can also be implemented in the frequency domain. The advantages of doing so is reduced computational complexity, and as each frequency bin can be processed separately, we can have separate filters per frequency bin. Using the short time Fourier transform (STFT), Equation (2.6) can be expressed as

$$Y\left(k,m\right) = X\left(k,m\right) + V\left(k,m\right), \tag{2.11}$$

where k is the frequency index ($k = 0, 1, \ldots, K-1$), m is the frame index ($m = 0, 1, \ldots, M-1$), and $Y\left(k,m\right)$, $X\left(k,m\right)$, and $V\left(k,m\right)$ are the STFTs of $y\left(n\right)$, $x\left(n\right)$, and $v\left(n\right)$ respectively.

Similarly to the time domain, we apply a delay τ to the observed signal $Y(k,m)$ and pass this delayed signal through filter $\mathbf{h}(k,m)$ of length L,

$$\mathbf{h}(k,m) = \left[H_0(k,m), H_1(k,m), \ldots, H_{L-1}(k,m)\right]^T, \tag{2.12}$$

where $\{H_j\}_{j=0}^{L-1}$ are the complex filter coefficients. And the filter output is

$$Z(k,m) = \mathbf{h}^H(k,m)\,\mathbf{y}(k,m-\tau), \tag{2.13}$$

where

$$\begin{aligned}\mathbf{y}(k,m-\tau) &= \left[Y(k,m-\tau), Y(k,m-\tau-1), \ldots, Y(k,m-\tau-L+1)\right]^T \\ &= \mathbf{x}(k,m-\tau) + \mathbf{v}(k,m-\tau),\end{aligned} \tag{2.14}$$

when $\mathbf{x}(k,m-\tau)$ and $\mathbf{v}(k,m-\tau)$ are defined in a similar fashion to the vector $\mathbf{y}(k,m-\tau)$. As a result, the error is defined as

$$\begin{aligned}E(k,m) &= Y(k,m) - Z(k,m) \\ &= Y(k,m) - \mathbf{h}(k,m)^H\,\mathbf{y}(k,m-\tau).\end{aligned} \tag{2.15}$$

The ANC and ALE adaptive processes can be implemented by different algorithms such as the Least Mean Squares (LMS), Normalized Least Squares (NLMS), or the Recursive Least Squares (RLS), though the LMS or NLMS are often used because of their robustness and simplicity. These algorithms require a step size which is traditionally a fixed value. The step size impacts the convergence rate and the performance of the algorithm. Using the NLMS, we can define

$$\mathbf{h}(k,m+1) = \mathbf{h}(k,m) + \frac{\mu}{\mathbf{y}^H(k,m-\tau)\,\mathbf{y}(k,m-\tau) + \delta}\mathbf{y}(k,m-\tau)\,E^*(k,m), \tag{2.16}$$

where $0 < \mu < 2$ is the step size parameter which should be smaller than 1 here for the algorithm to converge, and $\delta > 0$ is a regularization parameter.

Recently, Taghia and Martin [30] proposed a frequency domain ALE for harmonic noise reduction where the step size is based on mutual information (MI). The step size is frequency dependent and accounts for the presence of harmonic noise in different frequency bins. With such a step size, it is possible to improve on the signal distortion compared to the fixed step size where a filter is applied in all frequency bins, even those where the harmonic noise is absent.

2.1.1 Entropy and Mutual Information Definitions

We first define the concept of entropy, which is a measure of uncertainty of a random variable [34]. Let U be a discrete real random variable with probability $p(u)$, the entropy of U is

$$H(U) = -\sum_{u} p(u) \log p(u). \tag{2.17}$$

As the expectation of the random variable $g(U)$ is

$$E[g(U)] = \sum_{u} g(u) p(u), \tag{2.18}$$

the entropy of U can be expressed as the expected value of the random variable $-\log p(U)$,

$$H(U) = -E[\log p(U)]. \tag{2.19}$$

Given a pair of discrete real random variables (U, W) with the joint probability $p(u, w)$, the joint entropy $H(U, W)$ is defined as

$$\begin{aligned} H(U, W) &= -\sum_{u} \sum_{w} p(u, w) \log p(u, w) \\ &= -E[\log p(U, W)], \end{aligned} \tag{2.20}$$

and the conditional entropy is defined as

$$H(W|U) = -E[\log p(W|U)]. \tag{2.21}$$

The relationship between the joint entropy and the conditional entropy is given by the Chain rule

$$H(U, W) = H(U) + H(W|U). \tag{2.22}$$

We now define the mutual information, which is a measure of the amount of information that one random variable contains about another random variable. The mutual information $I(U; W)$ is

$$\begin{aligned} I(U; W) &= \sum_{u} \sum_{w} p(u, w) \log \frac{p(u, w)}{p(u) p(w)} \\ &= E\left[\log \frac{p(U, W)}{p(U) p(W)}\right]. \end{aligned} \tag{2.23}$$

The mutual information can be written in terms of the entropy as

$$I(U; W) = H(U) - H(U|W), \tag{2.24}$$

thus, we can interpret the mutual information as the reduction in the uncertainty of U given the knowledge of W. By symmetry

$$I(U;W) = H(W) - H(W|U). \tag{2.25}$$

Plugging (2.22) into (2.25) we get another definition for the mutual information:

$$I(U;W) = H(U) + H(W) - H(U,W). \tag{2.26}$$

Let us now consider the case where U and W are *complex* random variables. We can represent them in the polar coordinate system as

$$\begin{aligned} U &= U_{\|} e^{jU_{\angle}} \\ W &= W\| e^{jW_{\angle}}, \end{aligned} \tag{2.27}$$

where $U_{\|}, W_{\|} \in [0, \infty)$ denote the magnitudes and $U_{\angle}, W_{\angle} \in (-\pi, \pi]$ denote the phases of the complex random variables U and W respectively. By assuming that the phase and the magnitude are statistically independent, it can be shown [30] that the mutual information between the two complex variables can be approximated as

$$I(U;W) \approx I\left(U_{\|}; W_{\|}\right) + I\left(U_{\angle}; W_{\angle}\right). \tag{2.28}$$

Due to this decomposition of the MI into MI of magnitude and MI of phase, Taghia and Martin looked separately at the magnitude and phase spectrum of the noisy speech.

2.1.2 Mutual Information Estimation

The MI can be estimated either by parametric methods, which assume the data has a certain distribution and fit the model to the data, or by non-parametric methods, which do not assume a known distribution. In their study, Taghia and Martin proposed several approaches, both parametric and non-parametric, and they found that the best results were achieved with the non-parametric k-nearest neighbor (KNN) estimator [35].

Kraskov et al. in [35] proposed two different estimates for the MI. Let $z_i = (u_i, w_i), i = 1, \ldots, N$ be a sample of independent and identically distributed observations of the random joint variable $Z = (U, W)$. For each point $z_i = (u_i, w_i)$, we rank its neighbors by a distance metric $d_{i,j} = \|z_i - z_j\|$ so that $d_{i,j_1} \leq d_{i,j_2} \leq \ldots \leq d_{i,j_N}$. We perform similar ranking in the subspaces U and W. For the space Z we use the maximum norm, i.e.

$$\|z_i - z_j\| = \max\left\{\|u_i - u_j\|, \|w_i - w_j\|\right\}. \tag{2.29}$$

For U and W subspaces any norm can be used, such as the L_2 norm. We denote by $\epsilon(i)/2$ the distance from z_i to its kth neighbor and by $\epsilon_u(i)/2$ and $\epsilon_w(i)/2$ the distances between the same points projected into the U and W subspaces, where per Equation (2.29) we have

$$\epsilon(i) = \max\{\epsilon_u(i), \epsilon_w(i)\}. \tag{2.30}$$

We denote the number $n_u(i)$ as the number of points u_j whose distance to u_i is less than $\epsilon(i)/2$, and similarly for w we denote the number $n_w(i)$ as the number of points w_j whose distance to w_i is less than $\epsilon(i)/2$. The estimate for the MI is then

$$I^{(1)}(U;W) = \psi(k) - \frac{1}{N}\sum_{i=1}^{N}[\psi(n_u(i)+1) + \psi(n_w(i)+1)] + \psi(N), \tag{2.31}$$

where $\psi(x) = \Gamma^{-1}(x)\, d\Gamma(x)/dx$ is the digamma function.

Alternatively, if we define $n_u(i)$ as the number of points where the distances $\|u_i - u_j\| \leq \epsilon_u(i)/2$ and similarly $n_w(i)$ as the number of points where the distances $\|w_i - w_j\| \leq \epsilon_w(i)/2$, the estimate for the MI is then

$$I^{(2)}(U;W) = \psi(k) - \frac{1}{k} - \frac{1}{N}\sum_{i=1}^{N}[\psi(n_u(i)) + \psi(n_w(i))] + \psi(N). \tag{2.32}$$

According to Kraskov et al. both estimates provide very similar results. For the same k, the estimate in Equation (2.31) gives slightly smaller statistical errors but has larger systematic errors. Only if we are interested in very high dimensions, do the system errors become so severe that the estimate in Equation (2.32) is preferable.

2.1.3 Step Size Control Using Mutual Information

The method that Taghia and Martin proposed for controlling the adaptive filter step size is as follows. They implement the frequency domain ALE shown in Figure 2.2 in a blockwise manner. First, the time domain signal is divided into several overlapping blocks of 3 s length with overlap of 25%. Next, for each block the adaptive filter coefficients are initiated to zero and then derived and applied on the signal per the ALE system. Finally, the outputs of the ALE per block are windowed by a Tukey window and are concatenated using the overlap-add procedure to obtain the full length output signal. They consider the block length of 3 s to be long enough for robust computation of the MI and by blockwise implementation it is possible to process the blocks in a parallel fashion.

They define the de-correlation delay parameter of the ALE system as τ_1, and propose the following control algorithm for the filter step size $\mu = \mu(k)$ in Equation (2.16).

Let

$$\mathbf{y}_A(k) = \left[\log|Y(k,1)|, \ldots, \log\left|Y\left(k,\bar{N}\right)\right|\right]^T \tag{2.33}$$

and

$$\mathbf{y}_P(k) = \left[\angle|Y(k,1)|, \ldots, \angle\left|Y\left(k,\bar{N}\right)\right|\right]^T \tag{2.34}$$

respectively denote the vectors of the log-magnitude and phase of the noisy speech coefficients at frequency k, and $\bar{N}$ denotes the number of frames that compose a block of 3 s length. The delayed version of the vectors $\mathbf{y}_A(k)$ and $\mathbf{y}_P(k)$, while keeping the same vector length, is defined as

$$\tilde{\mathbf{y}}_A(k) = \left[\mathbf{0}_{\tau_2}, \log|Y(k,1)|, \ldots, \log\left|Y\left(k,\bar{N}-\tau_2\right)\right|\right]^T \tag{2.35}$$

and

$$\tilde{\mathbf{y}}_P(k) = \left[\mathbf{0}_{\tau_2}, \angle|Y(k,1)|, \ldots, \angle\left|Y\left(k,\bar{N}-\tau_2\right)\right|\right]^T, \tag{2.36}$$

where $\mathbf{0}_{\tau_2}$ is a vector of zeros with the length τ_2 (τ_2 is in frames). Note that two time delay parameters are used: τ_1 and τ_2, where τ_1 is the decorrelation delay for the ALE, and τ_2 is the delay for the MI calculation. The MI for the log-magnitude spectrum is then defined as

$$I^A(k) = I\left(\mathbf{y}_A(k); \tilde{\mathbf{y}}_A(k)\right), \tag{2.37}$$

where the KNN estimator approach described above is used. Similarly, the MI for the phase spectrum is defined as

$$I^P(k) = I\left(\mathbf{y}_P(k); \tilde{\mathbf{y}}_P(k)\right). \tag{2.38}$$

Next, the MI for the log magnitude and the phase are normalized so that they range between 0 and 1

$$\bar{I}^P(k) = \frac{I^P(k) - \min\left(I^P\right)}{\max\left(I^P\right) - \min\left(I^P\right)} \tag{2.39}$$

and similarly $\bar{I}^A(k)$ is defined for $I^A(k)$. Motivated by Equation (2.28), the total MI is defined as

$$\bar{I}^{total}(k) = \bar{I}^A(k) + \bar{I}^P(k). \tag{2.40}$$

Then median filters are applied on $\bar{I}^{total}(k)$ and $\bar{I}^P(k)$ to further emphasize the MI peaks that indicate frequency bins with large temporal dependency

and to remove the bias so that the values of the MI are reduced in other frequency bins to zero.

$$\hat{I}^P(k) = \max\left\{\bar{I}^P(k) - \text{medianfilt}\left(\bar{I}^P(k), n_1\right), 0\right\}, \tag{2.41}$$

and

$$\hat{I}^{total}(k) = \max\left\{\bar{I}^{total}(k) - \text{medianfilt}\left(\bar{I}^{total}(k), n_2\right), 0\right\}, \tag{2.42}$$

where medianfilt $(\cdot)$ is a median filter, and n_1 and n_2 are the orders of the filters applied on $\bar{I}^P(k)$ and $\bar{I}^{total}(k)$ respectively. The coefficient $\alpha(k)$ is then defined as

$$\alpha(k) = \begin{cases} \hat{I}^P(k), & k \leq k^* \\ \hat{I}^{total}(k), & k > k^* \end{cases} \tag{2.43}$$

where the frequencies are divided into two bands, with the frequency boundary between the bands k^*. The reason for this division is due to the nature of the clean speech signals, which has a quasi-periodic structure in the low frequencies. Taghia and Martin observed that the MI of the log magnitude of the speech signals is relatively high in frequencies below 1 kHz while the phase MI was small. Since the desire is to detect the periodic structure of the noise and not the clean speech, in the low frequencies only the MI of the phase is considered.

Finally, the step size is defined as

$$\mu(k) = Q_\mu \mu_0 \alpha(k), \tag{2.44}$$

where μ_0 is a constant, $\alpha(k)$ is defined based on the MI values in Equation (2.43), and Q_μ is a decision coefficient defined as

$$Q_\mu = \begin{cases} 1, & \text{if} \quad \sum_{k=0}^{K-1} \left(I^P(k)\right)^2 \geq I_{\text{thr}} \\ 0, & \text{otherwise} \end{cases} \tag{2.45}$$

where I_{thr} is a constant. The coefficient Q_μ is used to determine whether the clean speech is corrupted by harmonic noise. If it is determined that the speech is not corrupted with harmonic noise, then $Q_\mu = 0$, hence, the step size $\mu(k) = 0$ for all frequency bins, meaning that the noisy signal is not processed and is passed as is to the output.

In the study based on experiments, the following parameter selection was made: ALE filter length $L = 5$, ALE de-correlation delay $\tau_1 = 1$, delay for MI calculation $\tau_2 = 2$, KNN parameter $k = 100$, threshold $I_{thr} = 1$, phase MI median filter order $n_1 = 12$, total MI median filter order $n_2 = 16$, frequency boundry k^* is set to frequency bin equivalent to 1Khz, and μ_0, which should be less than 0.5 to ensure the step size $\mu(k)$ is less than 1, is set to 0.1.

2.1.4 ALE Performance Measures

The narrow-band and full-band input SNR are

$$\text{iSNR}\left(k,m\right)=\frac{\phi_X\left(k,m\right)}{\phi_V\left(k,m\right)}, \tag{2.46}$$

$$\text{iSNR}\left(m\right)=\frac{\sum_{k=0}^{K-1}\phi_X\left(k,m\right)}{\sum_{k=0}^{K-1}\phi_V\left(k,m\right)}, \tag{2.47}$$

where $\phi_X\left(k,m\right)=E\left[\left|X\left(k,m\right)\right|^2\right]$ and $\phi_V\left(k,m\right)=E\left[\left|V\left(k,m\right)\right|^2\right]$.

We can rewrite Equation (2.15) using Equation (2.14)

$$\begin{aligned}E\left(k,m\right)&=X\left(k,m\right)-\mathbf{h}\left(k,m\right)^H\mathbf{x}\left(k,m-\tau\right)+V\left(k,m\right)\\&\quad-\mathbf{h}\left(k,m\right)^H\mathbf{v}\left(k,m-\tau\right)\\&=X_{fd}\left(k,m\right)+V_{rn}\left(k,m\right),\end{aligned} \tag{2.48}$$

where

$$X_{fd}\left(k,m\right)=X\left(k,m\right)-\mathbf{h}\left(k,m\right)^H\mathbf{x}\left(k,m-\tau\right) \tag{2.49}$$

is the filtered desired signal and

$$V_{rn}\left(k,m\right)=V\left(k,m\right)-\mathbf{h}\left(k,m\right)^H\mathbf{v}\left(k,m-\tau\right) \tag{2.50}$$

is the residual noise.

We define the narrow-band and full-band output SNR as the ratio of the variance of the filtered desired signal over the variance of the residual noise,

$$\text{oSNR}\left[\mathbf{h}\left(k,m\right)\right]=\frac{\phi_{X_{fd}}\left(k,m\right)}{\phi_{V_{rn}}\left(k,m\right)}, \tag{2.51}$$

$$\text{oSNR}\left[\mathbf{h}\left(m\right)\right]=\frac{\sum_{k=0}^{K-1}\phi_{X_{fd}}\left(k,m\right)}{\sum_{k=0}^{K-1}\phi_{V_{rn}}\left(k,m\right)}, \tag{2.52}$$

where $\phi_{X_{fd}}\left(k,m\right)=E\left[\left|X_{fd}\left(k,m\right)\right|^2\right]$ and $\phi_{V_{rn}}\left(k,m\right)=E\left[\left|V_{rn}\left(k,m\right)\right|^2\right]$.

We define the noise reduction factor as the ratio of the power of the noise at the sensor over the power of the noise remaining at the filter output. The noise reduction factor is usually expected to be lower bounded by 1. The higher the value, the more the noise is rejected. The narrow-band and full-band noise reduction factors are then

$$\xi_{nr}\left[\mathbf{h}\left(k,m\right)\right]=\frac{\phi_V\left(k,m\right)}{\phi_{V_{rn}}\left(k,m\right)} \tag{2.53}$$

$$\xi_{nr}\left[\mathbf{h}\left(m\right)\right]=\frac{\sum_{k=0}^{K-1}\phi_{V}\left(k,m\right)}{\sum_{k=0}^{K-1}\phi_{V_{rn}}\left(k,m\right)}. \tag{2.54}$$

To evaluate the level of distortion, we define the narrow-band and full-band speech reduction factor as the variance of the desired signal over the variance of the filtered desired signal at the output,

$$\xi_{sr}\left[\mathbf{h}\left(k,m\right)\right]=\frac{\phi_{X}\left(k,m\right)}{\phi_{X_{fd}}\left(k,m\right)} \tag{2.55}$$

$$\xi_{sr}\left[\mathbf{h}\left(m\right)\right]=\frac{\sum_{k=0}^{K-1}\phi_{X}\left(k,m\right)}{\sum_{k=0}^{K-1}\phi_{X_{fd}}\left(k,m\right)} \tag{2.56}$$

Thus the speech reduction factor is equal to 1 if there is no distortion and is greater than 1 when distortion occurs.

We also define the desired signal distortion index, which is another way to measure the distortion. The desired signal distortion index is defined as the MSE between the desired signal and the filtered desired signal normalized by the variance of the desired signal. The closer the distortion index is to 0, the less the distortion. The narrow-band and full-band desired signal distortion indexes are then

$$\begin{aligned}v_{sd}\left[\mathbf{h}\left(k,m\right)\right]&=\frac{E\left[\left(X\left(k,m\right)-X_{fd}\left(k,m\right)\right)^{2}\right]}{\phi_{X}\left(k,m\right)}\\&=\mathbf{h}^{H}\left(k,m\right)\mathbf{\Phi}_{x}\left(k,m-\tau\right)\mathbf{h}\left(k,m\right),\end{aligned} \tag{2.57}$$

$$\begin{aligned}v_{sd}\left[\mathbf{h}\left(m\right)\right]&=\frac{\sum_{k=0}^{K-1}E\left[\left(X\left(k,m\right)-X_{fd}\left(k,m\right)\right)^{2}\right]}{\sum_{k=0}^{K-1}\phi_{X}\left(k,m\right)}\\&=\frac{\sum_{k=0}^{K-1}\phi_{X}\left(k,m\right)\mathbf{h}^{H}\left(k,m\right)\mathbf{\Phi}_{x}\left(k,m-\tau\right)\mathbf{h}\left(k,m\right)}{\sum_{k=0}^{K-1}\phi_{X}\left(k,m\right)},\end{aligned} \tag{2.58}$$

where $\mathbf{\Phi}_{x}\left(k,m-\tau\right)=E\left[\mathbf{x}\left(k,m-\tau\right)\mathbf{x}^{H}\left(k,m-\tau\right)\right]$.

By making the appropriate substitutions, one can derive the following relationships:

$$\frac{\mathrm{oSNR}\left[\mathbf{h}\left(k,m\right)\right]}{\mathrm{iSNR}\left(\mathrm{k,m}\right)}=\frac{\xi_{nr}\left[\mathbf{h}\left(k,m\right)\right]}{\xi_{sr}\left[\mathbf{h}\left(k,m\right)\right]} \tag{2.59}$$

$$\frac{\mathrm{oSNR}\left[\mathbf{h}\left(m\right)\right]}{\mathrm{iSNR}\left(\mathrm{m}\right)}=\frac{\xi_{nr}\left[\mathbf{h}\left(m\right)\right]}{\xi_{sr}\left[\mathbf{h}\left(m\right)\right]} \tag{2.60}$$

2.2 Statistical Model Based Approach

This section contains the required background for chapter 4. As discussed in the introduction chapter, one approach to the single-channel noise reduction is the statistical model based methods. Given a set of measurements (the noisy speech signal), we aim to find an estimator of the parameter of interest (the clean speech signal). Different techniques exist where the difference between them is primarily the assumptions made about the parameter of interest and the form of the optimization criterion used. The maximum likelihood approach, first applied to speech enhancement by McAulay and Malpass [14], is a popular approach, but it assumes that the unknown parameter of interest is deterministic. If the parameter is assumed to be a random variable, we need to estimate the realization of that random variable. This approach is called the Bayesian approach as it is based on Bayes's theorem. The Bayesian estimators typically perform better than the ML estimators as they incorporate prior knowledge in the estimator and improve on the estimation accuracy.

2.2.1 Bayesian MMSE Estimator

Let $y(n) = x(n) + d(n)$ be the observed noisy speech signal composed of the clean speech signal $x(n)$ and the uncorrelated additive noise $d(n)$, where n is a discrete time index. In the STFT domain, we have

$$Y_\ell(k) = X_\ell(k) + D_\ell(k) \tag{2.61}$$

where k is the frequency bin index ($k = 0, 1, \ldots, K-1$), and ℓ is the time frame index ($\ell = 0, 1, \ldots, M-1$). $X_\ell(k)$ and $D_\ell(k)$ are the respective STFT of $x(n)$ and $d(n)$.

One approach was to find an optimal estimator that would minimize the MSE between the magnitude and estimate magnitude of the speech signal,

$$e\left(\hat{A}_\ell(k), A_\ell(k)\right) = E\left[\left(\hat{A}_\ell(k) - A_\ell(k)\right)^2\right], \tag{2.62}$$

where we define $A_\ell(k) = |X_\ell(k)|$ the spectral magnitude, and $\hat{A}_\ell(k)$ the estimate of the spectral amplitude of the clean speech signal, respectively. The MMSE estimator of $\hat{A}_\ell(k)$ of $A_\ell(k)$ is obtained as a conditional expectation

$$\begin{aligned}\hat{A}_\ell(k) &= \int A_\ell(k)\, p\left(A_\ell(k) | Y_\ell(0), Y_\ell(1), \ldots, Y_\ell(M-1)\right) dA_\ell(k) \\ &= E\left[A_\ell(k) | Y_\ell(0), Y_\ell(1), \ldots, Y_\ell(M-1)\right].\end{aligned} \tag{2.63}$$

The MMSE estimator requires knowledge about the probability distributions of the speech and noise spectral coefficients. Ephraim and Malah proposed [15] the Gaussian statistical model, which makes the following two assumptions:

1. The noise spectral coefficients $D_\ell(k)$ are zero mean statistically independent Gaussian random variables. The real and imaginary parts of $D_\ell(k)$ are independent and identically distributed (iid) random variables $\sim \mathcal{N}\left(0, \frac{\sigma_\ell^2(k)}{2}\right)$

2. The speech spectral coefficients $X_\ell(k)$ are zero mean statistically independent Gaussian random variables. The real and imaginary parts of $X_\ell(k)$ are iid random variables $\sim \mathcal{N}\left(0, \frac{\lambda_\ell(k)}{2}\right)$

Using these assumptions, the MMSE estimator in Equation (2.63) can be expressed as

$$\begin{aligned}\hat{A}_\ell &= E\left[A_\ell | Y_\ell\right] \\ &= \frac{\sqrt{\pi}}{2}\frac{\sqrt{\nu_\ell}}{\gamma_\ell}\exp\left(-\frac{\nu_\ell}{2}\right)\left[(1+\nu_\ell) I_0\left(\frac{\nu_\ell}{2}\right) + \nu_\ell I_1\left(\frac{\nu_\ell}{2}\right)\right]|Y_\ell|, \end{aligned} \tag{2.64}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, and we have defined

$$\gamma_\ell = \frac{|Y_\ell|^2}{\sigma_\ell^2}, \tag{2.65}$$

$$\xi_\ell = \frac{\lambda_\ell}{\sigma_\ell^2}, \tag{2.66}$$

and

$$\nu_\ell = \frac{\xi_\ell}{1+\xi_\ell}\gamma_\ell. \tag{2.67}$$

Note that the estimator can be found independently for each frequency bin index, hence we have dropped the frequency index k for readability. The terms ξ_ℓ and γ_ℓ are referred to respectively as the a-priori SNR and the a-posteriori SNR. We can express Equation (2.64) in terms of a gain function, i.e.

$$\hat{A}_\ell = G_{STSA}(\xi_\ell, \gamma_\ell)|Y_\ell|, \tag{2.68}$$

where $G_{STSA}(\xi_\ell, \gamma_\ell)$ is the spectral gain function for the short time spectral amplitude (STSA)

$$G_{STSA}(\xi_\ell, \gamma_\ell) = \frac{\sqrt{\pi}}{2}\frac{\sqrt{\nu_\ell}}{\gamma_\ell}\exp\left(-\frac{\nu_\ell}{2}\right)\left[(1+\nu_\ell) I_0\left(\frac{\nu_\ell}{2}\right) + \nu_\ell I_1\left(\frac{\nu_\ell}{2}\right)\right]. \tag{2.69}$$

In addition to the optimal spectral magnitude, we also need a phase estimator, as the magnitude alone is not sufficient to reconstruct the signal. As we would like to retain the amplitude estimation which was optimal, the MMSE estimator for the phase is derived subject to a constraint that the modulus of the resulting estimator is 1. The optimal solution is the phase of the noisy signal. Hence, the estimator for the speech signal is

$$\hat{X}_\ell = G_{STSA}\left(\xi_\ell, \gamma_\ell\right) Y_\ell. \tag{2.70}$$

Another approach, also suggested by Ephariam and Malah [16], is to minimize the squared error of the log of the spectral magnitude

$$e\left(\hat{A}_\ell, A_\ell\right) = E\left[\left(\log \hat{A}_\ell - \log A_\ell\right)^2\right], \tag{2.71}$$

for which the optimal estimator is

$$\log \hat{A}_\ell = E\left[\log A_\ell | Y_\ell\right], \tag{2.72}$$

hence

$$\hat{A}_\ell = \exp\left\{E\left[\log A_\ell | Y_\ell\right]\right\}. \tag{2.73}$$

Using the same statistical model as in the STSA case, we obtain the optimal log-MMSE estimator

$$\hat{X}_\ell = G_{LSA}\left(\xi_\ell, \gamma_\ell\right) Y_\ell, \tag{2.74}$$

where $G_{LSA}\left(\xi_\ell, \gamma_\ell\right)$ is the spectral gain function for the log spectral amplitude (LSA)

$$G_{LSA}\left(\xi_\ell, \gamma_\ell\right) = \frac{\xi_\ell}{\xi_\ell + 1} \exp\left[\frac{1}{2}\int_{\nu_\ell}^{\infty} \frac{e^{-t}}{t} dt\right] \tag{2.75}$$

Similarly, it is possible to derive the pth power magnitude spectrum estimator that minimizes the error

$$e\left(\hat{A}_\ell, A_\ell\right) = E\left[\left(\hat{A}_\ell^p - A_\ell^p\right)^2\right]. \tag{2.76}$$

The estimator can be obtained by taking the pth root of the conditional expectation of the pth power magnitude

$$\hat{A}_\ell = \left(E\left[A_\ell^p | Y_\ell\right]\right)^{1/p}, \tag{2.77}$$

and the derived speech estimator

$$\hat{X}_\ell = G_p\left(\xi_\ell, \gamma_\ell\right) Y_\ell, \tag{2.78}$$

when $G_p(\xi_\ell, \gamma_\ell)$ is the spectral gain function

$$G_p(\xi_\ell, \gamma_\ell) = \frac{\sqrt{\nu_\ell}}{\gamma_\ell}\left[\Gamma\left(\frac{p}{2}+1\right)\Phi\left(-\frac{p}{2}, 1; -\nu_\ell\right)\right], \tag{2.79}$$

where $\Gamma(\cdot)$ denotes the gamma function and $\Phi(a, b; c)$ denotes the confluent hypergeometric function. Some cases of note

- For $p = 1$ we get the STSA-MMSE
- For $p = 2$ we get the spectral power (SP) estimate

$$G_{SP}(\xi_\ell, \gamma_\ell) = \sqrt{\frac{\xi_\ell}{\xi_\ell + 1}\left(\frac{1}{\gamma_\ell} + \frac{\xi_\ell}{\xi_\ell + 1}\right)} \tag{2.80}$$

In Figure 2.3 we plot the MMSE gain functions as a function of the a-priori SNR ξ_ℓ for fixed values of $\gamma_\ell - 1$ referred to as the instantaneous SNR. We see that for large values of the instantaneous SNR the MMSE gain functions are similar to the Wiener gain function, which is given by

$$G_W(\xi_\ell) = \frac{\xi_\ell}{\xi_\ell + 1}. \tag{2.81}$$

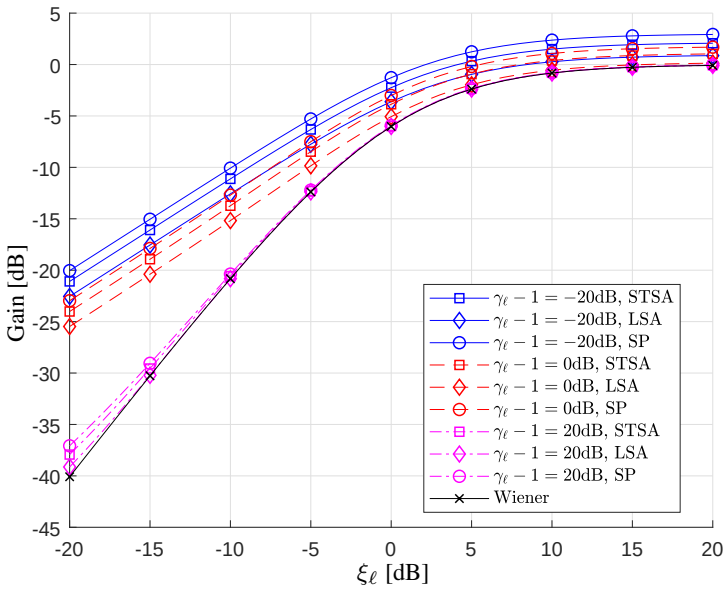

Figure 2.3: Attenuation curves of the different spectral gain estimators as a function of the a-priori SNR ξ_ℓ, the STSA (square), LSA (diamond), SP (circle), and Wiener (cross) for different values of the instantaneous SNR $\gamma_\ell - 1$: -20 dB (solid blue), 0 dB (dashed red), and 20 dB (dot-dash magenta).

Alternatively, in Figure 2.4 we plot the MMSE gain functions as a function of the instantaneous SNR for fixed values of ξ_ℓ. The curves are relatively flat for a large range of the instantaneous SNR for $\xi_\ell \geq -10$ dB, indicating that the a-posteriori SNR γ_ℓ has a small impact on the reduction, and it is evident for very low values of the a-priori SNR ξ_ℓ. The a-priori SNR ξ_ℓ is the main parameter influencing the reduction.

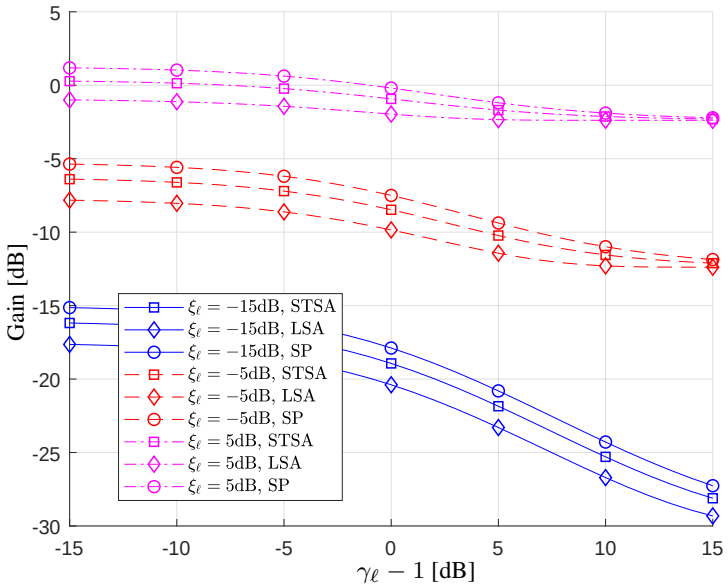

Figure 2.4: Attenuation curves of the different spectral gain estimators as a function of the instantaneous SNR $\gamma_\ell - 1$, the STSA (square), LSA (diamond), SP (circle), for different values of ξ_ℓ: -15 dB (solid blue), -5 dB (dashed red), and 5 dB (dot-dash magenta).

A key assumption in the derivations of the above spectral gains was the Gaussian distribution. The Gaussian assumption holds asymptotically for long duration analysis frames, but does not hold for the speech coefficients which are estimated using relatively short durations of 20-30ms. Hence, researches have proposed to use other distributions such as the gamma [36] and Laplacian [37] distributions.

2.2.2 Decision-Directed

For the various distributions both the a-priori SNR ξ_ℓ and the noise variance σ_ℓ^2 are required, while in practice we only have access to the observed noisy signal. Perhaps the most well-known approach to obtain the a-priori SNR is the "decision-directed" approach by Ephraim and Malah [15], which is based on the definition of the a-priori SNR and its relationship to the a-posteriori

SNR.

$$\begin{aligned}\xi_\ell &= \frac{E\left[A_\ell^2\right]}{\sigma_\ell^2} \\ &= \frac{E\left[|Y_\ell|^2 - |D_\ell|^2\right]}{\sigma_\ell^2} \\ &= E\left[\gamma_\ell\right] - 1 \end{aligned} \tag{2.82}$$

Using Equation (2.66) and Equation (2.82) we can write

$$\xi_\ell = E\left[\frac{1}{2}\frac{A_\ell^2}{\sigma_\ell^2} + \frac{1}{2}\left(\gamma_\ell - 1\right)\right]. \tag{2.83}$$

The estimator is then a recursive version of Equation (2.83)

$$\hat{\xi}_\ell = \alpha\frac{\hat{A}_{\ell-1}^2}{\sigma_{\ell-1}^2} + (1-\alpha)\,P\left[\left(\gamma_\ell - 1\right)\right], \quad 0 \le \alpha < 1. \tag{2.84}$$

where $P(x) = x$ if $x \ge 0$ and $P(x) = 0$ otherwise. The estimator is obtained by dropping the expectation operator and using $\hat{A}_{\ell-1}$ the amplitude estimator of the previous frame instead of the amplitude of the current frame, and introducing a weighing factor α between the two terms of ξ_ℓ. The operator P is used to ensure that the estimator is non-negative. This estimator requires initial conditions for the first frame, which were recommended to be:

$$\hat{\xi}_0 = \alpha + (1-\alpha)\,P\left[\left(\gamma_0 - 1\right)\right]. \tag{2.85}$$

Following this proposal, several improvements were made, where the main one was to limit the smallest value the estimator can get [38]

$$\hat{\xi}_\ell = \max\left\{\alpha\frac{\hat{A}_{\ell-1}^2}{\sigma_{\ell-1}^2} + (1-\alpha)\,P\left[\left(\gamma_\ell - 1\right)\right], \xi_{\min}\right\} \quad 0 \le \alpha < 1, \tag{2.86}$$

where $\xi_{\min}$ is the minimum value allowed for the estimator. This flooring of the indicator is important for reducing musical noise.

2.2.3 Incorporating Speech Presence Uncertainty

For the estimators derived above, it was assumed that the speech was present at all times and frequencies. In practice though, speech contains pauses and may not be present in all frequencies. Taking this into account in the reduction of the noise could improve speech quality.

Speech presence can be modeled by a binary hypothesis:

$$\begin{aligned} H_0^{(\ell,k)} &: \text{Speech absent}, |Y_\ell(k)| = |D_\ell(k)| \\ H_1^{(\ell,k)} &: \text{Speech present}, |Y_\ell(k)| = |X_\ell(k) + D_\ell(k)|, \end{aligned} \tag{2.87}$$

where $H_0^{(\ell,k)}$ denotes the hypothesis that speech is absent in frame ℓ and frequency bin k, and $H_1^{(\ell,k)}$ denotes the hypothesis that speech is present. Once more, we will drop the frequency bin index k for readability. Incorporating the speech presence model into the MMSE estimator for the spectral magnitude, the estimator is

$$\hat{A}_\ell = E\left[A_\ell|Y_\ell, H_1^\ell\right] p\left(H_1^\ell|Y_\ell\right) + E\left[A_\ell|Y_\ell, H_0^\ell\right] p\left(H_0^\ell|Y_\ell\right), \tag{2.88}$$

where $p\left(H_1^\ell|Y_\ell\right)$ is the conditional probability that speech is present given the noisy speech and similarly $p\left(H_1^\ell|Y_\ell\right)$ is the conditional probability that speech is absent given the noisy speech. The term $E\left[A_\ell|Y_\ell, H_0^\ell\right]$ is zero, hence,

$$\hat{A}_\ell = E\left[A_\ell|Y_\ell, H_1^\ell\right] p\left(H_1^\ell|Y_\ell\right). \tag{2.89}$$

To calculate the term $p\left(H_1^\ell|Y_\ell\right)$, we use Bayes's rule

$$\begin{aligned} p\left(H_1^\ell|Y_\ell\right) &= \frac{p\left(Y_\ell|H_1^\ell\right) p\left(H_1^\ell\right)}{p\left(Y_\ell|H_{1,\ell}\right) p\left(H_1^\ell\right) + p\left(Y_\ell|H_0^\ell\right) p\left(H_0^\ell\right)} \\ &= \frac{\Lambda\left(Y_\ell, q_\ell\right)}{1 + \Lambda\left(Y_\ell, q_\ell\right)} \end{aligned} \tag{2.90}$$

where $\Lambda\left(Y_\ell, q_\ell\right)$ is defined as

$$\Lambda\left(Y_\ell, q_\ell\right) = \frac{1 - q_\ell}{q_\ell} \frac{p\left(Y_\ell|H_1^\ell\right)}{p\left(Y_\ell|H_0^\ell\right)} \tag{2.91}$$

and $q_\ell = p\left(H_0^\ell\right)$. Assuming the Gaussian statistical model, the conditional PDFs of the observed signals given the hypothesis are

$$p\left(Y_\ell|H_0^\ell\right) = \frac{1}{\pi\sigma_\ell^2} \exp\left(\frac{-|Y_\ell|^2}{\sigma_\ell^2}\right) \tag{2.92}$$

$$p\left(Y_\ell|H_1^\ell\right) = \frac{1}{\pi\left[\sigma_\ell^2 + \lambda_\ell\right]} \exp\left(\frac{-|Y_\ell|^2}{\sigma_\ell^2 + \lambda_\ell}\right) \tag{2.93}$$

and plugging them into Equation (2.91) we get

$$\Lambda\left(Y_\ell, q_\ell, \xi_\ell'\right) = \frac{1 - q_\ell}{q_\ell} \frac{\exp\left[\frac{\xi_\ell'}{1+\xi_\ell'}\gamma_\ell\right]}{1 + \xi_\ell'}, \tag{2.94}$$

where ξ'_ℓ is the conditional a-priori SNR

$$\xi'_\ell = \frac{E\left[A_\ell^2 | H_1^\ell\right]}{\sigma_\ell^2}. \tag{2.95}$$

The conditional a-priori SNR ξ'_ℓ can be expressed in terms of the a-priori SNR ξ_ℓ:

$$\begin{aligned}\xi'_\ell &= \frac{E\left[A_\ell^2 | H_1^\ell\right]}{\sigma_\ell^2} \\ &= \frac{E\left[A_\ell^2\right]}{p\left(H_1^\ell\right)\sigma_\ell^2} \\ &= \frac{\xi_\ell}{(1-q_\ell)}.\end{aligned} \tag{2.96}$$

By using $\Lambda\left(Y_\ell, q_\ell, \xi'_\ell\right)$ from Equation (2.94) in Equation (2.90), we get

$$p\left(H_1^\ell | Y_\ell\right) = \frac{1-q_\ell}{1-q_\ell + q_\ell\left(1+\xi'_\ell\right)\exp\left(-\nu'_\ell\right)}, \tag{2.97}$$

where

$$\nu'_\ell = \frac{\xi'_\ell}{\xi'_\ell + 1}\gamma_\ell. \tag{2.98}$$

We can see that when ξ'_ℓ is large, suggesting that speech is present, indeed $p\left(H_1^\ell | Y_\ell\right) \approx 1$, and when ξ'_ℓ is really small $p\left(H_1^\ell | Y_\ell\right) \approx 1 - q_\ell = p\left(H_1^\ell\right)$ which is the probability of the speech presence.

And finally, the spectral amplitude estimator incorporating the signal presence uncertainty is

$$\begin{aligned}\hat{A}_\ell &= E\left[A_\ell | Y_\ell, H_1^\ell\right] p\left(H_1^\ell | Y_\ell\right) \\ &= G_{STSA}\left(\xi_\ell, \gamma_\ell\right)\Big|_{\xi_\ell = \xi'_\ell} \cdot |Y_\ell| \cdot p\left(H_1^\ell | Y_\ell\right) \\ &= \frac{1-q_\ell}{1-q_\ell + q_\ell\left(1+\xi'_\ell\right)\exp\left(-\nu'_\ell\right)} G_{STSA}\left(\xi'_\ell, \gamma_\ell\right)|Y_\ell|,\end{aligned} \tag{2.99}$$

where $G_{STSA}\left(\xi_\ell, \gamma_\ell\right)$ is the gain function in Equation (2.69).

Similarly, it is possible to derive the log-MMSE estimator that takes into account the signal presence probability. Combining Equation (2.72) and Equation (2.89), we can write

$$\log \hat{A}_\ell = E\left[\log A_\ell | Y_\ell, H_1^\ell\right] p\left(H_1^\ell | Y_\ell\right), \tag{2.100}$$

hence

$$\begin{aligned}\hat{A}_\ell &= e^{E\left[\log A_\ell|Y_\ell,H_1^\ell\right]p\left(H_1^\ell|Y_\ell\right)} \\ &= \left(e^{E\left[\log A_\ell|Y_\ell,H_1^\ell\right]}\right)^{p\left(H_1^\ell|Y_\ell\right)} \\ &= \left(G_{LSA}\left(\xi'_\ell,\nu'_\ell\right)|Y_\ell|\right)^{p\left(H_1^\ell|Y_\ell\right)}, \end{aligned} \tag{2.101}$$

where G_{LSA} is the gain function in Equation (2.75), ξ'_ℓ and ν'_ℓ are defined in Equation (2.95) and Equation (2.98) respectively. However, this estimator does not result in improved results compared to the original LSA estimator [16]. Consequently, Malah et al. [39] proposed to modify the estimator so that it is multiplicative, i.e.

$$\begin{aligned}\hat{A}_\ell &= G_{LSA}\left(\xi'_\ell,\nu'_\ell\right)|Y_\ell|\, p\left(H_1^\ell|Y_\ell\right) \\ &= G_{MMLSA}\left(\xi'_\ell,\nu'_\ell\right)|Y_\ell|, \end{aligned} \tag{2.102}$$

where $G_{MMLSA}\left(\xi'_\ell,\nu'_\ell\right)$ is the multiplicative modified gain function for the log spectral amplitude (MMLSA)

$$G_{MMLSA}\left(\xi'_\ell,\nu'_\ell\right) = G_{LSA}\left(\xi'_\ell,\nu'_\ell\right) p\left(H_1^\ell|Y_\ell\right). \tag{2.103}$$

Due to the multiplication change, this estimator is not optimal. Hence, Cohen [40] proposed an optimally modified estimator:

$$\log\hat{A}_\ell = E\left[\log A_\ell|Y_\ell,H_1^\ell\right] p\left(H_1^\ell|Y_\ell\right) + E\left[\log A_\ell|Y_\ell,H_0^\ell\right] p\left(H_0^\ell|Y_\ell\right), \tag{2.104}$$

where we don't assume that $E\left[\log A_\ell|Y_\ell,H_0^\ell\right]$ is zero, rather that it is very small. So now we get

$$\begin{aligned}\hat{A}_\ell &= e^{E\left[\log A_\ell|Y_\ell,H_1^\ell\right]p\left(H_1^\ell|Y_\ell\right)} e^{E\left[\log A_\ell|Y_\ell,H_0^\ell\right]p\left(H_0^\ell|Y_\ell\right)} \\ &= \left(e^{E\left[\log A_\ell|Y_\ell,H_1^\ell\right]}\right)^{p\left(H_1^\ell|Y_\ell\right)} \left(e^{E\left[\log A_\ell|Y_\ell,H_0^\ell\right]}\right)^{p\left(H_0^\ell|Y_\ell\right)} \\ &= \left(e^{E\left[\log A_\ell|Y_\ell,H_1^\ell\right]}\right)^{p\left(H_1^\ell|Y_\ell\right)} \left(e^{E\left[\log A_\ell|Y_\ell,H_0^\ell\right]}\right)^{1-p\left(H_1^\ell|Y_\ell\right)}. \end{aligned} \tag{2.105}$$

The first exponential term is the LSA estimator. The second exponential term is assumed to be small and is set to $G_{\min}|Y_\ell|$ where $G_{\min}$ is a lower bound threshold. The estimator can be expressed as:

$$\begin{aligned}\hat{A}_\ell &= \left(G_{LSA}\left(\xi'_\ell,\nu'_\ell\right)|Y_\ell|\right)^{p\left(H_1^\ell|Y_\ell\right)} \left(G_{\min}|Y_\ell|\right)^{1-p\left(H_1^\ell|Y_\ell\right)} \\ &= \left[G_{LSA}\left(\xi'_\ell,\nu'_\ell\right)^{p\left(H_1^\ell|Y_\ell\right)} G_{\min}^{1-p\left(H_1^\ell|Y_\ell\right)}\right]|Y_\ell| \\ &= G_{OMLSA}\left(\xi'_\ell,\nu'_\ell\right)|Y_\ell|, \end{aligned} \tag{2.106}$$

where $G_{OMLSA}\left(\xi'_\ell, \nu'_\ell\right)$ is the optimally modified gain function for the log spectral amplitude (OMLSA)

$$G_{OMLSA}\left(\xi'_\ell, \nu'_\ell\right) = G_{LSA}\left(\xi'_\ell, \nu'_\ell\right)^{p\left(H_1^\ell|Y_\ell\right)} G_{\min}^{1-p\left(H_1^\ell|Y_\ell\right)}. \tag{2.107}$$

The OMLSA compared to the original LSA and the MMLSA was shown to have improved performance [40, 41].

2.3 Microphone Arrays

This section contains the required background for chapter 5. Microphone arrays consist of multiple microphones organized in a structure to spatially sample the sound field. Spatial sampling and thence filtering can be beneficial to the noise reduction problem.

2.3.1 Problem Formulation and Definitions

In a typical microphone array environment, the desired speech signal is coming from the talker's position, while the corrupting noise signals generally originate from other space locations. The beamformer, which is a spatial filter operating on the outputs of the microphone array to form a desired beam, i.e. directivity pattern, reduces or enhances signals based on the location of the sources of the signals. In most beamforming applications, the far-field assumption is used, where the signal sources are located far enough away so that the wave fronts impinging on the microphone array can be modeled as a plane wave. The geometry of the array can be important to the processing, depending on the application. The conventional regular geometry is the uniform linear array (ULA), which consists of M omnidirectional microphones with distance δ between successive microphones. The direction of the desired source signal to the array is the angle θ. Neglecting the propagation attenuation, the signal received at the mth microphone is

$$\begin{aligned} y_m(n) &= x_m(n) + v_m(n) \\ &= x(n - t - \tau_m) + v_m(n), \quad m = 1, 2, \ldots, M, \end{aligned} \tag{2.108}$$

where n is the discrete time index, t is the propagation time from the source to microphone 1 which will be used as our reference point, τ_m is the relative time delay between the mth microphone to microphone 1, and $x_m(n)$ and $v_m(n)$ are the desired signal and the noise signal at the mth microphone, respectively. Given the geometry of the ULA shown in Figure 2.5, the delay τ_m can be expressed as

$$\tau_m = (m-1)\,\tau_0 \cos\theta, \quad m = 1, 2, \ldots, M, \tag{2.109}$$

where $\tau_0 = \delta/c$ and c is the speed of sound through air.

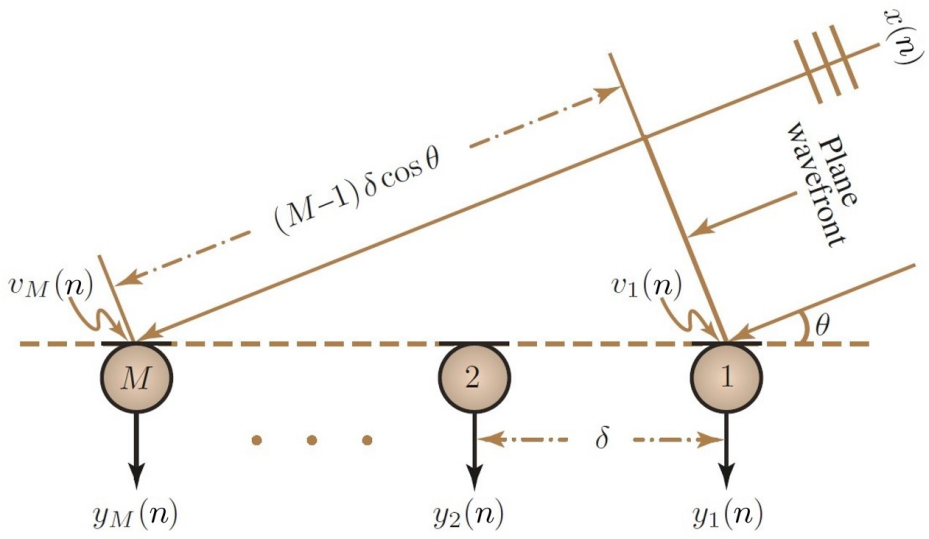

Figure 2.5: Illustration of a uniformly spaced linear additive microphone array for sound capture in the far-field [1]

Beamforming can be done either in the time domain or in the frequency domain. We will focus on the frequency domain, where the signal model in Equation (2.108) becomes

$$\begin{aligned} Y_m(\omega) &= X_m(\omega) + V_m(\omega) \\ &= X(\omega)\, e^{-j\omega(t+\tau_m)} + V_m(\omega) \\ &= X_1(\omega)\, e^{-j\omega\tau_m} + V_m(\omega), \quad m = 1, 2, \ldots, M, \end{aligned} \tag{2.110}$$

where $j = \sqrt{-1}$ is the imaginary unit, $\omega = 2\pi f$ is the angular frequency, f is the temporal frequency, and $Y_m(\omega)$, $X_m(\omega)$, $V_m(\omega)$, $X(\omega)$, and $X_1(\omega)$ are the frequency domain representation of $y_m(n)$, $x_m(n)$, $v_m(n)$, $x(n)$, and $x_1(n)$ (the signal at microphone 1), respectively.

In vector form we can write Equation (2.110) as

$$\begin{aligned} \mathbf{y}(\omega) &= \begin{bmatrix} Y_1(\omega) & Y_2(\omega) & \cdots & Y_M(\omega) \end{bmatrix}^T \\ &= \mathbf{x}(\omega) + \mathbf{v}(\omega) \\ &= \mathbf{d}(\omega, \theta)\, X_1(\omega) + \mathbf{v}(\omega), \end{aligned} \tag{2.111}$$

where the superscript $(\cdot)^T$ is the transpose operator, $\mathbf{v}(\omega)$ is defined similarly to $\mathbf{y}(\omega)$, and $\mathbf{x}(\omega) = \mathbf{d}(\omega, \theta)\, X_1(\omega)$ when we define the steering vector

$$\mathbf{d}(\omega, \theta) = \begin{bmatrix} 1 & e^{-j\omega\cos\theta\tau_0} & \cdots & e^{-j(M-1)\omega\cos\theta\tau_0} \end{bmatrix}^T. \tag{2.112}$$

For $\theta = 0$, known as the end-fire direction, the steering vector can be written as

$$\mathbf{d}(\omega) = \mathbf{d}(\omega, 0) = \begin{bmatrix} 1 & e^{-j\omega\tau_0} & \cdots & e^{-j(M-1)\omega\tau_0} \end{bmatrix}^T. \tag{2.113}$$

The beamformer output is formed by applying a complex weights vector

$$\mathbf{h}(\omega) = \begin{bmatrix} H_1(\omega) & H_2(\omega) & \cdots & H_M(\omega) \end{bmatrix}^T \tag{2.114}$$

to the microphone inputs,

$$\begin{aligned} Z(\omega) &= \sum_{m=1}^{M} H_m^*(\omega) Y_m(\omega) \\ &= \mathbf{h}^H(\omega) \mathbf{y}(\omega) \\ &= \mathbf{h}^H(\omega) \mathbf{d}(\omega, \theta) X_1(\omega) + \mathbf{h}^H(\omega) \mathbf{v}(\omega) \end{aligned} \tag{2.115}$$

where $Z(\omega)$ is supposed to be the estimate of the desired signal up to delay t, and the superscripts $(\cdot)^*$ and $(\cdot)^H$ are the conjugate and conjugate-transpose operators, respectively. The process of finding the appropriate filter so that $Z(\omega)$ is a good estimate is called beamforming, when the coefficients $H_m(\omega)$ are known as the beamforming coefficients. The beamforming is distortionless when

$$\mathbf{h}^H(\omega) \mathbf{d}(\omega, \theta) = 1. \tag{2.116}$$

Each beamformer has a pattern of directional sensitivity, called the beam pattern or directivity pattern, which describes the sensitivity of the beamformer to a plane wave coming from a source signal impinging on the array from direction θ_s. The beam pattern is defined as:

$$\begin{aligned} \mathcal{B}[\mathbf{h}(\omega), \theta_s] &= \mathbf{d}^H(\omega, \theta_s) \mathbf{h}(\omega) \\ &= \sum_{m=1}^{M} H_m(\omega) e^{j(m-1)\omega\tau_0 \cos\theta_s} \end{aligned} \tag{2.117}$$

We define the input signal to noise (SNR) ratio as the SNR at the reference microphone, microphone 1:

$$\text{iSNR}(\omega) = \frac{\phi_{X_1}(\omega)}{\phi_{V_1}(\omega)} = \frac{\phi_X(\omega)}{\phi_{V_1}(\omega)}, \tag{2.118}$$

where $\phi_{X_1}(\omega) = E\left[|X_1(\omega)|^2\right]$, $\phi_X(\omega) = E\left[|X(\omega)|^2\right]$, and $\phi_{V_1}(\omega) = E\left[|V_1(\omega)|^2\right]$ are the variances of $X_1(\omega)$, $X(\omega)$, and $V_1(\omega)$, respectively,

with $E[\cdot]$ denoting mathematical expectation. The output SNR of the beamformer is defined as

$$\begin{aligned}\text{oSNR}\left[\mathbf{h}\left(\omega\right)\right] &= \frac{\phi_X\left(\omega\right)\left|\mathbf{h}^H\left(\omega\right)\mathbf{d}\left(\omega,\theta\right)\right|^2}{\mathbf{h}^H\left(\omega\right)\mathbf{\Phi}_{\mathbf{v}}\left(\omega\right)\mathbf{h}\left(\omega\right)} \\ &= \frac{\phi_X\left(\omega\right)}{\phi_{V_1}\left(\omega\right)} \times \frac{\left|\mathbf{h}^H\left(\omega\right)\mathbf{d}\left(\omega,\theta\right)\right|^2}{\mathbf{h}^H\left(\omega\right)\mathbf{\Gamma}_{\mathbf{v}}\left(\omega\right)\mathbf{h}\left(\omega\right)},\end{aligned} \tag{2.119}$$

where $\mathbf{\Phi}_{\mathbf{v}}\left(\omega\right) = E\left[\mathbf{v}\left(\omega\right)\mathbf{v}^H\left(\omega\right)\right]$ and $\mathbf{\Gamma}_{\mathbf{v}}\left(\omega\right) = \frac{\mathbf{\Phi}_{\mathbf{v}}(\omega)}{\phi_{V_1}(\omega)}$ are the correlation and the pseudo-coherence matrix of $\mathbf{v}\left(\omega\right)$, respectively. From the definitions in Equation (2.118) and Equation (2.119) we can easily derive the gain in SNR:

$$\begin{aligned}\mathcal{G}\left[\mathbf{h}\left(\omega\right)\right] &= \frac{\text{oSNR}\left[\mathbf{h}\left(\omega\right)\right]}{\text{iSNR}\left(\omega\right)} \\ &= \frac{\left|\mathbf{h}^H\left(\omega\right)\mathbf{d}\left(\omega,\theta\right)\right|^2}{\mathbf{h}^H\left(\omega\right)\mathbf{\Gamma}_{\mathbf{v}}\left(\omega\right)\mathbf{h}\left(\omega\right)}.\end{aligned} \tag{2.120}$$

Assuming the matrix $\mathbf{\Gamma}_{\mathbf{v}}\left(\omega\right)$ is nonsingular, we have

$$\left|\mathbf{h}^H\left(\omega\right)\mathbf{d}\left(\omega,\theta\right)\right|^2 \leq \left[\mathbf{h}^H\left(\omega\right)\mathbf{\Gamma}_{\mathbf{v}}\left(\omega\right)\mathbf{h}\left(\omega\right)\right]\left[\mathbf{d}^H\left(\omega,\theta\right)\mathbf{\Gamma}_{\mathbf{v}}^{-1}\left(\omega\right)\mathbf{d}\left(\omega,\theta\right)\right], \tag{2.121}$$

in which case we can obtain an upper limit for the gain in SNR:

$$\mathcal{G}\left[\mathbf{h}\left(\omega\right)\right] \leq \mathbf{d}^H\left(\omega,\theta\right)\mathbf{\Gamma}_{\mathbf{v}}^{-1}\left(\omega\right)\mathbf{d}\left(\omega,\theta\right). \tag{2.122}$$

For the specific identity filter

$$\mathbf{h}\left(\omega\right) = \mathbf{i}_1 = \left[\begin{array}{cccc} 1 & 0 & \cdots & 0 \end{array}\right]^T \tag{2.123}$$

the output SNR is the same as the input SNR, i.e. $\mathcal{G}\left[\mathbf{h}\left(\omega\right)\right] = 1$. The desire it to find a filter $\mathbf{h}\left(\omega\right)$ so that $\mathcal{G}\left[\mathbf{h}\left(\omega\right)\right] > 1$.

There are two types of noise that we are commonly interested in:

- White noise
  The noise signals at the different microphones are uncorrelated with each other and have the same variance. This is a good model for the microphones' self-noise. In this case $\mathbf{\Gamma}_{\mathbf{v}}\left(\omega\right) = \mathbf{I}_M$, where $\mathbf{I}_M$ is the $M \times M$ identity matrix, and we can define the white noise gain (WNG) as
  $$\mathcal{W}\left[\mathbf{h}\left(\omega\right)\right] = \frac{\left|\mathbf{h}^H\left(\omega\right)\mathbf{d}\left(\omega,\theta\right)\right|^2}{\mathbf{h}^H\left(\omega\right)\mathbf{h}\left(\omega\right)}. \tag{2.124}$$

From Equation (2.122) we easily deduce the upper limit for the WNG

$$\mathcal{W}\left[\mathbf{h}\left(\omega\right)\right] \leq M, \quad \forall \mathbf{h}\left(\omega\right), \tag{2.125}$$

with the maximum $\mathcal{W}_{\max} = M$, which is frequency and angle independent.

- Diffuse noise
  Spherically isotropic noise. Many practical noise environments can be characterized by such a noise, for example office or car noise. In this case

$$\left[\mathbf{\Gamma}_{\mathbf{v}}\left(\omega\right)\right]_{ij} = \left[\mathbf{\Gamma}_{\mathrm{d}}\left(\omega\right)\right]_{ij} = \frac{\sin\left[\omega(j-i)\tau_0\right]}{\omega(j-i)\tau_0} = \operatorname{sinc}\left[\omega(j-i)\tau_0\right]. \tag{2.126}$$

  and the gain in SNR, called the directivity factor (DF), is

$$\mathcal{D}\left[\mathbf{h}\left(\omega\right)\right] = \frac{\left|\mathbf{h}^H\left(\omega\right)\mathbf{d}\left(\omega\right)\right|^2}{\mathbf{h}^H\left(\omega\right)\mathbf{\Gamma}_{\mathrm{d}}\left(\omega\right)\mathbf{h}\left(\omega\right)}. \tag{2.127}$$

  From Equation (2.122) we deduce the upper limit for the DF

$$\mathcal{D}\left[\mathbf{h}\left(\omega\right)\right] \leq \mathbf{d}^H\left(\omega,\theta\right)\mathbf{\Gamma}_{\mathrm{d}}^{-1}\left(\omega\right)\mathbf{d}\left(\omega,\theta\right), \quad \forall \mathbf{h}\left(\omega\right). \tag{2.128}$$

  And the maximum DF is

$$\begin{aligned}\mathcal{D}_{\max}\left[\mathbf{h}\left(\omega\right),\theta\right] &= \mathbf{d}^H\left(\omega,\theta\right)\mathbf{\Gamma}_{\mathrm{d}}^{-1}\left(\omega\right)\mathbf{d}\left(\omega,\theta\right) \\ &= \mathrm{tr}\left[\mathbf{\Gamma}_{\mathrm{d}}^{-1}\left(\omega\right)\mathbf{d}\left(\omega,\theta\right)\mathbf{d}^H\left(\omega,\theta\right)\right] \\ &\leq M\mathrm{tr}\left[\mathbf{\Gamma}_{\mathrm{d}}^{-1}\left(\omega\right)\right],\end{aligned} \tag{2.129}$$

  where $\mathrm{tr}\left(\cdot\right)$ denotes the trace of a square matrix. The maximum DF is frequency and angle dependent. When the maximum DF is close to M^2, it is referred to as supergain.

  One of the challenges in the fixed beamformer design is the compromise between the WNG and the DF. We would like to maximize the DF without amplifying the WNG.

2.3.2 Conventional Beamformers

In this section we provide an overview of some important conventional fixed beamformers:

- Delay-and-sum (DS)
  This is the simplest and the most well-known beamformer. It is derived

by maximizing the WNG subject to the distortionless constraint in Equation (2.116):

$$\min_{\mathbf{h}} \mathbf{h}^H(\omega)\,\mathbf{h}(\omega) \quad \text{subject to} \quad \mathbf{h}^H(\omega)\,\mathbf{d}(\omega,\theta) = 1 \tag{2.130}$$

The optimal filter is easily obtained:

$$\begin{aligned}\mathbf{h}_{DS}(\omega,\theta) &= \frac{\mathbf{d}(\omega,\theta)}{\mathbf{d}^H(\omega,\theta)\,\mathbf{d}(\omega,\theta)} \\ &= \frac{\mathbf{d}(\omega,\theta)}{M}.\end{aligned} \tag{2.131}$$

The WNG and the DF for this beamformer are therefore:

$$\mathcal{W}[\mathbf{h}_{DS}(\omega,\theta)] = M = \mathcal{W}_{\max}, \tag{2.132}$$

and

$$\mathcal{D}[\mathbf{h}_{DS}(\omega,\theta)] = \frac{M^2}{\mathbf{d}^H(\omega,\theta)\,\mathbf{\Gamma}_{\mathrm{d}}(\omega)\,\mathbf{d}(\omega,\theta)}. \tag{2.133}$$

As

$$\mathbf{d}^H(\omega,\theta)\,\mathbf{\Gamma}_{\mathrm{d}}(\omega)\,\mathbf{d}(\omega,\theta) \leq M\mathrm{tr}[\mathbf{\Gamma}_{\mathrm{d}}(\omega)] = M^2 \tag{2.134}$$

we have $\mathcal{D}[\mathbf{h}_{DS}(\omega,\theta)] \geq 1$. This means that while the DS beamformer maximizes the WNG it never amplifies the diffuse noise. In reverberant and noisy environments it is essential that the DF is large for good speech enhancement, but the DS beamformer performs poorly even with a large number of microphones. The beampattern is

$$\begin{aligned}\mathcal{B}[\mathbf{h}_{DS}(\omega,\theta),\theta_s] &= \frac{1}{M}\mathbf{d}^H(\omega,\theta_s)\,\mathbf{d}(\omega,\theta) \\ &= \frac{1}{M}\sum_{m=1}^{M} e^{j(m-1)\omega\tau_0(\cos\theta_s-\cos\theta)} \\ &= \frac{1}{M}\frac{\sin\left[\frac{M}{2}\omega\tau_0(\cos\theta_s-\cos\theta)\right]}{\sin\left[\frac{1}{2}\omega\tau_0(\cos\theta_s-\cos\theta)\right]} e^{j\frac{M-1}{2}\omega\tau_0(\cos\theta_s-\cos\theta)},\end{aligned} \tag{2.135}$$

which is clearly frequency dependent.

- Maximum DF
  As the name implies, this beamformer is derived by maximizing the DF subject to the distortionless constraint Equation (2.116):

$$\min_{\mathbf{h}} \mathbf{h}^H(\omega)\,\mathbf{\Gamma}_{\mathrm{d}}(\omega)\,\mathbf{h}(\omega) \quad \text{subject to} \quad \mathbf{h}^H(\omega)\,\mathbf{d}(\omega,\theta) = 1 \tag{2.136}$$

The optimal filter is then

$$\mathbf{h}_{mDF}(\omega,\theta)=\frac{\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}(\omega)\,\mathbf{d}(\omega,\theta)}{\mathbf{d}^{H}(\omega,\theta)\,\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}(\omega)\,\mathbf{d}(\omega,\theta)}, \tag{2.137}$$

and WNG and the DF for this beamformer are respectively

$$\mathcal{W}\left[\mathbf{h}_{mDF}(\omega,\theta)\right]=\frac{\left[\mathbf{d}^{H}(\omega,\theta)\,\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}(\omega)\,\mathbf{d}(\omega,\theta)\right]^{2}}{\mathbf{d}^{H}(\omega,\theta)\,\boldsymbol{\Gamma}_{\mathrm{d}}^{-2}(\omega)\,\mathbf{d}(\omega,\theta)}, \tag{2.138}$$

and

$$\mathcal{D}\left[\mathbf{h}_{mDF}(\omega,\theta)\right]=\mathbf{d}^{H}(\omega,\theta)\,\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}(\omega)\,\mathbf{d}(\omega,\theta)=\mathcal{D}_{max}\left[\mathbf{h}(\omega),\theta\right]. \tag{2.139}$$

The WNG for the maximum DF beamformer can be smaller than 1, meaning that the white noise can be amplified.

Some interesting relationships between the DS and maximum DF beamformers:

$$\frac{1}{\mathbf{h}_{mDF}^{H}(\omega,\theta)\,\mathbf{h}_{DS}(\omega,\theta)}=\mathcal{W}_{\max}, \tag{2.140}$$

$$\frac{1}{\mathbf{h}_{mDF}^{H}(\omega,\theta)\,\boldsymbol{\Gamma}_{\mathrm{d}}(\omega)\,\mathbf{h}_{DS}(\omega,\theta)}=\mathcal{D}_{\max}\left[\mathbf{h}(\omega),\theta\right], \tag{2.141}$$

and

$$\mathcal{D}_{\max}\left[\mathbf{h}(\omega),\theta\right]\boldsymbol{\Gamma}_{\mathrm{d}}(\omega)\,\mathbf{h}_{mDF}^{H}(\omega,\theta)=\mathcal{W}_{\max}\mathbf{h}_{DS}(\omega,\theta). \tag{2.142}$$

- Superdirective (SD)
  The well-known superdirective beamformer is a specific case of the maximum DF beamformer for $\theta=0$ and small δ.

$$\begin{aligned}\mathbf{h}_{SD}(\omega)&=\frac{\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}(\omega)\,\mathbf{d}(\omega,0)}{\mathbf{d}^{H}(\omega,0)\,\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}(\omega)\,\mathbf{d}(\omega,0)}\\&=\frac{\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}(\omega)\,\mathbf{d}(\omega)}{\mathbf{d}^{H}(\omega)\,\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}(\omega)\,\mathbf{d}(\omega)}.\end{aligned} \tag{2.143}$$

  The superdirective beamformer is derived from maximizing the maximum DF. In fact, it can be shown that [42]:

$$\lim_{\delta\to 0}\mathcal{D}_{\max}\left[\mathbf{h}_{SD}(\omega)\right]=M^{2}. \tag{2.144}$$

  As a specific case of the maximum DF, while maximizing the DF the superdirective beamformer can amplify the white noise, i.e. have WNG smaller than 1, especially at low frequencies.

- Robust Superdirective
  Since the superdirective is sensitive to white noise, Cox et al. [33, 32] proposed to maximize the DF subject to a constraint on the WNG. Using the distortionless constraint in Equation (2.116), the optimal filter is

$$\mathbf{h}_{R,\epsilon}(\omega) = \frac{\left[\mathbf{\Gamma}_{\mathrm{d}}(\omega) + \epsilon \mathbf{I}_M\right]^{-1} \mathbf{d}(\omega)}{\mathbf{d}^H(\omega)\left[\mathbf{\Gamma}_{\mathrm{d}}(\omega) + \epsilon \mathbf{I}_M\right]^{-1} \mathbf{d}(\omega)}, \tag{2.145}$$

  where $\epsilon \geq 0$ is a Lagrange multiplier [43]. Clearly Equation (2.145) is a regularized or robust form of Equation (2.143) with ϵ as the regularization parameter. The parameter ϵ enables a compromise between the DF and the WNG. Small values of ϵ lead to large DF and low WNG, while large values of ϵ lead to low DF and large WNG. For the two extreme cases we have

$$\mathbf{h}_{R,0}(\omega) = \mathbf{h}_{SD}(\omega) \tag{2.146}$$

$$\mathbf{h}_{R,\infty}(\omega) = \mathbf{h}_{DS}(\omega). \tag{2.147}$$

  Defining a regularized from of the pseudo-coherence matrix of the diffuse noise

$$\mathbf{\Gamma}_{\epsilon}(\omega) = \mathbf{\Gamma}_{\mathrm{d}}(\omega) + \epsilon \mathbf{I}_M \tag{2.148}$$

  we can express Equation (2.145) as an ϵ regularized superdirective beamformer

$$\mathbf{h}_{R,\epsilon}(\omega) = \frac{\mathbf{\Gamma}_{\epsilon}(\omega)^{-1} \mathbf{d}(\omega)}{\mathbf{d}^H(\omega)\mathbf{\Gamma}_{\epsilon}(\omega)^{-1} \mathbf{d}(\omega)}. \tag{2.149}$$

  The WNG, DF, and beampattern for the robust superdirective beamformer are respectively

$$\mathcal{W}\left[\mathbf{h}_{R,\epsilon}(\omega)\right] = \frac{\left[\mathbf{d}^H(\omega)\mathbf{\Gamma}_{\epsilon}^{-1}(\omega)\mathbf{d}(\omega)\right]^2}{\mathbf{d}^H(\omega)\mathbf{\Gamma}_{\epsilon}^{-2}(\omega)\mathbf{d}(\omega)}, \tag{2.150}$$

$$\mathcal{D}\left[\mathbf{h}_{R,\epsilon}(\omega)\right] = \frac{\left[\mathbf{d}^H(\omega)\mathbf{\Gamma}_{\epsilon}^{-1}(\omega)\mathbf{d}(\omega)\right]^2}{\mathbf{d}^H(\omega)\mathbf{\Gamma}_{\epsilon}^{-1}(\omega)\mathbf{\Gamma}_{\mathrm{d}}(\omega)\mathbf{\Gamma}_{\epsilon}^{-1}(\omega)\mathbf{d}(\omega)}, \tag{2.151}$$

  and

$$\mathcal{B}\left[\mathbf{h}_{R,\epsilon}(\omega), \theta_s\right] = \frac{\mathbf{d}^H(\omega,\theta_s)\mathbf{\Gamma}_{\epsilon}^{-1}(\omega)\mathbf{d}(\omega)}{\mathbf{d}^H(\omega)\mathbf{\Gamma}_{\epsilon}^{-1}(\omega)\mathbf{d}(\omega)}. \tag{2.152}$$

  Since the white noise amplification is worse at low frequencies than at high frequencies, it is better to take ϵ to be frequency dependent. While the robust superdirective beamformer has control on the white noise amplification, it is not easy to find a closed form expression for ϵ for a desired value of the WNG.

2.3.3 Combined Beamformer

As the DS beamformer maximizes the WNG and the regularized superdirective enables control of the DF, Berkun et al. [44] proposed to combine the two beamformers into the beamformer

$$\mathbf{h}_{\alpha,\epsilon}(\omega) = \frac{\left[\mathbf{\Gamma}_{\epsilon}^{-1}(\omega) + \alpha(\omega)\mathbf{I}_M\right]\mathbf{d}(\omega)}{\mathbf{d}^H(\omega)\left[\mathbf{\Gamma}_{\epsilon}^{-1}(\omega) + \alpha(\omega)\mathbf{I}_M\right]\mathbf{d}(\omega)}, \tag{2.153}$$

where $\alpha(\omega)$ is a real number. If we define

$$\mathcal{D}_{\max,\epsilon}(\omega) = \mathbf{d}^H(\omega)\mathbf{\Gamma}_{\epsilon}^{-1}(\omega)\mathbf{d}(\omega) \tag{2.154}$$

then Equation (2.153) becomes

$$\begin{aligned}\mathbf{h}_{\alpha,\epsilon}(\omega) &= \frac{\mathcal{D}_{\max,\epsilon}(\omega)}{\mathcal{D}_{\max,\epsilon}(\omega) + \alpha(\omega)M} \cdot \frac{\mathbf{\Gamma}_{\epsilon}^{-1}(\omega)\mathbf{d}(\omega)}{\mathbf{d}^H(\omega)\mathbf{\Gamma}_{\epsilon}^{-1}(\omega)\mathbf{d}(\omega)} \\ &\quad + \frac{\alpha(\omega)M}{\mathcal{D}_{\max,\epsilon}(\omega) + \alpha(\omega)M} \cdot \frac{\mathbf{d}(\omega)}{M} \\ &= \frac{\mathbf{h}_{R,\epsilon}(\omega)}{1+\alpha_{\epsilon}(\omega)} + \frac{\mathbf{h}_{DS}(\omega)}{1+\alpha_{\epsilon}^{-1}(\omega)},\end{aligned} \tag{2.155}$$

where

$$\alpha_{\epsilon}(\omega) = \alpha(\omega)\frac{\mathcal{W}_{\max}}{\mathcal{D}_{\max,\epsilon}(\omega)}. \tag{2.156}$$

This beamformer is distortionless, i.e. $\mathbf{h}_{\alpha,\epsilon}^H(\omega)\mathbf{d}(\omega) = 1$. The WNG and inverse DF corresponding to the combined beamformer are respectively

$$\mathcal{W}\left[\mathbf{h}_{\alpha,\epsilon}(\omega)\right] = \frac{\left[1+\alpha_{\epsilon}(\omega)\right]^2 \mathcal{W}\left[\mathbf{h}_{DS}(\omega)\right]\mathcal{W}\left[\mathbf{h}_{R,\epsilon}(\omega)\right]}{\mathcal{W}\left[\mathbf{h}_{DS}(\omega)\right] + \left\{\left[1+\alpha_{\epsilon}(\omega)\right]^2 - 1\right\}\mathcal{W}\left[\mathbf{h}_{R,\epsilon}(\omega)\right]}, \tag{2.157}$$

and

$$\begin{aligned}\mathcal{D}^{-1}\left[\mathbf{h}_{\alpha,\epsilon}(\omega)\right] &= \left[1+\alpha_{\epsilon}(\omega)\right]^{-2} \cdot \{\mathcal{D}^{-1}\left[\mathbf{h}_{R,\epsilon}(\omega)\right] \\ &\quad + 2\alpha_{\epsilon}(\omega)\left[\mathcal{D}_{\max,\epsilon}^{-1}(\omega) - \epsilon\frac{1}{M}\right] \\ &\quad + \alpha_{\epsilon}^2(\omega)\mathcal{D}^{-1}\left[\mathbf{h}_{DS}(\omega)\right]\}.\end{aligned} \tag{2.158}$$

The WNG of the combined beamformer depends on the WNGs of the DS and the regularized superdirective, and respectively the DF of the beamformer on the DFs of the DS and the regularized superdirective. We can clearly see that for $\alpha = 0$ we have

$$\mathcal{W}\left[\mathbf{h}_{0,\epsilon}(\omega)\right] = \mathcal{W}\left[\mathbf{h}_{R,\epsilon}(\omega)\right] \tag{2.159}$$

$$\mathcal{D}\left[\mathbf{h}_{0,\epsilon}(\omega)\right] = \mathcal{D}\left[\mathbf{h}_{R,\epsilon}(\omega)\right], \tag{2.160}$$

and for $\alpha \to \infty$

$$\mathcal{W}\left[\mathbf{h}_{\infty,\epsilon}(\omega)\right] = \mathcal{W}\left[\mathbf{h}_{DS}(\omega)\right] \tag{2.161}$$

$$\mathcal{D}\left[\mathbf{h}_{\infty,\epsilon}(\omega)\right] = \mathcal{D}\left[\mathbf{h}_{DS}(\omega)\right]. \tag{2.162}$$

Also, we have the following relationships

$$\mathcal{W}\left[\mathbf{h}_{\alpha,\epsilon}(\omega)\right] \leq \mathcal{W}_{\max} \tag{2.163}$$

$$\mathcal{W}\left[\mathbf{h}_{\alpha,\epsilon}(\omega)\right] \geq \mathcal{W}\left[\mathbf{h}_{R,\epsilon}(\omega)\right], \quad \forall \alpha_\epsilon(\omega) \geq 0 \tag{2.164}$$

$$\mathcal{D}\left[\mathbf{h}_{\alpha,\epsilon}(\omega)\right] \leq \mathcal{D}\left[\mathbf{h}_{R,\epsilon}(\omega)\right] \tag{2.165}$$

$$\mathcal{D}\left[\mathbf{h}_{\alpha,\epsilon}(\omega)\right] \geq \mathcal{D}\left[\mathbf{h}_{DS}(\omega)\right], \quad \forall \alpha_\epsilon(\omega) \geq 0 \tag{2.166}$$

suggesting that $\alpha(\omega)$ should be chosen so that $\alpha(\omega) \geq 0$.

Once ϵ value is set, Berkun et al. provide a closed form solution for $\alpha_\epsilon(\omega)$ and thus for $\alpha(\omega)$, given a desired fixed WNG or desired fixed DF. For a fixed WNG, i.e. $\mathcal{W}\left[\mathbf{h}_{\alpha,\epsilon}(\omega)\right] = \mathcal{W}_0$, with $\mathcal{W}\left[\mathbf{h}_{R,\epsilon}(\omega)\right] < \mathcal{W}_0 < M, \forall\omega$:

$$\alpha_\epsilon(\omega) = \sqrt{\frac{\mathcal{W}_0}{\mathcal{W}_{\max} - \mathcal{W}_0}}\left|\tan\varphi_\epsilon(\omega)\right| - 1, \tag{2.167}$$

where $\varphi_\epsilon(\omega)$ is the angle between the two vectors $\mathbf{d}(\omega)$ and $\mathbf{\Gamma}_\epsilon^{-1}(\omega)\,\mathbf{d}(\omega)$. Then $\alpha(\omega)$ is derived using Equation (2.156).

For a fixed DF, i.e. $\mathcal{D}\left[\mathbf{h}_{\alpha,\epsilon}(\omega)\right] = \mathcal{D}_0$, with $\mathcal{D}\left[\mathbf{h}_{DS,\epsilon}(\omega)\right] < \mathcal{D}_0 < \mathcal{D}\left[\mathbf{h}_{R,\epsilon}(\omega)\right], \forall\omega$:
Equation (2.158) is expressed as a quadratic equation of $\alpha_\epsilon(\omega)$

$$\begin{aligned}\alpha_\epsilon^2(\omega)\left\{\mathcal{D}^{-1}\left[\mathbf{h}_{DS}(\omega)\right] - \mathcal{D}_0^{-1}\right\} + 2\alpha_\epsilon(\omega)\left\{\mathcal{D}^{-1}\left[\mathbf{h}_{\max,\epsilon}(\omega)\right] - \epsilon\frac{1}{M} - \mathcal{D}_0^{-1}\right\} \\ + \mathcal{D}^{-1}\left[\mathbf{h}_{DS}(\omega)\right] - \mathcal{D}_0^{-1} = 0\end{aligned} \tag{2.168}$$

and the positive solution of this equation is taken to derive $\alpha(\omega)$ using Equation (2.156).

The proposed method, while finding a closed form solution for the parameter $\alpha(\omega)$, which enables control of the trade-off in performance between the WNG and the DF, does not address finding the regularization parameter ϵ and assumes it is user determined.

Chapter 3

Adaptive Line Enhancer for Nonstationary Harmonic Noise Reduction

3.1 Introduction

Acoustic background noise is an important factor that degrades both the perceptual speech quality/intelligibility and, unfortunately, it is common in all practical situations, such as telecommunications, teleconferencing, and human-machine interfaces. The process of suppressing this additive noise is known as noise reduction or alternatively as speech enhancement, and it is a fundamental problem that has been researched extensively over the past few decades, e.g., [45, 46, 47] and references therein. Many methods have been proposed for noise reduction, among them spectral subtraction techniques [3, 48], optimal filtering including the notorious Wiener filter [17, 49, 50], statistical-model-based algorithms [15, 16, 41], subspace methods [8, 51], binary mask methods [9, 10], and data-driven supervised learning methods namely deep learning (DL) based on deep neural networks which have recently gained much popularity [52, 53, 54, 55].

The noise reduction problem is traditionally formulated as a linear filtering problem, where we pass the observed noisy signal through a filter to obtain an estimate of the clean speech signal. In the design of this filter, the aim is to achieve maximal noise suppression without introducing noticeable speech distortion. When a single microphone is used to capture the noisy speech, studies [56, 57] have shown that traditional approaches generally do not provide improvement to the intelligibility while they do improve the speech perceptual quality. For DL algorithms it can also be challenging to

improve on both intelligibility and quality; though improving intelligibility, they can suffer from poor sound quality [58, 59]. Recent studies [59, 60, 61] include the processing of the phase spectra in order to achieve better speech quality, alternatively, others [62, 63] incorporate perceptual measures into the training method.

An adaptive line enhancer (ALE) [27, 64] modifies both the magnitude and phase, and so can potentially improve both intelligibility and quality. The ALE is a degenerate form of adaptive noise canceling (ANC), both first introduced in [26]. They work on the principle of correlation cancellation, suppressing noise from a signal using adaptive filters that self adjust their parameters. They involve producing an estimate of the noise by filtering a reference signal and then subtracting this noise estimate from the primary input containing both the speech signal and noise. They have the advantage of updating the filter coefficients automatically with no need for a-priori knowledge of noise and speech signals. While the ANC requires a reference noise signal that is highly correlated with the noise signal, the ALE consists of a single microphone and a delay element to produce a delayed version of the noisy signal to be used as the reference signal. The delay enables separation of the periodic and stochastic components in a signal; it de-correlates the stochastic components between the input and the delayed input, while leaving the periodic components correlated. The periodic components are extracted by the adaptive filter, usually using a normalized least-mean-square (NLMS) algorithm [65]. Depending on the characteristics of the noise, two different approaches are used. The first is to directly estimate the speech signal from wide-band background noise. In this case, the ALE estimates the speech signal using the pitch period for the delay [66]. The second case is when the noise is harmonic, e.g., ventilation fan noise and vehicle engine noise. The filter estimates the noise which is then subtracted from the input to obtain the desired speech signal. In this case, the ALE is typically used in conjunction with a traditional noise estimator or even with another ALE to remove any additional wide-band background noise present [28]. The ALE can also be implemented in the short-time Fourier transform (STFT) domain [29], where adaptive filters have less computational cost and can be calculated separately for each frequency bin. The performance of the ALE is highly dependent on the de-correlation delay parameter and the step size of the adaptive filter, which are either fixed or heuristically optimized; hence noise residuals remain, particularly for nonstationary noise.

Recently, Taghia and Martin [30] have shown that an ALE with a frequency-dependent step size based on mutual information (MI) can detect the presence of harmonic noise and reduce it. To deal with nonstationary noise, they

implemented the algorithm in a block-wise manner and assumed the span of the stationarity of the noise signal is at least as large as the block length. This is not an effective solution for highly nonstationary noise signals such as medical equipment beeps or alarm sounds.

In this chapter, we propose using a combination of both a forward linear predictor (FLP) and a non-causal backward linear predictor (BLP), both implemented with the NLMS algorithm, to better address the nonstationarity of the harmonic noise. The FLP on its own (this is the standard NLMS implementation) reduces the noise transient after a delay determined by the adaptive filter step size and de-correlation delay parameters, and hence residuals of the noise remain. In a similar manner, for the BLP on its own, residuals would remain at the end of the transient. By using a combination of the FLP and the BLP, we are able to increase the reduction span of the noise transient, thus reducing the amount of noise residuals. We use an indicator for noise presence to apply the filters only when noise is present, to reduce the amount of distortion to the speech signal. We also apply a set of changing filter lengths, taking the maximal filter length available according to the indicator, to ensure the combined filter spans throughout the noise transient. With this approach, we are able to achieve higher noise reduction and lower signal distortion, improving the speech quality and intelligibility compared to other methods.

The rest of the chapter is organized as follows. In Section 3.2, we describe the signal model and the ALE system, where we use the orthogonal decomposition introduced in [67, 68, 69]. In Section 3.3, we derive the relevant performance measures. In Section 3.4, we develop the Wiener filter and present our proposed combined filter. The experimental results are then presented in Section 3.5, and finally the conclusions are drawn in Section 3.6.

3.2 Problem Formulation

We consider the following signal model:

$$y(n) = x(n) + v(n), \tag{3.1}$$

where n is the discrete time index, $y(n)$ is the observed noisy signal, $x(n)$ is the zero-mean desired clean speech signal, and $v(n)$ is the zero-mean noise signal. We assume that $x(n)$ and $v(n)$ are real and uncorrelated signals. Using the STFT, (3.1) can be expressed as

$$Y(k,m) = X(k,m) + V(k,m), \tag{3.2}$$

where k (for $k = 0, 1, \ldots, K-1$) is the frequency index, m (for $m = 0, 1, \ldots, M-1$) is the frame index, and $Y(k,m)$, $X(k,m)$, and $V(k,m)$ are the STFTs of $y(n)$, $x(n)$, and $v(n)$, respectively. As $x(n)$ and $v(v)$ are uncorrelated per the assumption, the variance of $Y(k,m)$ is

$$\begin{aligned}\phi_Y(k,m) &= E\left[|Y(k,m)|^2\right] \\ &= E\left[|X(k,m)|^2\right] + E\left[|V(k,m)|^2\right] \\ &= \phi_X(k,m) + \phi_V(k,m),\end{aligned} \tag{3.3}$$

where $E[\cdot]$ denotes mathematical expectation. We apply a delay τ to the observed signal $Y(k,m)$ and pass this delayed signal through a complex-valued filter $\mathbf{h}(k,m)$ of length L:

$$Z(k,m) = \mathbf{h}^H(k,m)\,\mathbf{y}(k,m-\tau), \tag{3.4}$$

where

$$\mathbf{h}(k,m) = [H_0(k,m), H_1(k,m), \ldots, H_{L-1}(k,m)]^T, \tag{3.5}$$

superscripts ^{H} and ^{T} are the conjugate-transpose and transpose operators, respectively, and

$$\mathbf{y}(k,m-\tau) = [Y(k,m-\tau), Y(k,m-\tau-1), \ldots, Y(k,m-\tau-L+1)]^T. \tag{3.6}$$

The vectors $\mathbf{x}(k,m-\tau)$ and $\mathbf{v}(k,m-\tau)$ are defined in a similar fashion to $\mathbf{y}(k,m-\tau)$, so we get

$$\mathbf{y}(k,m-\tau) = \mathbf{x}(k,m-\tau) + \mathbf{v}(k,m-\tau). \tag{3.7}$$

As a result, the error signal is defined as

$$\begin{aligned}E(k,m) &= Y(k,m) - Z(k,m) \\ &= Y(k,m) - \mathbf{h}^H(k,m)\,\mathbf{y}(k,m-\tau).\end{aligned} \tag{3.8}$$

We consider decomposing the signal $X(k,m-\tau)$ into two orthogonal components: a part which is correlated to the desired signal, $X(k,m)$, and another part which is uncorrelated to the desired signal, $X(k,m)$, and hence will be considered as an interference component, i.e.,

$$X(k,m-\tau) = \Gamma_X^*(k,m,\tau)\,X(k,m) + X'(k,m,\tau), \tag{3.9}$$

where

$$\Gamma_X(k,m,\tau) = \frac{E\left[X(k,m)\,X^*(k,m-\tau)\right]}{E\left[\left|X(k,m)\right|^2\right]} \tag{3.10}$$

is the inter-frame correlation coefficient of the signal $X(k,m)$ and the interference is

$$X'(k,m,\tau) = X(k,m-\tau) - \Gamma_X^*(k,m,\tau)\,X(k,m), \tag{3.11}$$

with $E\left[X(k,m)\,X'^*(k,m,\tau)\right] = 0$ and the superscrit * being the complex-conjugate operator. In vector form we can write

$$\mathbf{x}(k,m-\tau) = \boldsymbol{\gamma}_X^*(k,m,\tau)\,X(k,m) + \mathbf{x}'(k,m,\tau), \tag{3.12}$$

where the signal correlation vector is

$$\begin{aligned}\boldsymbol{\gamma}_X(k,m,\tau) &= \left[\Gamma_X(k,m,\tau), \Gamma_X(k,m,\tau+1), \ldots, \Gamma_X(k,m,\tau+L-1)\right]^T \\ &= \frac{E\left[X(k,m)\,\mathbf{x}^*(k,m-\tau)\right]}{E\left[\left|X(k,m)\right|^2\right]}\end{aligned} \tag{3.13}$$

and the signal interference vector is

$$\begin{aligned}\mathbf{x}'(k,m,\tau) &= \left[X'(k,m,\tau), X'(k,m,\tau+1), \ldots, X'(k,m,\tau+L-1)\right]^T \\ &= \mathbf{x}(k,m-\tau) - \boldsymbol{\gamma}_X^*(k,m,\tau)\,X(k,m).\end{aligned} \tag{3.14}$$

We can implement a similar decomposition for the noise signal $V(k,m)$ to get

$$\mathbf{v}(k,m-\tau) = \boldsymbol{\gamma}_V^*(k,m,\tau)\,V(k,m) + \mathbf{v}'(k,m,\tau), \tag{3.15}$$

where $\boldsymbol{\gamma}_V(k,m,\tau)$ is the noise correlation vector and $\mathbf{v}'(k,m,\tau)$ is the noise interference. Plugging these vectors into $Z(k,m)$, we get four components which are uncorrelated, i.e.,

$$\begin{aligned}Z(k,m) &= \mathbf{h}^H(k,m)\,\mathbf{y}(k,m-\tau) \\ &= \mathbf{h}^H(k,m)\,\mathbf{x}(k,m-\tau) + \mathbf{h}^H(k,m)\,\mathbf{v}(k,m-\tau) \\ &= \mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_X^*(k,m,\tau)\,X(k,m) + \mathbf{h}^H(k,m)\,\mathbf{x}'(k,m,\tau) \\ &\quad + \mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_V^*(k,m,\tau)\,V(k,m) + \mathbf{h}^H(k,m)\,\mathbf{v}'(k,m,\tau),\end{aligned} \tag{3.16}$$

and the error signal can now be written as

$$\begin{aligned} E(k,m) =& Y(k,m) - Z(k,m) \\ =& X(k,m)\left[1 - \mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_X^*(k,m,\tau)\right] - \mathbf{h}^H(k,m)\,\mathbf{x}'(k,m,\tau) \\ &+ V(k,m)\left[1 - \mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_V^*(k,m,\tau)\right] - \mathbf{h}^H(k,m)\,\mathbf{v}'(k,m,\tau). \end{aligned} \tag{3.17}$$

There are two scenarios; $Z(k,m)$ can either be an estimate of the noise or the desired signal, depending on which signal is still correlated given the delay τ. For the case we are investigating, where the noise is nonstationary harmonic, we will assume that the speech signal is no longer correlated, while the noise is still correlated, meaning that $Z(k,m)$ is the estimate of the noise and $E(k,m)$ is the estimate of the speech signal. The described system is shown in Fig. 3.1. The ideal conditions for this system are

$$\mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_X^*(k,m,\tau) = 0, \tag{3.18}$$

$$\mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_V^*(k,m,\tau) = 1, \tag{3.19}$$

and the inherent estimation error is

$$\epsilon_{\text{inherent}}(k,m) = \mathbf{h}^H(k,m)\,\mathbf{x}'(k,m,\tau) + \mathbf{h}^H(k,m)\,\mathbf{v}'(k,m,\tau). \tag{3.20}$$

We can write the speech estimate as follows:

$$\begin{aligned} X_{\text{est}}(k,m) =& E(k,m) \\ =& X_{\text{fd}}(k,m) + X'_{\text{ri}}(k,m) + V_{\text{rn}}(k,m), \end{aligned} \tag{3.21}$$

where the filtered desired signal is

$$X_{\text{fd}}(k,m) = X(k,m)\left[1 - \mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_X^*(k,m,\tau)\right], \tag{3.22}$$

the residual interference is

$$X'_{\text{ri}}(k,m) = -\mathbf{h}^H(k,m)\,\mathbf{x}'(k,m,\tau), \tag{3.23}$$

and the residual noise is

$$V_{\text{rn}}(k,m) = V(k,m)\left[1 - \mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_V^*(k,m,\tau)\right] - \mathbf{h}^H(k,m)\,\mathbf{v}'(k,m,\tau). \tag{3.24}$$

The ideal conditions (3.18)-(3.19) are met when the desired signal is no longer correlated with the delayed desired signal, while the noise remains correlated. If $\mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_X^*(k,m,\tau) \neq 0$, we get distortion of the desired signal, while if $\mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_V^*(k,m,\tau) \neq 1$, we get less noise reduction. This shows the importance of the delay parameter τ.

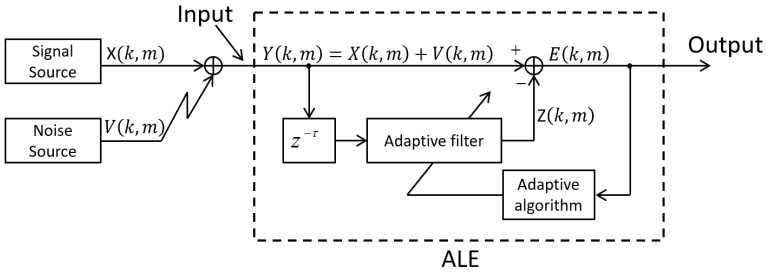

Figure 3.1: STFT domain ALE system.

3.3 Performance Measures

The narrow-band and full-band input SNRs are, respectively,

$$\text{iSNR}\left(k,m\right)=\frac{\phi_X\left(k,m\right)}{\phi_V\left(k,m\right)} \tag{3.25}$$

and

$$\text{iSNR}\left(m\right)=\frac{\sum_{k=0}^{K-1}\phi_X\left(k,m\right)}{\sum_{k=0}^{K-1}\phi_V\left(k,m\right)}. \tag{3.26}$$

The narrow-band output SNR is defined as the ratio of the variance of the filtered desired signal over the variance of the residual interference-plus-noise, i.e.,

$$\text{oSNR}\left[\mathbf{h}\left(k,m\right)\right]=\frac{\phi_{X_{\text{fd}}}\left(k,m\right)}{\phi_{X'_{\text{ri}}}\left(k,m\right)+\phi_{V_{\text{rn}}}\left(k,m\right)} \tag{3.27}$$

where

$$\begin{aligned}\phi_{X_{\text{fd}}}\left(k,m\right)&=E\left[\left|X_{\text{fd}}\left(k,m\right)\right|^2\right] \\ &=\phi_X\left(k,m\right)\left|1-\mathbf{h}^H\left(k,m\right)\boldsymbol{\gamma}_X^*\left(k,m,\tau\right)\right|^2,\end{aligned} \tag{3.28}$$

$$\begin{aligned}\phi_{X'_{\text{ri}}}\left(k,m\right)&=E\left[\left|X'_{\text{ri}}\left(k,m\right)\right|^2\right] \\ &=\mathbf{h}^H\left(k,m\right)\boldsymbol{\Phi}_{\mathbf{x}'}\left(k,m,\tau\right)\mathbf{h}\left(k,m\right),\end{aligned} \tag{3.29}$$

and

$$\begin{aligned}&\phi_{V_{\text{rn}}}\left(k,m\right)=E\left[\left|V_{\text{rn}}\left(k,m\right)\right|^2\right] \\ &=\phi_V\left(k,m\right)\left|1-\mathbf{h}^H\left(k,m\right)\boldsymbol{\gamma}_V^*\left(k,m,\tau\right)\right|^2+\mathbf{h}^H\left(k,m\right)\boldsymbol{\Phi}_{\mathbf{v}'}\left(k,m,\tau\right)\mathbf{h}\left(k,m\right),\end{aligned} \tag{3.30}$$

where we have defined the matrices:

$$\mathbf{\Phi}_{\mathbf{x}'}(k,m,\tau) = E\left[\mathbf{x}'(k,m,\tau)\,\mathbf{x}'^{H}(k,m,\tau)\right], \tag{3.31}$$

$$\mathbf{\Phi}_{\mathbf{v}'}(k,m,\tau) = E\left[\mathbf{v}'(k,m,\tau)\,\mathbf{v}'^{H}(k,m,\tau)\right]. \tag{3.32}$$

Expanding the term:

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{x}'}(k,m,\tau) &= E\left[\mathbf{x}'(k,m,\tau)\,\mathbf{x}'^{H}(k,m,\tau)\right] \\ &= E\Big[\left(\mathbf{x}(k,m-\tau) - \boldsymbol{\gamma}_X^*(k,m,\tau)\,X(k,m)\right)\times \\ &\qquad \left(\mathbf{x}^H(k,m-\tau) - X^*(k,m)\,\boldsymbol{\gamma}_X^T(k,m,\tau)\right)\Big] \\ &= \mathbf{\Phi}_{\mathbf{x}}(k,m-\tau) - \phi_X(k,m)\,\boldsymbol{\gamma}_X^*(k,m,\tau)\,\boldsymbol{\gamma}_X^T(k,m,\tau).\end{aligned} \tag{3.33}$$

Similarly,

$$\mathbf{\Phi}_{\mathbf{v}'}(k,m,\tau) = \mathbf{\Phi}_{\mathbf{v}}(k,m-\tau) - \phi_V(k,m)\,\boldsymbol{\gamma}_V^*(k,m,\tau)\,\boldsymbol{\gamma}_V^T(k,m,\tau). \tag{3.34}$$

Plugging these into (3.27), we get

$$\begin{aligned}&\mathrm{oSNR}\left[\mathbf{h}(k,m)\right] = \\ &\quad \phi_X(k,m)\left|1-\mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_X^*(k,m,\tau)\right|^2\times \\ &\quad \Big[\mathbf{h}^H(k,m)\,\mathbf{\Phi}_{\mathbf{x}'}(k,m,\tau)\,\mathbf{h}(k,m) + \\ &\qquad \phi_V(k,m)\left|1-\mathbf{h}^H(k,m)\,\boldsymbol{\gamma}_V^*(k,m,\tau)\right|^2 + \\ &\qquad \mathbf{h}^H(k,m)\,\mathbf{\Phi}_{\mathbf{v}'}(k,m,\tau)\,\mathbf{h}(k,m)\Big]^{-1}.\end{aligned} \tag{3.35}$$

For the ideal conditions (3.18)-(3.19), we have

$$\begin{aligned}&\mathrm{oSNR}_{\mathrm{ideal}}\left[\mathbf{h}(k,m)\right] = \\ &\frac{\phi_X(k,m)}{\mathbf{h}^H(k,m)\,\mathbf{\Phi}_{\mathbf{x}'}(k,m,\tau)\,\mathbf{h}(k,m) + \mathbf{h}^H(k,m)\,\mathbf{\Phi}_{\mathbf{v}'}(k,m,\tau)\,\mathbf{h}(k,m)}.\end{aligned} \tag{3.36}$$

We define the full-band output SNR as

$$\mathrm{oSNR}\left[\mathbf{h}(m)\right] = \frac{\sum_{k=0}^{K-1}\phi_{X_{\mathrm{fd}}}(k,m)}{\sum_{k=0}^{K-1}\phi_{X'_{\mathrm{ri}}}(k,m) + \sum_{k=0}^{K-1}\phi_{V_{\mathrm{rn}}}(k,m)}. \tag{3.37}$$

It can be verified [47] that

$$\mathrm{iSNR}\left[\mathbf{h}(m)\right] \le \sum_{k=0}^{K-1}\mathrm{iSNR}\left[\mathbf{h}(k,m)\right], \tag{3.38}$$

$$\mathrm{oSNR}\left[\mathbf{h}(m)\right] \le \sum_{k=0}^{K-1}\mathrm{oSNR}\left[\mathbf{h}(k,m)\right]. \tag{3.39}$$

The noise reduction factor quantifies the amount of noise being rejected by the filter. It is defined as the ratio of the power of the noise at the sensor over the power of the noise remaining at the filter output. The noise reduction factor is usually expected to be lower bounded by 1, however, as we are adding the residual interference this is not necessarily the case. The higher the value, the more the noise is rejected. The narrow-band and full-band noise reduction factors are then

$$\begin{aligned}\xi_{nr}\left[\mathbf{h}\left(k,m\right)\right] &= \frac{\phi_V\left(k,m\right)}{\phi_{X'_{\mathrm{ri}}}\left(k,m\right)+\phi_{V_{\mathrm{rn}}}\left(k,m\right)} \\ &= \phi_V\left(k,m\right)\times\Big[\mathbf{h}^H\left(k,m\right)\boldsymbol{\Phi}_{\mathbf{x}'}\left(k,m,\tau\right)\mathbf{h}\left(k,m\right)+ \\ &\quad \phi_V\left(k,m\right)\left|1-\mathbf{h}^H\left(k,m\right)\boldsymbol{\gamma}_V^*\left(k,m,\tau\right)\right|^2+ \\ &\quad \mathbf{h}^H\left(k,m\right)\boldsymbol{\Phi}_{\mathbf{v}'}\left(k,m,\tau\right)\mathbf{h}\left(k,m\right)\Big]^{-1},\end{aligned} \tag{3.40}$$

and

$$\xi_{\mathrm{nr}}\left[\mathbf{h}\left(m\right)\right]=\frac{\sum_{k=0}^{K-1}\phi_V\left(k,m\right)}{\sum_{k=0}^{K-1}\phi_{X'_{\mathrm{ri}}}\left(k,m\right)+\sum_{k=0}^{K-1}\phi_{V_{\mathrm{rn}}}\left(k,m\right)}. \tag{3.41}$$

For the ideal conditions (3.18)-(3.19):

$$\xi_{\mathrm{nr,ideal}}\left[\mathbf{h}\left(k,m\right)\right]= \frac{\phi_V\left(k,m\right)}{\mathbf{h}^H\left(k,m\right)\boldsymbol{\Phi}_{\mathbf{x}'}\left(k,m,\tau\right)\mathbf{h}\left(k,m\right)+\mathbf{h}^H\left(k,m\right)\boldsymbol{\Phi}_{\mathbf{v}'}\left(k,m,\tau\right)\mathbf{h}\left(k,m\right)}. \tag{3.42}$$

In practice, the filter might distort the signal. To evaluate the level of this distortion, we define the narrow-band and full-band speech reduction factors, respectively, as

$$\begin{aligned}\xi_{\mathrm{sr}}\left[\mathbf{h}\left(k,m\right)\right] &= \frac{\phi_X\left(k,m\right)}{\phi_{X_{\mathrm{fd}}}\left(k,m\right)} \\ &= \frac{1}{\left|1-\mathbf{h}^H\left(k,m\right)\boldsymbol{\gamma}_X^*\left(k,m,\tau\right)\right|^2}\end{aligned} \tag{3.43}$$

and

$$\begin{aligned}\xi_{\mathrm{sr}}\left[\mathbf{h}\left(m\right)\right] &= \frac{\sum_{k=0}^{K-1}\phi_X\left(k,m\right)}{\sum_{k=0}^{K-1}\phi_{X_{\mathrm{fd}}}\left(k,m\right)} \\ &= \frac{\sum_{k=0}^{K-1}\phi_X\left(k,m\right)}{\sum_{k=0}^{K-1}\phi_X\left(k,m\right)\left|1-\mathbf{h}^H\left(k,m\right)\boldsymbol{\gamma}_X^*\left(k,m,\tau\right)\right|^2}.\end{aligned} \tag{3.44}$$

Thus the speech reduction factor is equal to 1 if there is no distortion and is greater than 1 when distortion occurs. We can clearly see from this ratio

that the design of a filter that does not distort the speech signal is dependent on there being no remaining correlation.

Another way to measure the distortion of the desired signal due to the filtering is via the desired signal distortion index, which is defined as the mean-squared error (MSE) between the desired signal and the filtered desired signal normalized by the variance of the desired signal. The closer the distortion index is to 0, the less the distortion. The narrow-band and full-band desired signal distortion indexes are then

$$\begin{aligned} v_{\text{sd}}\left[\mathbf{h}\left(k,m\right)\right] &= \frac{E\left[\left|X\left(k,m\right)-X_{\text{fd}}\left(k,m\right)\right|^2\right]}{\phi_X\left(k,m\right)} \\ &= \frac{\phi_X\left(k,m\right)\left|\mathbf{h}^H\left(k,m\right)\boldsymbol{\gamma}_X^*\left(k,m,\tau\right)\right|^2}{\phi_X\left(k,m\right)} \\ &= \left|\mathbf{h}^H\left(k,m\right)\boldsymbol{\gamma}_X^*\left(k,m,\tau\right)\right|^2 \end{aligned} \tag{3.45}$$

and

$$v_{\text{sd}}\left[\mathbf{h}\left(m\right)\right] = \frac{\sum_{k=0}^{K-1} E\left[\left|X\left(k,m\right)-X_{\text{fd}}\left(k,m\right)\right|^2\right]}{\sum_{k=0}^{K-1}\phi_X\left(k,m\right)}, \tag{3.46}$$

where for the ideal conditions there is no distortion. By making the appropriate substitutions, one can derive the following relationships:

$$\frac{\text{oSNR}\left[\mathbf{h}\left(k,m\right)\right]}{\text{iSNR}\left(k,m\right)} = \frac{\xi_{\text{nr}}\left[\mathbf{h}\left(k,m\right)\right]}{\xi_{\text{sr}}\left[\mathbf{h}\left(k,m\right)\right]}, \tag{3.47}$$

$$\frac{\text{oSNR}\left[\mathbf{h}\left(m\right)\right]}{\text{iSNR}\left(m\right)} = \frac{\xi_{\text{nr}}\left[\mathbf{h}\left(m\right)\right]}{\xi_{\text{sr}}\left[\mathbf{h}\left(m\right)\right]}. \tag{3.48}$$

3.4 Optimal Filters

We define the narrow-band MSE as

$$\begin{aligned} J\left[\mathbf{h}\left(k,m\right)\right] &= E\left[\left|E\left(k,m\right)\right|^2\right] \\ &= E\left[\left|X_{\text{fd}}\left(k,m\right)\right|^2\right] + E\left[\left|V_{\text{rn}}\left(k,m\right)\right|^2\right] + E\left[\left|X'_{\text{ri}}\left(k,m\right)\right|^2\right] \\ &= \phi_X\left(k,m\right)\left|1-\mathbf{h}^H\left(k,m\right)\boldsymbol{\gamma}_X^*\left(k,m,\tau\right)\right|^2 + \\ &\quad + \mathbf{h}^H\left(k,m\right)\boldsymbol{\Phi}_{\mathbf{x}'}\left(k,m,\tau\right)\mathbf{h}\left(k,m\right) + \\ &\quad + \phi_V\left|1-\mathbf{h}^H\left(k,m\right)\boldsymbol{\gamma}_V^*\left(k,m,\tau\right)\right|^2 + \\ &\quad + \mathbf{h}^H\left(k,m\right)\boldsymbol{\Phi}_{\mathbf{v}'}\left(k,m,\tau\right)\mathbf{h}\left(k,m\right) \\ &= J_{\text{d}}\left[\mathbf{h}\left(k,m\right)\right] + J_{\text{r}}\left[\mathbf{h}\left(k,m\right)\right], \end{aligned} \tag{3.49}$$

where

$$\begin{aligned} J_{\mathrm{d}}\left[\mathbf{h}\left(k,m\right)\right] =& E\left[\left|X_{\mathrm{fd}}\left(k,m\right)\right|^{2}\right] \\ &= \phi_{X}\left(k,m\right)\left|1-\mathbf{h}^{H}\left(k,m\right)\boldsymbol{\gamma}_{X}^{*}\left(k,m,\tau\right)\right|^{2} \end{aligned} \tag{3.50}$$

and

$$\begin{aligned} J_{\mathrm{r}}\left[\mathbf{h}\left(k,m\right)\right] =& E\left[\left|V_{\mathrm{rn}}\left(k,m\right)\right|^{2}\right]+E\left[\left|X_{\mathrm{ri}}^{\prime}\left(k,m\right)\right|^{2}\right] \\ &= \phi_{V}\left|1-\mathbf{h}^{H}\left(k,m\right)\boldsymbol{\gamma}_{V}^{*}\left(k,m,\tau\right)\right|^{2}+ \\ &\quad+\mathbf{h}^{H}\left(k,m\right)\boldsymbol{\Phi}_{\mathbf{v}^{\prime}}\left(k,m,\tau\right)\mathbf{h}\left(k,m\right)+ \\ &\quad+\mathbf{h}^{H}\left(k,m\right)\boldsymbol{\Phi}_{\mathbf{x}^{\prime}}\left(k,m,\tau\right)\mathbf{h}\left(k,m\right). \end{aligned} \tag{3.51}$$

We can easily see the relation between the MSEs and some of the performance measures:

$$\mathrm{oSNR}\left[\mathbf{h}\left(k,m\right)\right]=\frac{J_{\mathrm{d}}\left[\mathbf{h}\left(k,m\right)\right]}{J_{\mathrm{r}}\left[\mathbf{h}\left(k,m\right)\right]} \tag{3.52}$$

and

$$\xi_{\mathrm{nr}}\left[\mathbf{h}\left(k,m\right)\right]=\frac{\phi_{V}\left(k,m\right)}{J_{\mathrm{r}}\left[\mathbf{h}\left(k,m\right)\right]}. \tag{3.53}$$

We can define the full-band MSE as

$$\begin{aligned} J\left[\mathbf{h}\left(m\right)\right] &= \frac{1}{K}\sum_{k=0}^{K-1}J\left[\mathbf{h}\left(k,m\right)\right] \\ &= \frac{1}{K}\sum_{k=0}^{K-1}J_{\mathrm{d}}\left[\mathbf{h}\left(k,m\right)\right]+\frac{1}{K}\sum_{k=0}^{K-1}J_{\mathrm{r}}\left[\mathbf{h}\left(k,m\right)\right] \\ &= J_{\mathrm{d}}\left[\mathbf{h}\left(m\right)\right]+J_{\mathrm{r}}\left[\mathbf{h}\left(m\right)\right]. \end{aligned} \tag{3.54}$$

It is clear that minimization of the narrow-band MSE for each index k is equivalent to minimization of the full-band MSE.

3.4.1 Wiener

We derive the Wiener filter by minimizing the narrow-band MSE, $J\left[\mathbf{h}\left(k,m\right)\right]$:

$$\mathbf{h}_{\mathrm{W}}\left(k,m\right)=\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\left(k,m-\tau\right)\left[\phi_{X}\left(k,m\right)\boldsymbol{\gamma}_{X}^{*}\left(k,m,\tau\right)+\phi_{V}\left(k,m\right)\boldsymbol{\gamma}_{V}^{*}\left(k,m,\tau\right)\right], \tag{3.55}$$

where

$$\boldsymbol{\Phi}_{\mathbf{y}}\left(k,m-\tau\right)=\boldsymbol{\Phi}_{\mathbf{x}}\left(k,m-\tau\right)+\boldsymbol{\Phi}_{\mathbf{v}}\left(k,m-\tau\right). \tag{3.56}$$

When the clean speech signal is no longer correlated after delay τ, i.e. $\boldsymbol{\gamma}_X(k,m,\tau)=0$, the Wiener filter reduces to

$$\mathbf{h}_{\mathrm{W}}(k,m)=\phi_V(k,m)\,\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(k,m-\tau)\,\boldsymbol{\gamma}_V^{*}(k,m,\tau)\,. \tag{3.57}$$

If we define another error signal:

$$E_V(k,m)=V(k,m)-\mathbf{h}_V^H(k,m)\,\mathbf{y}(k,m-\tau)\,, \tag{3.58}$$

where $\mathbf{h}_V(k,m)$ is another filter of length L, we can clearly see that this filter estimates the noise. The appropriate MSE is

$$\begin{aligned}
J\left[\mathbf{h}_V(k,m)\right] &= E\left[\left|E_V(k,m)\right|^2\right] \\
&= E\left[\left|V(k,m)\right|^2\right]+ \\
&\quad+\mathbf{h}_V^H(k,m)\,E\left[\mathbf{y}(k,m-\tau)\,\mathbf{y}^H(k,m-\tau)\right]\mathbf{h}_V^H(k,m)+ \\
&\quad-\mathbf{h}_V^H(k,m)\,E\left[\mathbf{y}(k,m-\tau)\,V^{*}(k,m)\right]+ \\
&\quad-E\left[V(k,m)\,\mathbf{y}^H(k,m-\tau)\right]\mathbf{h}_V(k,m) \\
&= \phi_V(k,m)+\mathbf{h}_V^H(k,m)\,\boldsymbol{\Phi}_{\mathbf{y}}(k,m-\tau)\,\mathbf{h}_V(k,m)+ \\
&\quad-\phi_V(k,m)\,\mathbf{h}_V^H(k,m)\,\boldsymbol{\gamma}_V^{*}(k,m,\tau)+ \\
&\quad-\phi_V(k,m)\,\boldsymbol{\gamma}_V^{T}(k,m,\tau)\,\mathbf{h}_V(k,m)\,.
\end{aligned} \tag{3.59}$$

Minimizing $J\left[\mathbf{h}_V(k,m)\right]$, we get the Wiener filter for the noise estimate:

$$\mathbf{h}_{V,\mathrm{W}}(k,m)=\phi_V(k,m)\,\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(k,m-\tau)\,\boldsymbol{\gamma}_V^{*}(k,m,\tau)\,, \tag{3.60}$$

which is the same filter we obtained previously in (3.57) when the clean speech is not correlated. This means that $Z(k,m)$ is a good estimate for the noise, and as a result $E(k,m)$ is a good estimate for the desired speech signal.

The interesting thing about this method is that we can solve (3.55) adaptively for each frequency bin. We can, for example, use the simple NLMS algorithm [65]:

$$\mathbf{h}(k,m+1)=\mathbf{h}(k,m)+\frac{\mu(k)}{\mathbf{y}^H(k,m-\tau)\,\mathbf{y}(k,m-\tau)+\delta}\mathbf{y}(k,m-\tau)\,E^{*}(k,m)\,, \tag{3.61}$$

where $0<\mu(k)<2$ is a step-size parameter which should be smaller than 1 here for the algorithm to converge, and $\delta>0$ is the regularization parameter, which can be quite large depending on the amount of noise.

3.4.2 Proposed Combined Approach

The conventional ALE is a forward linear predictor typically implemented by an adaptive NLMS algorithm, where a filter of length L is used and is updated across all frames, i.e.,

$$\begin{aligned} E(k,m) &= Y(k,m) - \sum_{l=0}^{L-1} H_l(k,m)\, Y(k,m-l-\tau) \\ &= Y(k,m) - \mathbf{h}^H(k,m)\, \mathbf{y}(k,m-\tau) \end{aligned} \tag{3.62}$$

and the filter $\mathbf{h}(k,m)$ is found using (3.61). We propose using a combination of both a forward linear predictor (FLP) and a backward linear predictor (BLP), to get the combined linear predictor (CLP). The BLP is a non-causal predictor; however, we deem the use of it acceptable as a few frames delay is reasonable in most applications, while we get the added advantage of using future information for better noise estimation. The combination of the FLP and the BLP is done by applying on each frame the filter (either FLP or BLP) that provides the smallest spectral error.

In addition, we propose estimating the noise and updating the filter only when the nonstationary harmonic noise is present by using a noise presence detector. The detector can be developed, for example, by using a DL algorithm on a database containing medical equipment beeping sounds, similar to the implementation in [70, 71]. The development of such a detector is outside the scope of this chapter, and we will assume an ideal detector. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be two hypotheses denoting noise presence and absence, respectively, and let $\mathcal{I}(k,m)$ be a noise indicator, given by

$$\mathcal{I}(k,m) = \begin{cases} 1, & V(k,m) \in \mathcal{H}_0 \\ 0, & V(k,m) \in \mathcal{H}_1 \end{cases}. \tag{3.63}$$

Another option we propose for the noise indicator would be to use the MI approach calculation for step size $\mu(k)$ to identify the frequencies that contain harmonic noise [30]. For filter length $L > 1$, when the noise transients are only starting, not enough samples of noise are present for the entire filter length. Instead of taking the samples where noise is not present to zero (pre-windowing), we use a set of filters with changing length, based on the available amount of noise samples, until the maximal filter length of L. We denote the forward predictor per filter length as

$$\begin{aligned} E_f^{\ell}(k,m) &= Y(k,m) - \mathcal{I}(k,m) \cdot \sum_{i=0}^{\ell} H_{\ell,i}^{*}(k,m)\, Y(k,m-i-\tau) \\ &= Y(k,m) - \mathcal{I}(k,m) \cdot \mathbf{h}_{\ell}^{H}(k,m)\, \mathbf{y}_{\ell}(k,m-\tau), \end{aligned} \tag{3.64}$$

where

$$\mathbf{h}_\ell(k,m) = [H_{\ell,0}(k,m),\ldots,H_{\ell,\ell}(k,m)]^T, \tag{3.65}$$
$$\mathbf{y}_\ell(k,m-\tau) = [Y(k,m-\tau),Y(k,m-\tau-1),\ldots,Y(k,m-\tau-\ell)]^T,$$

and

$$\ell = \left\{ \ell' \mid \sum_{i=0}^{\ell'} \mathcal{I}(k,m-i-\tau) = \ell'+1, \quad \ell' = 0,1,\ldots,L-1 \right\}. \tag{3.66}$$

The forward predictor, the FMLNLMS (forward-mapped-L-NLMS), is then

$$E_f(k,m) = E_f^{\max \ell}(k,m). \tag{3.67}$$

Similarly, we define the backward predictor per filter length as

$$\begin{aligned} E_b^p(k,m) &= Y(k,m-L) - \mathcal{I}(k,m-L)\cdot\sum_{i=0}^{p} G_{p,i}^*(k,m)\,Y(k,m-i+\tau) \\ &= Y(k,m-L) - \mathcal{I}(k,m-L)\cdot\mathbf{g}_p^H(k,m)\,\mathbf{y}_p(k,m+\tau), \end{aligned} \tag{3.68}$$

where

$$\mathbf{g}_p(k,m) = [G_{p,0}(k,m),\ldots,G_{p,p}(k,m)]^T, \tag{3.69}$$
$$\mathbf{y}_p(k,m+\tau) = [Y(k,m+\tau),Y(k,m+\tau-1),\ldots,Y(k,m+\tau-p)]^T,$$

$$p = \left\{ \ell' \mid \sum_{i=0}^{\ell'} \mathcal{I}(k,m-i+\tau) = \ell'+1, \quad \ell' = 0,1,\ldots,L-1 \right\}, \tag{3.70}$$

and the backward predictor, the BMLNLMS (backward-mapped-L-NLMS), is then

$$E_b(k,m) = E_b^{\max p}(k,m). \tag{3.71}$$

Finally, the signal estimate for the combined linear predictor, the CMLNLMS (combined-mapped-L-NLMS), is the filter which provides the smallest spectral error per frame:

$$E_c(k,m) = \begin{cases} E_b(k,m+L), & |E_b(k,m+L)|^2 \le |E_f(k,m)|^2 \text{ and} \\ & \quad |E_b(k,m+L)|^2 \le |Y(k,m)|^2 \\ E_f(k,m), & |E_b(k,m+L)|^2 > |E_f(k,m)|^2 \text{ and} \\ & \quad |E_f(k,m)|^2 \le |Y(k,m)|^2 \\ Y(k,m), & else. \end{cases} \tag{3.72}$$

Note that if the spectral error of either the FMLNLMS or BMLNLMS is larger than the noisy signal spectra, we don't use either of the filters' results, rather we use the noisy signal itself. This is done to avoid over-estimation of the noise.

3.5 Experimental Results

The evaluation is done on speech signals taken from the TIMIT database [72], including 20 different utterances from 20 different speakers, half male and half female. The signals are sampled at 16 kHz and degraded by nonstationary harmonic noise with overall SNR in the range $[0, 20]$ dB. The noisy signals are transformed to the time-frequency domain using STFT, which is implemented with overlapping Hamming analysis. The overlap-add method is used for the signal reconstruction in the time domain. For the noise, different nonstationary harmonic noise signals such as heart monitor beeping, train door beeping, house alarm, smoke alarm, and rail road crossing bells were collected to form a database of 26 different signals [73, 74, 75, 76]. The database contains both real-life recorded signals and synthesized signals generated for sound effects. The signals were converted from stereo to mono and down-sampled to 16 kHz for our use.

To implement the noise detector, we define a threshold value relative to the maximum spectrum of the nonstationary harmonic noise. If the noise spectrum is above this threshold, we consider the noise to be present; if it is equal or below it, we consider the noise to be absent. By using different threshold values and the frame only detector based on the MI approach, we can emulate the impact of a non-ideal noise detector. We evaluate the performance of our proposed approach using the distortion index and noise reduction factor, where we investigate the measures as defined in Section 3.3, which we designate the orthogonal decomposition (OD) measures where the residual interference is treated as part of the noise. We compare the OD measures to the traditional performance measures where the residual interference is part of the filtered desired signal [47]. In addition, we use the perceptual evaluation of speech quality (PESQ) measure [77] and the short-time objective intelligibility (STOI) measure [78], where larger values of PESQ and STOI indicate better quality and intelligibility, respectively.

3.5.1 Correlation Vector

To evaluate the correlation vector $\boldsymbol{\gamma}_X(k,m,\tau)$, we define the vector:

$$\mathbf{x}_1(k,m,\tau) \in \mathbb{C}^{(L+\tau)\times 1} = \tag{3.73}$$
$$[X(k,m), X(k,m-1), \ldots, X(k,m-\tau), \ldots, X(k,m-\tau-L+1)]^T$$

and its correlation matrix:

$$\boldsymbol{\Phi}_{\mathbf{x}_1}(k,m,\tau)_{(L+\tau)\times(L+\tau)} = E\left[\mathbf{x}_1(k,m,\tau)\,\mathbf{x}_1^H(k,m,\tau)\right]. \tag{3.74}$$

The correlation vector of speech is then the transpose of the first row vector of $\boldsymbol{\Phi}_{\mathbf{x}_1}(k,m,\tau)$ from location τ, normalized by its first element:

$$\boldsymbol{\gamma}_X(k,m,\tau) = \frac{E\left[X(k,m)\,\mathbf{x}^*(k,m-\tau)\right]}{E\left[|X(k,m)|^2\right]} = \tag{3.75}$$
$$= \frac{\boldsymbol{\Phi}_{\mathbf{x}_1}(k,m,\tau)(1,\tau:(\tau+L))}{\boldsymbol{\Phi}_{\mathbf{x}_1}(k,m,\tau)(1,1)} \tag{3.76}$$

and the delayed correlation matrix is the matrix block from row and column $\tau+1$:

$$\boldsymbol{\Phi}_{\mathbf{x}}(k,m-\tau) = \boldsymbol{\Phi}_{\mathbf{x}_1}(k,m,\tau)(\tau+1:\tau+L,\tau+1:\tau+L). \tag{3.77}$$

We evaluate in a similar fashion the noise correlation vector $\boldsymbol{\gamma}_V(k,m,\tau)$ and the delayed correlation matrix $\boldsymbol{\Phi}_{\mathbf{v}}(k,m-\tau)$.

We examine the correlation vectors behavior in Fig. 3.2. We see that the absolute value of the correlation vector of the speech signal is reduced compared to that of the noise and it decays faster per delay τ. Hence, $Z(k,m)$ would be a better estimate of the noise than of the speech signal, though we can expect distortion as the delayed speech signal correlation is clearly not zero. This demonstrates the existing tradeoff between the noise reduction and distortion. The noise reduction will be better for smaller delay τ as the noise correlation is higher, but the speech signal distortion will be higher as also the speech correlation is higher. For smaller window lengths or for larger percentage overlaps, the correlations can be expected to decay more slowly. We will continue the analysis with overlap of 75% to retain high correlation, and use window length $K = 512$.

3.5.2 Least Squares

For illustrative purposes, we start the analysis with the least-squares (LS) filter [65, 79] and a synthetic nonstationary harmonic noise corrupting a

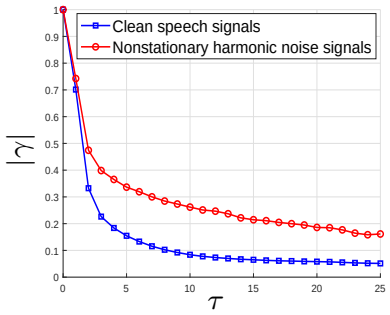

Figure 3.2: Absolute value of the correlation vectors vs. delay parameter τ for all database clean speech signals (blue square) and for all database nonstationary harmonic noise signals (red circle), for filter length $L = 1$, window length $K = 512$ and 75% overlap.

3.4 s long female speech signal from the database. The noisy speech signal is shown in Fig. 3.3. The noise is generated by adding harmonics in different frequencies modulated by trapezoidal trains in the time domain to white Gaussian noise (WGN). We use a threshold value of -25 dB for the noise detector, and compare between the OD performance measures in Section 3.3, and the traditional performance measures.

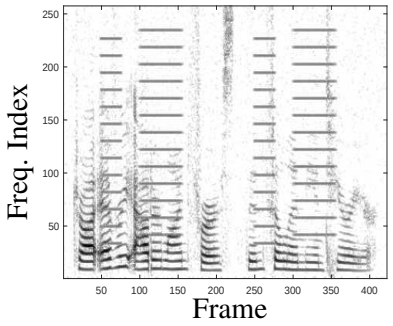

Figure 3.3: Spectrogram of a 3.4 s female speech signal corrupted by a synthetic nonstationary harmonic noise at 10 dB SNR.

Observing Fig. 3.4, which shows the resulting performance measures for the different filters, we see that, as expected, the speech signal is still correlated to its delayed version. As we increase delay τ, the correlation and hence the signal distortion decrease, and as the noise signal also becomes less correlated as τ increases, the noise reduction also decreases. These trends are true for all filter types: the FLP, BLP, and the proposed CLP. The residual interference is considered part of the distortion for the traditional definition compared to the OD definition. Accordingly, we see that the traditional distortion index decreases as τ increases to a higher level than the OD

distortion index. While on the other hand, the noise reduction for the OD definition is lower than the noise reduction for the traditional definition, as we compare the original noise level to the residual noise plus residual interference. In fact, in Fig. 3.5, which demonstrates the distortion index and noise reduction factor for the CLP filter at different filter lengths, we see that the OD noise reduction factor can go below 0 dB. This means that from the OD perspective we are enhancing the noise and not suppressing it. This does not reflect in the perception of quality of the enhanced signal seen in the PESQ level, or in the intelligibility seen in the STOI level, where we have improved levels compared to the unprocessed noisy signal. The results for PESQ and STOI at the different filter lengths are very similar, so only the result for a single filter length is shown in Fig. 3.4. The general trend for PESQ and STOI is also a decrease in value as we increase the delay; hence the noise reduction decrease is more dominant than the distortion decrease.

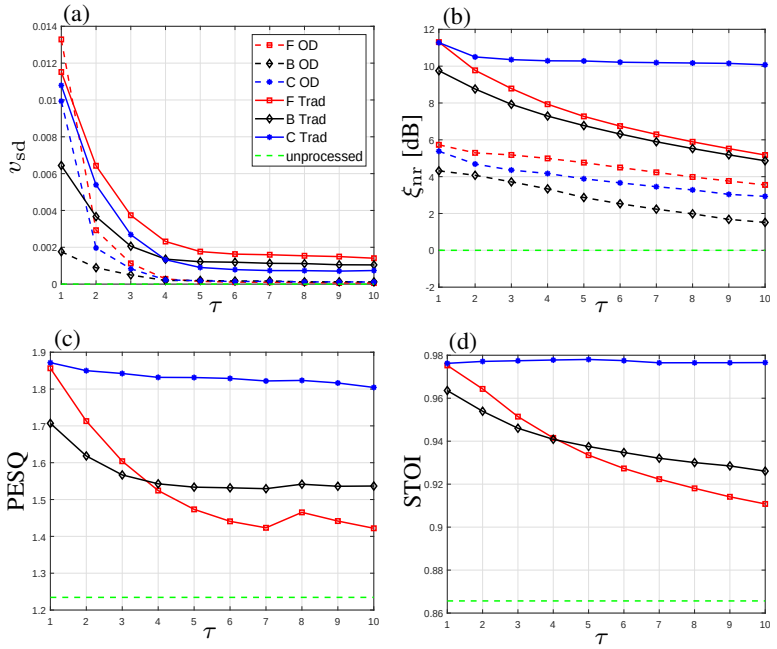

Figure 3.4: (a) Distortion index, (b) noise reduction factor, (c) PESQ, and (d) STOI for LS ALE filter applied on the 3.4 s female speech signal corrupted by synthetic nonstationary harmonic noise at 10 dB SNR for varying delay τ and filter length $L = 3$, for the different LS filters: FLP (red square), BLP (black diamond), and CLP (blue asterisk). Solid lines are for traditional performance measures definition, the dashed lines are for the OD definition in Section 3.3. Dashed green is for the unprocessed noisy signal.

We observe the dependence on filter length in Fig. 3.5. For the OD distortion index, longer filter length means we are deducting more correlated

components from the current speech signal, so the distortion increases as the filter length increases. For the traditional distortion index, both the amount of correlated speech being deducted from the speech and the amount of residual interference increase as we increase the filter length; therefore the distortion index also increases. For the OD noise reduction factor, for longer filter lengths we get more residual interference which is more dominant than the improvement in the noise modeling, so we get lower noise reduction. For the CLP filter, the traditional noise reduction dependence on the delay τ becomes smaller. This will be explained presently. In Fig. 3.4 we see that for the traditional definition the CLP has better noise reduction than the FLP and the BLP. For the OD definitions though, the CLP has lower noise reduction. Again, this does not reflect the PESQ and STOI levels which are better for the CLP.

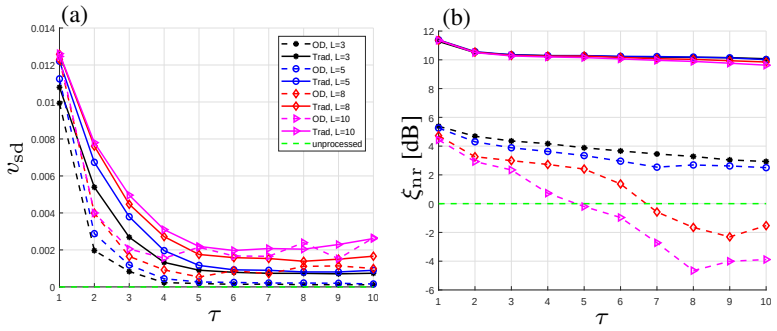

Figure 3.5: (a) Distortion index and (b) noise reduction factor for CLP LS ALE filter applied on the 3.4 s female speech signal corrupted by synthetic nonstationary harmonic noise at 10 dB SNR for varying delay τ and different filter lengths $L \in \{3, 5, 8, 10\}$. Solid lines are for traditional performance measures definition, the dashed lines are for the OD definition in Section 3.3. Dashed green is for the unprocessed noisy signal.

Figure 3.6 illustrates the reduction of the noise transients (noise only) for each filter, given the delay parameter and filter length. As expected, the forward and backward filters do not reduce the entire transient. The FLP reduces the transient starting after delay τ until the end of the transient, while the BLP starts from the beginning until $\tau + 1$ from the end of the transient. Using the combination of both FLP and BLP, the CLP reduction of the transient spans the entire length of the transient. At the edges of the transient, the noise level is low; however the LS filter uses the same average filter coefficients to estimate it, so it is possible to overestimate the noise. One way to mitigate this, as was implemented here, is to use the filtered results only if the spectral error is smaller than the noisy signal spectra. Another case where the noise can be overestimated is when the transient

edge overlaps speech components, and the noise is estimated erroneously based on the higher-leveled speech. This becomes more pronounced with longer delay, as can be seen in Fig. 3.6. In this we see the clear impact of the noise detector. For a higher threshold used, we will get less overestimation of the level of the noise; however, as we can expect, we will get also less noise reduction since we are taking into account less noise to be reduced. For the CLP filter, since we span the entire transient regardless of τ, the traditional noise reduction dependence on the delay τ becomes smaller, as we saw earlier. Some dependence remains, as the correlation of the noise does reduce as τ increases, and therefore we see that for larger delay τ we still have less noise reduction.

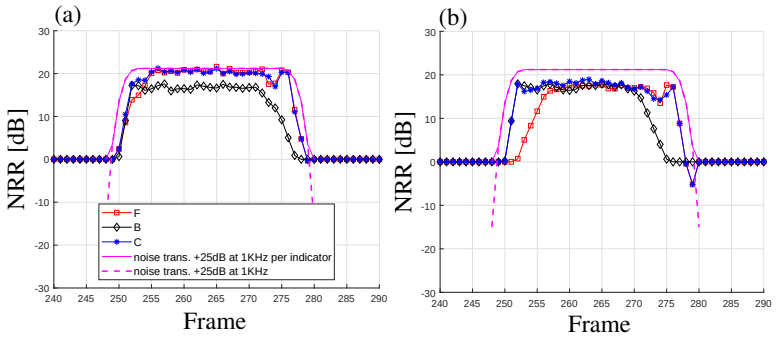

Figure 3.6: Average noise reduction ratio (NRR) per frame for the LS filters applied on a 3.4 s female speech signal corrupted by synthetic nonstationary harmonic noise at 10 dB SNR, filter length $L = 3$ and with delay (a) $\tau = 1$ and (b) $\tau = 3$, where red square is for the forward filter, black diamond is for the backward filter, blue asterisk is for the combined filter, solid magenta is the spectral noise level at 1 KHz given the specified indicator shifted by 25 dB, and dashed magenta is the spectral noise level at 1 KHz shifted by 25 dB.

Given the results for the LS filters using the OD definitions, we continue the analysis with the traditional performance measures only.

3.5.3 Adaptive Filtering

All other experiments are done with noise signals from the collected database. We use the adaptive NLMS filter, where we compare between the forward filter (FMLNLMS) in (3.67), the backward filter (BMLNLMS) in (3.71), and the proposed combined filter (CMLNLMS) in (3.72). In addition, we compare the performance of our proposed approach to the conventional forward NLMS with fixed step size, as well as to the MI approach in [30]. We also propose joining the MI approach and the combined filter by using the MI approach calculation for step size $\mu(k)$ and the identification of the frames

that contain noise, i.e. use the MI as the noise indicator with the calculated step size. We compare this to the previously mentioned methods as well.

We should note that for the MI calculation in [30] a decision block was included to determine whether the clean signal is corrupted by harmonic noise, represented by coefficient Q_μ which would get 1 if it is corrupted by harmonic noise and 0 otherwise. This coefficient multiplies the step size, so if $Q_\mu = 0$, the step size is zero and the noisy signal does not get processed. It is set according to a threshold level, which was decided upon based on a comparison of random noise types such as babble and WGN to harmonic noises such as vehicle engine and traffic. For the highly nonstationary harmonic noises that we evaluate, the resulting coefficient Q_μ is typically zero. Hence, we disregard this coefficient in our analysis and always process the noisy signal.

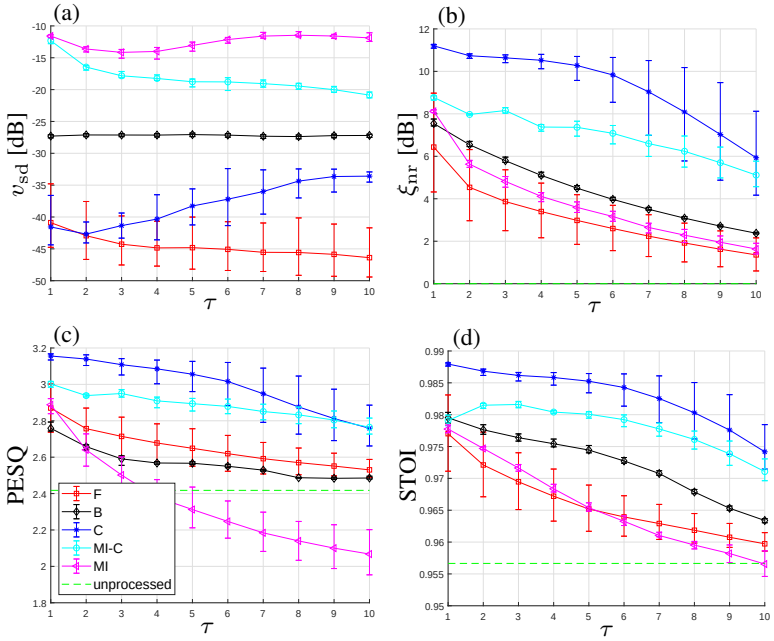

Figure 3.7: (a) Distortion index, (b) noise reduction factor, (c) PESQ, and (d) STOI, for the adaptive filters applied on a 3.4 s female speech signal corrupted by a hospital beeping noise at 10 dB SNR for varying delay τ and filter length L, where red square is for the forward filter, black diamond is for the backward filter, blue asterisk is for the proposed combined filter, cyan circle is for the combined filter using MI, and magenta triangle is for the standard MI. The error bar indicates the results for different $L \in \{3, 5, 8, 10\}$, where the marker is the mean result, the top cap is the max result, and the bottom cap is the min result. Dashed green is the result for the unprocessed noisy signal.

In Fig. 3.7 we show the performance measures for the different adaptive

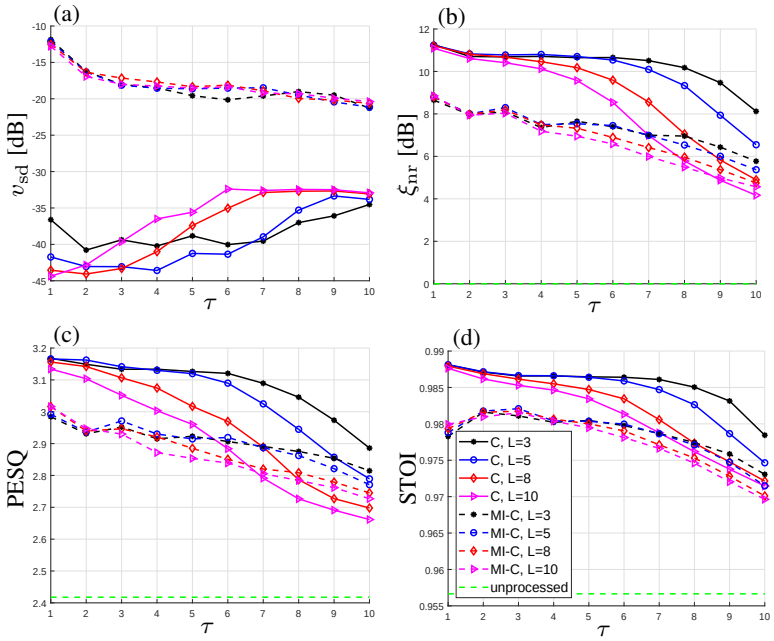

Figure 3.8: (a) Distortion index, (b) noise reduction factor, (c) PESQ, and (d) STOI, for the adaptive filters applied on a 3.4 s female speech signal corrupted by a hospital beeping noise at 10 dB SNR for varying delay τ and different filter lengths L, where solid lines are for the combined filter and dashed lines are for the MI-combined filter. Dashed green is the result for the unprocessed noisy signal.

filters applied on a speech signal corrupted by a real nonstationary harmonic signal, specifically a hospital beeping noise. Similar to the LS, we see that using the combined filter we are improving on all the measures compared to the forward or backward filters standalone, except for additional distortion in some cases. We clearly see that the proposed combined filter outperforms the MI approach. As expected, the proposed combined approach achieves less distortion than the MI approach, since it employs an indicator on frames as well as frequencies. It is worth noting that also for the MI-combined approach which implements the MI indicator while using the combination of forward and backward NLMS filters we get an improvement over the standard MI approach. In fact, with the MI-combined we can get better results for some of the measures compared to the combined. This occurs for large values of the delay parameter and a long filter, as demonstrated in Fig. 3.8, where we compare the combined and the MI-combined for varying delay parameter and different filter lengths. For the proposed combined approach, an appropriate selection of step size μ is still required for optimal results. This remains unchanged from the conventional NLMS. Larger values of the step size lead to faster convergence which results in better noise reduction;

however, we also get more distortion. From $\mu = 0.5$ and larger there is no significant improvement in the noise reduction, but as mentioned the distortion increases, hence, we continue the analysis with $\mu = 0.5$. Another approach not investigated here, would be to use the maximum between the MI calculated step to a fixed value such as the selected 0.5.

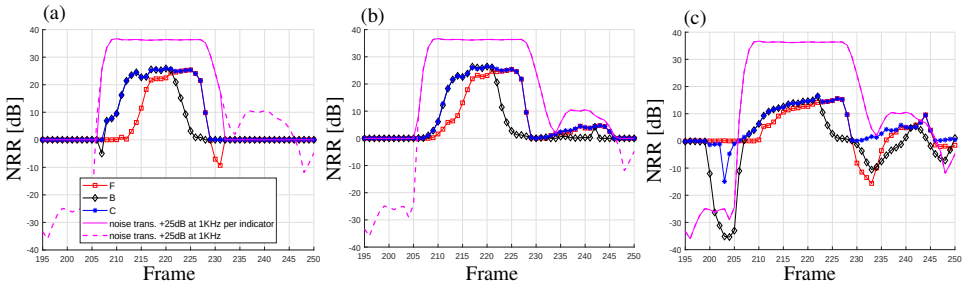

Figure 3.9: Average NRR per frame for the adaptive filters applied on a 3.4 s female speech signal corrupted by a hospital beeping noise at 10 dB SNR, filter length $L = 3$ and with delay $\tau = 3$ for (a) MLNLMS with $\mu = 0.5$ and indicator threshold=-25 dB, (b) MLNLMS with $\mu = 0.5$ and indicator threshold=-40 dB, and (c) MI-MLNLMS. Where red square is for the forward filter, black diamond is for the backward filter, blue asterisk is for the combined filter, solid magenta is the spectral noise level at 1 KHz given the specified indicator shifted by 25 dB, and dashed magenta is the spectral noise level at 1 KHz shifted by 25 dB.

Figure 3.9 illustrates the adaptive filters behavior for the noise transients reduction. We get similar behavior to the LS filter, where the combined filter has increased reduction span of the transient compared to the forward or backward filters standalone. We see a similar improvement using the combined filter with the MI approach compared to the standard MI approach which uses a forward NLMS (shown by the red square curve). Figure 3.9 also demonstrates the reduction of the transient given different indicators: an indicator with threshold value of -25 dB, an indicator with threshold value of -40 dB, and an indicator based on the MI, as used in the MI-CMLNLMS described above. Table 3.1 shows that the range of results for the CMLNLMS between an almost ideal detector (-40 dB threshold) and a degenerate detector that only detects harmonic frequencies (MI) is not large, when the optimal results are not necessarily achieved for the better detector. Therefore, a reasonable amount of misdetection or false-detection in an implemented indicator is not expected to have a large impact.

In (3.67) and (3.71) we calculate the filters for a set of filters, from length 1 to length L per the noise presence and use the maximal filter length available. The advantage of the filter length set compared to a single fixed length for each standalone forward or backward filter is clear. We can

start filtering when the samples where noise is present are few. For the combined filter, the main advantage is for the scenario when there is no overlap between the forward and the backward filters, which can happen when the transient is short compared to the delay. From Fig. 3.8 we see that at larger delays shorter filters have better performance. This can be expected, as at large delay the correlation of both the noise and the signal to their delayed version is reduced. Therefore, we recommend keeping the filter length short, to reduce the required computing resources, to reduce the distortion which generally increases as the filter length is longer, and to achieve overall better performance. As we do not want to introduce a large delay into the system, and given the shown results, we recommend to choose τ to be 1 or at least no more than a few frames. For smaller window size, this should scale up, i.e. a larger delay could be used.

Method	PESQ	STOI	v_{sd}	$\xi_{\text{nr}}[dB]$	oSNR$[dB]$
Unprocessed	2.4177	0.9567	0	0	10
MI-CMLNLMS	2.9843	0.9783	0.065	8.6493	17.8414
CMLNLMS-0.5 -25dB	**3.1677**	**0.9881**	**0.0002**	**11.2678**	**21.2662**
CMLNLMS-0.5 -40dB	3.1569	0.9875	0.0378	10.6965	20.5634

Table 3.1: Results for a 3.4 s female speech signal degraded by hospital beeping noise at 10 dB SNR, with delay $\tau = 1$ and filter length $L = 3$. The performance of the CMLNLMS with different noise indicators is presented, in addition to the unprocessed noisy condition, in terms of quality measure PESQ, intelligibility measure STOI, distortion index v_{sd}, noise reduction factor ξ_{nr}, and overall oSNR.

Table 3.2 shows the mean results for the database of noise signals degrading the database of speech signals at different SNR conditions (20 dB, 10 dB, and 0 dB respectively), for the different methods: the conventional NLMS with fixed step size, the MI approach, the proposed combined filter, and the joint MI-combined filter as well as the unprocessed noisy condition. We can clearly see that with the proposed approach of the CMLNLMS it is possible to achieve the best results for the different performance measures. Second to the CMLNLMS would be the proposed MI-CMLNLMS approach. We should note that with a smaller step size for the conventional NLMS with fixed step size it is possible to decrease the distortion so that it is lower than the combined filter; however, we would get even lower noise reduction and oSNR which are already lower than the combined filter.

To conclude this section, we present in Fig. 3.10 the resulting spectrograms of the enhanced signals for the MI approach and for the proposed

Method	PESQ	STOI	v_{sd}	$\xi_{\text{nr}}[dB]$	oSNR$[dB]$
iSNR 20 [dB]					
Unprocessed	3.1405	0.986	0	0	20
NLMS-0.05	2.8187	0.961	0.1263	2.6729	21.1417
MI	3.208	0.9778	0.0471	3.4674	22.9966
MI-CMLNLMS	3.231	0.9782	0.0465	3.6651	23.1704
CMLNLMS-0.5	**3.4202**	**0.9885**	**0.024**	**5.6781**	**25.5897**
iSNR 10 [dB]					
Unprocessed	2.3409	0.9458	0	0	10
NLMS-0.05	2.17	0.9307	0.1294	4.606	13.0758
MI	2.5134	0.9551	0.0494	5.6614	15.1939
MI-CMLNLMS	2.5519	0.9568	0.0487	6.0447	15.553
CMLNLMS-0.5	**2.6913**	**0.9712**	**0.0231**	**8.2737**	**18.194**
iSNR 0 [dB]					
Unprocessed	1.662	0.8663	0	0	0
NLMS-0.05	1.6107	0.8669	0.136	6.0182	4.4684
MI	1.8383	0.8988	0.0561	7.1133	6.6275
MI-CMLNLMS	1.8618	0.9019	0.055	7.6637	7.1537
CMLNLMS-0.5	**1.8957**	**0.9185**	**0.0216**	**9.7319**	**9.678**

Table 3.2: Results for speech degraded by nonstationary harmonic noise at 0,10, and 20 dB SNR, with delay $\tau = 1$ and filter length $L = 3$. The performance of the conventional NLMS with fixed step size $\mu = 0.05$, the MI approach, the proposed the joint MI-CMLNLMS, and the proposed CMLNLMS with step size $\mu = 0.5$ are presented, in addition to the unprocessed noisy condition, in terms of quality measure PESQ, intelligibility measure STOI, distortion index v_{sd}, noise reduction factor ξ_{nr}, and overall oSNR.

CMLNLMS with $\mu = 0.5$ and -25 dB threshold. By applying the proposed approach, we are able to remove most of the highly nonstationary harmonics with little residuals remaining, while with the MI approach clearly more residuals remain.

3.6 Conclusions

ALE is commonly implemented with a forward NLMS adaptive filter. Here we proposed using an additional non-causal backward NLMS adaptive filter and taking a combination of the forward and backward filters according to the minimal spectral error to reduce nonstationary harmonic noise. The combination of these two filters enables better reduction of the noise tran-

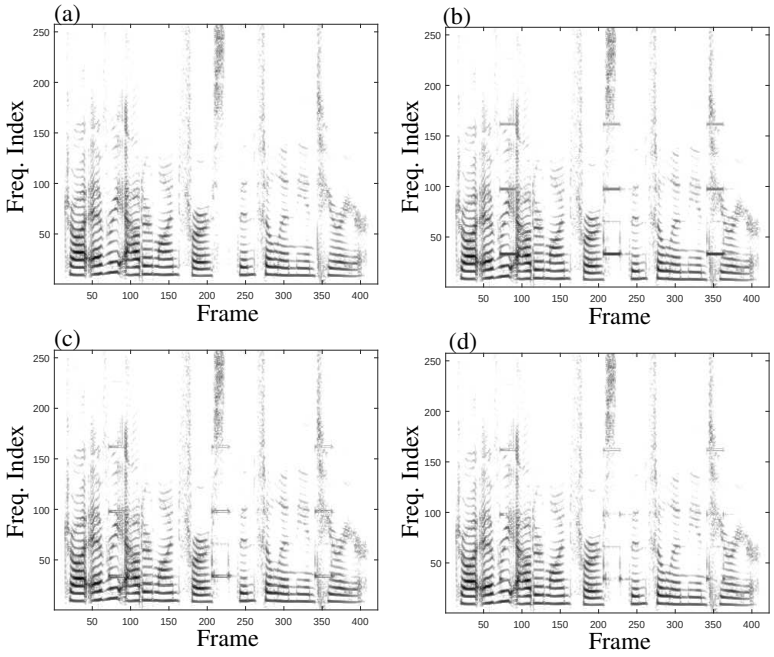

Figure 3.10: Spectrograms of (a) 3.4 s female clean speech signal from the database, (b) speech signal corrupted by nonstationary harmonic hospital beeping noise at 10 dB SNR, (c) enhanced signal obtained from the MI approach, and (d) enhanced signal obtained from the proposed CMLNLMS with $\mu = 0.5$. The results are obtained with filter length $L = 3$, de-correlation delay $\tau = 1$, and indicator threshold of -25 dB for the combined filter.

sients, compared to using only the forward filter, which would only start reducing each noise transient after the de-correlation delay. We showed that our proposed approach is effective compared to conventional forward NLMS with fixed step size and the MI approach; it improves the noise reduction and improves the speech quality and the speech intelligibility, while introducing little distortion. In addition, we showed that the improvement in results holds for different SNR conditions.

We used the combined filter with a noise indicator allowing for lower distortion and better noise reduction. Though the development of the indicator was outside the scope of this chapter, we showed that by using different indicators, including a degenerate indicator that only detects the harmonic frequencies, we were able to improve on the results. Finally, we explained the behavior of the filter for the different parameters, the step size, the de-correlation delay, and the filter length, and provided recommendations on the selection of these parameters. The proposed ALE system was shown to reduce nonstationary harmonic noise. For reduction of the random wide-

band components of the noise, the ALE can be used as the first stage of a two stage system, where it is followed by a traditional spectral speech enhancement method, which we discuss in the next chapter.

Chapter 4

Speech Enhancement Using ARCH model

4.1 Introduction

Spectral domain estimators for speech enhancement, such as the minimum mean-square error (MMSE) estimator [46], the short-time spectral amplitude (STSA) estimator [15], and the log-spectral amplitude (LSA) estimator [16], require an estimate of the a-priori signal-to-noise ratio (SNR). One of the most commonly-used approaches for this estimate is the decision-directed method [15]. Due to non-linearity of the processing methods in these algorithms, artificial noise distortion may be introduced at the output, causing what is known as musical noise. Cappé [38] has shown that using the decision-directed estimator it is possible to reduce this artifact, though at the expense of some distortion to the estimated speech signal and a higher level of residual background noise, which masks the musical noise.

Recently, it was proposed in [80] to use *generalized ARCH* (GARCH) for statistically modeling the speech signals in the time-frequency domain. The GARCH model is extensively used in financial applications where it is necessary to model time varying volatility while taking into account the time-series heavy tailed behavior and volatility clustering. As the short-time Fourier transform (STFT) expansion coefficients exhibit such heavy-tailed behavior and "volatility clustering" (in the sense that large magnitudes tend to follow large magnitudes and small magnitudes tend to follow small magnitudes, while the phase is uncorrelated), the GARCH is a reasonable model.

In this chapter, we investigate the use of a simplified case of the GARCH model, the ARCH model, as part of a spectral domain noise reduction algorithm. We define three measures representing different components of the

sound quality: speech distortion, noise reduction, and musical noise, and we explain the effect that the ARCH model parameters have on these measures. We present a similar evaluation of the decision-directed estimator, with a comparison to the ARCH estimator.

This chapter is organized as follows. In Section 4.2, we review the signal estimation problem, along with the decision-directed and ARCH estimators. In Section 4.3, we present the performance measures employed in our evaluation of the algorithms. Experimental results and discussion are presented in Section 4.4. Finally, a summary is given in Section 4.5.

4.2 Signal Estimation

We consider an observed speech signal $y(n) = x(n) + d(n)$ composed of a clean speech signal $x(n)$ which is corrupted by uncorrelated additive noise $d(n)$, where n is a discrete time index. In the time-frequency domain, applying the STFT, the observed signal is:

$$Y_\ell(k) = X_\ell(k) + D_\ell(k) \tag{4.1}$$

where k is the frequency bin index ($k = 0, 1, \ldots, K-1$), and ℓ is the time frame index ($\ell = 0, 1, \ldots, M-1$). $X_\ell(k)$ and $D_\ell(k)$ are the respective STFT of $x(n)$ and $d(n)$. Our objective is to find an estimate $\hat{X}_\ell(k)$ for the STFT of the clean speech signal.

Given an error function between the STFT of the clean signal and its estimate $e\left(X_\ell(k), \hat{X}_\ell(k)\right)$, the estimated signal is derived from:

$$\hat{X}_\ell(k) = \arg\min_{\hat{X}} E\left[e\left(X_\ell(k), \hat{X}(k)\right) | Y_0(k), \ldots, Y_{\ell'}(k)\right]. \tag{4.2}$$

We shall consider the causal case in which $\ell' \leq \ell$, and the LSA error function

$$e_{\mathrm{LSA}}\left(X_\ell(k), \hat{X}_\ell(k)\right) = \left(\log|X_\ell(k)| - \log\left|\hat{X}_\ell(k)\right|\right)^2. \tag{4.3}$$

Using the statistical signal model in [81], an estimate $\hat{X}_\ell(k)$ can be found independently for each frequency bin index k. Hence, for simplicity, the frequency bin index k will be omitted from this point forward. Also, similarly to [81], we assume knowledge of the noise probability density distribution, which in practice can be estimated using the *Minima Controlled Recursive Averaging* method [24]. The estimate for X_ℓ is obtained by applying a spectral gain function to each noisy spectral component of the observed signal:

$$\hat{X}_\ell = G\left(\xi_{\ell|\ell'}, \gamma_\ell\right) \cdot Y_\ell, \tag{4.4}$$

where the *a-priori* and *a-posteriori* SNRs are defined, respectively, by:

$$\xi_{\ell|\ell'} \triangleq \frac{\lambda_{\ell|\ell'}}{\sigma_\ell^2}, \quad \gamma_\ell \triangleq \frac{|Y_\ell|^2}{\sigma_\ell^2}. \tag{4.5}$$

$\sigma_\ell^2 \triangleq E\left[|D_\ell|^2\right]$ denotes the short-term spectrum of the noise, and $\lambda_{\ell|\ell'} \triangleq E\left[|X_\ell|^2 |Y_0(k), \ldots, Y_{\ell'}\right]$ denotes the short-term spectrum of the speech signal.

For the LSA error function in (4.3), the gain function is [16]

$$G_{\mathrm{LSA}}\left(\xi_{\ell|\ell'}, \gamma_\ell\right) = \frac{\xi_{\ell|\ell'}}{\xi_{\ell|\ell'}+1} \exp\left(0.5 \int_{\nu_\ell}^{\infty} \frac{e^{-t}}{t} dt\right), \tag{4.6}$$

where ν_ℓ is defined by $\nu_\ell \triangleq \dfrac{\gamma_\ell \xi_{\ell|\ell'}}{\xi_{\ell|\ell'}+1}$. Hence, the problem in this case reduces to finding an estimator for the a-priori SNR.

4.2.1 Decision-Directed Estimator

The widely used decision-directed (DD) estimator of Ephraim and Malah [15] is given by:

$$\hat{\xi}_{\ell|\ell} = \max\left\{\alpha \frac{\left|\hat{X}_{\ell-1}\right|^2}{\sigma_\ell^2} + (1-\alpha) P\left(\gamma_\ell - 1\right), \xi_{\min}\right\}, \tag{4.7}$$

where $P(x) = x$ if $x \geq 0$ and $P(x) = 0$ otherwise, and α is a smoothing parameter, $\alpha \in [0, 1)$. As Cappé clearly explained [38], there is a trade-off in the choice of the parameter α. To reduce the musical noise, it is necessary to choose α close to 1; however, the closer α gets to 1, the higher the distortion introduced into the signal is. A typical value for α that has been found to provide a good compromise is 0.98. The parameter $\xi_{\min}$ is the noise floor, essentially controlling the noise reduction and the perceptual masking of the residual musical noise. Applying a lower limit of $\xi_{\min}$ is an additional option for this algorithm which is typically set to -15 dB.

4.2.2 ARCH Model

Instead of the DD estimator, we use a two step estimator [81], composed of a propagation step and an update step to recursively update the estimate of the conditional a-priori SNR.

Suppose we are given an estimate of the *one-frame-ahead a-priori SNR* $\hat{\xi}_{\ell|\ell-1}$ and a new noisy spectral component Y_ℓ, then the estimate $\hat{\xi}_{\ell|\ell}$ can be updated by computing the conditional a-priori SNR:

$$\hat{\xi}_{\ell|\ell} = E\left[\left.\frac{|X_\ell|^2}{\sigma_\ell^2}\right| \hat{\xi}_{\ell|\ell-1}, Y_\ell\right]. \tag{4.8}$$

The gain function

$$G_{\mathrm{SP}}\left(\xi_{\ell|\ell'}, \gamma_\ell\right) = \sqrt{\frac{\xi_{\ell|\ell'}}{\xi_{\ell|\ell'}+1}\left(\frac{1}{\gamma_\ell} + \frac{\xi_{\ell|\ell'}}{\xi_{\ell|\ell'}+1}\right)} \tag{4.9}$$

minimizes the expected spectral power distortion [82], yielding:

$$\hat{\xi}_{\ell|\ell} = G_{\mathrm{SP}}^2\left(\hat{\xi}_{\ell|\ell-1}, \gamma_\ell\right) \cdot \frac{|Y_\ell|^2}{\sigma_\ell^2} = G_{\mathrm{SP}}^2\left(\hat{\xi}_{\ell|\ell-1}, \gamma_\ell\right) \cdot \gamma_\ell. \tag{4.10}$$

Using (4.9) and (4.10), we can write

$$\hat{\xi}_{\ell|\ell} = \alpha_\ell \hat{\xi}_{\ell|\ell-1} + (1-\alpha_\ell)\left(\gamma_\ell - 1\right), \tag{4.11}$$

where

$$\alpha_\ell = 1 - \left(\frac{\hat{\xi}_{\ell|\ell-1}}{\hat{\xi}_{\ell|\ell-1}+1}\right)^2, \quad \alpha_\ell \in [0,1]. \tag{4.12}$$

Computation of the update step requires an estimate of $\hat{\xi}_{\ell|\ell-1}$. According to the GARCH (p,q) model presented in [80],

$$\hat{\xi}_{\ell|\ell-1} = \kappa + \sum_{i=1}^{q} \mu_i \hat{\xi}_{\ell-i|\ell-i} + \sum_{j=1}^{p} \delta_j \hat{\xi}_{\ell-j|\ell-j-1}, \tag{4.13}$$

where the values of the parameters are constrained by

$$\kappa > 0,\ \mu_i \geq 0,\ \delta_j \geq 0, \quad i = 1,\ldots,q, j = 1,\ldots,p$$
$$\sum_{i=1}^{q} \mu_i + \sum_{j=1}^{p} \delta_j < 1.$$

Using the special case GARCH $(0,1)$, also known as ARCH (1), we get the propagation step:

$$\hat{\xi}_{\ell|\ell-1} = \kappa + \mu \hat{\xi}_{\ell-1|\ell-1}, \quad \kappa > 0,\ 0 \leq \mu < 1. \tag{4.14}$$

Since the a-priori SNRs need to be equal to $\xi_{\min}$ when speech is absent, we obtain from (4.14) a condition on κ, $\kappa = (1-\mu)\,\xi_{\min}$, implying

$$\hat{\xi}_{\ell|\ell-1} = (1-\mu)\,\xi_{\min} + \mu \hat{\xi}_{\ell-1|\ell-1}. \tag{4.15}$$

The effect that $\xi_{\min}$ and μ have on the processed signal is investigated in Section 4.4.

Note that from (4.7) and (4.11) we can observe that an a-priori SNR estimator, which is based on a GARCH model, has a similar form of the decision-directed estimator but with a *time-varying frequency-dependent* weighting factor α_ℓ.

4.3 Performance Measures

We employ three performance measures commonly used for the quality assessments of a speech enhancement algorithm. First is the speech distortion, which is used to assess the quality of the estimated speech component. The second measure is the noise reduction, and the third measure is the artificial distortion of the noise, i.e, the musical noise.

4.3.1 Distortion and Noise Reduction Ratio

Combining (4.1) and (4.4) we obtain:

$$\hat{X}_\ell = G\left(\xi_{\ell|\ell'}, \gamma_\ell\right) X_\ell + G\left(\xi_{\ell|\ell'}, \gamma_\ell\right) D_\ell, \tag{4.16}$$

where the first element is the filtered desired signal, and the second element is the residual noise. With the same gain applied to both the signal and the noise, the trade-off between the distortion and the noise reduction when speech is present is clear. From this expression we derive the distortion measure

$$\mathrm{J_X} \triangleq E\left[\left(\log|X_\ell| - \log\left|G\left(\xi_{\ell|\ell'}, \gamma_\ell\right) X_\ell\right|\right)^2\right], \tag{4.17}$$

which is evaluated only in time-frequency bins containing the speech signal. From (4.16) we also derive the noise reduction ratio (NRR), which is evaluated across all time-frequency bins,

$$\mathrm{NRR} \triangleq \frac{E\left[|D_\ell|^2\right]}{E\left[\left|G\left(\xi_{\ell|\ell'}, \gamma_\ell\right) D_\ell\right|^2\right]}. \tag{4.18}$$

4.3.2 Musical Noise via Higher Order Statistics

The attenuated noise, as a result of the processing, will be composed of isolated spectral components, also known as *tonal* components, or musical noise. These tonal components can be quantified, as they are related to the weight of the tail of the noise components' probability density function (pdf). Hence the kurtosis, defined as kurtosis $= \mu_4/\mu_2^2$, where μ_m is the mth order moment of the signal, can be used to evaluate the percentage of components which are tonal. However, as the original noise could also contain some tonal components, and we are interested in the amount of tonal components caused by the processing, we use the ratio of the kurtosis before and after the processing. Hence, we define the third measure, the log

of the kurtosis ratio (LKR) as:

$$\text{LKR} \triangleq \log_{10}\left(\frac{\text{kurtosis}_{\text{proc}}}{\text{kurtosis}_{\text{org}}}\right), \tag{4.19}$$

which is evaluated on noise only frames. $\text{kurtosis}_{\text{org}}$ is the kurtosis of the input noise and $\text{kurtosis}_{\text{proc}}$ is the kurtosis of the processed noise. The LKR increases as the musical noise increases [83], and the absence of musical noise corresponds to LKR of zero and below.

Analytical calculation of the kurtosis ratio requires the use of a specific noise reduction method or assumptions about the statistical spectral components [83, 84]. Here, we use the sample kurtosis [85]:

$$\text{kurtosis} = \frac{1}{M}\sum_{\ell=0}^{M-1}\left[\frac{\frac{1}{K}\sum_{k=0}^{K-1}\left(|D_\ell(k)|^2 - \overline{|D_\ell(k)|^2}\right)^4}{\left(\frac{1}{K}\sum_{k=0}^{K-1}\left(|D_\ell(k)|^2 - \overline{|D_\ell(k)|^2}\right)^2\right)^2}\right] \tag{4.20}$$

where $\overline{|D_\ell(k)|^2} = \frac{1}{K}\sum_{k=0}^{K-1}|D_\ell(k)|^2$. The sample kurtosis is calculated for the input signal and the processed signal separately and then plugged into the LKR (4.19).

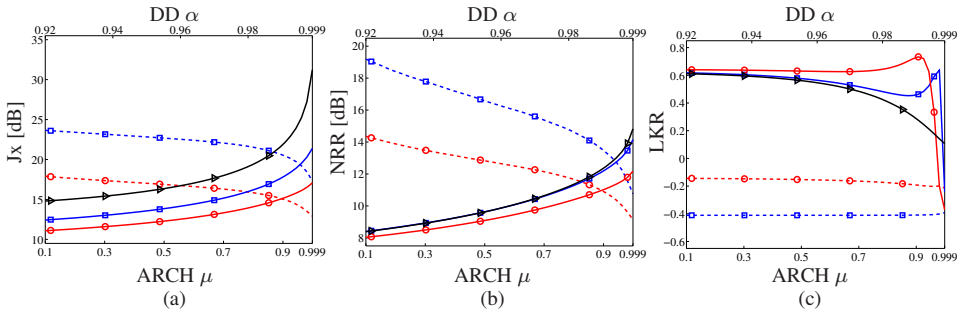

Figure 4.1: Comparison of DD (solid lines) and ARCH (dashed lines) estimators for 5 dB SNR: (a) Distortion, (b) NRR and, (c) LKR, with varying α (upper axis) and μ (lower axis) respectively per estimator, and $\xi_{\min}$ of -20 dB (square), -15 dB (circle), and for DD method only $\xi_{\min} = 0$ (triangle).

4.4 Experimental Results and Discussion

The evaluation of the signal distortion was done on speech signals taken from the TIMIT database [72], including 20 different utterances from 20 different speakers, half male and half female. The signals are sampled at 16kHz and degraded by white Gaussian noise with SNR in the range

$[0, 20]$ dB. The noisy signals are transformed to the time-frequency domain using STFT, with 75% overlapping Hamming analysis windows of 32 ms length. To calculate the distortion (4.17) the time-frequency bins containing speech were defined as $\mathcal{H}_1 = \{\ell, k \mid 20\log_{10}|X_\ell(k)| > \epsilon\}$, where $\epsilon = \max_{\ell,k}\{20\log_{10}|X_\ell(k)|\} - 50$, confining the dynamic range of the log-spectrum to 50 dB.

The evaluation of the musical noise (represented by the LKR) was done separately on a complex white Gaussian noise in the time-frequency domain, to emulate performance in noise only frames.

Figure 4.1(a)–(c) clearly demonstrates Cappé's explanation of the influence of the parameters α and $\xi_{\min}$ on the distortion of the signal, the musical noise, and the residual noise level for the decision-directed estimator. The distortion increases as α increases, while the musical noise decreases and is in fact almost eliminated for high enough α. By taking a higher noise floor $\xi_{\min}$ there is more residual noise, and the value of α required to reduce or eliminate the musical noise decreases. It is interesting to note that for α close to 1 and without a noise floor limit $\xi_{\min}$ (i.e. $\xi_{\min} = 0$), the musical noise monotonically decreases as α increases, but with $\xi_{\min} > 0$ there is actually a slight increase in the amount of musical noise before it drops. When we add a noise floor limit, intervals in which the a-priori SNR was lower than this limit are replaced by that value, thus distorting the spectral noise components pdf.

From Figure 4.1 we can observe that for the ARCH estimator, increasing the value of μ actually decreases the distortion. This can be easily understood from (4.11) and (4.15). As μ increases, the one-frame-ahead a-priori SNR $\hat{\xi}_{\ell|\ell-1}$ becomes more dependent on the previous frame's a-priori SNR $\hat{\xi}_{\ell|\ell-1} \approx \hat{\xi}_{\ell-1|\ell-1}$. Hence, when a signal component appears abruptly ($\gamma_\ell \gg 1$), after a one frame delay, α_ℓ will *decrease* approximately according to the previous frame's a-priori SNR, and the a-priori SNR will follow the a-posteriori SNR of the current frame $\hat{\xi}_{\ell|\ell} \approx \gamma_\ell$ (whereas for the DD after a frame delay it is approximately $\gamma_{\ell-1}$). It is important to note that when the signal component *disappears*, the conditional a-priori SNR immediately drops and there is no frame delay, as opposed to the DD which has one frame delay. This strong dependence on the one-frame-ahead a-priori SNR when μ increases also causes the NRR to decrease, as the a-priori SNR does not drop to $\xi_{\min}$ immediately after frames that contain speech. Clearly, the lower we take the noise floor $\xi_{\min}$, the more noise reduction we get.

While the distortion and NRR strongly depend on the value of μ, the musical noise (LKR) mainly depends on the noise floor $\xi_{\min}$. Lower $\xi_{\min}$

means higher α_ℓ, resulting in a smoother a-priori SNR around $\xi_{\min}$, thus reducing the musical noise. This is in contrast to the DD estimator where for low $\xi_{\min}$ less variation is being masked, resulting in more musical noise and less attenuation, until α is large enough to smooth out the variations.

For the DD estimator we have to compromise between the amount of distortion and amount of musical noise, while for the ARCH estimator, the musical noise can be eliminated by choosing an appropriate value of $\xi_{\min}$. However, for the ARCH estimator we need to compromise between the amount of distortion and the amount of residual noise. In comparison to the typical DD estimator values of $\alpha = 0.98$ and $\xi_{\min} = -15$ dB, where we still have musical noise, we can see that by using the ARCH model we can eliminate the musical noise almost completely if we choose the same $\xi_{\min} = -15$ dB. If we choose, for example, $\mu = 0.98$ we also get less distortion; however, we get slightly less noise reduction than the DD estimator. Maintaining the same $\xi_{\min} = -15$ dB, if α for the DD estimator is increased so that we perceive no musical noise, the speech distortion and noise reduction also increase. Using the ARCH model, we can either use the same values of parameters as before, obtaining less distortion than the DD estimator but also less noise reduction, or we can decrease μ to get higher noise reduction at the expense of higher speech distortion. In both cases we can perceptually avoid musical noise. On the other hand, if we can tolerate a slightly higher distortion than the DD estimator, we can choose smaller μ so that more noise reduction can be obtained compared to DD estimator (with no impact on the musical noise).

4.5 Summary

We have demonstrated and compared the use of the DD and ARCH estimators for spectral speech enhancement. We investigated the influence of the ARCH parameters on the different measures of the sound quality of the processed signal, and demonstrated that by using the ARCH model it is possible to achieve better results than the DD method for some of those measures, specifically the musical noise, while compromising between the speech distortion and noise reduction.

Chapter 5

Robust Superdirective Beamformer with Optimal Regularization

5.1 Introduction

Most fixed conventional beamformers are optimally designed for a given noise field. The most well-known beamformers are the delay-and-sum (DS) [31], which maximizes the signal-to-noise ratio (SNR) gain under white noise conditions, and the superdirective beamformer [32], which does the same only for diffuse noise. Realistic environments, however, are likely to impose several types of noises all at once. Unfortunately, it turns out that beamformers designed to operate solely under white noise perform poorly under diffuse noise, and vice-versa. Hence, extensive work has been done to find a superdirective beamformer with increased robustness to the white noise.

Cox et al. [32, 33] introduced an optimal beamformer which is derived when the white noise gain is constrained. Other methods suggested different variations on optimization problems, e.g., diagonal loading [86], or addressing microphone characteristics mismatch [87, 88], to solve this trade-off. Recently, Berkun et al. [44, 89] proposed robust approaches, which use a closed-form expression that enables tuning the beamformer's performance under various noise types. However, in almost all of these designs, the regularization factor, which is necessary for obtaining optimal results, is not easy to find. Often, the regularization factor is set by some heuristic considerations or some prior knowledge regarding the signal and the interference.

In this chapter, we address the trade-off between the beamformer performances under white noise and diffuse noise by taking a slightly different

approach. In Section 5.2, we present the signal model and the array setup as well as some basic performance measures. Section 5.3 summarizes the properties of conventional beamformers: DS, superdirective, and regularized superdirective. Next, in Section 5.4, we propose the usage of a combined noise field, composed of both white and diffuse noise. Considering this new noise model, we define the relevant SNR gain criterion and find the respective optimal beamformer. We then present a simple and computationally efficient search algorithm for calculating the optimal regularization factor. Section 5.5 shows simulation results, which demonstrate our design method and its improved performance compared to the combined beamformers method described in [44]. Finally, a short conclusion is given in Section 5.6.

5.2 Signal Model and Array Setup

We consider a plane wave, in the farfield, that propagates in an anechoic acoustic environment at the speed of sound in air and impinges on a uniform linear sensor array consisting of M omnidirectional microphones. The distance between two successive sensors is equal to δ and the direction of the source signal to the array is parameterized by the azimuth angle θ. The steering vector (of length M) is therefore given by

$$\mathbf{d}(\omega,\theta) = \left[\begin{array}{cccc} 1 & e^{-j\omega\tau_0\cos\theta} & \cdots & e^{-j(M-1)\omega\tau_0\cos\theta} \end{array}\right]^T, \tag{5.1}$$

where the superscript ^{T} is the transpose operator, $j=\sqrt{-1}$ is the imaginary unit, $\omega = 2\pi f$ is the angular frequency, $f > 0$ is the temporal frequency, and $\tau_0 = \delta/c$ is the delay between two successive sensors at the angle $\theta = 0$. We are interested in superdirective [32, 33] or differential beamforming [90, 69], where the inter-element spacing, δ, is small, the main lobe is at the angle $\theta = 0$ (endfire direction), and the desired signal propagates from the same angle. With the conventional signal model [69], the observation signal vector (of length M) is

$$\begin{aligned} \mathbf{y}(\omega) &= \left[\begin{array}{cccc} Y_1(\omega) & Y_2(\omega) & \cdots & Y_M(\omega) \end{array}\right]^T \\ &= \mathbf{x}(\omega) + \mathbf{v}(\omega) = \mathbf{d}(\omega) X(\omega) + \mathbf{v}(\omega), \end{aligned} \tag{5.2}$$

where $Y_m(\omega)$ is the mth microphone signal, $\mathbf{x}(\omega) = \mathbf{d}(\omega) X(\omega)$, $X(\omega)$ is the desired signal, $\mathbf{d}(\omega) = \mathbf{d}(\omega, 0)$ is the steering vector at $\theta = 0$ (direction of the source), and $\mathbf{v}(\omega)$ is the additive noise signal vector. By applying

a complex-valued linear filter, $\mathbf{h}(\omega)$, to the observation signal vector, we obtain the beamformer output [91]:

$$\begin{aligned} Z(\omega) &= \mathbf{h}^H(\omega)\,\mathbf{y}(\omega) \\ &= \mathbf{h}^H(\omega)\,\mathbf{d}(\omega)\,X(\omega) + \mathbf{h}^H(\omega)\,\mathbf{v}(\omega), \end{aligned} \tag{5.3}$$

where $Z(\omega)$ is an estimate of the desired signal, $X(\omega)$, and the superscript $(\cdot)^H$ is the conjugate-transpose operator. In our context, the distortionless constraint is desired, i.e., $\mathbf{h}^H(\omega)\,\mathbf{d}(\omega) = 1$.

5.3 Performance Measures and Conventional Beamformers

The first important measures are the input and output SNRs. Taking the first microphone as a reference, we can define the input SNR as

$$\mathrm{iSNR}(\omega) = \frac{\phi_X(\omega)}{\phi_{V_1}(\omega)}, \tag{5.4}$$

where $\phi_X(\omega) = E\left[|X(\omega)|^2\right]$ and $\phi_{V_1}(\omega) = E\left[|V_1(\omega)|^2\right]$ are the variances of $X(\omega)$ and $V_1(\omega)$, respectively, with $E[\cdot]$ denoting mathematical expectation. The output SNR is defined as

$$\mathrm{oSNR}[\mathbf{h}(\omega)] = \frac{\phi_X(\omega)}{\phi_{V_1}(\omega)} \times \frac{\left|\mathbf{h}^H(\omega)\,\mathbf{d}(\omega)\right|^2}{\mathbf{h}^H(\omega)\,\mathbf{\Gamma}_{\mathbf{v}}(\omega)\,\mathbf{h}(\omega)}, \tag{5.5}$$

where $\mathbf{\Gamma}_{\mathbf{v}}(\omega) = \frac{E\left[\mathbf{v}(\omega)\mathbf{v}^H(\omega)\right]}{\phi_{V_1}(\omega)}$ is the pseudo-coherence matrix of $\mathbf{v}(\omega)$. From the two previous definitions, we deduce the gain in SNR:

$$\mathcal{G}[\mathbf{h}(\omega)] = \frac{\mathrm{oSNR}[\mathbf{h}(\omega)]}{\mathrm{iSNR}(\omega)} = \frac{\left|\mathbf{h}^H(\omega)\,\mathbf{d}(\omega)\right|^2}{\mathbf{h}^H(\omega)\,\mathbf{\Gamma}_{\mathbf{v}}(\omega)\,\mathbf{h}(\omega)}. \tag{5.6}$$

The most convenient way to evaluate the sensitivity of the array to some of its imperfections such as sensor noise is via the so-called white noise gain (WNG), which is defined by plugging $\mathbf{\Gamma}_{\mathbf{v}}(\omega) = \mathbf{I}_M$ ($\mathbf{I}_M$ is the $M \times M$ identity matrix) into (5.6):

$$\mathcal{W}[\mathbf{h}(\omega)] = \frac{\left|\mathbf{h}^H(\omega)\,\mathbf{d}(\omega)\right|^2}{\mathbf{h}^H(\omega)\,\mathbf{h}(\omega)} \leq M. \tag{5.7}$$

It is easy to see that $\mathcal{W}[\mathbf{h}(\omega)]$ is maximized with the well-known DS beamformer:

$$\mathbf{h}_{\mathrm{DS}}(\omega) = \frac{\mathbf{d}(\omega)}{\mathbf{d}^H(\omega)\,\mathbf{d}(\omega)} = \frac{\mathbf{d}(\omega)}{M}. \tag{5.8}$$

Another important measure, which quantifies how the microphone array performs in the presence of reverberation, is the directivity factor (DF). Considering the spherically isotropic (diffuse) noise field, the DF is defined as

$$\mathcal{D}\left[\mathbf{h}\left(\omega\right)\right] = \frac{\left|\mathbf{h}^{H}\left(\omega\right)\mathbf{d}\left(\omega\right)\right|^{2}}{\mathbf{h}^{H}\left(\omega\right)\mathbf{\Gamma}_{\mathrm{d}}\left(\omega\right)\mathbf{h}\left(\omega\right)} \leq M^{2}, \tag{5.9}$$

where $\mathbf{\Gamma}_{\mathrm{d}}\left(\omega\right) = \frac{1}{2}\int_{0}^{\pi}\mathbf{d}\left(\omega,\theta\right)\mathbf{d}^{H}\left(\omega,\theta\right)\sin\theta d\theta$. It can be verified that the elements of the $M \times M$ matrix $\mathbf{\Gamma}_{\mathrm{d}}\left(\omega\right)$ are

$$\left[\mathbf{\Gamma}_{\mathrm{d}}\left(\omega\right)\right]_{ij} = \frac{\sin\left[\omega(j-i)\tau_{0}\right]}{\omega(j-i)\tau_{0}} = \mathrm{sinc}\left[\omega(j-i)\tau_{0}\right]. \tag{5.10}$$

It can be shown that $\mathcal{D}\left[\mathbf{h}\left(\omega\right)\right]$ is maximized with the conventional superdirective (SD) beamformer [33]:

$$\mathbf{h}_{\mathrm{SD}}\left(\omega\right) = \frac{\mathbf{\Gamma}_{\mathrm{d}}^{-1}\left(\omega\right)\mathbf{d}\left(\omega\right)}{\mathbf{d}^{H}\left(\omega\right)\mathbf{\Gamma}_{\mathrm{d}}^{-1}\left(\omega\right)\mathbf{d}\left(\omega\right)}. \tag{5.11}$$

This filter is a particular form of the celebrated minimum variance distortionless response (MVDR) beamformer [92, 93]. While the DS beamformer maximizes the WNG and never amplifies the diffuse noise since $\mathcal{D}\left[\mathbf{h}_{\mathrm{DS}}\left(\omega\right)\right] \geq 1$, it performs poorly in reverberant and noisy environments, even with a large number of microphones, because its DF is relatively low. On the other hand, with the superdirective beamformer we can obtain a DF close to M^{2}, which is good for speech enhancement (i.e., dereverberation and noise reduction), but the WNG can be much smaller than 1, especially at low frequencies, implying a severe problem of white noise amplification, which is the most serious issue with the SD beamformer.

Hence, one of the most important aspects in practice is how to compromise between $\mathcal{W}\left[\mathbf{h}\left(\omega\right)\right]$ and $\mathcal{D}\left[\mathbf{h}\left(\omega\right)\right]$. Ideally, we would like $\mathcal{D}\left[\mathbf{h}\left(\omega\right)\right]$ to be as large as possible with $\mathcal{W}\left[\mathbf{h}\left(\omega\right)\right] \geq 1$. To achieve this goal, the authors in [32, 33] proposed to maximize the DF, subject to a constraint on the WNG. Using the distortionless constraint, we find the robust superdirective beamformer:

$$\mathbf{h}_{\mathrm{R},\epsilon}\left(\omega\right) = \frac{\left[\epsilon\mathbf{I}_{M} + \mathbf{\Gamma}_{\mathrm{d}}\left(\omega\right)\right]^{-1}\mathbf{d}\left(\omega\right)}{\mathbf{d}^{H}\left(\omega\right)\left[\epsilon\mathbf{I}_{M} + \mathbf{\Gamma}_{\mathrm{d}}\left(\omega\right)\right]^{-1}\mathbf{d}\left(\omega\right)}, \tag{5.12}$$

where $\epsilon \geq 0$ is a Lagrange multiplier. Note that (5.12) is a regularized (or robust) version of (5.11), where ϵ can be seen as the regularization parameter. This parameter aims to compromise between supergain and white noise amplification. A small ϵ leads to a large DF and a low WNG, while a large ϵ yields low DF and large WNG. Two interesting cases of (5.12) are

$\mathbf{h}_{\mathrm{R},0}(\omega) = \mathbf{h}_{\mathrm{SD}}(\omega)$ and $\mathbf{h}_{\mathrm{R},\infty}(\omega) = \mathbf{h}_{\mathrm{DS}}(\omega)$. While $\mathbf{h}_{\mathrm{R},\epsilon}(\omega)$ has some control on white noise amplification, it is certainly not easy to find a closed-form expression for ϵ, given a desired value of the WNG.

5.4 New Noise Field and Proposed Beamformer

We assume that the sensed signal is corrupted both by some additive diffuse noise and by some additive white noise. Therefore, the input SNR is now

$$\begin{aligned}\mathrm{iSNR}(\omega) &= \frac{\mathrm{tr}\left[\phi_X(\omega)\,\mathbf{d}(\omega)\,\mathbf{d}^H(\omega)\right]}{\mathrm{tr}\left[\phi_{\mathrm{d}}(\omega)\,\mathbf{\Gamma}_{\mathrm{d}}(\omega) + \phi_{\mathrm{w}}(\omega)\,\mathbf{I}_M\right]} = \\ &= \frac{\phi_X(\omega)}{\phi_{\mathrm{d}}(\omega) + \phi_{\mathrm{w}}(\omega)},\end{aligned} \tag{5.13}$$

where $\mathrm{tr}[\cdot]$ denotes the trace of a square matrix, and $\phi_{\mathrm{d}}(\omega)$ and $\phi_{\mathrm{w}}(\omega)$ are the variances of the diffuse and white noises, respectively. We deduce that the output SNR is

$$\mathrm{oSNR}[\mathbf{h}(\omega)] = \frac{\phi_X(\omega)\left|\mathbf{h}^H(\omega)\,\mathbf{d}(\omega)\right|^2}{\phi_{\mathrm{d}}(\omega)\,\mathbf{h}^H(\omega)\,\mathbf{\Gamma}_{\mathrm{d}}(\omega)\,\mathbf{h}(\omega) + \phi_{\mathrm{w}}(\omega)\,\mathbf{h}^H(\omega)\,\mathbf{h}(\omega)}. \tag{5.14}$$

As a result, the gain in SNR is

$$\mathcal{G}[\mathbf{h}(\omega)] = \frac{\left|\mathbf{h}^H(\omega)\,\mathbf{d}(\omega)\right|^2}{[1-\alpha(\omega)]\,\mathbf{h}^H(\omega)\,\mathbf{\Gamma}_{\mathrm{d}}(\omega)\,\mathbf{h}(\omega) + \alpha(\omega)\mathbf{h}^H(\omega)\,\mathbf{h}(\omega)}, \tag{5.15}$$

where $\alpha(\omega) = \frac{\phi_{\mathrm{w}}(\omega)}{\phi_{\mathrm{d}}(\omega)+\phi_{\mathrm{w}}(\omega)}$, with $0 \leq \alpha(\omega) \leq 1$. It is easy to check that the beamformer that maximizes $\mathcal{G}[\mathbf{h}(\omega)]$ is

$$\mathbf{h}_\alpha(\omega) = \frac{\mathbf{\Gamma}_{\mathrm{d},\alpha}^{-1}(\omega)\,\mathbf{d}(\omega)}{\mathbf{d}^H(\omega)\,\mathbf{\Gamma}_{\mathrm{d},\alpha}^{-1}(\omega)\,\mathbf{d}(\omega)}, \tag{5.16}$$

where $\mathbf{\Gamma}_{\mathrm{d},\alpha}(\omega) = [1-\alpha(\omega)]\,\mathbf{\Gamma}_{\mathrm{d}}(\omega) + \alpha(\omega)\mathbf{I}_M$. Then, the maximum gain in SNR is

$$\mathcal{G}[\mathbf{h}_\alpha(\omega)] = \mathbf{d}^H(\omega)\,\mathbf{\Gamma}_{\mathrm{d},\alpha}^{-1}(\omega)\,\mathbf{d}(\omega). \tag{5.17}$$

The problem is that $\phi_{\mathrm{d}}(\omega)$ and $\phi_{\mathrm{w}}(\omega)$ are not known in practice. In fact, we can express (5.16) as (5.12) with a frequency dependent regularizer $\epsilon(\omega) = \alpha(\omega)/(1-\alpha(\omega))$, showing that our beamformer is equivalent to (5.12). However, our robust superdirective beamformer (5.16) is preferred for two reasons. First, $\alpha(\omega)$ varies only from 0 to 1 while ϵ in (5.12) varies from 0 to ∞. The second reason is the simple dependence between the gain and $\alpha(\omega)$, which allows us to efficiently find the appropriate $\alpha(\omega)$ values, as

Algorithm 1 MAS - Minimize and Search

Input: Desired gain $\mathcal{G}_0$, and tolerance

Output: Optimal regularization α

1: Find $\alpha_{\min}$ that minimizes the gain (e.g., using gradient descent).
2: Divide the range $[0, 1]$ into 2 sections in which the gain is monotonic: $[0, \alpha_{\min}]$ and $[\alpha_{\min}, 1]$.
3: For each section, apply the following continuous binary search:
4: Divide the section into 2 sub-sections.
5: Calculate the gain $\mathcal{G}_k$ in the middle of each sub-section.
6: Choose the gain $\mathcal{G}_k$ and its respective sub-section for which $|\mathcal{G}_k - \mathcal{G}_0|$ is minimal
7: **if** $|\mathcal{G}_k - \mathcal{G}_0| \leq$ tolerance **then**
8: $\alpha \leftarrow$(middle of chosen sub-section) and stop.
9: **else**
10: update range to be the chosen sub-section and go back to 4
11: **end if**
12: Compare results from $[0, \alpha_{\min}]$ and $[\alpha_{\min}, 1]$ and choose the best result.

will be shown later. Finding the value of $\alpha(\omega)$ that corresponds to a fixed gain of $\mathcal{G}_0$ ($M \leq \mathcal{G}_0 \leq M^2$) can be expressed using the following optimization problem:

$$\min_{\alpha} \left| \mathbf{d}^H(\omega) \, \mathbf{\Gamma}_{\mathrm{d},\alpha}^{-1}(\omega) \, \mathbf{d}(\omega) - \mathcal{G}_0 \right| \text{ s. t. } 0 \leq \alpha \leq 1 \ . \tag{5.18}$$

From simulations not presented here, it can be seen that the gain is continuous and has a single minimum point in the range $\alpha \in [0, 1]$, denoted here as $\alpha_{\min}(\omega)$. The gain will monotonically decrease in the range $[0, \alpha_{\min}(\omega)]$ and monotonically increase in the range $[\alpha_{\min}(\omega), 1]$. This property enables us to calculate α simply by conducting a binary-like search for each monotonic section. This method is described in Algorithm 1, which numerically solves (5.18), i.e., finds α for which the beamformer's SNR gain is closest to $\mathcal{G}_0$ for each frequency independently.

This approach can be used to constrain and optimize other gain properties as well. Instead of fixing the SNR gain, many applications require maximizing it while fixing the WNG or the DF. Since both the WNG and DF are monotonic in $\alpha \in [0, 1]$, as can also be seen in simulations, Algorithm 1 can be used here as well. The computational complexity of the binary-like search is $\mathcal{O}\left\{|\omega| \log_2\left[\left(M^2 - M\right)/\sigma\right]\right\}$, where $\sigma > 0$ is the acceptable tolerance from the desired gain. This is the only step necessary for a fixed WNG/DF. When fixing the SNR gain, we need to add the complexity of finding the initial minimum point, e.g., using gradient descent method with exact line search which requires $\mathcal{O}\left\{\log\left(1/\epsilon_0\right)\right\}$ iterations to converge up to

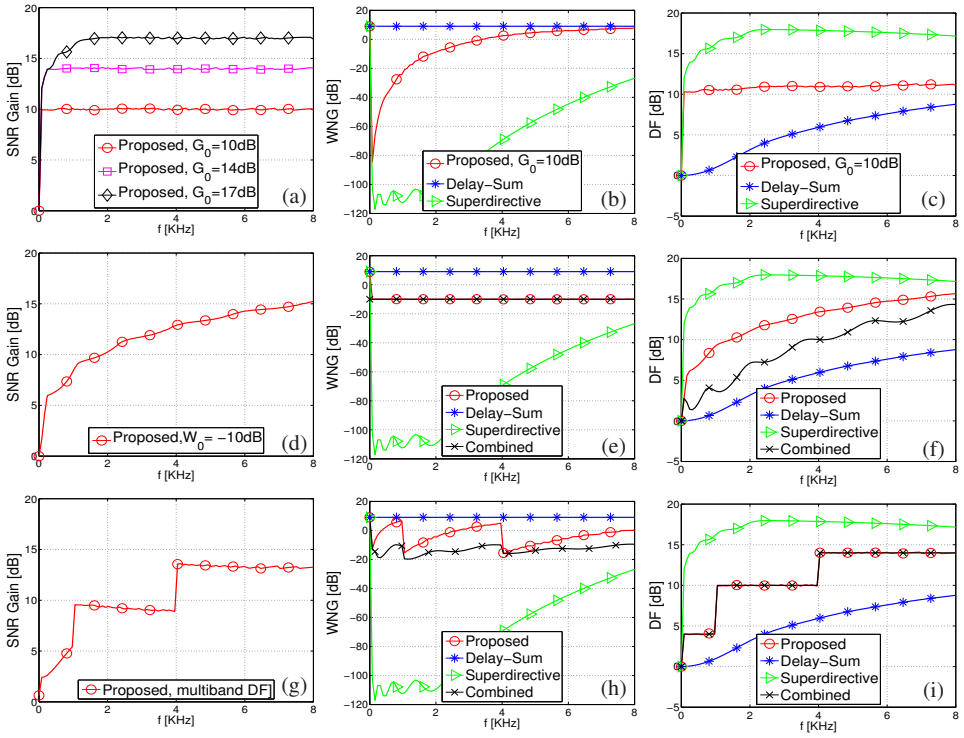

Figure 5.1: Array gains of the proposed beamformer for three cases; fixed SNR gain (a)-(c), fixed WNG (d)-(f) and fixed DF in multi-bands (g)-(i). All three cases are compared to the DS and superdirective beamformers, the latter two are compared also to the combined beamformer with $\epsilon = 10^{-4}$. (a) SNR gain, (b) WNG and (c) DF for fixed SNR gain. (d) SNR gain, (e) WNG and (f) DF with desired WNG set to -10 dB. (h) SNR gain, (g) WNG and (i) DF with desired DF gain set to 4, 10, and 14 dB in multi-bands.

tolerance $\epsilon_0 > 0$ [94].

5.5 Simulation Results

We simulated the proposed robust superdirective beamformer (5.16) for several different gain values, where the regularization parameter $\alpha(\omega)$ was found using Algorithm 1. All the presented simulations were performed for a linear microphone array, with $M = 8$ microphones and $\delta = 1$ cm. However, the results are general, and can be repeated for other configurations. In Figure 5.1(a)–(c) we show the SNR gain alongside the DF and WNG of the proposed beamformer, when set to a constant desired gain level. It can be seen that the gain value can be set as desired within the appropriate range. Although the algorithm converges for every frequency in the range,

the desired gain is not always reached at very low frequencies. This is due to the constant regularization (of 10^{-14}), which is added to avoid singularity issues while inverting $\mathbf{\Gamma}_{\mathrm{d}}$ at low frequencies. In those low frequencies, the constant regularization is more dominant than $\alpha(\omega)$, hence the desired gain is not reached. While achieving the fixed SNR gain under the combined noise field, the proposed beamformer also performs well under diffuse noise. However, white noise may still be amplified to intolerable levels.

To continue and improve the WNG, a modified optimization problem can be defined, which maximizes the SNR gain under a constant WNG. As depicted in Figure 5.1(d)–(f), our approach yields an accurate solution for this scenario as well (using Algorithm 1 from step 4). Furthermore, it can be seen that the proposed beamformer outperforms the combined beamformer [44] with $\epsilon = 10^{-4}$. That is, the proposed beamformer has a higher DF for a fixed WNG given a similar setup. Taking this approach one step further, we design a multi-band fixed beamformer. This way, we can constrain the DF to be piece-wise constant gradually increasing in steps, thus considering the WNG-DF trade-off at each frequency band separately. The proposed approach yields accurate results both for the fixed bands and transition areas, as can be seen in Figure 5.1(g)–(i). A similar analysis can be done to design a multi-band fixed WNG beamformer.

5.6 Conclusions

We have introduced an optimal robust beamformer and a computationally efficient algorithm for finding its regularization parameter. We showed that our approach facilitates the design of beamformers with fixed SNR gain, beamformers with maximal SNR gain for constant WNG or DF, and multi-band fixed beamformers. The proposed design method enables a fine tuning of the compromise between the DF and robustness against white noise.

Chapter 6

Conclusions

6.1 Research Summary

In this thesis we addressed the problem of additive noise reduction where the main focus was on the single channel case.

In Chapter 3 we developed an algorithm to reduce nonstationary harmonic noise by using a frequency domain ALE with a combination of forward and backward linear adaptive filters where we use either one or the other or neither of them according to the lowest spectral error. The ALE is applied based on a noise indicator to further reduce the amount of distortion introduced in the processing. Using simulations of both synthetic and real noise, we demonstrated that this approach has increased reduction span of the noise transients, hence reduced amount of noise residuals, and so outperforms the recent MI approach as well as the conventional fixed step size methods.

In Chapter 4 we demonstrated the use of the ARCH model for the LSA a-priori SNR estimator, a spectral domain noise reduction method. We defined three measures that represent different components of sound quality and analyzed the effect that the ARCH model parameters have on the measures. We conducted a similar evaluation for the well-known DD estimator and compared between the two. We showed that while for the DD estimator the compromise is between the amount of distortion and the amount of musical noise, for the ARCH estimator the musical noise effect can be eliminated. However, it is still necessary to compromise between the distortion and noise reduction, hence the ARCH estimator can outperform the DD estimator for some of the measures.

In Chapter 5 we touched upon the multi-channel case, specifically the design of a microphone array and fixed beamforming, where there is a constant

compromise between performance under white noise conditions and performance under diffuse noise conditions. We proposed to assume a combined noise field that contains both types of noise and based on this assumption developed a robust superdirective beamformer with a regularization parameter that can be found using a simple and efficient one dimensional search algorithm. We demonstrated the design of beamformers with different gain properties, specifically fixed SNR gain, maximal SNR under fixed WNG, and maximal SNR under fixed DF in multi-bands which enables treating the WNG-DF trade-off at each frequency band separately. We showed that the proposed approach has improved performance compared to the competing method of the combined beamformer.

6.2 Future Research

The methods we have proposed in this thesis open a number of options for further study.

1. The method proposed in Chapter 3 for nonstationary harmonic noise reduction assumes a noise indicator is available. Though we discuss the impact the noise indicator has on the results, it would be interesting to implement such a noise indicator and test it together with the developed algorithm. In addition, the algorithm doesn't completely suppress all the noise, some residual noise remain besides the wide-band background noise. Future work could concentrate on jointly suppressing any such residuals along with the remaining background noise. Another approach to the nonstationary harmonic noise reduction could be to develop a deep learning algorithm on a database containing such noises to directly suppress the noise. It would be interesting to see if on the one hand we could utilize the noise structure to simplify the network, and on the other still be able to generalize so that it could be applied successfully also on a set containing stationary harmonic noise such as a vacuum cleaner noise, while of course improving on both the intelligibility and sound quality.

2. We used the ARCH(1) model for the a-priori SNR estimator in Chapter 4, which is a special case of the GARCH(0,1). It would be interesting to expand the model to a full GARCH(p,q) model and conduct a similar analysis, to understand if the full general model could provide additional advantages.

3. In Chapter 5, where we discussed robust superdirective beamforming,

there are a few topics which should be investigated further. First, the proposed beamformer should be tested for various angles of incidence, and not only in the end-fire direction. Also, it may be useful to incorporate additional considerations into the design process, such as side-lobe requirements and performance under other types of noise fields (besides the white noise and the diffuse noise).

Bibliography

[1] J. Benesty and C. Jingdong, *Study and Design of Differential Microphone Arrays*, vol. 6. Springer Science & Business Media, 2012.

[2] M. R. Schroeder, "Apparatus for suppressing noise and distortion in communication signals," filed 1 Dec. 1960, issued 27 Apr. 1965.

[3] S. F. Boll, "Supression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, April 1979.

[4] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. 208–211, April 1979.

[5] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 799–807, November 2001.

[6] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2002.

[7] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, pp. 45–47, February 1991.

[8] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 251–166, July 1995.

[9] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, New Jersey: Wiley intersciences, 2006.

[10] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, pp. 1486–1494, July 2009.

[11] Y. Wang and D. L. Wang, "Boosting classification based speech separation using temporal dynamics," in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, p. 1528–1531, September 2012.

[12] Y. Wang and D. L. Wang, "Cocktail party processing via structured prediction," in *Proceedings of Advanced Neural Information Processing Systems 25 (NIPS)*, p. 224–232, 2012.

[13] J. Roux, J. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70, 2015.

[14] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 137–145, April 1980.

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, pp. 1109–1121, December 1984.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-33, pp. 443–445, April 1985.

[17] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, pp. 1586–1604, December 1979.

[18] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System technical journal*, vol. 54, pp. 297–315, February 1975.

[19] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in *Proceedings of IEEE TENCON*, vol. 3, pp. 321–324, October 1993.

[20] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, vol. 3, pp. 1182–1185, September 1994.

[21] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 504–512, July 2001.

[22] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 2, pp. 1513–1516, September 1995.

[23] L. Lin, W. H. Holmes, and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement," *Electronics Letters*, vol. 39, pp. 754–755, May 2003.

[24] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 466–475, September 2003.

[25] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 153–156, May 1995.

[26] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R.Zeidler, E. Dong, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE*, vol. 63, December 1975.

[27] B. Widrow, J. M. McCool, and M. G. Larimore, "Stationary and nonstationary learning characteristics of the lms adaptive filter," *Proceedings of the IEEE*, vol. 64, August 1976.

[28] N. Sasaoka, K. Shimada, S. Sonobe, Y. Itoh, and K. Fujii, "Speech enhancement based on adaptive filter with variable step size for wideband and periodic noise," in *Proceedings of IEEE International Midwest Symposium on Circuits and Systems*, p. 648–652, August 2009.

[29] I. Nakanishi, H. Namba, and S. Li, "Speech enhancement based on frequency domain ale with adaptive de-correlation parameters," *International Journal of Computer Theory and Engineering*, vol. 5, April 2013.

[30] J. Taghia and R. Martin, "A frequency-domain adaptive line enhancer with step-size control based on mutual information for harmonic noise reduction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, June 2016.

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

[32] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[33] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 3, pp. 393–398, 1986.

[34] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley-Interscience, 2006.

[35] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, pp. 066138–1–066138–16, 2004.

[36] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 253–256, May 2002.

[37] B. Chen and P. C. Loizou, "A laplacian-based mmse estimator for speech enhancement," *Speech Communication*, vol. 49, pp. 134–143–256, 2007.

[38] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 2, pp. 345–349, April 1994.

[39] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary environments," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 789–792, 1999.

[40] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectra amplitude estimator," *IEEE signal processing letters*, vol. 9, pp. 113–116, april 2002.

[41] I. Cohen and B. Berdugo, "Speech enhancement for non-stationarynoise environments," *Signal Processing*, vol. 81, pp. 2403–2418, November 2001.

[42] A. I. Uzkov, "An approach to the problem of optimum directive antenna design," *Comptes Rendus (Doklady) de l'Academie des Sciences de l'URSS*, vol. LIII, no. 1, pp. 35–38, 1946.

[43] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*. John Wiley & Sons, 2003.

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

[45] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Berlin: Springer, 2005.

[46] P. C. Loizou, *Speech Enhancement Theory and Practice*. Boca Raton: CRC Press, 2007.

[47] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.

[48] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 20, p. 2080–2094, September 2012.

[49] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, p. 1218–1234, July 2006.

[50] G. Huang, J. Benesty, T. Long, and J. Chen, "A family of maximum snr filters for noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, p. 2034–2047, December 2014.

[51] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 334–331, July 2003.

[52] Y. Xu, J. Du, L. R. Dai, , and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, pp. 65–68, January 2014.

[53] Y. Xu, J. Du, L. R. Dai, , and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7–19, January 2015.

[54] X. Zhang and D. L. Wang, "A deep ensemble learning method for nonaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 967–977, May 2016.

[55] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, October 2018.

[56] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *Journal of the Acoustical Society of America*, vol. 122, pp. 1777–1786, September 2007.

[57] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 47–56, January 2011.

[58] T. Zhang and A. K. Bhowmik, "Enhancing speech in noisy and reverberant environments using deep learning techniques," in *Proceedings of SID Symposium Digest of Technical Papers*, vol. 49, pp. p467–470, May 2018.

[59] J. Lee and H. G. Kang, "A joint learning algorithm for complex-valued t-f masks in deep learning-based single-channel speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1098–1108, June 2019.

[60] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, p. 483–492, March 2016.

[61] J. Lee, J. Skoglund, T. Shabestary, and H. G. Kang, "Phase-sensitive joint learning algorithms for deep learning-based speech enhancement," *IEEE Signal Processing Letters*, vol. 25, p. 1276–1280, August 2018.

[62] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, , and Y. Haneda, "Dnn-based source enhancement self-optimized by reinforcement learning using sound quality measurements," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 81–85, March 2017.

[63] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5074–5078, April 2018.

[64] R. M. Ramli, A. O. A. Noor, and S. A. Samad, "A review of adaptive line enhancers for noise cancellation," *Australian Journal of Basic and Applied Sciences*, vol. 6, no. 6, pp. 337–352, 2012.

[65] S. Haykin, *Adaptive Filter Theory*. International ed. Upper Saddle River: Pearson, 5 ed., 2014.

[66] N. Sasaoka, M. Watanabe, Y. Itoh, and K. Fujiit, "A study on step size control for noise reconstruction system with ale," in *Proceedings of IEEE International Symposium on Intelligent Signal Processing and Communications*, pp. 307–310, December 2006.

[67] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters—A Theoretical Study*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[68] J. Chen, J. Benesty, Y. Huang, and T. Gaensler, "On single-channel noise reduction in the time domain," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 277–280, May 2011.

[69] 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. 20, pp. 1109–1121, May 2012.

[70] X. L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, p. 697–710, April 2013.

[71] I. Ariav, D. Dov, and I. Cohen, "A deep architecture for audio-visual voice activity detection in the presence of transients," *Signal Processing*, vol. 142, pp. 69–74, May 2018.

[72] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database." National Institute of Standards and Technology (NIST), 1993.

[73] BBC Sound Effects. `http://bbcsfx.acropolis.org.uk`.

[74] Freesound. `https://freesound.org`.

[75] YouTube. `https://www.youtube.com`.

[76] SoundBible. `https://soundbible.com`.

[77] ITU-T, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," itu-t recommendation p.862.2, ITU-T, 2007.

[78] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 2125–2136, September 2011.

[79] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing*. Boston: Artech House, 2005.

[80] I. Cohen, "Modeling speech signals in the time-frequency domain using GARCH," *Signal Processing*, vol. 84, pp. 2453–2459, December 2004.

[81] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 870–881, September 2005.

[82] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement," *EURASIP Journal on Applied Signal Processing*, vol. 2003:10, pp. 1043–1051, February 2003.

[83] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, September 2008.

[84] S. Kanehara, H. Saruwatari, K. S. R. Miyazaki, and K. Kondo, "Theoretical analysis of musical noise generation in noise reduction methods with decision-directed a priori SNR estimator," in *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–4, September 2012.

[85] H. Yu and T. Fingscheidt, "Black box measurement of musical tones produced by noise reduction systems," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 4573–4576, March 2012.

[86] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *IEEE Transactions on Signal Processing*, vol. 51, no. 7, pp. 1702–1715, 2003.

[87] S. A. Vorobyov, A. B. Gershman, and Z. Q. Lou, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 313–324, 2003.

[88] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 617–631, 2007.

[89] R. Berkun, I. Cohen, and J. Benesty, "A tunable beamformer for robust superdirective beamforming," in *Proceedings of IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[90] G. W. Elko and J. Meyer, "Microphone arrays," in *Springer Handbook of Speech Processing*, pp. 1021–1041, Springer, 2008.

[91] J. Benesty, C. Jingdong, and Y. Huang, *Microphone Array Signal Processing*, vol. 1. Berlin: Springer, 2008.

[92] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[93] R. T. Lacoss, "Data adaptive spectral analysis methods," *Geophysics*, vol. 36, no. 4, pp. 661–675, 1971.

[94] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

שיפור אות דיבור בהתבסס על משפר קו אדפטיבי

אביבה אטקינס

שיפור אות דיבור בהתבסס על משפר קו אדפטיבי

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר מגיסטר למדעים בהנדסת חשמל

אביבה אטקינס

הוגש לסנט הטכניון – מכון טכנולוגי לישראל
שבט תש"ף חיפה פברואר 2020

המחקר נעשה בהנחיית פרופסור ישראל כהן בפקולטה להנדסת חשמל בטכניון.

תודות

בראש ובראשונה, ברצוני להביע את תודתי והערכתי הרבה לפרופסור ישראל כהן על הנחייתו, תמיכתו וסבלנותו לאורך כל שלבי המחקר.
אבקש להוקיר תודה גם לפרופסור ג'ייקוב בניסטי על התמיכה המקצועית, ותרומתו למחקר.
תודה לפרופסור רונן טלמון ולפרופסור אמיר אוורבוך על סקירת המחקר ובחינת המחקר.
היה לי גם העונג לעבוד עם יובל בן־חור על פרק 5. תודה על שיתוף הפעולה ועל השיחות המעניינות והממריצות.
תודה מיוחדת למנהלים שלי מקס פרי ובני פילטובסקי וכן לקולגות שלי ב־ MVS HW באינטל על ההבנה והתמיכה בזמן המחקר.
ולבסוף, ברצוני להודות למשפחתי, לבעלי ולחבריי על התמיכה, האהבה והעידוד. אתם שותפים להישג זה.

אני מודה לטכניון, לקרן הלאומית למדע (מענק מס' 576/16) ולתוכנית המחקר המשותפת של הקרן הלאומית למדע והקרן הלאומית למדעי הטבע של סין (מענק מס' 2514/17) על התמיכה הכספית הנדיבה בהשתלמותי.

פרסומים

החיבור על המחקר מבוסס על הפרסומים הבאים:

1. פרק 2 מבוסס על:
A. Atkins, I. Cohen and J. Benesty, Adaptive Line Enhancer for Non-Stationary Harmonic Noise Reduction, submitted to Computer Speech and Language, 2019.

2. פרק 3 מבוסס על:
A. Atkins and I. Cohen, Speech Enhancement Using ARCH model, in Proceedings of IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), 2016.

3. פרק 4 מבוסס על:
A. Atkins, Y. Ben-Hur, I. Cohen and J. Benesty, Robust superdirective beamformer with optimal regularization, in Proceedings of IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), 2016.

תקציר

רעש בסביבה אקוסטית הינו כל אות שאיננו אות הדיבור הרצוי. הרעש פוגם באיכות אות הדיבור וברמת המובנות שלו, ולמרבה הצער לא ניתן להימנע ממנו. מקורות הרעש הם מגוונים: רעש עצמי של המיקרופון הקולט את האות, רעש אדטיבי מהסביבה, רעש הנובע מהאות עצמו כאשר נמצאים בחלל סגור כגון חדר וישנם החזרים מהקירות ומעצמים שנקלטים חזרה במיקרופון בהשהיה מהאות ויוצרים את תופעת ההדהודים, או כתוצאה מצימוד בין מיקרופון לרמקול היוצר הד של האות. בנוסף, הרעש יכול להיות גם אות דיבור, אך ממקור אחר מהרצוי. במחקר זה אנו נדון בהרחבה על רעש שהינו אדטיבי באופיו.

הנחתת הרעש הינו תהליך שמטרתו לנקות את הרעש מאות הדיבור המורעש. תהליך זה ידוע גם בתור שיפור אות הדיבור. בעיית שיפור אות הדיבור נחקרה רבות, ולאורך השנים פותחו מגוון רב של שיטות הנמצאות בשימוש באפליקציות שונות כגון תקשורת טלפונית, ועידות טלפוניות, עזרי שמיעה, ממשקי אדם־מכונה ועוד. עם זאת, בעיית שיפור האות עדיין רלוונטית ומאתגרת. רוב השיטות שפותחו משפרות את איכות אות הדיבור ומנחיתות את הרעש, אך יוצרות עיוותים באות הדיבור הפוגעים במובנותו. למרות שכיום השימוש במערכי מיקרופונים נעשה נפוץ יותר בשל ההתקדמות ביכולות מזעור המערכים, עדיין נעשה שימוש רב במיקרופון בודד לקליטת אותות דיבור. מקרה זה של מיקרופון בודד הוא מקרה קשה ביותר בבעיית שיפור אות הדיבור, והוא ידוע בתור מקרה הערוץ הבודד.

במרבית השיטות לשיפור אות הדיבור נדרש שערוך של הרעש על מנת להנחיתו, כאשר טיב שיפור אות הדיבור יהיה תלוי בטיב השערוך. אם השערוך נמוך מידי יוותרו שאריות רעש, אך אם הוא גבוה מידי יווצר עיוות של אות הדיבור. המקרה של רעש לא סטציונרי, כלומר רעש שמשתנה מהר בזמן, הוא מקרה קשה לשערוך אשר נחקר רבות, ועם זאת השיטות הקיימות עדיין סובלות מהבעיה הבסיסית של שערוך נמוך מידי או שערוך גבוה מידי.

רעש הרמוני, הידוע גם בתור רעש מחזורי, הינו רעש המכיל הרמוניות דטרמיניסטיות, ולכן הוא בעל מבניות אשר ניתן לנצל על מנת להשיג שערוך טוב יותר של הרעש. צלילי שואב אבק, מנוע של כלי רכב, צפצופים של מכשור רפואי או צפצופי אזעקות הינם מספר דוגמאות לרעש הרמוני. משפר קו אדפטיבי (adaptive line enhancer) הינו אלגוריתם שפותח עוד בשנת 1976 לביטול רעשים מחזוריים בערוץ בודד. האלגוריתם כולל אלמנט השהיה שיוצר אות ייחוס מאות הכניסה הבודד, ומסנן אדפטיבי. המסנן מחלץ את החלק

הקורלטיבי בין אות הכניסה לאות המושהה. בהנחה שהרעש ההרמוני נשאר קורלטיבי ואות הדיבור לא נשאר קורלטיבי, פלט המסנן הוא שערוך של הרעש, ותוצאת חיסור פלט המסנן מאות הכניסה המורעש הינה שערוך של אות הדיבור, כלומר אות הדיבור המשופר. טיב השיפור תלוי בהשהיה ובגודל הצעד של המסנן האדפטיבי. עבור השיטה הקונבנציונלית המשתמשת בגודל צעד קבוע, קשה מאוד להשיג הנחתה טובה של הרעש מבלי לגרום לעיוות של אות הדיבור. לאחרונה, פותחה שיטה להורדת רעשים הרמוניים באמצעות משפר קו אדפטיבי, כאשר באמצעות גודל צעד המשתנה בתדר המחושב על ידי מידע הדדי (mutual information), מוצאים את תדרי הרמוניות הרעש ובתדרים אלו מפעילים את המסנן האדפטיבי. דבר זה מאפשר שימור טוב יותר של רכיבי אות הדיבור, כלומר פחות עיוותים. שיטה זו מנחיתה רעשים הרמוניים סטציונריים היטב, אך הטיפול שלה ברעשים הרמוניים לא סטציונריים אינו מיטבי. בשיטה זו מחלקים את האות לבלוקים ולכל בלוק מריצים את האלגוריתם כאשר ההנחה שהרעש בבלוק הוא סטציונרי. עבור רעשים הרמוניים מאוד לא סטציונריים כגון צפצופים של מכשור רפואי הנחה זו לא מתקיימת, או לחילופין אינה אפקטיבית.

בעבודת מחקר זו, אנו מפתחים שיטה להנחתת רעש הרמוני לא סטציונרי בערוץ בודד. אנו מציעים להשתמש במשפר קו אדפטיבי אשר משלב מסנן סטנדרטי סיבתי ומסנן לא סיבתי. אנו משווים את שגיאת השערוך הספקטרלית המתקבלת מכל מסנן, ומשתמשים במסנן הנותן את השגיאה המינימלית, כאשר השגיאה הספקטרלית צריכה להיות קטנה מספקטרום האות המורעש על מנת למנוע שערוך יתר של הרעש. השימוש במסנן המשולב מאפשר הנחתה נרחבת יותר של הרעש. על מנת לשמר ככל הניתן את רכיבי הדיבור ולמנוע עיוותים, אנו משתמשים באינדיקטור לנוכחות רעש הרמוני להפעלת המסנן. אנו מדגימים ומנתחים את השיטה באמצעות סימולציות על רעשים הרמוניים לא סטציונריים סינתטיים ואמיתיים ומראים שהשיטה מסוגלת להגיע לביצועים טובים יותר משיטת המידע ההדדי ושיטות אחרות, הן מבחינת איכות האות והן מבחינת מובנות האות, כפי שניתן להתרשם מהתוצאות.

השיטה שאנו מציעים בחלק הראשון של עבודת המחקר מנחיתה את הרעש ההרמוני. לרוב, לרעש ישנו גם רכיב רחב סרט, לכן שיטה זו מתאימה להיות שלב מקדים בעיבוד האות, ולאחריו יהיה שלב נוסף להנחתת רכיב הרעש רחב הסרט, כאשר ניתן להשתמש בשלב זה בשיטות הספקטרליות הקלאסיות.

בחלק השני של עבודת המחקר אנו חוקרים שיטה ספקטרלית קלאסית, לוגריתם האמפליטודה הספקטרלית (log spectral amplitude), בשילוב עם מודל ה־ Auto־Regressive Conditional Heteroscedasticity ובקצרה ARCH, למשערך האפריורי של יחס אות הדיבור לרעש. מודל זה הוא מקרה מנוון של מודל כללי יותר, ה־ Genralized ARCH (GARCH), אשר נמצא בשימוש באפליקציות פיננסיות למידול תנודתיות המשתנה בזמן עם פילוג זנב עבה (heavy tail) וקיבוץ לאשכולות (clustering). המוטיבציה לשימוש במודל זה הינה התנהגות דומה של קבועי התמרת הפוריה בתדר־זמן של אות הדיבור. אנו מנתחים את השפעת הפרמטרים השונים של המודל על שלושה מדדי ביצועים לאיכות האות ומשווים אותו למשערך מפורסם הנמצא בשימוש נרחב, ה־ Decision־Directed שפותח על ידי אפרים ומלאך.

בחלק האחרון של עבודת המחקר, אנו מתייחסים לבעיית הנחתת הרעש במקרה של

מערך מיקרופונים. מערך מיקרופונים מורכב ממספר מיקרופונים במבנה גאומטרי מסוים הדוגמים את שדה האות במרחב. טיב ביצועי המערך תלוי בגורמים רבים, ביניהם כמות המיקרופונים, איכותם, מבנה המערך והאלגוריתם שמעבד את נתוני המערך לפלט הנקרא עיצוב אלומה (beamforming), אשר יכול להיות קבוע או אדפטיבי. עבור עיצוב אלומה קבוע, תכנון האלומה מתבצע תחת הנחות מסוימות על מבנה המערך, כיוון אות הדיבור ביחס למערך וסטטיסטיקת הרעש, והמקדמים שחושבו נשארים קבועים גם אם בפועל המצב שונה. מעצבי אלומות קבועים קונבנציונליים מתוכננים תחת הנחת שדה רעש מסוים, ולכן הינם אופטימליים לשדה רעש זה ובעלי ביצועים נחותים לשדות רעש אחרים. למשל, מעצב האלומה מסוג השהיה־וסכימה (delay־and־sum) ממקסם את הגבר הרעש הלבן, אך משיג פקטור כיווניות (directivity factor) נמוך מאוד, כלומר הגבר נחות עבור רעש דיפוסיבי שהוא רעש ספרי איזוטרופי. בשונה ממנו, מעצב אלומה על־כיווני (superdirective) משיג פקטור כיווניות נעלה, כלומר הגבר מירבי עבור רעש דיפוסיבי, אך הוא בעל הגבר רעש לבן נחות, במיוחד בתדרים נמוכים. בסביבה אמיתית סביר שיהיו מספר שדות רעש שונים, ולפיכך רוצים לתכנן מעצב אלומה שיהיה חסין נגד סוגים שונים של רעש. על מנת לפתור בעיה זו פותחו שיטות שמבוססות על פתרון בעיית אופטימיזציה תחת אילוצים. בשיטות אלו קיים פקטור מסדיר (regularization factor) שחיוני לביצועים אופטימליים של האלומה, אך קשה למציאה. בשיטות הקיימות פקטור זה לעיתים נקבע באופן היוריסטי או על פי מידע אפריורי לגבי האות והרעש, מידע שלרוב אינו זמין בעת התכנון.

בעבודה זו אנו דנים בשיטה לעיצוב אלומה שמאפשרת שליטה על האיזון בין הגבר הרעש הלבן ופקטור הכיווניות. אנו משתמשים בשדה שהוא שילוב של שדה רעש לבן ושדה רעש דיפוסיבי ומפתחים מעצב אלומה על־כיווני מוסדר (regularized Superdirective) שגם כולל פקטור מסדיר, ומציעים אלגוריתם פשוט ואפקטיבי למציאתו. אנו מדגימים את ביצועי מעצב האלומה באמצעות סימולציות ומראים תוצאות מבטיחות ביותר.