Ego Noise Estimation for Robot Audition

Gökhan Ince

Department of Mechanical and Environmental Informatics

Graduate School of Information Science and Engineering

Tokyo Institute of Technology

September 2011

A thesis submitted for the degree of Doctor of Philosophy
Abstract

Robots should listen to their surrounding world by the microphones embedded in their bodies to recognize and understand the auditory environment. This artificial listening capability called robot audition is an important function to understand the surrounding auditory world including sounds such as human voices, music, and other environmental sounds. Robot audition can be improved by incorporating another modality, robot motion, so that the framework is extended to active robot audition. In that sense, active audition can be considered as the first step towards endowing the robot with intelligent behavior. It provides the robot with a processing architecture that will allow it to learn and reason about how to behave in response to complex acoustic environments and conditions.

The most important problem encountered in the active audition domain is ego noise, which can be described as the robot’s own noise generated during a motion of the robot. However, it cannot be solved effectively with conventional methods proposed in other signal processing domains. The basic problem with ego noise, like all types of noise in a robot audition system, is that it causes the Signal-to-Noise Ratio (SNR) to drop and it contaminates the spectrum of the recorded signal so that it is almost impossible to perform the fundamental applications of robot audition, such as Sound Source Localization (SSL), Sound Source Separation (SSS) and Automatic Speech Recognition (ASR), accurately. Because the complexity of the ego noise is enhanced by the number of motors in action, the negative effects of ego noise are even more severe for a moving robot with many degrees of freedom.

This thesis addresses the estimation problem of the ego noise of a robot in order to suppress it for various tasks. The aim of this thesis is to establish a real-time and online ego noise estimation system. To develop a framework for estimating ego noise and to integrate it into the general robot audition framework effectively, we have to consider the following three issues: (1) modeling the process of ego noise estimation, (2) online
processing and (3) general applicability of our ego noise estimation method for robot audition.

In order to address the modeling issue of ego noise estimation, we first have to resolve three important sub-issues we have determined: Knowledge gathering issue, representation issue and algorithm issue. The templates are good representations of motor noise when the same actions are performed over and over again. We model the ego noise using templates by associating discrete time series data representing the motion (i.e., the angular status of each joint of the robot) with another series of discrete time data representing the ego noise spectrum. The data is stored in a database so that later it can be estimated instantaneously. However, the necessity of offline training poses strict constraints. The new “online” scheme can distinguish between stationary noise (i.e., static fan noise, hardware noise of the robot and possibly changing background noise) and non-stationary ego-motion noise and treat both of them in separate processes. Furthermore, the proposed online training of the templates makes template-based noise estimation method more adequate to real-world applications because it can learn the ego noise of unknown motions on the fly. Whereas the proposed “template learning” mechanism can discriminate the new data entries from the existing templates in the database, the “template update” mechanism adaptively sustains the accuracy and precision of the templates. It also prevents the rapid growth of the size of database. The final issue is the confirmation of general applicability and compliance of the proposed ego noise estimation method on several robot audition applications. The established frameworks for ego noise reduction, noise robust feature extraction, ASR and SSL are presented.

In Chapter 1, we introduce our motivation, our goals, and the technical issues for this study. The problems and requirements for robot audition are explained, and we give the appropriate approaches to these issues.

Chapter 2 surveys the literature related to robot audition and signal processing. Since there are different noise sources in a robot environment and ego noise is strongly intertwined with all of them, our robot audition framework has diverse noise processing blocks. We explain the basic methods used in these blocks in a detailed way as existing work. The properties of all noise sources are explained along with a detailed analysis of the noise signals and robot motions. Also, related work is summarized in this chapter. We describe the technical differences between our approaches and conventional ones.

In Chapter 3, after specifying general criteria to be able to choose the optimal esti-
ABSTRACT

mation process for each noise type, we explain how to approach the modeling process of
ego noise estimation specifically. Later on, we propose an estimation method called pa-
parameterized template estimation. The performance of this original method is compared
with those of existing single-channel noise estimation methods.

Chapter 4 describes the developments we made on the basic parameterized template
estimation system so that it runs online. In order to cope with changing environmental
noise, we modify the abstract template concept to our needs. We generate the templates
in a way that they only represent the non-stationary noise. The stationary portion of
the ego noise with ambient noise is dealt with by a stationary noise estimation method.
We explain the details of this unified framework for noise estimation. Moreover, we
eliminate the necessity of human intervention in the training procedure by introducing
an incremental template learning scheme. Finally, we evaluate the performance of the
proposed methods in terms of estimation quality and noise reduction accuracy by using
objective performance criteria and discuss the results.

Chapter 5 delves into the question of how to suppress the whole-body motion noise of
a robot more robustly. For this purpose we integrate template-based ego noise estimation
with the already established works from the multi-channel noise reduction literature.
Microphone array-based sound source separation is adequate to cancel motor noise with
certain spatial properties, thus the performance of this hybrid noise reduction system
exceeds the individual performances of the template estimation and multi-channel noise
reduction methods. In this chapter, we discuss the implementation and its evaluation
in terms of ASR accuracy.

Chapter 6 describes Missing Feature Theory (MFT)-based integration of ego noise
reduction and ASR. We focus on two different ASR systems: single-talker ASR and
multi-talker ASR. Both systems rely on the single-channel and multi-channel noise re-
duction methods to generate spectro-temporal masks filtering the unreliable acoustic
features. We present detailed results regarding recognition accuracy to determine opti-
mal parameters of the mask generation process for each system.

In Chapter 7, we provide an extended version of the parameterized template esti-
mation to operate on multi-channel audio data. This feature enables an Sound Source
Localization (SSL) scheme to whiten the ego noise allowing to eliminate its interfer-
ing effect on the the spatio-temporal plane of Multiple Signal Classification (MUSIC)
method for SSL. We assess the performance in terms of localization accuracy and peak
detection rates for MUSIC.
Chapter 8 outlines the contributions of this thesis and gives an insight into the remaining issues and future work.

Chapter 9 summarizes and concludes this dissertation.
Acknowledgements

The journey that led to the realization of this doctoral thesis has been rewarding, instructive and exciting. A great part of this is due to the support and care of the countless people who influenced my life and this work. It is with great pleasure that I take this opportunity to express my gratitude for the support I have received.

First, I would like to express my deep and sincere gratitude to my supervisor, Prof. Jun-ichi Imura, for his guidance and support throughout my doctoral studies. I have always admired his professionalism and sense of commitment. His overall view on research and his high-quality work standards have deeply influenced my work. Most of all, I am grateful to him for giving me the opportunity to pursue my interest in Robotics and Robot Audition under his supervision and for his continuous care.

I would like to express my foremost gratitude to my daily supervisor and also my manager, Prof. Kazuhiro Nakadai for his constant guidance, and fruitful suggestions. He encouraged me to carry out my research for this dissertation through my current job. His patience, pedagogic abilities and profound knowledge in the field of Robotics and Signal Processing have contributed to making him a great mentor. Without his support, I could not have made this dissertation a reality.

I would also like to express my appreciation to all the members of my thesis committee, Prof. Jun-ichi Imura, Prof. Koichi Shinoda, Prof. Tohru Yagi, Prof. Kenji Amaya and Prof. Masaaki Okuma for their valuable comments on this dissertation.

My most sincere thanks go to Dr. Tobias Rodemann from Honda Research Institute Europe GmbH (HRI-EU) for his enthusiastic supervision and for his encouraging way of guiding me to a deeper understanding of research work. Not to mention his positive attitude and joy of life which have cheered me up countless times during the dark moments of PhD-student life. It was an absolute pleasure working with him. Actually, he kindly gave me many suggestions before flying to Japan and starting my doctoral course.

I am most grateful to Mr. Hiroshi Tsujino for giving me the opportunity of joining
Honda Research Institute Japan Co., Ltd. (HRI-JP) and for encouraging me to walk the path of research. I also thank all of the team members (including former members) belonging to HRI-JP for the intelligent discussions. Among them especially, Dr. Hirofumi Nakajima, Mr. Keisuke Nakamura and Mr. Yuji Hasegawa helped me a lot as the co-authors of many of my papers. I would like to express my gratitude to the director of HRI-JP, Mr. Yasunobu Kawakami and my chief coordinator Mr. Shuji Nitta, who continuously took care of my administrative and daily problems during my stay in Japan. They clearly emphasized the importance of a doctoral degree in my individual development plan, and encouraged me to tackle the challenge.

I wish to thank all the members in Imura-Hayakawa-Nakadai Laboratories, who accepted me with a warm and friendly atmosphere. I especially thank Dr. Kenji Kashima for giving me valuable hints in my very early days in Tokyo Institute of Technology and taking care of my enrollment procedure. Mr. Takami Yoshida, as my tutor, helped me in countless situations, where I got lost.

My thoughts go also to all my friends, especially in Japan, in Germany and in Turkey, who enriched my life with their presence. I would not be where I am today without their support, care and love.

To my family, words fail to express my gratitude. I owe my parents, Sidika and Cihan Ince, much of who I have become. They have raised me to set high goals for myself and have always encouraged me to put education as a first priority in my life. They taught me to value honesty, hard work, and humility above all other virtues. Without their care and encouragement in my daily life, I could not even have begun this dissertation, let alone manage to complete it. Lastly, I would like to thank my loving, supportive, encouraging, and patient sweetheart Elif whose faithful support of this Ph.D. is so much appreciated.
Contents

Abstract i
Acknowledgements v
Contents vii
List of Figures xi
List of Tables xv

1 Introduction 1
   1.1 Motivation ................................................. 1
   1.2 Goal and Technical Issues .................................. 5
   1.3 Overview of Our Approaches ................................ 7
   1.4 Thesis Organization ....................................... 9

2 Literature Review and Basic Methods 11
   2.1 Preliminaries .............................................. 12
       2.1.1 Setting of Microphones ............................... 12
       2.1.2 Acoustic Signal Model ............................... 12
       2.1.3 Sound Fields ......................................... 15
       2.1.4 Noise Problem in Robot Audition ................... 16
   2.2 Existing Work on Diffuse Noise Reduction ............... 24
       2.2.1 Single-channel Noise Estimation Methods .......... 26
       2.2.2 Gain Calculation and Spectral Filter ............. 29
   2.3 Existing Work on Directional Noise Reduction ........... 31
       2.3.1 Beamforming Techniques ............................ 31
       2.3.2 Blind Source Separation Techniques ............... 32
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.3 Geometric Source Separation</td>
<td>33</td>
</tr>
<tr>
<td>2.3.4 Supporting Microphone Array Technologies to GSS</td>
<td>34</td>
</tr>
<tr>
<td>2.4 Existing Work on Noise-robust Automatic Speech Recognition</td>
<td>39</td>
</tr>
<tr>
<td>2.4.1 Basic Automatic Speech Recognition System</td>
<td>39</td>
</tr>
<tr>
<td>2.4.2 Acoustic Model Adaptation</td>
<td>40</td>
</tr>
<tr>
<td>2.4.3 Missing Feature Theory-based ASR</td>
<td>41</td>
</tr>
<tr>
<td>2.5 Related Work on Ego Noise Reduction</td>
<td>42</td>
</tr>
<tr>
<td>2.6 Positioning of this Thesis towards Related Work</td>
<td>44</td>
</tr>
<tr>
<td>3 Template Based Ego Noise Estimation</td>
<td>47</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>47</td>
</tr>
<tr>
<td>3.2 Issues in Noise Estimation</td>
<td>48</td>
</tr>
<tr>
<td>3.3 Ego Noise Estimation using Templates</td>
<td>50</td>
</tr>
<tr>
<td>3.3.1 Biological Inspiration of Ego Noise Cancellation</td>
<td>51</td>
</tr>
<tr>
<td>3.3.2 Approaches for Estimating Ego Noise using Templates</td>
<td>52</td>
</tr>
<tr>
<td>3.4 Parameterized Template Estimation</td>
<td>54</td>
</tr>
<tr>
<td>3.4.1 Method for Knowledge-gathering Issue: Framewise Templates</td>
<td>54</td>
</tr>
<tr>
<td>3.4.2 Method for Representation Issue: Parameterized Templates based on the State of Robot Joints</td>
<td>55</td>
</tr>
<tr>
<td>3.4.3 Method for Algorithm Issue: Nearest Neighborhood Search in the Template Database</td>
<td>57</td>
</tr>
<tr>
<td>3.5 Parameterized Template Subtraction</td>
<td>58</td>
</tr>
<tr>
<td>3.6 Evaluation</td>
<td>59</td>
</tr>
<tr>
<td>3.6.1 Evaluation Criteria</td>
<td>59</td>
</tr>
<tr>
<td>3.6.2 Experimental Settings</td>
<td>60</td>
</tr>
<tr>
<td>3.6.3 Estimation Performance</td>
<td>61</td>
</tr>
<tr>
<td>3.6.4 Noise Reduction Performance</td>
<td>63</td>
</tr>
<tr>
<td>3.6.5 Discussion and Future Work</td>
<td>64</td>
</tr>
<tr>
<td>3.7 Summary</td>
<td>65</td>
</tr>
<tr>
<td>4 Online System for Ego Noise Estimation</td>
<td>67</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>67</td>
</tr>
<tr>
<td>4.2 Unified Framework for Noise Estimation</td>
<td>68</td>
</tr>
<tr>
<td>4.3 Incremental Template Learning</td>
<td>69</td>
</tr>
<tr>
<td>4.3.1 Extensions of the Basic Classification Model</td>
<td>69</td>
</tr>
</tbody>
</table>
## CONTENTS

4.3.2 Learning Algorithm ........................................ 70
4.4 Evaluation ..................................................... 72
  4.4.1 Evaluation Criteria ........................................ 72
  4.4.2 Evaluation of Unified Framework for Noise Estimation .... 74
  4.4.3 Evaluation of Incremental Learning ......................... 78
4.5 Summary ....................................................... 83

5 Whole Body Motion Noise Reduction of a Robot 85
  5.1 Introduction .................................................. 85
  5.2 Issues on Whole Body Motion Noise .......................... 85
  5.3 Hybrid Framework ........................................... 87
    5.3.1 Multi-channel Noise Reduction System .................... 88
    5.3.2 Switching Module for ASR Input Selection ................. 89
  5.4 System Architecture and Implementation ..................... 90
  5.5 Evaluation .................................................. 91
    5.5.1 Experimental Settings ................................... 91
    5.5.2 Speech Recognition with Arm and Leg Motion Noise ....... 92
    5.5.3 Speech Recognition with Head Motion Noise ............... 94
    5.5.4 Discussion and Future Work ............................ 96
  5.6 Summary ..................................................... 97

6 MFT-based Integration of Ego Noise Reduction and ASR 99
  6.1 Introduction .................................................. 99
  6.2 MFT-ASR System for a Single Talker ........................ 99
    6.2.1 System Overview ........................................ 100
    6.2.2 Selective White Noise Superimposition .................... 100
    6.2.3 Missing Feature Mask Generation ........................ 101
  6.3 MFT-ASR System for Multiple Talkers ......................... 102
    6.3.1 System Overview ........................................ 103
    6.3.2 Reliability Estimation for Multiple Talkers .............. 104
    6.3.3 Reliability Estimation for Ego Noise .................... 104
    6.3.4 MFM Generation ........................................ 105
    6.3.5 MFM Integration ........................................ 106
  6.4 Evaluation .................................................. 107
    6.4.1 Evaluation of MFT-ASR for Single Talker ................. 107
List of Figures

1.1 Robot audition ................................................. 2
1.2 Active audition ............................................. 2
1.3 Organization of this thesis ................................. 9

2.1 Signal model with a single sound source .................. 14
2.2 A typical robot audition environment ....................... 17
2.3 A typical robot with ego noise sources ..................... 19
2.4 Spectrograms of various sounds in a robot environment .... 20
2.5 The relationship between the state of motors and head motion noise ... 23
2.6 The relationship between the state of motors and arm motion noise ... 23
2.7 The relationship between the state of motors and head motion noise for a motion repeated three times ............................. 24
2.8 The relationship between the state of motors and arm motion noise for a motion repeated three times ......................... 25
2.9 General configuration for single-channel noise reduction ....... 26
2.10 A general scheme representing multi-channel noise reduction .......... 35
2.11 Architecture of the post-filter .............................. 39
2.12 Proposed structure of ASR: Black modules indicate the basic building blocks of the ASR system, whereas the red modules are the alternative blocks to deal with noise. ................................. 40

3.1 Electric fish radiating its EOD .............................. 51
3.2 Ego motion processes ....................................... 53
3.3 Template-based representation of ego noise ................. 53
3.4 Illustration of the two template types ....................... 55
3.5 Template-based representations of ego noise ................ 56
3.6 Generation of a template database .......................... 57
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.7</td>
<td>Ego motion noise estimation from database</td>
<td>58</td>
</tr>
<tr>
<td>3.8</td>
<td>Noisy spectrogram (Stationary noise: Background noise + hardware noise + fan noise)</td>
<td>62</td>
</tr>
<tr>
<td>3.9</td>
<td>Estimated noise spectra</td>
<td>62</td>
</tr>
<tr>
<td>4.1</td>
<td>Offline template database generation</td>
<td>68</td>
</tr>
<tr>
<td>4.2</td>
<td>Offline template estimation using a unified framework</td>
<td>69</td>
</tr>
<tr>
<td>4.3</td>
<td>Online template learning, update and estimation scheme in a single framework</td>
<td>72</td>
</tr>
<tr>
<td>4.4</td>
<td>Experimental setup</td>
<td>75</td>
</tr>
<tr>
<td>4.5</td>
<td>Refined spectrograms ($SNR = 3,dB$)</td>
<td>76</td>
</tr>
<tr>
<td>4.6</td>
<td>Refined spectrograms ($SNR = 3.2,dB$)</td>
<td>77</td>
</tr>
<tr>
<td>4.7</td>
<td>Estimation error in relation with the number of iterations</td>
<td>79</td>
</tr>
<tr>
<td>4.8</td>
<td>Number of templates in relation with the number of iterations</td>
<td>79</td>
</tr>
<tr>
<td>4.9</td>
<td>Estimated noise spectrograms</td>
<td>80</td>
</tr>
<tr>
<td>4.10</td>
<td>Smoothed NNEE distribution in time ($-3.0,dB$)</td>
<td>81</td>
</tr>
<tr>
<td>4.11</td>
<td>Smoothed NNEE distribution over frequency bins ($-3.0,dB$)</td>
<td>81</td>
</tr>
<tr>
<td>5.1</td>
<td>Typical spectrograms of three types of motor noise</td>
<td>87</td>
</tr>
<tr>
<td>5.2</td>
<td>Proposed noise cancellation system</td>
<td>90</td>
</tr>
<tr>
<td>5.3</td>
<td>Recognition performance during arm and leg motions</td>
<td>93</td>
</tr>
<tr>
<td>5.4</td>
<td>Recognition performance during head motion</td>
<td>94</td>
</tr>
<tr>
<td>5.5</td>
<td>SNR improvements of compared methods</td>
<td>96</td>
</tr>
<tr>
<td>6.1</td>
<td>Proposed noise cancellation system</td>
<td>100</td>
</tr>
<tr>
<td>6.2</td>
<td>Proposed multi-talker speech recognition system</td>
<td>103</td>
</tr>
<tr>
<td>6.3</td>
<td>Spectra of speech signal (utterance: ”Nan desu ka?” (What is this?)), noisy speech signals, refined speech signals and corresponding masks. In a)-h), the y-axis represents 256 frequency bins between 0 and 8kHz and in i)-j) the y-axis represents 13 static MSLS features. The x-axis in all panels is the index of frames.</td>
<td>109</td>
</tr>
<tr>
<td>6.4</td>
<td>Speech recognition performance of different processing stages</td>
<td>110</td>
</tr>
<tr>
<td>6.5</td>
<td>Speech recognition performance of different processing stages</td>
<td>111</td>
</tr>
<tr>
<td>6.6</td>
<td>Speech recognition performance of soft mask - hard mask comparisons for given parameters.</td>
<td>112</td>
</tr>
<tr>
<td>6.7</td>
<td>Recognition performance for different types of ego-motion noise</td>
<td>113</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

6.8 Spectrograms for preprocessing. y-axis represents frequency bins between 0 and 8kHz. x-axis represents the index of frames. .......................... 115
6.9 Masks generated with a mild threshold value ($T_e=0.1$). y-axis represents 13 static mel-features. x-axis represents the index of frames. .............. 116
6.10 Recognition performance for the speaker located at $[-80^\circ, 0^\circ, 80^\circ]$ ...... 116
6.11 Recognition performance for the speakers located at $[-20^\circ, 0^\circ, 20^\circ]$ ...... 117
6.12 Average WCRs for all speakers located at $[-80^\circ, 0^\circ, 80^\circ]$ ..................... 118
6.13 Average WCRs for all speakers located at $[-20^\circ, 0^\circ, 20^\circ]$ ..................... 118

7.1 Parameterized template prediction method and its application to SSL ...... 122
7.2 MUSIC spectrum obtained with SEVD during fan noise and arm motion (E2). ................................................................. 125
7.3 MUSIC spectrum obtained with GEVD (fixed) during fan noise and arm motion (E2). ................................................................. 125
7.4 MUSIC spectrum obtained with GEVD (estimated) during fan noise and arm motion (E2). ................................................................. 125
7.5 3-dimensional representation of Figure 7.2 ................................................................. 126
7.6 3-dimensional representation of Figure 7.3 ................................................................. 126
7.7 3-dimensional representation of Figure 7.4 ................................................................. 126
7.8 Prediction of positions based on the highest peak of MUSIC spectrum in each frame during random arm and head motion (E3). ................. 127
7.9 Prediction of positions based on the highest peak of MUSIC spectrum in each frame during random arm and head motion (E3). ................. 127
7.10 Prediction of positions based on the highest peak of MUSIC spectrum in each frame during random arm and head motion (E3). ................. 127
7.11 Peak Accuracy curves for all three methods ............................................................. 128

8.1 Ego motion processes .......................................................... 134
# List of Tables

3.1 Properties of noise modeling strategies depending on the knowledge gathering and representation issues ............................... 49
3.2 Parameter settings for MCRA and HRLE .............................................. 61
3.3 Noise estimation performance for all methods .................................. 63
3.4 Ego noise reduction performance for all methods ............................... 63
3.5 Ideal noise estimation and reduction performance for $SNR_{(2)} = -3dB$ 64
3.6 Ego noise reduction performance for all methods ............................... 64

4.1 Parameter settings for HRLE ................................................................. 74
4.2 Ego noise reduction performance for all methods ............................... 76
4.3 Ideal noise estimation and reduction performance for $SNR_{(2)} = -3dB$ 76
4.4 Ego noise reduction performance for all methods ............................... 77
4.5 Ego noise estimation performance for all methods ............................... 80
4.6 Ego noise reduction performance for all methods ............................... 82
4.7 Ego noise reduction performance for all methods ............................... 82

5.1 Recognition accuracy (%) for different ego-motions achieved by single-channel template subtraction and multi-channel noise reduction $(SNR=-5dB)$ ................................................................. 95

6.1 Recognition accuracies (% WCR) of all methods utilized in this study with realistic SNRs during a robot-human interaction $(SNR\in\{-5,0\}dB)$ 111

7.1 Mean localization error (MLE [°]) results for E1 .................................. 124
7.2 Mean localization error (MLE [°]) results for E2 .................................. 124
7.3 Mean localization error (MLE [°]) results for E3 .................................. 124
Chapter 1

Introduction

1.1 Motivation

As the technologies in electronics, mechanics, control theory, and computer science continue to advance, robotics also continues to evolve. The design and development of robots started mainly for industrial purposes, whereas nowadays the robots operate in almost all fields of service. Furthermore, the mobility and perceptual capabilities of the robots make them the perfect candidates to be deployed to house-hold environments to help humans. In order to benefit from the deployment of robots, it is essential that these robots are able to walk (relocate their body), make gestures and perform tasks, while still being aware of their environment.

In this respect, robots should listen to their surrounding world by their own ears (microphones) to recognize and understand the auditory environments. This artificial listening capability is called “robot audition.” Robot audition is considered as an important function to understand the surrounding auditory world such as human voices, music, and other environmental sounds. It comprises many subjects from mainly three academic fields: Artificial Intelligence (AI), signal processing and robotics (see Figure 1.1). One of its extensions called “active audition” [1] improves robot audition by integrating it with another modality, active motion, such as turning towards a target sound source, approaching the target or relocating itself to avoid noise sources. Active audition is basically inspired by the early works carried out in the area of active vision [2], which provides a framework for obtaining additional information by coupling vision with behaviors, such as control of optical parameters or actuating camera mount positions. In that sense, active audition can be considered as the first step towards
improving the auditory capabilities of the robot with intelligent behavior in complex acoustic environments and conditions (see Figure 1.2). However, the audition for robots is a relatively young field compared to the progress which has been achieved in the field of signal processing and electrical engineering using static mounted microphones, microphone arrays and hand-held communication devices since decades. Furthermore, some problems specific to the robot audition domain cannot be solved effectively with conventional methods proposed in other signal processing domains. These include changes in the environmental noise characteristics due to the motion of the robot/robot-embedded microphones, and the robot’s own noise called “ego noise.”

![Figure 1.1: Robot audition](image1)

**Figure 1.1: Robot audition**

![Figure 1.2: Active audition](image2)

**Figure 1.2: Active audition**

Ego noise of a robot can be defined as the sum of fan noise, hardware noise and motor
noise (i.e., ego-motion noise); in other words all the noises emitted from the physical embodiment of the robot. Fan noise comes from the fans that are located throughout the interior of the robot and help to dissipate the large amount of heat generated by the Central Processing Unit (CPU), the power supply and other components, while hardware noise stems from the electrical circuits. The static (steady-state) fan noise and hardware noise can be removed easily by applying spectral filtering operations. In contrast, ego-motion noise is of special interest because it occurs only when the robot is performing an action using its motors. This special type of mechanical noise has so far either been ignored or circumvented due to its non-stationarity and complex characteristics. The complexity is enhanced by the number of motors in action, making noise even more severe for a moving robot with many Degrees of Freedom (DoF). Nevertheless, mobility is a necessary condition for improving the perceptual capabilities of robots and endows the robot with intelligent behavior that will allow it to learn and reason about how to behave in response to complex goals in a complex world. Thus, an autonomous robot with active perception may require a highly robust ability to suppress ego noise at any cost.

An efficient design for noise estimation must fulfill several criteria, such as low computational cost and its compatibility to other robot audition methods. Computationally inexpensive methods are favored to be used in robotic platforms because the resources are rather limited and complex algorithms usually do not meet the real-time constraints. The compatibility is another important requirement in regards to the main aim of robot audition, which is an audio/speech based interaction between a robot and human. Therefore, ego noise estimation must serve to greater purposes, such as noise reduction, speech enhancement, Automatic Speech Recognition (ASR) etc.. A system designed to achieve a robot audition task can be considered useful only if the method is compatible to other components used in the system or if the method can be adopted to the system easily. In this respect, the proposed design must be flexible enough to be integrated into systems with other robot audition components.

Ego noise estimation and reduction consequently, pursue to remove the practical limitations of an active audition framework for a moving robot working in real world and real time. The ultimate goal of this research in a broad sense is to enhance a robot with a system that allows it to listen to its surrounding environment without being affected by its ego noise and thus improves its perceptual capabilities.

The main benefits of being able to mitigate the disturbance of ego noise are:
1. Uninterrupted audition of the robot (e.g., understanding human commands even while performing some other task).

2. Improved perceptual capabilities by applying intelligent motion strategies (e.g., screening of noise sources, moving to a less noisier spot, estimating the distance).

3. Improved naturalness of human-robot interaction (e.g., following the speaker, keeping optimal talking distance, using gestures).

Because we perform ego noise reduction in the pre-processing stage of robot audition applications on low-level audio data, the following applications of robot audition will strongly benefit from the enhanced audio signals explicitly:

- sound source localization
- sound source tracking
- sound source separation
- automatic speech recognition
- pitch extraction
- music recognition
- sound classification and identification
- computational auditory scene analysis
- sound/speech-based human-robot interaction

The implicit influence of ego noise reduction can be observed as improvement of performance in the following subjects:

- active audition (multi-modal integration of motion and audio)
- sound map generation/sound-based Simultaneous Localization And Mapping (SLAM)

This study on estimation and cancellation of ego noise can be applied to several fields in robotics besides personal robotics. Underwater robots are equipped with Sound Navigation And Ranging (SONAR) sensors that use sound propagation to navigate,
CHAPTER 1. INTRODUCTION

communicate with each other or detect other objects. Ego noise suppression will enhance the quality of the sensory data, while allowing the robots to move without stopping their motion. Security and rescue robots will also benefit from ego noise suppression because they require superior stealth, patrol and monitoring capabilities during motion. A robot audition system enhanced with ego noise estimation technique may not be limited to the robotics area, but it can be applied also to any field where motors are used to generate any type of motion, such as automobiles, airplanes or any other vehicle powered by engines. Therefore, the studies on ego noise estimation contribute to the achievement of user-friendly man machine interfaces.

1.2 Goal and Technical Issues

The objective of our research is to develop an ego noise estimation method for a robot. The basic problem with ego noise, like all types of noise in a robot audition system, is that it causes the Signal-to-Noise Ratio (SNR) to drop. In addition, it contaminates the spectrum of the recorded signal so that it is almost impossible to perform the fundamental applications of robot audition, such as Sound Source Localization (SSL), Sound Source Separation (SSS) and ASR, accurately. Therefore, the ego noise must be first estimated and then eliminated prior to those applications in the preprocessing steps. This section introduces four technical issues that challenge us in achieving our goal:

Issue 1: Modeling the process of ego noise estimation

In a typical environment, robots have to deal with various noise types such as diffuse background noise, directional interfering sound sources and reverberation. However, it is not possible to categorize ego noise distinctly into any of those types of noise. The first reason is that the overall ego noise consists of the contributions of various noise sources, such as fan and hardware noise, which are rather stationary and diffuse, and motor noise, which is rather non-stationary and directional. In addition, all these noise sources are embodied inside a shielding cover, which causes effects like multi-path propagation, dispersion, dissipation and diffusion, which are not negligible in reality. Moreover, the intensively-studied signal processing techniques, which are only valid for far-field sound sources (approximately 1.0-1.5 meters away from the robot), cannot be applied for the direct suppression of the ego noise perfectly because all these sound sources are located in the near-field of the microphones. The near-field models, on the other hand, can usually only deal with
stationary sound sources, thus it requires very sophisticated and computationally expensive models to localize and suppress the moving sound sources (motors) for a robot with a large number of DoFs. Therefore, we need a specific ego noise model and an estimation technique that can deal with these difficulties.

In order to address the modeling of this complicated type of noise and its estimation technique, we have to resolve three important issues also involved in machine-based modeling, identification or recognition tasks: Knowledge gathering, representation and algorithm.

Issue 2: Offline constraints for template-based ego noise estimation

Issue 2.1: Changing environmental noise conditions

Since the robots are intended to be deployed to realistic environments, the ego noise will be practically inseparable from the background noise. During a training session, the original templates represent both ego noise and ambient noise. Although we assume that the stationary portion of the ego noise concerning the static noise sources, i.e., fan and hardware noise, does not change over time, the background noise changes based on the acoustic conditions of the environment at any moment, e.g., the robot can move closer to the sound sources, noise emitting devices can be turned on/off, or the robot can move to another room with completely different acoustical environment. Therefore, the robot should clearly distinguish the stationary noise and non-stationary ego-motion noise in terms of its noise estimation processes, but still it should be able to suppress both of them at the same time.

Issue 2.2: Training constraints

The typical problem tightly-coupled to offline training procedures is that they are not adequate to real-world applications due to insufficiency of accurate data about the ego noise templates of unknown motions, i.e. missing templates. Databases created in an offline manner are also mostly not suitable to be used for new data-entries or updates of templates. Furthermore, in case of long training sessions, the data can grow rapidly and expand enormously unless there is an explicit template discarding/update algorithm.

Issue 3: General applicability of ego noise estimation
CHAPTER 1. INTRODUCTION

Ego noise estimates are expected to be used for various purposes. We should confirm the general applicability of the proposed ego noise estimation method on several robot audition applications for eliminating the ego noise.

1.3 Overview of Our Approaches

We deal with the above-mentioned issues through the following approaches:

Approach 1: Template-based ego noise estimation

Most of the audible noise produced by a motor originates from the speed of its rotation/translation. Therefore, we can assume that there is a correlation between the motion data (position/velocity/acceleration) of joints and the audio data (motor noise) implicitly. Because ego noise is a continuous process that is coupled with the continuous motions of the robot, we can easily capture the dynamic nature of the motion-related data represented by a sequence of discrete observations, thus we can solve the knowledge gathering issue. We tackle the representation issue by introducing templates that map the input data streams of audio and motion to a specific ego noise state space, in other words we model each noise state using templates. Based on the instantaneous observations on the input streams, it should be possible to associate a motion command or discrete time series data representing the motion (i.e., the angular status of each joint of the robot) with another series of discrete time data representing the total ego noise spectrum, thereby processing an arbitrary sequence of associated data. This approach relies on seamless synchronization between joint status data (i.e., angular position, velocity and acceleration) and audio data. Algorithm issue for the estimation of templates depends actually on how the first and second issues are solved. We should measure the correlation between the current template and the templates in the model to estimate the current ego noise. Since the data streams create an abundance of templates, we propose to utilize a template database for easy and rapid data storage/retrieval. A non-parametric, instance-based classification technique called the nearest neighborhood search is performed on the database to estimate the templates on the fly.

Approach 2: Design of an online system

Approach 2.1: Unified framework for noise estimation
1.3. OVERVIEW OF OUR APPROACHES

Template-based Estimation (TE) is fairly accurate in reconstructing any type of noise that can be reproduced. Ego noise falls into this noise category because the duration and spectral power of the motor noise signals do not change drastically when the same motion is performed again. The main drawback of this method is that it lacks generalization capability and cannot perform adaptation to ego noise in an environment with changing noise conditions. It can only reproduce the templates that exist in the database, thus it reflects the noise conditions in the training session only. Therefore, we propose to use a stationary noise estimation method based on recursive averaging of the power for certain noise periods, and a non-stationary noise estimation method in series for the generation of the template database, which will separate the noise estimation process into two parallel and independent processes. While recursive averaging takes care of the stationary noise (incl. background, fan and hardware noise) in this proposed scheme, template estimation tackles the remaining non-stationary noise portion belonging to the ego-motion noise only.

Approach 2.2: Incremental learning of the templates

To cope with real-world constraints, we extend the proposed method to an online and adaptive ego noise estimation method, which can interpolate the missing templates. We design a learning mechanism provided with a measure of performance, which enables the learning to continue over the entire lifespan of a robot without human intervention (also termed lifelong learning [3]). This incremental learning algorithm is a natural extension of our basic template estimation method and allows the robot to learn its ego noise not only in isolated training conditions, but also in daily environments even in the presence of humans. Moreover, it allows us to ease the curse of dimensionality problem caused by the large number of DoF of the robot.

Approach 3: Guidelines for applying the ego noise estimation method to robot audition

To evaluate the proposed ego noise estimation method, we create frameworks for various applications using the ego noise estimation as the core concept. The conducted experiments help us to evaluate the effectiveness of the proposed method for robot audition applications, such as noise reduction, automatic speech recognition and sound source localization.
1.4 Thesis Organization

This thesis consists of nine chapters. The organization of this thesis is shown in Fig. 1.3

Chapter 2 covers a survey of the research related to robot audition. We present the existing work on acoustical applications for robot audition, especially from signal processing point of view. Particularly, diffuse noise cancellation methods, directional noise separation techniques and automatic speech recognition are of special interest to us. After describing the related work about ego noise estimation and suppression, we finally discuss the role of our research by emphasizing its position towards the related work.

Chapter 3 describes a methodological framework for tackling the ego noise problem in a systematic way. We explain how to approach the modeling of ego noise estimation process. Later on, we elaborate the design of our estimation method called parameterized template estimation and assess its performance. The performance of this method is compared with those of existing single-channel noise estimation methods.

Chapter 4 describes the extensions we proposed to make the template-based estima-
1.4. THESIS ORGANIZATION

In order to cope with changing environmental noise, we modify the general template concept to our needs. We generate the templates in a way that they only represent the non-stationary noise. The stationary portion of the ego noise with ambient noise is dealt using a stationary noise estimation method. We explain the details of this unified framework for noise estimation. Moreover, we eliminate the necessity of an offline training procedure by introducing an incremental template learning scheme that can also create missing templates by interpolating the existing templates. Finally, we evaluate the performance of the proposed methods in terms of estimation quality and noise reduction accuracy by using objective performance criteria and discuss the results.

Chapter 5 delves into the question of how to suppress the whole-body motion noise of a robot more robustly. For this purpose we integrate template-based ego noise estimation with the already established works from the multi-channel noise reduction literature. Microphone array-based sound source separation is adequate to cancel directional noise of distinct motors which are and spatially separated from the speaker, thus the performance of this hybrid noise reduction system exceeds the individual performances of the single-channel and multi-channel noise reduction methods. In this chapter, we discuss the implementation and its evaluation in terms of ASR accuracy.

Chapter 6 describes Missing Feature Theory (MFT)-based integration of ego noise reduction and ASR. We focus on two different ASR systems: single-talker ASR and multi-talker ASR. Both systems rely on the single-channel and multi-channel noise reduction methods to generate spectro-temporal masks filtering the unreliable acoustic features. We present detailed results regarding recognition accuracy to determine optimal parameters of the mask generation process.

In Chapter 7, we provide an extended version of the parameterized template estimation to operate on multi-channel audio data. This feature enables a Sound Source Localization (SSL) scheme to whiten the ego noise allowing the elimination of its interfering effect on SSL in the spatio-temporal plane. We assess the performance in terms of localization accuracy and peak detection rates.

Chapter 8 outlines the contributions of this thesis and gives an insight to the remaining issues and future work.

Chapter 9 summarizes and concludes this dissertation.
Chapter 2

Literature Review and Basic Methods

A mathematical framework is required in order to describe the signal model used in robot audition and to perform signal processing. The mathematical framework typically comprises a model together with a set of a priori assumptions that describe the events happening in the physical domain. This chapter starts with an introduction into a general model of how acoustic sound waves are propagated and received by a set of microphones. This signal model is used as a common ground for all parts of the thesis. Having a signal model allows us to describe various types of noise with different characteristics.

The ultimate aim of noise suppression systems is to reduce the distorting components while keeping the useful signal as untouched as possible. Therefore, the theory of noise reduction also implies speech or sound enhancement. The speech enhancement techniques can be classified according to various criteria related to the objective of the application. A popular classification of the speech enhancement techniques is made whether the system is based on a single microphone (also referred as channel) or on multi-microphones. The main discrimination in this thesis, however, is done based on the types of noise they tackle, such as the single-channel techniques against diffuse noise, multi-channel techniques against directional noise, techniques to deal with the artifacts of noise reduction and finally, ego noise reduction techniques. In theory, this classification also implies a classification made depending on the information of a specific type of noise, such as the class of distribution or the spatial and/or temporal information which can be employed in statistical models.
Most speech enhancement techniques were originally designed for speech intelligibility rather than for ASR [4]. Whilst there have been numerous works where enhancement techniques are presented as an efficient preprocessing stage for robust speech recognition, some enhancement techniques also distort the speech signal in ways which can cause ASR performance to decrease. Any solutions derived from subjective criteria where humans are used as test subjects are therefore typically sub-optimal for ASR. Redesigning the existing enhancement techniques to optimize them for ASR and SSL under ego noise of a robot is one of the focal points of this dissertation.

2.1 Preliminaries

Throughout this thesis, we assume that the acquired signals are electrical representations of physical quantities. A microphone essentially measures variations in the air pressure level and translates them into a representative electrical voltage. Microphones are therefore used to sense acoustic sound. It is also assumed that a continuous time electrical sensor signal has been sampled with the sampling frequency $f_s$ [Hz] and that it is being represented by a discrete-time signal.

2.1.1 Setting of Microphones

The microphones can be mounted in different microphone array configurations. They can be positioned on a line, in a planar or even a 3-dimensional ordering. In general, the microphone array configuration has an influence on the performance of the multi-microphone signal enhancement algorithms. Microphones are generally considered to be perfect point sensors with ideal omni-directional properties. This is quite unrealistic, since generally the microphones also perform a spatial and a spectral filtering operation, but we neglect this effect in this dissertation.

2.1.2 Acoustic Signal Model

We use the following definitions to accompany our signal model:

- A source refers to an entity that generates an acoustic sound different from any other sound source from either temporal, spatial, or statistical point of view.
A clean source signal refers to the recorded signal that has been produced by a source in a non-dispersive, lossless, and homogeneous medium where no other sounds are present.

A received source signal of any spatial source carries spatial information that is unique from other spatial sources. (Assumption of uncorrelated sound sources)

Two main source classes are acknowledged in this work: desired sources and disturbing sources.

This final class of disturbing, unwanted signals that interfere with the desired signals can also be defined as noise. A noise itself is a signal that conveys information regarding the source of the noise. For example, the noise from an engine conveys information regarding the state of the engine. The possible causes of the noise are the properties of the source, such as motion, vibration, and collision. The success of a noise processing method depends on its ability to characterize and model the noise process, and to use the noise characteristics advantageously to differentiate the signal from the noise, and even a type of noise from another type of noise. For example, some of the noise sources are stationary (e.g. fans) and have a slowly varying spectral content, whereas other noise sources can be highly non-stationary (e.g. radio, alarm clock). The most difficult problem arises when the noise sources are also speech signals (e.g. concurrent speakers), which are similar in structure to the desired signal.

It is assumed that a number of $M$ microphones are being used. An electrical microphone signal is denoted as $y_m(t)$, where $t$ denotes a sample index, and the index $m$, $m \leq M$ is used to refer to the $m^{th}$ microphone signal. If $M = 1$, a short notation is used: $y(t) \equiv y_1(t)$. The microphone signal $y_m(t)$ comprises a superposition of several signal components such as, desired signals and disturbing signals. Let $s(t)$ denote the desired source signal at time sample $t$, let $a_m(t)$ represent the room impulse response of the $m^{th}$ sensor to the desired source, and let $n_m(t)$ denote the noise component at the $m^{th}$ sensor. The observed signal at the $m^{th}$ sensor ($m = 1, \ldots, M$) is given by

$$y_m(t) = x_m(t) + n_m(t)$$

$$= a_m(t) * s(t) + n_m(t)$$

$$= a_m(t) * s(t) + n_m^s(t) + n_m^n(t),$$
where \( x_m(t) \), \( n^s_m(t) \) and \( n^n_m(t) \) represent the desired signal, stationary and non-stationary noise components at the \( m^{th} \) sensor, respectively, and \( * \) denotes convolution. We assume that both noise components may be coherent (directional) noise component and diffuse noise component [5] (see Section 2.1.3 for details). It is acknowledged that the propagation model in Equation (2.3) is restricted to linear propagation channels since it only captures linear dynamics of the wave equation [6]. Nonlinear acoustic or electric effects that can arise in the wave propagation or the signal acquisition, e.g., due to a dispersive (multi-path) propagation [5] or sensor signal saturation, are not modeled here. Figure 2.1 shows our signal model for a circular type of array configuration with eight microphones.

![Signal model with a single sound source](image)

Figure 2.1: Signal model with a single sound source

The observed signals are divided in time into overlapping frames by the application of a window function and analyzed using the Short-Time Fourier Transform (STFT). The complex input spectrum \( Y(k, l) \) of frequency bin index \( k \) and time frame index \( l \) is obtained from:

\[
Y(k, l) = \sum_{t=0}^{t=W-1} y(t + lS)w(t) \exp\{j(2\pi/W)tk\},
\]

(2.4)

where \( W \) is the window length, \( S \) is the shift length and \( w(t) \) is the window function. Assuming time-invariant transfer functions, we have in the time-frequency domain
\[ Y_m(k, l) = X_m(k, l) + N_m(k, l) \] (2.5)
\[ = A_m(k) S(k, l) + N_m(k, l) \] (2.6)
\[ = A_m(k) S(k, l) + N_m^s(k, l) + N_m^n(k, l), \] (2.7)

where \( k = 1, 2, \ldots, K \) represents the frequency bin index, \( Y_m(k, l), X_m(k, l), S(k, l), N_m(k, l), N_m^s(k, l), N_m^n(k, l) \) are the STFT of the respective signals \( y_m(t), x_m(t), s(t), n_m(t), n_m^s(t), n_m^n(t) \), and \( A_m(k) \) is the acoustical transfer function relating the speech source with the \( m^{th} \) sensor.

### 2.1.3 Sound Fields

Wave theory describes an acoustic sound as traveling local variations in air pressure levels. Sound waves propagate through an air medium by producing a movement of the molecules in the direction of propagation. In general, waves propagate from their source as spherical waves, with the amplitude decaying at a rate proportional to the distance from the source [7]. These properties imply a rather complex mathematical analysis of propagating signals, which is a major issue in array processing of signals present in the near-field of the microphones. The typical rule of thumb is that far-field assumptions are only valid [8] when

\[ r \geq \frac{d_{tot}^2 f_{sig}}{c}, \] (2.8)

with \( r \) the radial distance of the source to the centre of the microphone array, \( d_{tot} \) the total length (aperture) of the microphone array, \( f_{sig} \) the signal frequency and \( c \) the speed of sound (\( c = 340 \, \text{m/s} \)). For example, when \( d_{tot} = 0.2 \, \text{m} \) and \( f_{sig} = 8 \, \text{kHz} \), the minimum source distance for the far-field assumption to be valid is \( r = 0.94 \, \text{m} \) (typical values for a humanoid robot). In far-field situations plane wave propagation is assumed, thus signal attenuation can be assumed to be equal for all microphones. In a near-field situation, spherical wave propagation and signal attenuation have to be taken into account, which is a rather tedious task for the design of an algorithm.

The acoustic field in the absence of information transmission is commonly referred to as a noise field. In general, it consists of the summation of unwanted or disturbing acoustic waves introduced into a common field by man-made and natural sources. Hence,
2.1. PRELIMINARIES

depending on the degree of correlation between noise signals at distinct spatial locations, two categories of noise fields can be defined for microphone array applications [6]: (1) diffuse noise field and (2) directional noise field

A diffuse (incoherent) noise field is characterized by spatially uncorrelated noise signals. A perfectly diffuse sound field is typically generated in an unbounded medium by distant, uncorrelated sources of random noise evenly distributed over all directions [9]. A directional (coherent) noise field, on the other hand, corresponds to noise signals propagating from their source without reflection, dispersion or dissipation. It is characterized by a high correlation between received signals at different spatial locations. A directional noise field results from a source in open air environments with no major obstacles to sound propagation.

2.1.4 Noise Problem in Robot Audition

The auditory perception system of a robot is essential to obtain knowledge from sound sources. ASR is one of the most natural extensions for a human-robot interface and one of the main targets of robot audition. As one of the earliest works, Nakadai et al. proposed Robot Audition, which realizes recognition of noisy speech such as simultaneous speech by using robot-embedded microphones [1]. There are also many studies for speech recognition systems in robots trying to mimic human auditory mechanisms [10, 11, 12]. However, a robot can extract reliable auditory information only if the acquired speech signals are not drastically corrupted by noise. In the robotics literature, there are numerous works utilizing close-talk headset microphones to achieve noise-robust human-robot interaction in auditory domain [13, 14]. However, those headsets limit and degrade the naturalness of human-robot interaction.

The main difficulty of acoustical applications in the area of robotics is that the acoustical domain of the robot is fairly complex. The acoustical conditions differ from environment to environment substantially. Even if the same types of noise sources are present in the environment, the auditory scene can change with time. Besides the properties of the environment, the remote distance between target sound source and the microphones of the robot poses another significant problem. If the signal of interest is farther away from the robot or the noise sources are closer to the robot, the robot suffers lower SNRs. It is also well known that the performance of sound source localization techniques [15, 16] and ASR systems [5, 17] rapidly degrades with decreasing SNR.

The target sound signal can be corrupted due to the presence of not only external
factors such as ambient noise, interfering sounds, and room reverberation, but also internal factors such as synthesized voice of the robot and ego noise. It is apparent that the internal noise sources, by definition, are the closest ones to the microphone array, therefore they require special attention. The modeling and removal of the effects of noise has been at the core of the theory and practice of signal processing. Figure 2.2 illustrates a typical robot audition scenario. In the rest of this section, we study the characteristics and modeling of several forms of noise except reverberation and self-voice. Dereverberation, which aims to extract the clean speech signal $s(t)$ from $y(t)$ without any prior knowledge about the acoustic impulse response $a(t)$ as in Equation (2.3). Because dereverberation techniques are not explored in this dissertation, readers are encouraged to refer to the cited publications for more technical details about the popular techniques such as inverse filtering [18, 19], matched filtering [20], and cepstrum-based techniques [18, 21, 22]. Self-voice cancellation is extensively studied in [23].

![Figure 2.2: A typical robot audition environment](image)

**Diffuse Noise**

In urban environments, diffuse noise (background noise) is omnipresent and generated by car traffic, engines, fans, audio equipment and background music in public places. Most of the acoustic disturbances in an in-door environment stem from air conditions, fans and computer noise, and they appear to arrive from all directions simultaneously with equal energy and low spatial correlation, giving an impression of surrounding noise sources. In practice, many indoor environments can be characterized by a diffuse noise
field, to some extent. In this thesis, diffuse-like noise was generated in a room by using the Heating, Ventilation, Air-Conditioning (HVAC) system, computers, power supply of the robot, which all emit uncorrelated noise. Background noise can be tackled using the methods described in Section 2.2.

**Directional Noise**

In public places such as schools, restaurants, as well as offices and other indoor environments, the desired speech may be corrupted by human babble noise from neighboring speakers also referred to as *cocktail party noise*. Other localized interferences may also disturb or alter the robot audition, such as alarm sounds, radio sounds, loudspeakers or the sound from a musical instrument. Furthermore, the noise sources can be smallband (e.g. siren) or wideband, intermittent or persistent, and they may have the same spectral characteristics and/or angle of arrival as the desired speech signal. In comparison to diffuse noise, these interfering directional sound signals are generated by spatially constrained sound sources. Mostly, the spatial separation between the desired source and the acoustic interference sources are exploited for speech enhancement, as well as signal separation, using multi-microphone-based systems as explained in Section 2.3.

**Noise Artifacts for Speech Recognition**

Many ASR applications in the presence of diffuse noise and directional interfering speakers are introduced for robot audition so far. For example, Hara *et al.* reported a humanoid robot, HRP-2, which uses a microphone array to localize and separate a mixture of sounds and is capable of recognizing speech commands in a noisy environment [24]. Nakadai *et al.* reported SIG, a humanoid robot which uses a pair of microphones to separate multiple speech signals through an active direction-pass filter, and recognizes each separated speech phrase [25]. Valin *et al.* developed a robot audition system with eight microphones to perform speaker tracking and separation [26, 27]. All these above-mentioned systems are examples of robot audition frameworks, which suffered from a mismatch between noise reduction and ASR systems. The so called *noise artifacts* present in the acoustic features, such as leakage noise between the simultaneous speech signals, and musical noise caused by over/under-subtraction of background noise, must be also eliminated. Therefore, the integration between the preprocessing and ASR is also a noise-related concern of robot audition, which will be discussed in Section 2.4.
CHAPTER 2. LITERATURE REVIEW AND BASIC METHODS

Ego Noise

Components of the Ego Noise: The main components of the ego noise are shown in Figure 2.3. Hardware noise comes mainly from the fans that are located throughout the interior of the robot and help to dissipate the large amount of heat generated by the CPU, the power supply and other components. There can be multiple fans positioned on multiple locations on the robot, which generate diffuse and stationary type of noise. The static (steady-state) hardware noise can be removed by applying spectral filtering operations, e.g. low pass filter, notch filter.

![Diagram of a typical robot with ego noise sources]

In contrast, ego-motion noise or motor noise is emitted only when the robot is performing an action using its motors. The audible noise produced by a motor originates from its stator core laminations [28]. The stator core is made of thin laminated metal sheets. When a motor is powered from an adjustable frequency drive using a Pulse Width Modulated (PWM) output waveform, a magnetic flux is induced in the stator core. This magnetic flux causes the stator to vibrate at the carrier frequency changing the pitch of the audible noise. Whether the actual power level of the noise is increased due to a PWM waveform will depend upon the level of the applied excitation voltage. The voltage and the duration of the pulses determine where the motor positions itself and how fast it rotates/translations. Because position/velocity range of a motor has a wide spectrum and different sets of motors operate during various tasks, the ego-motion noise is highly non-stationary.

Figure 2.4 depicts typical spectro-temporal representations of the sounds in a robot environment. The first two seconds is the background noise emitted by the environment.
2.1. PRELIMINARIES

of the room. Data between 2-4 seconds show the fan (hardware) noise, which is considerably lower compared to the fan noise. Their energy is mostly located in the lower frequency bands, and the peaks at certain frequencies indicate operating frequencies of some hardware (e.g. at 3kHz). On the other hand, the spectrogram between 4-6 seconds shows the frequency response of total ego noise where the robot moves its body. We see that the motor noise has a significant proportion of energy in the higher frequencies as well.

Comparison of Ego Noise To Speech Signals: Speech is also a wideband signal similar to ego noise, with frequency components ranging from 100 to 8000Hz. For voiced sounds, very little energy is present above 4 kHz. For unvoiced sounds however, the spectrum is much flatter. The frequencies between 300Hz and 3400Hz are especially important to understand speech. Although, a sampling rate of 8kHz is usually sufficient to obtain an acceptable speech quality, higher sampling rates (e.g. 16kHz) are used nowadays because of the demand for higher quality. Speech is also
CHAPTER 2. LITERATURE REVIEW AND BASIC METHODS

a non-stationary signal, with both time envelope and spectrum continuously changing. Short-time stationarity in the order of 10-20ms can be assumed for speech analysis. Furthermore, speech is an intermittent signal, i.e. silences exist between the words and in a typical conversation more than 50% of the time will consist of pauses. This on/off characteristic of speech signals are similar to the active/idle states of ego noise due to a moving/stopping robot. Therefore algorithms exploiting the speech pauses, e.g. VAD, have trouble classifying noise-only periods and speech-and-noise periods. Figure 2.4 shows overlapped speech and ego noise signals between 6-8 seconds. It is not easy to distinguish what portion of the signal belongs to ego noise in such case.

Acoustic Field Properties of the Ego Noise: One major difficulty we face while trying to model the motor noise is its noise field assumption. The motors of the robot are present at all joints of the robot, therefore the noise signals with different power levels come from several directions simultaneously. First of all, all components of ego noise are located in the near-field of the microphones. Because the motors can be considered as point sources, they can be categorized as directional sounds. However, since the motors are surrounded by external shielding or other electro-mechanical components, additional (non-linear) phenomena [29], such as diffraction, diffusion, dissipation, non-linear absorption and temperature-dependent effects are present, which mitigates the directional characteristics of the motor noise and causes diffuse-like characteristics. It is also noteworthy to mention about the reverberation properties of ego noise. Reverberation is caused by the fact that acoustic waves are reflected by the body covers or obstacle-like electro-mechanical components, such that the signals recorded by the microphone array consist of a direct path signal and multiple delayed/attenuated versions. However, we assume that due to the short distance between the motors and microphone array, the reverberation time of ego noise inside the body of the robot is shorter than one processing frame (< 10 ms) and thus negligible.

Spectro-temporal Characteristics of the Ego Noise: To model ego noise accurately, we need a structure for characterizing the patterns in the noise process using its temporal and spectral characteristics. Accurate modeling of noise statistics is the key to high-quality noisy signal classification and enhancement. Even a simple task of signal/noise classification is crucially dependent on the availability of good signal and noise models. One of the most useful tools for gaining insight into the structure of a noise process is the use of Fourier transform for frequency analysis.

21
The upper panels of Figure 2.5 and Figure 2.6 illustrate the noise signals recorded during a motion of the head and arm, respectively, whereas the lower panels show the instantaneous state of each motor regarding *angular position, velocity and acceleration* involved in the same motion. Note that the motor state are normalized to $[0, 1]$, e.g., 0 means minimum position, maximum velocity/acceleration in negative direction and 1 means maximum position, maximum velocity/acceleration in positive direction for the allowable trajectory of each individual motor. Some important findings are like following:

- The spectrum of the motion noise shown in Figure 2.5 reveals that most of the noise energy is concentrated in the lower-frequency part of the spectrum. In fact, most audio signals and noise have a predominantly low-frequency spectrum. However, it must be noted that the relatively “lower-energy, high-frequency” part of audio signals plays an important part in conveying sensation and quality, therefore the effect of the noise energy in higher frequencies must be also taken seriously.

- The red arrows between the panels in Figure 2.5 indicate the parts where the noise levels are high due to the increased velocities and especially accelerations, whereas the blue arrows show the parts of the spectrum with low energy in both high and low frequencies due to the stopped motion of the motors. In the latter case, only hardware and fan noise are active. By all means, there is a high correlation between the emitted ego noise and velocity/acceleration of each motor.

**Observation 1**

The regions of the spectrum where the noise power is densely distributed, correspond to the increased rotational velocities/accelerations of the motors involved in the motion.

Repeated motions for the same trajectories yield the spectra and motor noise states in Figure 2.7 and Figure 2.8. We observe that although the same commands for motion are sent to the controller, the generated trajectories and motion patterns with respect to instantaneous states of the motors are not exactly the same. Therefore, it is impossible to replicate a motion to be identical. Nonetheless, the spectra presented in both Figure 2.7 and Figure 2.8 are consistently similar in terms of repeated energy patterns. The exact moments of noise excitation in each repetition coincide based on the corresponding motor states. This implies that a time-series of motor states can easily suffer from small
CHAPTER 2. LITERATURE REVIEW AND BASIC METHODS

Figure 2.5: The relationship between the state of motors and head motion noise

Figure 2.6: The relationship between the state of motors and arm motion noise

23
deviations either in time or space, thus the noise energy may vary or the spectrum may shift in time.

Figure 2.7: The relationship between the state of motors and head motion noise for a motion repeated three times

Observation 2

The duration and the spectral pattern of ego noise signals are similar for repetitions of the same motion.

2.2 Existing Work on Diffuse Noise Reduction

Diffuse noise reduction techniques have attracted a great deal of interest in the last few decades [30]. They can be broadly classified in parametric and non-parametric techniques. Parametric techniques model the noisy speech signal as a stochastic Auto Regressive (AR) model embedded in Gaussian noise [31]. Speech enhancement then roughly consists of estimating the speech AR parameters and applying a (non-causal) Wiener filter [32] or Kalman filter [33] to the noisy signal, where the optimal filters are based on the estimated AR parameters. Non-parametric techniques do not estimate the speech parameters, but require a noise fingerprint in a transform domain,
CHAPTER 2. LITERATURE REVIEW AND BASIC METHODS

such as the Short-Time Fourier Transform (STFT) or the Karhunen-Loeve Transform (KLT) domain. This noise fingerprint is estimated during noise-only periods and used during subsequent speech-and-noise periods in order to obtain an estimate of the clean speech signal. Well-known non-parametric techniques include spectral subtraction [34, 35, 36, 37, 38, 39, 40, 41, 42], signal subspace-based techniques [43, 44, 45] and model-based compensation [46, 47]. By exploiting the a priori knowledge that the signal to be recovered is speech, methods can be developed to extract signals of a certain class of distribution. For example, background noise commonly displays a nearly Gaussian distribution, while speech is characterized by a Laplacian distribution [48].

Diffuse noise can be treated by using a single microphone. Figure 2.9 shows the general configuration for such a single-channel noise reduction. The basic idea is to use the STFT of the noisy speech input and recover an estimate of the STFT of the clean speech by removing the contribution of the additive noise. The input to the system is the noise-corrupted signal \( y(t) \). While there are many methods for the analysis-synthesis processing, the STFT of the signal with OverLap and Add (OLA) [18] is the most commonly used method. For restoration of time-domain speech signals, an estimate of the instantaneous spectrum of speech, \( \hat{X}(k, l) \) is transformed via an inverse discrete
Fourier transform to the time domain. We divide the intermediate noise reduction process into three successive processes: (1) noise estimation, (2) gain calculation and (3) spectral filter. Section 2.2.1 outlines the (1) noise estimation methods we used in our thesis, whereas Section 2.2.2 reviews the details about the (2) gain calculation and (3) spectral filter methods.

\[ y(t) = x(t) + n(t) \]

\[ |Y(k, l)|^2 = |\hat{N}(k, l)|^2 \]

2.2.1 Single-channel Noise Estimation Methods

Many speech enhancement systems [36, 45] have been developed to compute the estimates of the power spectrum of clean speech using the power spectrum of noisy speech and noise. Noise estimation techniques therefore play an important role in the effectiveness of these enhancement algorithms. For example, if the noise estimate is too low, the resulting clean speech estimate will still contain significant levels of noise. Alternatively, if the noise estimate is too high, the speech may be distorted. Another important consideration for applications in robot audition is their ability to track continually changing noise conditions.

In single-channel speech enhancement systems there is access only to noisy speech and hence the noise statistics have to be estimated from the noisy speech itself. Usually the noise spectrum estimate is obtained from the first few milli-seconds of noisy speech which are silence regions. This assumption is valid for the case of stationary noise in which the noise spectrum does not vary much over time. Traditional Voice Activity Detection (VAD) technologies [49, 50] also track the noise-only (speech absent) frames of the noisy speech to update the noise estimate. This is not enough in the case of non-stationary noise in which the power spectrum of noise varies even during speech activity. Hence there is a need to update the noise spectrum continuously over time and
this is done by noise estimation algorithms [51, 52]. Most of these algorithms can be classified broadly into two classes. The first class is based on updating the noise estimate by tracking the silence regions of speech (e.g., Minima Controlled Recursive Averaging) and the other class is based on updating noise estimate using the histogram of the noisy speech power spectrum (e.g., Histogram-based Recursive Level Estimation).

Minima Controlled Recursive Averaging (MCRA)

MCRA [53] calculates the estimated noise power as the averaged input power of noise periods detected by level-based VAD. First, MCRA calculates smoothed input power spectrum \( S(k, l) \) for level-based VAD as

\[
S_f(k, l) = \sum_{u=-w}^{w} b(u)|Y(k-u, l)|^2, \tag{2.9}
\]

\[
S(k, l) = \alpha_s S(k, l-1) + (1 - \alpha_s) S_f(k, l), \tag{2.10}
\]

where \( \alpha_s, w \) and \( b(u) \) are all smoothing parameters. With \( S(k, l) \), MCRA calculates the minimum noise spectrum \( S_{\min}(k, l) \) using the Minimum Tracking method [39], where \( S_{\min}(k, l) \) is updated for every \( L \) frames. The voice activity flag \( I(k, l) \), which equals 0 for noise periods and 1 for speech periods, is decided as

\[
I(k, l) = \begin{cases} 
1 & S_f(k, l) > \delta S_{\min}(k, l) \\
0 & \text{otherwise}, 
\end{cases} \tag{2.11}
\]

where \( \delta \) is a threshold parameter. The final estimated noise power \( \lambda_{MCRA}(k, l) \) is calculated as

\[
\lambda_{MCRA}(k, l + 1) = \lambda_{MCRA}(k, l)p(k, l) + [\alpha_d \lambda_{MCRA}(k, l) \\
+ (1 - \alpha_d)|Y(k, l)|^2(1 - p(k, l))], \tag{2.12}
\]

where \( \alpha_d \) is a smoothing parameter and \( p(k, l) \) shows the voice active probability. This \( p(k, l) \) is derived as

\[
p(k, l) = \alpha_p p(k, l-) + (1 - \alpha_p) I(k, l - 1), \tag{2.13}
\]

where \( \alpha_p \) is a smoothing parameter.
2.2. EXISTING WORK ON DIFFUSE NOISE REDUCTION

MCRA works well when the parameters are optimally adjusted to the acoustic conditions. However, this adjustment is especially difficult in non-stationary environments because the optimum parameters also change accordingly.

Histogram-based Recursive Level Estimation (HRLE)

HRLE [54] estimates input noise levels as an “x” percentile value $L_x$ value from an input power level histogram. Since HRLE uses recursive averages to obtain temporal histograms, HRLE can adapt smoothly and quickly to the environmental changes. The estimated noise spectrum $\lambda_{HRLE}$ is obtained as:

$$Y_L(k, l) = 20 \log_{10} |Y(k, l)|,$$  \hfill (2.14)

$$I_y(k, l) = \lfloor (Y_L(k, l) - L_{min})/L_{step} \rfloor,$$  \hfill (2.15)

$$N(k, l, i) = \alpha N(k, l - 1, i) + (1 - \alpha) \delta(i - I_y(k, l)),$$  \hfill (2.16)

$$S(k, l, i) = \sum_{j=0}^{i} N(k, l, j),$$  \hfill (2.17)

$$I_x(k, l) = \arg\min_l \left[ S(k, l, I_{max}) \frac{x}{100} - S(k, l, I) \right],$$  \hfill (2.18)

$$\lambda_{HRLE}(k, l) = L_{min} + L_{step} \cdot I_x(k, l).$$  \hfill (2.19)

$L_{min}$, $L_{step}$, and $I_{max}$ are the minimum level, the level width of 1 bin and the maximum index of the histogram, respectively; $x$ indicates the percentile position, $\alpha$ is the time decay parameter calculated from time constant $T_r$ and sampling frequency $F_s$ as $\alpha = 1 - 1/(T_r F_s)$, $\delta(t)$ shows the Dirac delta function and $\lfloor \cdot \rfloor$ is the flooring function. The parameters $L_{min}$, $L_{step}$ and $I_{max}$ do not affect the estimated results, as long as proper values are set to cover the input level range with few errors. On the other hand, $x$ and $\alpha$ are the main parameters that influence the estimated level and both are SNR-independent. Although, HRLE does not require fine-tuning of parameters according to the environmental SNR, $x$ value is of utmost importance to determine how aggressively the noise is estimated. Higher $x$ values are more appropriate to estimate non-stationary noise, whereas HRLE with lower $x$ values can capture only stationary noise.
CHAPTER 2. LITERATURE REVIEW AND BASIC METHODS

2.2.2 Gain Calculation and Spectral Filter

As shown in Figure 2.9, the spectral power $|Y(k, l)|^2$ of the noisy input signal $y(t)$ and the resulting $\lambda(k, l) = |\hat{N}(k, l)|^2$ from Section 2.2.1 are used to calculate the gain $G(k, l)$. The equation for computing $G(k, l)$ is derived based on the reduction method and is explained later in this section. The gain is used to modify the noisy signal $Y(k, l)$, also called spectral filter, as follows:

$$\hat{X}(k, l) = G(k, l)Y(k, l). \quad (2.20)$$

Since the noise is assumed to be uncorrelated with the speech signal, the weighted signal $X(k, l)$ is considered as an estimate $\hat{X}(k)$ of the original clean speech signal $x(t)$. The unprocessed phase of the noisy input signal $y(t)$ is used to synthesize the enhanced speech signal under the assumption that the human ear is not able to perceive the distortions in the phase of the speech signal for perceptual concerns and that the phase information is not important in the ASR process.

Principles of Spectral Subtraction

Spectral Subtraction (SS) is a well-known method used to compute $G(k, l)$. The basic power spectral subtraction technique, as proposed by Boll [34], is popular due to its simple concept and its effectiveness in enhancing speech degraded by additive noise. The basis for spectral subtraction is the assumption that the noise and speech signals are statistically independent [55]. In that case, noise can be regarded as being added acoustically to the clean speech signal as in Equation (2.6) and can be subtracted likewise.

$$|\hat{X}(k, l)|^\gamma = |Y(k, l)|^\gamma - |\hat{N}(k, l)|^\gamma, \quad (2.21)$$

where $\gamma$ is the order of the SS. The subtraction typically takes place in the magnitude ($\gamma = 1$) or power ($\gamma = 2$) spectrum, but it may also operate on filterbank energies.

Due to the errors in the estimated noise, we may have some negative values in the modified spectrum. These values are set to zero and this process is called half-wave rectification. With half-wave rectification the modified spectrum can be written as:

$$|\hat{X}(k, l)|^\gamma = \begin{cases} |\hat{X}(k, l)|^\gamma & |\hat{X}(k, l)|^\gamma > 0 \\ 0 & \text{otherwise}, \end{cases} \quad (2.22)$$

The noise suppression is implemented as a time-varying gain calculation and spectral
2.2. EXISTING WORK ON DIFFUSE NOISE REDUCTION

filter process by rewriting the spectral subtraction method as in Equation (2.20), where \( G(k, l) \) is a gain function represented by Equation (2.23):

\[
G(k, l) = \left[ \max \left( 1 - \frac{|\hat{N}(k, l)|^\gamma}{|Y(k, l)|^\gamma}, 0 \right) \right]^{\frac{1}{\gamma}}.
\]  

(2.23)

Please note that Figure 2.9 depicts a configuration, where \( \gamma \) is equal to 2.

Modifications to Spectral Subtraction

The spectral subtraction method is easily implemented, effectively reduces the noise present in the corrupted signal, and computationally convenient [34, 56]. However, single-microphone speech enhancement techniques generally have problems to reduce the background noise without introducing noticeable artifacts called musical noise [57] or they cause speech distortion because the speech and the noise signals usually occupy overlapping frequency bands. There are some significant variations between the estimated noise spectrum and the actual noise content present in the instantaneous speech spectrum. The subtraction of these quantities results in generating isolated residual noise levels of large variance, producing a metallic noise sounding like the sum of tone generators with random fundamental frequencies which are turned on and off constantly [34]. The processing distortion becomes more noticeable as the signal-to-noise ratio decreases. Also, it has been shown that intelligibility is decreased at low input SNR levels (below 5 dB SNR) [5] and with rapid variations in noise characteristics.

Using only half-wave rectification it is not possible to remove this noise completely, without compromising the quality of the enhanced speech. An aggressiveness factor or overestimation factor, \( \alpha \), can be introduced allowing a compromise between perceptual signal distortion and noise reduction level. Moreover, the elimination of negative estimates of attenuations are avoided using a maximum operation with a parameter called spectral floor, \( \beta \). In practice, applying a spectral floor substitutes the usage of half-wave rectification and the final extended formula for gain calculation can be given such as follows:

\[
G(k, l) = \left[ \max \left( 1 - \frac{|\hat{N}(k, l)|^\gamma}{|Y(k, l)|^\gamma}, \beta \right) \right]^{\frac{1}{\gamma}}.
\]  

(2.24)

Spectral floor parameter reduces the effect of the musical noise, but introduces a
2.3 Existing Work on Directional Noise Reduction

Single-channel speech enhancement technologies against diffuse noise are restricted to the available and estimated spectro-temporal information that is provided by the input signal. If this information is insufficient, e.g., if the SNR is too low, the single-channel noise reduction methods are destined to fail. A method that is able to also utilize the spatial domain has, therefore, the possibility of further improving the speech enhancement performance. In order to operate in the spatial domain, several microphones (i.e., a microphone array) are required to be used. Speech enhancers that are based on microphone arrays provide a high degree of noise and interference reduction due to the spatial selectivity. In addition, it is possible, using a microphone array, to construct a distortionless, linear speech enhancement method, i.e., a speech enhancer that reduces disturbing signals without compromising the target speech signals. This is not possible in single-channel methods that distort the speech signals during the non-linear noise reduction process.

A large diversity of multi-microphone-based algorithms derived from classical signal processing methods can be found in the literature [5, 58]. In the last decade, the separation of acoustic signals has been investigated using BeamForming (BF) methods and Blind Source Separation (BSS) methods. While beamformers exploit the spatial separation of the source signals, most solutions for blind signal separation rely on a contrast in properties of the source signals, such as independence, non-stationarity and non-Gaussianity [58, 59]. In Section 2.3.1 and 2.3.2, common approaches for both methods are briefly described in order to provide a complete review of state-of-the-art speech enhancement techniques. One of the hybrid methods incorporating both methods called Geometric Source Separation (GSS) and two supporting technologies including multi-channel Post-Filter (PF) and SSL are introduced in Section 2.3.3.

2.3.1 Beamforming Techniques

When the speech and the noise sources are physically located at different positions, spatial diversity can be exploited by using multi-microphone noise reduction techniques, such that both spectral and spatial characteristics of the signal sources can be used. If the geometry of the microphones with respect to the target source is known a priori, a
beam can be formed (thus called beamforming), which includes the target source but excludes the noise source. The most basic beamformer is called Delay and Sum (DS) beamformer [60, 61]. This sums up the signals for all of the channels with their own delays so that all of the signals are synchronized to the target sound source. This operation causes cancellation of noise and other signal sources outside the target direction as they are assumed to be uncorrelated in each microphone.

An Adaptive BeamFormer (ABF) also sums up signals with individual delays and gains so that the residual noise may be minimized [62, 63]. When a noise source is directional and the room is non-reverberant, ABF will have a null-beamforming pattern for each noise source. ABF is also known as a Minimum Variance (MV) beamformer [64]. The ABF can achieve higher performance than the DSBF. However, the ABF requires a priori information, e.g., the target direction and speech break interval. These requirements are due to the fact that the conventional ABF is based on supervised adaptive filtering, which significantly limits its applicability of ABF to source separation in practical applications. In addition, the ABF cannot work well when the interfering signal is non-stationary noise.

The Generalized Sidelobe Canceller (GSC) is another widely-used classical beamformer [65]. Other advanced beamforming techniques using post-filtering, optimal filtering or signal subspace concepts have been suggested [66, 67]. Many of these algorithms rely on VADs. This is needed in order to avoid source signal cancellation effects, which may result in unacceptable levels of speech distortion.

### 2.3.2 Blind Source Separation Techniques

BSS techniques aim to distinguish a set of signals which have been mixed with some unknown model [5, 59, 68, 69]. Examples of mixing scenarios include adding background noise to clean speech, or several speakers talking at the same time. In order to perform the separation, it is generally required to have at least the same number of channels as sources to separate. BSS assumes that the signals that have been mixed are uncorrelated, and can therefore be considered statistically independent, but it makes no assumption about the spatial locations of the sources, which is in contrast to the beamforming techniques described in the previous section. BSS also has opened the path to several speech separation algorithms [70, 71].

In particular, BSS based on Independent Component Analysis (ICA) [72], in which the independence among source signals is mainly used for the separation, has been
studied actively [73, 74, 75, 76, 77]. Conventional ICA works particularly by using the statistical independence between the signals as in the case of speech-speech mixing, i.e., all sources can be regarded as point sources, but such a mixing condition is very rare and unrealistic because real noises are often widespread sources. Other microphone array techniques have been proposed using multi-channel spectral subtraction [78], Kurtosis maximization [79] and time-frequency masking [80].

### 2.3.3 Geometric Source Separation

Recently, BSS methods incorporating BF concepts can be found in the literature for improved separation performance [81]. GSS, too, is one of those hybrid Sound Source Separation (SSS) algorithms [82]. In general, BSS has a number of limitations such as permutation and scaling problems, which can be relaxed in GSS by the introduction of “geometric constraints”. These are obtained from the locations of microphones and sound sources. Unlike the Linearly Constrained Minimum Variance (LCMV) beamformer which minimizes the output power subject to a distortion-less constraint, GSS explicitly minimizes crosstalk, leading to faster adaptation. The method is also interesting for use in the mobile robotics context because it allows easy addition and removal of sources.

Geometric Source Separation is an adaptive algorithm that can process the input data incrementally and makes use of the locations of the sources explicitly. These locations are estimated with a sound source localization method, which will be discussed in Section 2.3.4. It requires lower computational cost compared to ICA-based BSS algorithms. To formulate GSS, we extend Equation (2.7) to support bold notations to represent vectors in the formulae, where the non-stationary noise sources are also treated as source signals \( S_m(k, l) \) represented in a source signal vector \( \mathbf{S}(k, l) \). Let matrix \( \mathbf{A}(k) \), \( \mathbf{N}^*(k, l) \), \( \mathbf{Y}(k, l) \) denote the transfer function, stationary noise and observed spectrum at the microphone array, respectively.

\[
\mathbf{Y}(k, l) = \mathbf{A}(k) \mathbf{S}(k, l) + \mathbf{N}^*(k, l). \tag{2.25}
\]

Suppose \( \mathbf{W}(k) \) is the separation matrix, then source separation is defined as:

\[
\mathbf{Z}(k, l) = \mathbf{W}(k, l) \mathbf{Y}(k, l). \tag{2.26}
\]

To estimate \( \mathbf{W}(k, l) \) properly, GSS introduces two cost functions that must be min-
2.3. EXISTING WORK ON DIRECTIONAL NOISE REDUCTION

The first cost function comes along with the constraint of decorrelation of the separated signals.

\[ J_1(W(k, l)) = \| R_{yy}(k, l) - \text{diag}[R_{yy}(k, l)] \|_2^2. \]  

(2.27)

The second cost function is derived from the geometric constraint that ensures unity gain in the direction of the source of interest and places zeros in the direction of interferences [26].

\[ J_2(W(k, l)) = \| W(k, l) A(k) - I \|_2^2. \]  

(2.28)

Finally, calculating the gradient of the cost functions with respect to \( W(k, l) \) [82] and providing an iterative update formulation for the separation matrix results to:

\[
W(k, l) = W(k, l - 1) - \mu \left[ \alpha(k) \frac{\partial J_1(W(k, l - 1))}{\partial W(k, l - 1)} + \frac{\partial J_2(W(k, l - 1))}{\partial W(k, l - 1)} \right],
\]

(2.29)

where \( \alpha(k) \) is an energy normalization factor and \( \mu \) is defined as adaptation rate. Moreover, adaptive step-size control that provides fast convergence of the separation matrix [83] can be used. Besides, GSS implementation also can exploit a method called Optima Controlled Recursive Averaging [84], which controls window size adaptively causing a smoother convergence and thus better separation results [85]. The weight initialization corresponds to a DS beamformer [82]. Such an initialization ensures that the worst performance prior to adaptation is equal to a DS beamformer. In addition, in case of single sound source, the algorithm is equivalent to a DS beamformer.

2.3.4 Supporting Microphone Array Technologies to GSS

GSS needs to know the location of the target signal. In addition, it can only deal with non-stationary sound sources, whereas the stationary noise sources need to be taken care of in a separate speech enhancement stage also called Post-Filter (PF) in the context of SSS. Figure 2.10 shows a typical multi-channel noise reduction scheme used in the signal processing community. In this section, we will discuss the remaining elements of SSL and PF.
CHAPTER 2. LITERATURE REVIEW AND BASIC METHODS

Sound Source Localization

Speaker localization is of particular interest in the development of speech enhancement methods requiring information of the speaker position. Based on the localized speaker position, the microphone array can be steered towards the corresponding direction for effective speech acquisition. This approach is appropriate for speech enhancement applications with a moving speaker, such as in video-conferencing [58] or robot audition [85]. A localization system may also be used in a multi-speaker scenario to enhance speech from a particular speaker with respect to others or with respect to noise sources. The beamforming principle may be used as foundation for source localization by steering the array to various spatial points to find the peak in the output power. Localization methods based on the maximization of the Steered Response Power (SRP) of a beamformer have been shown to be robust [58]. However, they present a high dependency on the spectral content of the source signal, which in most practical situations is unknown. Other popular SSL methods are based on Head Related-Transfer Function (HRTF) [86], Cross-power Spectrum Phase (CSP) [87] and on one of the most popular adaptive beamforming algorithms called MUltiple Signal Classification (MUSIC), which we will focus on next.

**Standard Eigenvalue Decomposition-based MUSIC:** MUSIC [88] was successfully applied to robotics by Asano et al. [89] to detect the Direction of Arrival (DoA) of sound sources. Standard Eigenvalue Decomposition (SEVD)-based MUSIC performs an eigenvalue decomposition on the correlation matrix of the noisy signal such as following:

$$R_{yy}(k, \phi) = Y(k, \phi)Y^*(k, \phi), \quad (2.30)$$

where $Y(k, \phi) = [Y_1(k, \phi), Y_2(k, \phi), \ldots, Y_M(k, \phi)]^T$ with $Y_m(k, \phi)$ is the vector of spectrum of the signal captured by the $m$-th microphone, $^T$ is the transpose operator, $\phi$
2.3. EXISTING WORK ON DIRECTIONAL NOISE REDUCTION

denotes the orientation of the microphones, and \((\cdot)^*\) represents complex conjugate transpose operator. For the sake of clarity, we omitted the time-frame \(l\). Eigen decomposition of \(R_{yy}(k, \phi)\) leads to

\[
R_{yy}(k, \phi) = Q(k, \phi) \Lambda Q^{-1}(k, \phi), \tag{2.31}
\]

where \(\Lambda\) is the matrix, whose diagonal elements are the corresponding eigenvalues, i.e. \(\Lambda_{ii} = \lambda_i\) and \(Q\) is the square matrix, whose \(i\)-th column is the eigenvector \(q_i\). Moreover, we assume that the \(\lambda_i\) and \(q_i\) belong to the sound sources of interest for \(1 \leq i \leq N\) and to the undesired noise sources for \(N + 1 \leq i \leq M\), where \(N\) represents the number of sound sources.

Prior to the localization, steering vectors of the microphone array, \(G(k, \psi)\), are determined, which are measured as impulse responses for a certain orientation of \(\psi\).

\[
P(k, \psi) = \frac{|G^*(k, \psi)G(k, \psi)|}{\sum_{m=N+1}^{M} |G^*(k, \psi)q_m|}. \tag{2.32}
\]

The whole area of interest is scanned over the range of \(\psi\) using \(G(k, \psi)\). If \(G(k, \psi)\) matches the true location vectors during scanning, the denominator in Equation (2.32) goes to zero, resulting in peaks in the spatial spectrum and indicating the estimated source locations. Moreover, a consequent source tracker system, which actually performs a temporal integration of the source directions in a given time window, runs to ensure the reliability of the location estimations. The decision on the source locations is made by comparing the power of the peaks of \(P(k, \psi)\) to a threshold value \(T\) and if the power of the source is less than the threshold, the source is eliminated. The threshold is usually set manually.

**Generalized Eigenvalue Decomposition-based MUSIC:** SEVD-MUSIC only works if target sources have stronger power than noise sources. If the robot emits noise with high power, undesired peaks occur in the spatial spectrum unavoidably according to Equation (2.32). In order to solve the problem of SEVD-MUSIC, the Generalized Eigenvalue Decomposition (GEVD) \([90]\) method is applied to the MUSIC algorithm called GEVD-based MUSIC \([16]\). Contrary to SEVD-MUSIC, it utilizes a noise correlation matrix in order to suppress environmental noise sources. The way to solve the problem is to determine the correlation matrix in terms of noises \(N(k, \phi)\). Let \(K(k, \phi)\)
be the correlation matrix derived by noise sources, which is described as

$$K(k, \phi) = N(k, \phi) N^*(k, \phi),$$  \hspace{1cm} (2.33)

where $N, \phi(\omega) = [N_1(k, \phi), N_2(k, \phi), \ldots, N_M(k, \phi)]^T$ with $N_m(k, \phi)$ is the vector of spectrum of the noise captured by the $m$-th microphone. The noise can be both stationary or non-stationary.

$$K^{-1}(k, \phi) R(k, \phi) = \hat{Q}(k, \phi) \hat{\Lambda} \hat{Q}^{-1}(k, \phi),$$  \hspace{1cm} (2.34)

where $\hat{\Lambda}$ is the new eigenvalue matrix with $\hat{\Lambda}_{ii} = \lambda_i$ and $\hat{Q}$ is the new regular matrix, whose $i$-th column is the new eigenvector $\hat{q}_i$. Moreover, we assume that the new $\hat{\lambda}_i$ and $\hat{q}_i$ correspond to the sound sources of interest for $1 \leq i \leq N$ and to the undesired noise sources for $N + 1 \leq i \leq M$. $K^{-1}(k, \phi)$ has an effect of whitening the overall noise. The new GEVD spatial spectrum is described as

$$\hat{P}(k, \psi) = \frac{|G^*(k, \psi) . G(k, \psi)|}{\sum_{m=N+1}^{M} |G^*(k, \psi) . \hat{q}_m|}.$$  \hspace{1cm} (2.35)

Again, the peaks occurring in the MUSIC spatial spectrum yield the source locations. The decision on the source locations is made by comparing the sum of the peak powers, $\sum_k \hat{P}(k, \psi)$ to a threshold value $T$. So far, GEVD-MUSIC was used to detect stationary noise only [16].

**Multi-channel Post-filter**

Many methods of integrating microphone array signal processing and nonlinear signal processing such as SS have been studied [91, 92, 93]. It has been well demonstrated that such integration methods can achieve higher noise reduction performance than that obtained using conventional microphone array noise reduction. Therefore, after a separation process a multi-channel post-filter operation is applied so that the sounds can be enhanced even further. Most of these post-filters address reduction of stationary background noise [94, 95].

We explain a post-filter concept, which is based on the optimal estimator proposed by Ephraim and Malah [36]. Since their method takes temporal and spectral continuities into consideration, it generates less distortion compared to the conventional spectral subtraction based noise reduction methods. By extending their idea further, a multichannel post-filter is proposed by Cohen [96], which can cope with non-stationary interferences.
as well as stationary noise. This module treats the transient components in the spectrum as if they are caused by the leakage energies that may occasionally arise due to poor separation performance caused by the reverberation, localization error, near-field effects.

The main aim of post-filter is to find the weighting coefficients $G_m(k)$ and estimate the clean audio signal that is represented by $\hat{X}_m(k)$ by attenuating the separated signal $Z_m(k)$ as in Equation (2.36). This equation is similar to Equation (2.20) used in spectral filter for SS except it is applied to all separated sound sources.

$$\hat{X}_n(k, l) = G_n(k, l)Z_n(k, l), \quad (2.36)$$

where $n$ is the index for separated sound source ($1 \leq n \leq N$). To be able to calculate the gain $G_n(k, l)$, noise variances of both stationary noise $\lambda_{n}^{stat}(k, l)$ and source leakage $\lambda_{n}^{leak}(k, l)$ must be predicted. Whereas the former one is computed using the MCRA [96] method, to estimate the latter $\lambda_{n}^{leak}(k, l)$ the following formulation is proposed [26]:

$$\lambda_{n}^{leak}(k, l) = \eta \sum_{i=0, i \neq n}^{N-1} H_i(k, l), \quad (2.37)$$

where $\eta$ defines the leakage factor, an empirical value representing a suppression factor of the interferences from the desired source achieved by the source separation step. $H_n(k, l)$ is the smoothed spectrum of $Z_n(k, l)$ calculated according to Equation (2.38).

$$H_n(k, l + 1) = \alpha H_n(k, l) + (1 - \alpha)|Z_n(k, l + 1)|^2, \quad (2.38)$$

where $\alpha$ is an pre-selected smoothing factor. Finally, total variance can be calculated by:

$$\lambda_{n}^{total}(k, l) = \lambda_{n}^{stat}(k, l) + \lambda_{n}^{leak}(k, l). \quad (2.39)$$

Figure 2.11 depicts the architecture of this post-filter. It is possible either to use the same gain calculation and spectral filter operations as explained in Section 2.2.2 or another attenuation rule called Minimum Mean Square Estimation (MMSE). This rule involves speech presence probability calculations such as given in [96] and is based on minimum mean-square error estimation of the spectral amplitude [37].
2.4 Existing Work on Noise-robust Automatic Speech Recognition

The main motivation for using speech enhancement for an ASR task arises primarily due to the time and expense required to collect significant amounts of data on which to train/adapt acoustic models for a wide range of noise conditions. The noise types and repertoire may not only depend on environmental factors, but also robot’s own noise. The noise reduction techniques such as explained in Section 2.2 and Section 2.3 can handle the constantly changing noise conditions, allowing noisy speech signals to be transformed into a clean speech representation (either waveform or feature vectors). Eventually, it enables the use of well-trained clean speech acoustic models for ASR. However, in practice noise reduction involves also some side effects called “noise artifacts”. This section starts with preliminary background information on ASR system and consequently introduces two important methods to tackle the noise artifacts.

2.4.1 Basic Automatic Speech Recognition System

Common speech recognition systems are built by matching frequency patterns of input speech features with frequency patterns of phonemes or syllables. The principle of ASR is that input speech \( X \) is decoded to a sequence of words \( W \) using an Acoustic Model (AM) and a The Language Model (LM). Figure 2.12 shows a diagram of a standard ASR system. The feature extraction part converts the input signal to feature vectors. The acoustic model contains statistical data trained from a large corpus of speech sounds with phonetic labels. The AM is often modeled as a Hidden Markov Model (HMM) using Expectation Maximization (EM) algorithm and calculates the probability that the
pattern $X$ is generated from the sequence of words $W$, namely $P(X|W)$. Language model part contains statistical data about word sequences. For transcription, it is often modeled using an N-Gram approach. For command input, it is modeled using a constrained grammar. In both cases, the LMs are often compiled into a Finite State Machine (FSM) to be used in decoding and calculates $P(W)$. The decoder part searches for the most likely word sequences $W$ using the AM and LM data, which gives the maximum posterior probability $P(X|W)$. Bayes law lets $P(W|X)$ to transform into:

$$P(W|X) = \frac{P(W)P(X|W)}{P(X)},$$

where $P(X)$ can be ignored since $W$ does not involve $P(X)$.

Figure 2.12: Proposed structure of ASR: Black modules indicate the basic building blocks of the ASR system, whereas the red modules are the alternative blocks to deal with noise.

To allow noise robustness, a noise reduction part is placed in the front-end part to preprocess the input data. This dissertation focuses on this noise reduction part including single-channel and multi-channel signal processing methods explained in Section 2.2 and 2.3 as well as an alternative front-end design proposed in Section 2.4.3.

### 2.4.2 Acoustic Model Adaptation

Performance degradation in the presence of noise and reverberation is caused by the mismatch between the received noisy signals and the acoustic models used by speech recognizers. Major ASR algorithms are based on statistical pattern recognition ap-
proaches making them sensitive to reduced input speech quality. Therefore, a common approach to realize noise-robust ASR is the use of an acoustic model trained with noise adaptation techniques \[97, 98, 99\]. These techniques are also called \textit{multi-style training}. Because these techniques train an acoustic model using noise-added speech data, the robustness of ASR against the trained noises improves. In optimal Bayesian signal processing methods, a set of probability models are trained for the signal and the noise processes. The models are then used for the decoding of the underlying states of the signal and noise, thus for noisy signal recognition. Even if noise reduction is applied, the acoustic model could be retrained with data that was processed with the noise reduction method.

In practice, however, acoustic model adaptation is not an elegant solution for recognizing noisy speech captured by a robot, such as shown in the work of Nakadai \textit{et al.} [25] whose system used 51 acoustic models trained under different conditions at the same time, but the performance still deteriorated in an environment with unexpected and/or dynamically changing noises.

### 2.4.3 Missing Feature Theory-based ASR

\textit{Missing Feature Theory (MFT)} is a technique used to integrate pre-processing and ASR. The underlying idea of MFT is to mask out unreliable acoustic features using a so-called \textit{Missing Feature Mask (MFM)} \[100, 101, 102\]. Several studies in connected digit recognition for telephony applications \[103, 104\], speaker verification \[105, 106\], dereverberation \[107\] have shown that it helps to improve the reliability of fundamental sound processing applications.

Yamamoto introduced MFT to integrate ASR with a binaural robot audition system [108]. In this system, the reliability of each \textit{Time-Frequency (TF)} component was estimated by comparing separated speech with the corresponding clean speech. Thereby, a \textit{hard mask} consisting of 0 or 1 for each TF component was generated based on the reliability using a threshold. Word recognition accuracy improved significantly, since the generated MFMs are considered a priori MFM due to the presence of the pilot signals. Yamamoto and Valin \textit{et al.} developed an automatic MFM generation process based on microphone-array processing [109]. They showed that unreliable features generated by pre-processing are mainly caused by energy leakage from other sound sources. A microphone-array-based technique was developed to estimate the reliability of each time-frequency component from this energy leakage, by considering the properties of a
2.5 Related Work on Ego Noise Reduction

Traditional acoustic noise control uses passive techniques such as enclosures, barriers and silencers to attenuate undesired noise, however they are relatively large, heavy, costly and ineffective at low frequencies [5]. Especially for robots, the additional bulk of these silencers can be a hindrance. Another technique called Active Noise Control (ANC) uses an electro-acoustic or electro-mechanical system to cancel the unwanted noise, based on the principle of superposition. An antinoise of equal amplitude, but opposite phase is generated through a secondary source and combined with the primary noise, thus resulting in the cancellation of both noises. Usually, noise stemming from large engines in vehicles such as aircrafts [114], helicopters [115], ships [116], and heavy machinery in industry [117, 118] is treated using ANC. In these kind of environments,
the listener is not acoustically isolated from the environment and the ambient noise has a very large amplitude, therefore listeners use devices equipped with ANC applications, such as hearing protectors, headsets, etc. Although ANC methods show excellent performance for low frequency periodic noise, their main drawbacks are the requirements of additional microphone/s (for feedforward control), loudspeaker/s, power source/s and, most important of all, their poor performance of adaptability to time-varying noise.

Nakadai et al. [119] proposed a noise cancellation method using two pairs of microphones. One pair, in the inner part of the shielding body, records only internal motor noise and helps the sound localizer to distinguish between spectral subbands that are noisy and not noisy, and to ignore the subbands in which noise is dominant. This technique was not designed to remove the noise and obtain refined speech. Its major drawback was that, by filtering out the noise, it also eliminates useful signals. The ego-motion noise problem has also been addressed by predicting and removing ego-motion noise using templates recorded in advance. For example, Nishimura et al. [120] estimated the ego-motion noise of distinct gestures of the robot. Using motion commands, the correct noise template matching to the current motion was selected from the template database and subtracted. Furthermore, a blockwise template prediction, as in [120], which uses templates recorded from the onset until the offset time of motor noises, fails completely when the exact onset of the template is not detected properly or the trajectory/duration of the motion changes slightly. Ito et al. [121] used an Artificial Neural Network (ANN) to develop a new frame-by-frame based prediction to cope with unstable walking noise. This ANN also solved the synchronization problem of Nishimura’s template based approach. The trained network was designed to predict the noise spectrum from the angular velocities of the joints of the robot. However, they concentrated on a small robot with the limited degrees of freedom. For a huge dataset, ANN will have a slow training speed and its online adaptation will be difficult. Previous works [120, 121] were based mainly on estimating templates for different motions, but did not focus on the possibility of quality improvement resulting from spectral enhancement optimization factors.

In the field of “Robot Audition,” noise is suppressed primarily by using sound source separation techniques with microphone arrays (see Section 2.3) or single-channel noise reduction methods (see Section 2.2). In contrast to these methods, neither a directional noise model, such as that utilized for interfering speakers, nor a diffuse background noise model, is an entirely appropriate representation for ego noise. Because the motors are
located in the near-field of the microphones, they produce sounds that have both diffuse and directional characteristics. In a related study involving a microphone array, Even et al. [91] proposed to use semi-blind signal separation to obtain both external and internal noise by attaching additional sensors (e.g., microphones or vibration sensors) inside the robot. The predictions were used to compute Wiener coefficients. After this suppression step, a delay-and-sum beamformer enhances the refined speech. Although it improves speech recognition accuracy considerably, the additional sensors inside the robot cover pose constraints on implementation. For certain types of sensors, this method may require a body cover made of high-quality or thick material so that external noise is definitely not recorded by these additional sensors (i.e., microphones) or an accurate correspondence model between different sensor signals may be required (i.e., vibration sensors). Besides, semi-blind signal separation methods demonstrate good performance only if the interfering signal is known, however even slightest inaccurate estimations may lead to deteriorated performance.

Several studies have focused on specific conditions for near-field sound sources. For example, Mizumachi et al. described a model for sound sources in the near-field with spherical wave propagation and line sound sources in contrast to conventional far field assumptions like plane wave propagation and point-shaped sound sources [122]. Zheng et al. proposed a spherically isotropic noise model for near-field objects that achieves stronger reverberation suppression and reduced beampattern variations for broadband signals like our motor noise signals [123]. However, these proposed models are computationally expensive, can only deal with single sound sources, and, more importantly, are designed for stationary sound sources. In a standard task with robot motions, acoustic properties of the noise, such as the power and frequencies of the motor noise spectrum as well as the location and number of the active motors dynamically change over time. Thus, the performance of near-field sound processing will definitely deteriorate.

2.6 Positioning of this Thesis towards Related Work

In our study, to achieve ego noise reduction, we model the process of ego noise and propose an effective estimation method. Furthermore, we deal with general constraints of an online ego noise estimation system and design ego noise robust applications for robot audition.

In the robot audition literature, the ego noise problem is probably the most complex
noise problem due to its non-stationarity and constraints related to its proximity to the microphones. Throughout the review in this chapter, numerous examples of speech enhancement methods have been cited, however none of them work properly under the circumstances posed by ego noise. Our study, representing ego noise with parameterized templates and estimating it from a template database recorded in advance, is original. In contrast to attaching additional sensors our method has several advantages, including its ability to be easily implemented on any mobile robot regardless of the physical constraints about the external shielding. By exploiting only existing microphones, it is also cost effective and applicable without any hardware modifications. Some studies using single-channel templates pursued to represent these templates as whole blocks of noise data, as from the onset to the offset time of the motion, and some researches aimed at using neural networks to learn and estimate the templates, but their performance was not high and their system was unable to cope with all possible templates covering the whole motion space of a robot with a high number of DoF. Furthermore, approximate search strategies for selecting the appropriate templates make our method more suitable for online learning. In addition, we enhance the accuracy of the templates further by incorporating more information related to the joints, such as angular acceleration.

In our study, we update the template database in an online manner, which enables the robot to learn its own noise without human intervention and on the fly. Being the first of its kind, this strategy of template learning makes use of previously learned knowledge about the templates to speed up the learning. It also makes the noise estimation module more robust because errors in the training set can be corrected during operation and it enables the system to adapt to partially-known or dynamic environments.

Automatic speech recognition has been chosen as the primary application environment for design and evaluation of ego noise reduction techniques in the robotics community. Neither many other application domains have been addressed, nor the integration of noise reduction and ASR has been discussed much. To demonstrate the general applicability of the proposed approach, we extend it to other applications such as sound source localization, and MFT-based integration of noise reduction and ASR.
Chapter 3

Template Based Ego Noise Estimation

3.1 Introduction

Noise characteristics differ based on the cause and properties of the noise source. In this chapter, we first provide a framework for formalizing (1) the modeling of any given complex type of noise and (2) the “estimation” process of this noise. In that sense, we pursue a methodology for categorizing noise models and consequently choose a suitable noise estimation method. The ideal description of the noise estimation process can be determined only if we can resolve the issues within this framework regarding “knowledge gathering,” “representation” and “algorithm.”

In order to be able to provide solutions to these issues, we have to consider the main causes of a noise type and find its most dominant properties. This is crucial because all noise reduction methods operate under different conditions under different constraints. An analysis of the noise sources specific to the problem domain is, therefore, the first step. For example, in car applications, noise generated by the engine or the car radio is considered directional, and noise from the wind passing around the car cabin or from the contact between the road and the tires is considered diffuse noise [124]. For hearing aid applications noise sources have to be classified into “clean speech,” “speech in noise,” “noise,” and “music” based on their properties [125]. The observations we made on the ego noise generation process has shown us that the noise follows a similar pattern each time the respective motion is performed. Therefore, the templates are good representations of motor noise when the same actions are performed over and over again.
3.2 Issues in Noise Estimation

We pursue a methodology to find a simplified representation of noise with an objective of its explanation and later its estimation. For a general noise modeling and the estimation of the noise, this methodology entails three problems that we need to address: “knowledge gathering,” “representation” and “algorithm.” Whereas the first two are basically concerned with modeling the noise, the algorithm issue focuses on the noise estimation process itself.

1. Knowledge Gathering:

This issue concerns what kind of data is available and how we acquire a knowledge from it. In general, we can categorize two broad methods for knowledge gathering:

- *Data-driven (inductive) model* that connects the system variables (input, output and internal variables) with only a limited knowledge of the details about the physical behavior of the noise
- *Knowledge-driven (deductive) model* based on a good understanding of the underlying processes of noise

For example, we can use statistics knowledge to construct models or rules of the interesting noise states, and then use these to identify the noise. Alternatively, we can use labeled data to build models of noise states that can help in identification/estimation and apply these models to our input data stream.

2. Representation:

This issue focuses on how we represent our data and knowledge, and in what space we model the noise. The noise estimation problem is defined in terms of searching through a state space, where we know in advance what state constitutes our noise
CHAPTER 3. TEMPLATE BASED EGO NOISE ESTIMATION

model. For example, we can operate directly in the observable state space, which is typically considered the spectra of the noisy speech itself and define the properties of noise in those terms. Alternatively, we can construct abstractions on top of the observable space, and perform the estimation in an abstract state space using a terminology that brings a higher degree of meaning. To exemplify, we can extract relational information, thus representing noise models in terms of spatio-spectro-temporal information.

Table 3.1: Properties of noise modeling strategies depending on the knowledge gathering and representation issues

<table>
<thead>
<tr>
<th></th>
<th>Observable state space</th>
<th>Abstract state space</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data-driven</strong></td>
<td><strong>Category 1</strong></td>
<td><strong>Category 2</strong></td>
</tr>
<tr>
<td></td>
<td>● Requires large quantities of low-level data</td>
<td>● Requires complicated pre-processing</td>
</tr>
<tr>
<td></td>
<td>● Hard to capture complex situations</td>
<td>● Requires domain knowledge reld. representations</td>
</tr>
<tr>
<td><strong>Knowledge-driven</strong></td>
<td><strong>Category 3</strong></td>
<td><strong>Category 4</strong></td>
</tr>
<tr>
<td></td>
<td>● Requires domain-specific knowledge</td>
<td>● Requires extensive formalization of the domain</td>
</tr>
<tr>
<td></td>
<td>● Hard to model complex situations</td>
<td>● Increases the complexity of simple tasks</td>
</tr>
</tbody>
</table>

Table 3.1 explains the characteristics of each modeling strategy considering the knowledge gathering and representation issues. All four categories of approaches for noise modeling have their strengths and weaknesses. We briefly want to categorize and exemplify the existing solutions for noise models according to the proposed chart.

Data-driven approaches are suitable when we have access to large amounts of data, and knowledge-driven approaches are suitable when we have access to high quality domain knowledge. For example, models that operate directly on the low-level data, such as STFT or waveforms of noise signals, are data-driven models. In case of knowledge-driven models, we use higher-level domain knowledge such as labels or identifiers for the noise data. Choosing an approach that operates directly in the observable state-space can be suitable when data is open-access, and when there is no need for extracting deep information from the process. Neither do we need to have much domain knowledge concerning what to represent. If instantaneous observations have sufficiently high resolution, are frequent enough and available all the time, it is advantageous to keep the model simple. Abstract representations however require that we have some degree of domain knowledge concerning what to represent. To name a few, stationary, diffuse or directional types of noises
have special characteristics that can be abstracted by using physical/mathematical models. However, complex and non-stationary noise types are rather cumbersome to be modeled that way. The ideal category to be applied to ego noise problem will be proposed in Section 3.3

3. Algorithm:

Algorithm issue concentrates on how to perform the actual estimation. This issue depends on how we solve the first and second issues. For example, we can assume that the noise follows certain patterns or has distinguishable properties and treat it with established noise estimation methods. Another example is a rule-based system, in which if-then rules are used for identifying certain noise states. Moreover, by building a statistical model over the input space, we could measure the correlation between the current state and the state in the model.

3.3 Ego Noise Estimation using Templates

By considering the observations about the ego noise in Section 2.1.4, we make the following assumptions to model the noise:

<table>
<thead>
<tr>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>• The superposition of single joint motor noises, fan noise and hardware noise at an arbitrary time point is equal to the whole-body noise at the corresponding time point.</td>
</tr>
<tr>
<td>• The noise of a motor is dependent on the position, velocity and acceleration of that motor.</td>
</tr>
<tr>
<td>• Similar combinations of joint status will result in similar motor noise spectral vectors at any instant of time.</td>
</tr>
</tbody>
</table>

After explaining the biological findings about how some animals deal with ego noise, we describe the proposed approaches for solving the issues raised earlier in Section 3.2, namely knowledge-gathering, representation and algorithm consequently in Section 3.3.2.
3.3.1 Biological Inspiration of Ego Noise Cancellation

Weakly electric fishes (see Figure 3.1) are an exceptional model system to study sensory acquisition, neuronal information processing and sensory-motor integration. They produce electric signals with a specialized organ in their tail. This discharge is called Electric Organ Discharge (EOD), which is typically less than one volt in amplitude. In addition, they are electrosensitive and can perceive their self-generated signals for navigation, object detection (electrolocation) and electric signals of other electric fishes for communication purposes (electrocommunication). They can encode useful information in the waveforms of their EODs, which can then be detected and interpreted by other weakly electric fish [126].

They have evolved sensory systems that make use of copies of their own EOD to decode the temporal characteristics of incoming sensory signals. The reception and localization follows as the fish gets the image of objects in its environment, which is coded at the skin as a 2-D representation of impedances distorting its own electric field. Then, this image is processed further in the Electrosensory Lateral Line (ELL) of the electric fish. EEL is a cerebellum-like structure specialized for processing sensory input and it has adaptive sensory processors that subtract out predictable features of the sensory inflow after a period of association between predictive signals and particular patterns of sensory input [127]. The electroreceptors are driven by large electric fields produced by the swimming and respiratory movements of the animal. By adjusting the gain of parallel fiber synapses, the responses to self-generated fields are cancelled in the principal cells, so that only electrical signals from external sources, which are not predictable from the parallel fiber activity, are transmitted. The spatial map of
sensory expectations computed from recent inputs, is subtracted from the total input. This negative image (current spatial map related to ego-EOD) is updated with each EOD [127]. In this context, the fish learns sensory effects of motor actions (e.g. fin movement) as well.

It is hypothesized also that the Dorsal Cochlea Nucleus (DCN) used by the mammals has the ability to cancel the neural activity associated with self-generated noise. This occurs when the animal moves, especially when the pinna moves. The self-noise cancellation could be accomplished by autonomous mechanisms similar to the electrosensory system of the electric fishes, so that the animal learns eventually what kind of a noise template it has to subtract in case of the presence of this specific ego noise [128].

### 3.3.2 Approaches for Estimating Ego Noise using Templates

**Approach for Knowledge-gathering Issue: Observing the Motion and Ego Noise**

We have shown that the process of motion generation is in a strong relationship with ego noise generation. Both processes can be described in terms of time and change [129]. Although they are both continuous processes by nature, we approximate them by discrete processes that are observed in some state space. A state can be described as a vector of properties that we are able to observe and infer. A discrete process is therefore a sequence of states. Changes in the states of both processes are directly observable in this case because motion states and corresponding audio frames are accessible any time. There is one point that should not confuse the reader. As an outside observer we lack the means of directly observing the state space and actions that are used in planning the robot motion. We do not intend to have a control on the motion plan of the robot. We rather observe this process through the observed process pair as in Figure 3.2. It illustrates a process of a motion plan being executed in a state space through actions. We observe this process in a different space from that in which it is carried out, and events denote change in the observed state space.

**Approach for Representation Issue: Ego Noise Modeling based on Templates**

A template is the representation of an actual state of “ego noise” regarded as a discrete process pair. As shown in Figure 3.3, it consists of

- the content of spectral data regarding noise
The elegance of template based noise representation is that the noise related information such as its distribution, directivity/diffuseness and spatial position of its source is not needed to be modeled unlike in most mathematical/physical models. Noise observations are not used to train probabilistic noise models, but rather the “model” is the preprocessed data itself. Since we use knowledge-driven time variables of joint state to represent ego noise, we can do it directly in the observable state space. This will put our approach called “template-based noise modeling” into the Category 3 of the chart given back in Table 3.1.

Particular demerits of creating templates are the requirement of large amounts of memory and computing power. However, the amounts of available processing and memory capacity have increased exponentially over the past decades, as stated by Moores Law: The number of transistors on an integrated circuit doubles every twenty months.

---

**Figure 3.2: Ego motion processes**

- the label with meta-information regarding motion data

**Figure 3.3: Template-based representation of ego noise**
Moores law is very general and applies to the available RAM memory, hard disk capacity as well as to the computing performance of processors.

**Approach for Algorithm Issue: Similarity measurement**

*Signal classification* is used in detection, pattern recognition and decision making systems. Our aim is to design a minimum-error system for selecting a template from one of a number of likely classes of noise data (i.e., templates). Designing a classifier involves creation of a set of models for each class of noise, i.e. templates. The simplest form that the models can assume is a bank, or code book of noise data, each representing the prototype for one template. (A more complete model for each class of signals takes the form of a probability distribution function.) Instead of using some statistical models, new templates are directly compared with available examples of the already existing noise data. In the final phase, current signal is labeled with the nearest or the most likely template.

### 3.4 Parameterized Template Estimation

This section discusses the proposed methods based on the approaches explained in the previous section. We introduce the following concepts to tackle the knowledge gathering, knowledge representation and algorithm issues respectively:

#### 3.4.1 Method for Knowledge-gathering Issue: Framewise Templates

A conventional blockwise template represents an actual segment of noise as a single block from the onset until the offset time of a certain motor noise as in Figure 3.4(a). This method, however, has several shortcomings; e.g., it could be performed properly only after the detection of the exact starting moment of the template, which is a very hard task to achieve. Otherwise, it suffers from misalignments in time throughout the entire timespan of the template. Furthermore, the necessity of templates for each possible motion makes it very cumbersome to implement for a real-time system. This is a matter of the time granularity of each state in the observed process. The more coarse-grained (longer) the state is, the more the possibility of errors due to small variations in the robot motion. Therefore, we introduce instantaneous *framewise templates* to tackle this
problem. We can fragment the blockwise template into smaller pieces for each frame, i.e. framewise templates as in Figure 3.4(b), thus increasing the reusability of the templates and eliminating possible errors caused by small fluctuations in the trajectory.

![Image of templates](image)

**Figure 3.4:** Illustration of the two template types

### 3.4.2 Method for Representation Issue: Parameterized Templates based on the State of Robot Joints

**Representation of Templates**

A **blockwise template** fundamentally consists of two parts:

1. Label: Motion command (e.g., “wave right hand” or “turn head from 0° to 40°”)
2. Data: Whole ego noise spectrum recorded during the motion (i.e., spectral matrix).

It associates the ego noise spectrum to the corresponding motion commands as representative labels. This method requires a huge amount of data for each possible motion. Considering the impossibility of collecting and producing templates for each joint of different combinations of origin, target, position, velocity and acceleration parameters, this approach is simply not feasible in a realistic scenario. Finally, this rather primitive representation of motion labels cannot deal with deviations in motion trajectories.

To overcome these deficits, we developed a technique that parameterizes a discrete audio segment under consideration using motor status, obtaining a spectral vector that represents the ego noise at that instant of time. This **parameterized template** has a different structure than the blockwise template:
1. Label: Instantaneous joint status of the robot (i.e., feature vector),

2. Data: Instantaneous ego noise spectrum of one frame (i.e., spectral vector).

The feature vectors consist of the angular position, velocity and acceleration of each joint. To create so-called parameterized templates, we need a robot with joint angle sensors (encoders) that measure the angular positions of each of its joints separately. Figure 3.5 depicts the differences between both template structures.

**Figure 3.5: Template-based representations of ego noise**

**Generation of a Template Database**

During the motion of the robot, the actual position ($\theta$) information for each motor is gathered regularly. Using the difference between consecutive sensor outputs, velocity ($\dot{\theta}$) and acceleration ($\ddot{\theta}$) can be calculated. If $J$ joints are active, $3J$ attributes are generated. Each feature is normalized to $[-1, 1]$, so that all features have the same contribution to the prediction. The resulting feature vector has the form, $\mathbf{F}(l) = [\theta_1(l), \dot{\theta}_1(l), \ddot{\theta}_1(l), \ldots, \theta_J(l), \dot{\theta}_J(l), \ddot{\theta}_J(l)]$. At the same time, motor noise is recorded as in Figure 3.6. The spectrum of the noise, $D(l) = [D(1,l), D(2,l), \ldots, D(F,l)]$, where $F$ represents the number of frequency bins is calculated by the sound processing branch running in parallel. Both feature vectors and spectra are continuously labeled with time tags so that corresponding templates can be generated when their time tags match. Each parameterized template is in the format of a data block consisting of two concatenated vectors, $[\mathbf{F}(l), D(l)]$. Finally, a large noise template database, consisting of short noise templates for many joint configurations, can be created.
3.4.3 Method for Algorithm Issue: Nearest Neighborhood Search in the Template Database

Suppose we have a robot with 30 DoF. We gather templates in every frame (i.e. 10 milliseconds), which contain feature vectors $\mathbf{F}(l)$ consisting of 90 features and spectral vectors $\mathbf{D}(l)$ consisting of 128 spectrotemporal values, all represented in floating point values. It is easy to imagine that in several minutes huge streams of continuous data will be stored that must be processed, learned and updated in an online fashion. For high dimensional learning tasks, the performance of the machine learning algorithm plays a crucial role. One alternative is using ANNs with sigmoidal activation functions, which learn slowly in high dimensional spaces and are vulnerable to unlearning of relevant knowledge when trained on new data points [130]. Instead, we prefer to use a non-parametric, instance-based classification technique like the Nearest Neighbor (NN) algorithm because it is easy to implement, does not need any a priori knowledge about the data and the output of the NN algorithm can be interpreted as an a posteriori probability of the input pattern being the estimated pattern [131]. The last point is especially important because it provides us a measure of performance allowing to update existing templates in our incremental learning algorithm based on the relative template confidence levels.

The estimation phase starts with a search of the database for the best matching template of motor noise at that time. Finding the correct template involves a search of all the templates in the database for most similar joint configuration (See Figure 3.7). We utilized a 1-Nearest Neighbor (1-NN) search to accomplish this task: For a given
3.5. PARAMETERIZED TEMPLATE SUBTRACTION

database $F$ of template feature vector in $3J$-dimensional feature space and a query feature vector $Q(l)$, we find the closest feature vector in $F$ to $Q(l)$. The distance is measured by the Euclidean distance between two feature vectors $Q(l) = (Q_1(l), Q_2(l), ..., Q_{3J}(l))$ and $\tilde{F}(l) = (\tilde{F}_1(l), \tilde{F}_2(l), ..., \tilde{F}_{3J}(l))$, where $\tilde{F}(l)$ is an element $F$.

$$d(Q(l), \tilde{F}(l)) = ||Q(l) - \tilde{F}(l)|| = \sqrt{\sum_{j=1}^{3J} (Q_j(l) - \tilde{F}_j(l))^2} \quad (3.1)$$

The spectral vector $|N_n(l)|^2$ stored in the template $T(l)$ with $\tilde{F}(l)$ having the shortest distance to $Q(l)$ is selected as the ego noise estimate $\lambda(l)$. This prediction process can be applied to every frame. In that sense, a template for a single arbitrary motion of an arbitrary duration can be regarded as the concatenation of smaller templates predicted according to the above-mentioned approach on a frame-by-frame basis.

![Figure 3.7: Ego motion noise estimation from database](image)

3.5 Parameterized Template Subtraction

Parameterized template subtraction is basically a magnitude spectral subtraction operation where the a template substitutes the estimated signal in Equation (2.24), thus gains are calculated by using Equation (3.2):

$$G(k,l) = max \left( 1 - \alpha \frac{|D(k,l)|}{|Y(k,l)|}, \beta \right). \quad (3.2)$$
A spectral filtering operation of the signal $Y(\omega, k)$ with this coefficient finalizes the template subtraction as in Equation (2.20).

Unlike in blockwise template processing, our prediction, generation and subtraction methods do not require any starting or ending signals, indicating that no abrupt blockwise templates are applied to the noisy signals discontinuously. Our method continuously process the data, even when the robot does not move. Therefore, our template database does not only consist of recordings of motor noise, but also of recordings of the joints in resting positions. We conduct training sessions of uninterrupted sound recording for a single continuous motion sequence consisting of hundreds of individual motions with short ($< 1$ sec.) pauses between each motion.

### 3.6 Evaluation

#### 3.6.1 Evaluation Criteria

**Normalized Noise Estimation Error (NNEE)**

NNEE computes the error of the noise estimate normalized by the energy of the actual noise using the following formula:

$$
\epsilon(l) = 10 \cdot \log_{10} \left( \frac{\sum_{k=0}^{M} ||N(k, l)||^2 - ||\hat{N}(k, l)||^2)}{\sum_{k=0}^{M} ||N(k, l)||^2} \right),
$$

$$
\bar{\epsilon} = \frac{1}{L} \sum_{l=1}^{L} \epsilon(l),
$$

where $L$ is the number of frames. Smaller $\epsilon$ indicates a more accurate noise estimate.

**Segmental SNR**

The average of the SNR values is calculated for segments of audio data such as:

$$
SNR = \frac{1}{L} \sum_{l=1}^{L} 10 \cdot \log_{10} \left( \frac{\sum_{t} x^2(t)}{\sum_{t} (x(t) - \hat{x}(t))^2} \right),
$$
Log-Spectral Distortion

This evaluation measure computes the reconstruction error of the clean speech by comparing the enhanced speech signal $\hat{X}(k,l)$ with the original speech $X(k,l)$ in the log domain as follows:

$$LSD = \frac{1}{L} \sum_{l=1}^{L} \left( \frac{1}{K} \sum_{k=1}^{K} \left[ L X(k,l) - L \hat{X}(k,l) \right]^2 \right)^{1/2}, \tag{3.6}$$

where $L X(k,l) \triangleq \max\{20\log_{10}|X(k,l)|, \delta\}$ is the log spectrum confined to about 50 dB dynamic range, hence $\delta = \max_{k,l}\{20\log_{10}|X(k,l)|\} - 50$ [5].

Ideal Estimation

In order to evaluate the accuracy of the noise spectrum estimation, ideal gain $G_i(k,l)$ is computed from the original noise spectrum in the training session and used in Equation (2.20).

Automatic Speech Recognition

The noise signals are mixed with clean speech utterances used in a typical human-robot interaction dialog and recorded by us. This Japanese word dataset includes 236 words for 4 female and 4 male speakers. We used a clean acoustic model trained with Japanese Newspaper Article Sentences (JNAS) corpus, 60-hour of speech data spoken by 306 male and female speakers, hence the speech recognition is a word and speaker-open test. We used 13 static Mel-Scale Log Spectrum (MSLS) features, 13 delta MSLS features and 1 delta power feature. Speech recognition results are given as average Word Correct Rates (WCR) of instances from the noisy test set.

3.6.2 Experimental Settings

The experiments are designed to assess the estimation and suppression capabilities of single-channel noise estimation methods (MCRA, HRLE, TE) by individually applying them to the noise signals consisting of ego noise and environmental background noise. One set of noise data (200 seconds long) for training and three sets of noise data (100 seconds long) for testing are collected during a continuous head motion of 2 Degree of Freedoms (DoF) and arm motion of 4 DoF. The recording environment is a room with
the dimensions of 4.0 m × 7.0 m × 3.0 m with a reverberation time ($RT_{20}$) of 0.2 sec. The implementation runs on the robot audition platform called HARK [132]. The performance of all methods are compared under 4 different SNR conditions for the same signal segments as in Figure 3.8.

- Condition (1)-(2): Noise energy is fixed, speech signals are amplified to yield $SNR_{(1)} = 3dB$ and $SNR_{(2)} = -3dB$.
- Condition (3)-(4): Gaussian white noise is added to (2) to represent changing conditions of static BGN with $SNR_{(3)} = -3.1dB$ and $SNR_{(4)} = -3.2dB$.

The parameters of the HRLE and MCRA are selected appropriately for non-stationary noise estimation [133],[54] and are given in Table 3.2. Since the musical noise [34] due to nonlinear signal processing causes discomfort to listening as well as damage to ASR features, it is desirable that musical noise is controlled through signal processing. However, in almost all nonlinear noise reduction methods, parameters to mitigate musical noise in nonlinear signal processing are determined heuristically. Unfortunately, evaluations were conducted on subjective tests by humans, and no objective evaluations have been performed effectively. In this work, too, a minor spectral floor $\beta = 0.1$ and a regular overestimation factor $\alpha = 1.0$ are selected heuristically for the SS stage.

Table 3.2: Parameter settings for MCRA and HRLE

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MCRA</th>
<th>HRLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_d$</td>
<td>0.95, $\alpha_p = 0.2$</td>
<td>$L_{min} = -200dB$</td>
</tr>
<tr>
<td>$\alpha_s$</td>
<td>0.8, $L = 125$</td>
<td>$L_{step} = 0.2dB$, $x = 50%$</td>
</tr>
<tr>
<td>$\lambda_{MCRA}$</td>
<td>5</td>
<td>$I_{max} = 2000$, $T_r = 1 sec$</td>
</tr>
</tbody>
</table>

3.6.3 Estimation Performance

Mobile robots are intended to be deployed to environments with (possibly changing) background noise, ego noise and speech such as depicted in Figure 3.8. Although noise estimation methods based on recursive averaging like MCRA or HRLE show a good performance in stationary noise, they cannot adapt to the ego-motion noise rapidly (see Figure 3.9(a)) because motions of different joint combinations produce drastically different noise spectra in every frame. On the other hand, the frame-by-frame based
3.6. EVALUATION

estimation method, TE, is fairly accurate in reconstructing any type of noise that can be reproduced. Ego-motion noise falls into this noise category because the duration and spectral power of the motor noise signals do not change drastically for the same type of motions when the motion is performed again (Figure 3.9(b)).

![Figure 3.8: Noisy spectrogram (Stationary noise: Background noise + hardware noise + fan noise)](image)

**Figure 3.8:** Noisy spectrogram (Stationary noise: Background noise + hardware noise + fan noise)

![Figure 3.9: Estimated noise spectra](image)

(a) Spectrum obtained using HRLE  
(b) Spectrum obtained using TE

**Figure 3.9:** Estimated noise spectra

The estimation performance of all methods are given in Table 3.3. In all conditions of stationary noise, TE performed worse than other methods because this method is not suitable for estimating the stationary noise. It yields the lowest estimation error for (1) and (2) in the presence of EN because the test data was recorded in the same room. However, unfamiliar background noise conditions such as in (3) and (4) degrade the accuracy of the TE because the portion of the stationary noise becomes more dominant in the overall noise energy compared to the portion of the ego noise. Furthermore, HRLE outperforms MCRA especially in the presence of ego noise.

62
### CHAPTER 3. TEMPLATE BASED EGO NOISE ESTIMATION

Table 3.3: Noise estimation performance for all methods

<table>
<thead>
<tr>
<th>SNR</th>
<th>$\bar{\theta}$ for given segment</th>
<th>HRLE</th>
<th>MCRA</th>
<th>TE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Stationary noise</td>
<td>-5.81</td>
<td>-6.26</td>
<td>-5.08</td>
</tr>
<tr>
<td>3dB</td>
<td>Non-stationary + stationary noise</td>
<td>-4.61</td>
<td>-4.38</td>
<td>-4.89</td>
</tr>
<tr>
<td></td>
<td>Total noise + speech</td>
<td>-4.92</td>
<td>-4.68</td>
<td>-5.06</td>
</tr>
<tr>
<td>(2)</td>
<td>Stationary noise</td>
<td>-5.81</td>
<td>-6.26</td>
<td>-5.08</td>
</tr>
<tr>
<td>-3dB</td>
<td>Non-stationary + stationary noise</td>
<td>-4.61</td>
<td>-4.38</td>
<td>-4.89</td>
</tr>
<tr>
<td></td>
<td>Total noise + speech</td>
<td>-4.84</td>
<td>-4.63</td>
<td>-5.06</td>
</tr>
<tr>
<td>(3)</td>
<td>Stationary noise</td>
<td>-7.96</td>
<td>-7.12</td>
<td>-4.95</td>
</tr>
<tr>
<td>-3.1dB</td>
<td>Non-stationary + stationary noise</td>
<td>-6.3</td>
<td>-5.75</td>
<td>-4.95</td>
</tr>
<tr>
<td></td>
<td>Total noise + speech</td>
<td>-6.71</td>
<td>-5.63</td>
<td>-5.11</td>
</tr>
<tr>
<td>(4)</td>
<td>Stationary noise</td>
<td>-8.87</td>
<td>-8.03</td>
<td>-4.52</td>
</tr>
<tr>
<td>-3.2dB</td>
<td>Non-stationary + stationary noise</td>
<td>-7.1</td>
<td>-6.64</td>
<td>-4.76</td>
</tr>
<tr>
<td></td>
<td>Total noise + speech</td>
<td>-7.42</td>
<td>-6.45</td>
<td>-4.92</td>
</tr>
</tbody>
</table>

#### 3.6.4 Noise Reduction Performance

We evaluate the *noise reduction performance* by using the system depicted in Figure 2.9. This time, we compare the results of all methods to the baseline results (i.e., No Processing, NP). As Table 3.4 demonstrates, TE achieves the smallest LSD and largest WCRs among all methods in the conditions (1) and (2). A substantial improvement of 30.4 points in WCR is achieved especially for $-3dB$. In terms of SNR, too, TE outperforms the other methods. In general, HRLE creates less distortion in speech (LSD), thus, achieves higher recognition rates (WCR) compared to MCRA.

Table 3.4: Ego noise reduction performance for all methods

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>$SNR_{(1)} = 3dB$</th>
<th>$SNR_{(2)} = -3dB$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNR</td>
<td>LSD</td>
</tr>
<tr>
<td>NP</td>
<td>3.00</td>
<td>9.7</td>
</tr>
<tr>
<td>MCRA</td>
<td>3.90</td>
<td>9.49</td>
</tr>
<tr>
<td>HRLE</td>
<td>3.96</td>
<td>8.94</td>
</tr>
<tr>
<td>TE</td>
<td>5.49</td>
<td>8.51</td>
</tr>
</tbody>
</table>

Table 3.5 demonstrates the performance of the template estimation for the ideal noise and is a good measure to show the best case scenario when the ideal templates are applied.
3.6. EVALUATION

Table 3.5: Ideal noise estimation and reduction performance for $SNR(2) = -3dB$

<table>
<thead>
<tr>
<th>Method</th>
<th>$SNR$</th>
<th>$LSD$</th>
<th>$WCR$</th>
<th>$\epsilon_{P1}$</th>
<th>$\epsilon_{P2}$</th>
<th>$\epsilon_{P3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE</td>
<td>2.67</td>
<td>7.32</td>
<td>92.6</td>
<td>-12.1</td>
<td>-86.5</td>
<td>-88.0</td>
</tr>
</tbody>
</table>

Table 3.6 shows the results for the simulation of changing ambient noise. We observe that the higher the contribution of the background noise, the more effective MCRA and HRLE methods are. Under conditions (3) and (4), especially HRLE contributes more to canceling the overall noise by eliminating the background noise.

Table 3.6: Ego noise reduction performance for all methods

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>$SNR_{(3)} = -3.1dB$</th>
<th>$SNR_{(4)} = -3.2dB$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$SNR$</td>
<td>$LSD$</td>
</tr>
<tr>
<td>NP</td>
<td>-3.1</td>
<td>12.6</td>
</tr>
<tr>
<td>MCRA</td>
<td>-1.31</td>
<td>11.2</td>
</tr>
<tr>
<td>HRLE</td>
<td>-0.75</td>
<td>10.0</td>
</tr>
<tr>
<td>TE</td>
<td>1.97</td>
<td>9.75</td>
</tr>
</tbody>
</table>

3.6.5 Discussion and Future Work

By inspecting the spectrograms of the refined signal, $\hat{X}(k,l)$, we see clearly that TE has difficulty in estimating the stationary noise segment because it relies only on one single template representing the fixed stationary position of the robot. Furthermore, the difference between the ideal values and real performance gives us a possible room for improvement. This indicates that in its current form, our system has difficulties in achieving precise prediction of templates because they rely only simple insertion of the instantaneous templates into the database.

An important advantage of parameterized approach would be that it is possible to update the database on the fly making the prediction more accurate in case any change in the characteristics of the overall noise occurs at any time. Therefore, more sophisticated online learning and indexing techniques will endow the system with a capability of adaptation. Moreover, it can run online on the background while the robot is performing its duties and tasks. In order to improve the robustness further, we plan to embed the current single-channel ego noise reduction stage into a general multi-channel
CHAPTER 3. TEMPLATE BASED EGO NOISE ESTIMATION

microphone array processing framework for speech recognition that utilizes geometric source separation and post filtering.

3.7 Summary

We tackled the problem of estimating and removing ego noise from useful signals. This problem is general and important in human-robot interaction because microphones capture the disturbing noise and the system cannot differentiate if the total signal is a clean speech or contaminated speech. In the framework of computational auditory scene analysis and robot audition, we developed a method for effectively overcoming the ego noise problem.

We modeled the ego noise using templates which are well-suited to capture the dynamic nature of the motion data represented by a sequence of observations. Based on these observations, it was possible to associate discrete time series data representing the motion (i.e., the angular status of each joint of the robot) with another series of discrete time data representing the total ego noise spectrum. This approach relied on seamless synchronization between data on joint status (i.e., angular position, velocity and acceleration) and audio data. The high estimation quality achieved by this approach allows us to suppress noise accurately by applying a template-based spectral subtraction. The method is evaluated using several objective criteria such as normalized noise estimation error, segmental SNR, log-spectral distortion and word correct rates obtained by an ASR. Results are very promising in the sense that high suppression rates are achieved while keeping the speech as untouched as possible.
Chapter 4

Online System for Ego Noise Estimation

4.1 Introduction

The strengths of the template-based estimation method introduced in Chapter 3 are that it is not SNR-dependent, it is not prone to VAD errors and adaptation latency is zero theoretically. However, the template estimation method cannot perform adaptation to the overall (ego and background) noise in an environment with changing noise conditions. It is only capable of reproducing the templates that exist in the database, thus it can reflect only the ambient noise conditions during the offline training session. Furthermore, we intend to build mobile/humanoid robots, which obtain information about the environment, accomplish useful tasks without human intervention, move either all or part of their bodies in a real-world environment and meeting these goals under real-time constraints particularly. Since some tasks are complex to hand-code and human-involved operations become costly in time or money, learning is required by robots in those circumstances. Especially when the robot does not have sufficiently accurate data about the task/environment or the environment changes, the robot must acquire the necessary information by itself. A task that must be handled in a similar fashion is the prediction of robot (ego) noise.

The typical problem tightly coupled to offline training for TE is that they are not well suited to real-time, real-world applications due to insufficiently accurate data in ego noise templates of unknown motions (i.e., missing templates) or the differences in environmental conditions from the conditions in the training session. Creating databases
4.2 Unified Framework for Noise Estimation

Noise estimation methods based on recursive averaging like MCRA or HRLE cannot adapt to the ego-motion noise rapidly, because motions of different joint combinations produce different noise spectra in every frame. On the other hand, the frame-by-frame based estimation method, TE, is fairly accurate in reconstructing ego noise. However, it can only reproduce the templates that exist in the database, thus it reflects the noise conditions in the training session only. Considering that

$$N(k, l) = N_s(k, l) + N_n(k, l),$$

(4.1)

where $N_s(k, l)$ and $N_n(k, l)$ denote the stationary and non-stationary portions of the overall noise, we propose to use the stationary and non-stationary noise estimation methods in series as in Figure 4.1 for the training of the template database, which allows us to create a unified framework for noise estimation consisting of two parallel and independent processes as in Figure 4.2. While recursive averaging takes care of $N_s(k, l)$ the background and stationary portion of ego noise (i.e. fan/hardware noise), TE with a template representation of $T(l) = [\hat{F}(l) : |\hat{N}_n(l)|^2]$ tackles the remaining non-stationary noise portion of motor noise $N_n(k, l)$. It is noteworthy to mention that in such an offline scheme, the data transfer to/from the database is simplex (unidirectional) during the separate training/estimation phases.

![Figure 4.1: Offline template database generation](image-url)
4.3 Incremental Template Learning

We already explained the basic offline classification model incorporated into the core of template learning and estimation algorithm in the Section 3.4.3. On top of it, Section 4.3.1 proposes methods to enhance the classification model and finally the template learning algorithm is discussed in Section 4.3.2.

4.3.1 Extensions of the Basic Classification Model

In order to improve the robustness of template prediction, we made the following modifications on the classification method.

**Inverse distance weighted average (IDWA)**

Instead of using 1-NN, we assign confidence-based weights $\omega^k$ to $K$ nearest templates $\mathbf{3}^1(l) \ldots \mathbf{3}^K(l)$ by giving the highest weight to the closest neighbor, and compute the final spectral vector $\hat{\lambda}_{TE}(l)$ from their Inverse Distance Weighted Average (IDWA) of all nearest candidates such as in Equation 4.2.

$$
\hat{\lambda}_{TE}(l) = \sum_{n=1}^{K} \left( \frac{1}{\sum_{m=1}^{K} \frac{d(Q(l), \mathbf{3}^m(l))}{\sum_{m=1}^{K} \frac{d(Q(l), \mathbf{3}^m(l))}} \cdot \omega^{n}} \cdot \lambda_{TE}^{n}(l) \right)
$$

Figure 4.2: Offline template estimation using a unified framework
IDWA assumes that templates closer to the query point are more representative than templates further away. The denominator term is used for normalization of the $\omega^k$ such that $\sum_{k=1}^{K} \omega^k = 1$.

K-dimensional trees

To increase the speed of K-NN, we suggest to utilize tree structures, such as K-Dimensional (KD) trees [134]. So, the search is conducted more efficiently by using the tree properties that quickly eliminate large portions of the search space. Still, the curse of dimensionality causes the algorithm to visit more branches than in lower dimensional spaces in high dimensional spaces. In particular, when the number of points is only slightly higher than the number of dimensions, the algorithm is only slightly better than a linear exhaustive search of all of the points.

Approximate K-NN search

Arya et al. [135] showed that tolerating a small amount of error in the search (returning a point that may not be the nearest neighbor, but is not significantly further away from the query point than the true nearest neighbor) can cause an approximate search strategy to achieve significant improvements in running time. Because the computation cost is also related to the code implementation quality, the reader is advised to address to reliable machine learning software\textsuperscript{1}, which was able to provide real-time computation for our system implementation and also offers wide variety of computation options.

4.3.2 Learning Algorithm

Incremental learning is essential for online generation of the template database because it makes use of previously learned knowledge about the templates to speed up the learning. It makes the noise estimation module more robust because errors in the training set can be corrected during operation and it enables the system to adapt to partially-known or dynamic environments. Therefore, it is expected that the performance will gradually improve in time.

In the proposed system, the task of the learning system is to autonomously extract and learn templates. The system checks whether the acquired audio signal is mixed with a speech signal based on the decision of VAD and discards it if it is not only ego noise.

\textsuperscript{1}e.g., http://www.cs.umd.edu/˜mount/ANN/
This continuous VAD loop determines the onset and offset times of the template update interval. During this interval, the system also decides if each observed template is a known template or a new template to be learned. The observed template is searched in the trained database and its similarity with other templates in the database is computed using the same distance metric as in Equation 3.1. Based on the comparison of \( d_{\text{min}}(l) \), the smallest distance \( d(Q(l), \mathcal{F}(l)) \) in \( \mathcal{F} \), with a given fixed distance threshold, \( T \), the current template is either used to update the old template or it is inserted into the database as a new template. When the similarity is low, the template is treated as missing template and inserted into \( \mathcal{F} \); otherwise the adaptive update mechanism is active, which computes the weighted sum of the old and current template by laying the focus more on recently-acquired templates and less on earlier observations. The contribution of past templates are reduced by introducing a forgetting factor \( \eta \) with \( 0 \leq \eta \leq 1 \), which helps to provide a moderate balance between adaptivity (learning quality) and stability (robustness against errors). The former is achieved by using lower \( \eta \), whereas higher \( \eta \) causes stability. The pseudo-code of the incremental learning algorithm is shown below.

```
Incremental Learning Algorithm

while (state(VAD)=NON-SPEECH) do
    if \( d_{\text{min}}(Q(l), \mathcal{F}(l)) \geq T \) then
        \[
        [\mathcal{F}_{\text{new}} : |N_{n}^{\text{new}}|^2] \leftarrow [\mathcal{F}(l) : |N_{n}(l)|^2]
        \]
    else
        \[
        [\mathcal{F}_{\text{old}} : |N_{n}^{\text{old}}|^2] \leftarrow [\mathcal{F}_{\text{old}} : |N_{n}^{\text{old}}|^2 + (1 - \eta)|N_{n}(l)|^2]
        \]
    end if
    if (timer=$\tau$) then
        Rebuild the tree and reset the timer
    end if
end while
```

One key aspect of this incremental learning algorithm is the rebuilding of the KD-tree due to practical concerns. Insertion-based incremental construction of a KD-tree for a long time is problematic, because the tree becomes unbalanced eventually. The rebalancing task is tedious and repeating it at each insertion is also costly. Therefore, we re-build the data structure in constant time intervals determined by \( \tau \), which is clearly a negligible compromise between the retrieval time and computational cost of the learning algorithm. Until the next rebuilding phase, new templates to be inserted are stored in
4.4. Evaluation

4.4.1 Evaluation Criteria

Normalized Noise Estimation Error (NNEE)

NNEE computes the error of the noise estimate normalized by the energy of the actual noise using the following formula:

$$
\epsilon(l) = 10 \cdot \log_{10} \left( \frac{\sum_{k=0}^{M} |N(k, l)|^2 - |\hat{N}(k, l)|^2}{\sum_{k=0}^{M} |N(k, l)|^2} \right),
$$

$$
\bar{\epsilon} = \frac{1}{L} \sum_{l=1}^{L} \epsilon(l),
$$

where $L$ is the number of frames. Smaller $\epsilon$ indicates a more accurate noise estimate.
CHAPTER 4. ONLINE SYSTEM FOR EGO NOISE ESTIMATION

Segmental SNR

The average of the SNR values is calculated for segments of audio data such as:

\[
SNR = \frac{1}{L} \sum_{l=1}^{L} 10 \cdot \log_{10} \left( \frac{\sum_t x^2(t)}{\sum_t (x(t) - \hat{x}(t))^2} \right),
\]  

(4.5)

Log-Spectral Distortion

This evaluation measure computes the reconstruction error of the clean speech by comparing the enhanced speech signal \( \hat{X}(k, l) \) with the original speech \( X(k, l) \) in the log domain as follows:

\[
LSD = \frac{1}{L} \sum_{l=1}^{L} \left( \frac{1}{K} \sum_{k=1}^{K} \left[ \mathcal{L}X(k, l) - \mathcal{L}\hat{X}(k, l) \right] \right)^{1/2},
\]  

(4.6)

where \( \mathcal{L}X(k, l) \triangleq \max\{20\log_{10}|X(k, l)|, \delta\} \) is the log spectrum confined to about 50 dB dynamic range, hence \( \delta = \max_{k,l}\{20\log_{10}|X(k, l)|\} - 50 \) [5].

Ideal Estimation

In order to evaluate the accuracy of the noise spectrum estimation, ideal gain \( G_i(k, l) \) is computed from the original noise spectrum in the training session and used in Equation (2.20). Note that \( G_i(k, l) \) is still subject to negligible errors caused by approximation strategy of the fast NN search\(^2\).

Automatic Speech Recognition

The noise signals are mixed with clean speech utterances used in a typical human-robot interaction dialog and recorded by us. This Japanese word dataset includes 236 words for 4 female and 4 male speakers. We used a clean acoustic model trained with Japanese Newspaper Article Sentences (JNAS) corpus, 60-hour of speech data spoken by 306 male and female speakers, hence the speech recognition is a word and speaker-open test. We used 13 static Mel-Scale Log Spectrum (MSLS) features, 13 delta MSLS features and 1 delta power feature. Speech recognition results are given as average Word Correct Rates (WCR) of instances from the noisy test set.

---

\(^2\)http://www.cs.umd.edu/~mount/ANN/
4.4.2 Evaluation of Unified Framework for Noise Estimation

Experimental Settings

We evaluate the individual estimation and suppression capabilities of MCRA, HRLE, TE where we applied them to the noise signals consisting of ego noise and environmental background noise as a benchmark. Then the performance of proposed method with different combinations (i.e., HRLE+TE, MCRA+TE) and with different settings (i.e., HRLE+TE, TE+HRLE) is evaluated. We also test the performance of stationary noise reduction after applying TE-based SS to obtain comparative results. To get consistent and comparable results, we used the same dataset as in the previous chapter. One set of noise data (200 seconds long) for training and three sets of noise data (100 seconds long) for testing are collected during a continuous head motion of 2 Degree of Freedoms (DoF) and arm motion of 4 DoF (see Figure 4.4). The recording environment is a room with the dimensions of 4.0 m × 7.0 m × 3.0 m with a reverberation time ($RT_{20}$) of 0.2 sec. The implementation runs on the robot audition platform called HARK [132]. The performance of all methods are compared under 4 different SNR conditions for the same signal segments as in Figure 3.8:

- Condition (1)-(2): Noise energy is fixed, speech signals are amplified to yield $SNR_{(1)} = 3\, dB$ and $SNR_{(2)} = −3\, dB$.
- Condition (3)-(4): Gaussian white noise is added to (2) to represent changing conditions of static BGN with $SNR_{(3)} = −3.1\, dB$ and $SNR_{(4)} = −3.2\, dB$.

This time, the parameters of the HRLE are selected appropriately for stationary noise estimation [54] and are given in Table 4.1. A minor spectral floor $\beta = 0.1$ is used in the SS stage.

Table 4.1: Parameter settings for HRLE

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{step}$</td>
<td>0.2 dB</td>
</tr>
<tr>
<td>$L_{min}$</td>
<td>$−200, dB$</td>
</tr>
<tr>
<td>$I_{max}$</td>
<td>2000</td>
</tr>
<tr>
<td>$T_r$</td>
<td>10 sec</td>
</tr>
<tr>
<td>$x$</td>
<td>20%</td>
</tr>
</tbody>
</table>

Results

We evaluate the noise reduction performance by using the systems depicted in Figure 4.1 and 4.2. This time, we also use the combinations of several estimators in series (labeled
as method A + method B, e.g. HRLE+TE) for database generation and compare their results to the baseline results (i.e., No Processing, NP). As Table 4.2 demonstrates, TE achieves the smallest LSD and largest WCRs among all methods in the conditions (1) and (2). A substantial improvement of 30.4 points in WCR is achieved especially for $-3dB$. In terms of SNR, only HRLE+TE can overperform TE. In general, HRLE creates less distortion in speech (LSD), thus, achieves higher recognition rates (WCR) compared to MCRA. We also observe that using the stationary noise estimation techniques rather as a secondary step after TE, such as, TE+HRLE or TE+MCRA, does not improve the quality of the refined speech any better than when they are used as a primary step.

Table 4.3 provides not only ideal results like, $SNR$, $LSD$ or $WCR$, but also normalized noise spectrum errors for stationary noise, stationary+non-stationary noise, total noise+speech, resp. $\bar{\epsilon}_{P1}$, $\bar{\epsilon}_{P2}$, $\bar{\epsilon}_{P3}$. The performance reduction for the combined methods is due to errors in the nonlinear noise reduction operation prior to database generation (see Figure 4.1). By making just a small compromise in the accuracies as shown in both Tables 4.2 and 4.3, the framework of HRLE+TE will later provide adaptivity to the system and achieve even better results in changing background noise.

By inspecting the spectrograms of the refined signal, $\hat{X}(k,l)$, in Figure 4.5(a) and Figure 4.5(b), we see clearly that TE alone has difficulty in estimating the stationary noise segment because it relies only on one single template representing the fixed stationary position of the robot. In addition, it creates sharp vertical valleys and peaks in the spectrum, which are caused by the smaller attenuations of the frequencies compared to relatively larger attenuations of their neighboring frequencies due to the incorrect estimations or missing templates in the database. This so-called musical noise effect is reduced by the HRLE+TE combination, because HRLE attenuates the spectrum more smoothly and the characteristics of the residual noise is similar to less harmful salt-and-
4.4. EVALUATION

Table 4.2: Ego noise reduction performance for all methods

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR</th>
<th>LSD</th>
<th>WCR</th>
<th>SNR</th>
<th>LSD</th>
<th>WCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP</td>
<td>3.00</td>
<td>9.7</td>
<td>78</td>
<td>-3.0</td>
<td>11.2</td>
<td>28.3</td>
</tr>
<tr>
<td>MCRA</td>
<td>3.90</td>
<td>9.49</td>
<td>83.2</td>
<td>-1.38</td>
<td>10.8</td>
<td>44.5</td>
</tr>
<tr>
<td>HRLE</td>
<td>3.96</td>
<td>8.94</td>
<td>84.1</td>
<td>-1.2</td>
<td>10.2</td>
<td>47.2</td>
</tr>
<tr>
<td>TE</td>
<td>5.49</td>
<td>8.51</td>
<td>87.4</td>
<td>2.05</td>
<td>8.73</td>
<td>58.7</td>
</tr>
<tr>
<td>MCRA+TE</td>
<td>5.85</td>
<td>8.61</td>
<td>79.7</td>
<td>1.85</td>
<td>8.95</td>
<td>51.6</td>
</tr>
<tr>
<td>HRLE+TE</td>
<td>6.02</td>
<td>8.66</td>
<td>86.2</td>
<td>2.74</td>
<td>8.88</td>
<td>54.8</td>
</tr>
<tr>
<td>TE+MCRA</td>
<td>5.28</td>
<td>8.74</td>
<td>84.5</td>
<td>1.44</td>
<td>9.48</td>
<td>53.9</td>
</tr>
<tr>
<td>TE+HRLE</td>
<td>5.22</td>
<td>8.91</td>
<td>85.6</td>
<td>1.41</td>
<td>9.3</td>
<td>55.6</td>
</tr>
</tbody>
</table>

Table 4.3: Ideal noise estimation and reduction performance for $SNR_{(2)} = -3dB$

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR</th>
<th>LSD</th>
<th>WCR</th>
<th>$\bar{\epsilon}_{P1}$</th>
<th>$\bar{\epsilon}_{P2}$</th>
<th>$\bar{\epsilon}_{P3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE</td>
<td>2.67</td>
<td>7.32</td>
<td>92.6</td>
<td>-12.1</td>
<td>-86.5</td>
<td>-88.0</td>
</tr>
<tr>
<td>HRLE+TE</td>
<td>3.84</td>
<td>7.43</td>
<td>92.3</td>
<td>-5.80</td>
<td>-9.22</td>
<td>-8.36</td>
</tr>
<tr>
<td>TE+HRLE</td>
<td>2.56</td>
<td>8.19</td>
<td>89.4</td>
<td>-6.06</td>
<td>-12.42</td>
<td>-12.54</td>
</tr>
</tbody>
</table>

pepper noise.

Figure 4.5: Refined spectrograms ($SNR = 3dB$)

Table 4.4 shows the results for the simulation of changing ambient noise. We observe that the higher the contribution of the background noise, the more effective MCRA and HRLE methods are. Especially HRLE contributes more to cancelling the overall noise by eliminating the background noise. Hence, in its combination with TE, TE deals only with the non-stationary part of the overall noise regarding the ego-motion noise. This kind of configuration improves the robustness of the noise suppression system and makes
CHAPTER 4. ONLINE SYSTEM FOR EGO NOISE ESTIMATION

Table 4.4: Ego noise reduction performance for all methods

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>$SNR_{(3)} = -3.1dB$</th>
<th>$SNR_{(4)} = -3.2dB$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$SNR$</td>
<td>$LSD$</td>
</tr>
<tr>
<td>NP</td>
<td>-3.1</td>
<td>12.6</td>
</tr>
<tr>
<td>MCRA</td>
<td>-1.31</td>
<td>11.2</td>
</tr>
<tr>
<td>HRLE</td>
<td>-0.75</td>
<td>10.0</td>
</tr>
<tr>
<td>TE</td>
<td>1.97</td>
<td>9.75</td>
</tr>
<tr>
<td>MCRA+TE</td>
<td>1.93</td>
<td>9.75</td>
</tr>
<tr>
<td>HRLE+TE</td>
<td><strong>2.63</strong></td>
<td><strong>9.03</strong></td>
</tr>
<tr>
<td>TE+MCRA</td>
<td>1.44</td>
<td>10.3</td>
</tr>
<tr>
<td>TE+HRLE</td>
<td>1.52</td>
<td>9.74</td>
</tr>
</tbody>
</table>

it less susceptible to any change in the environmental noise condition. Besides ASR, the high $SNR$ and low $LSD$ results indicate that the estimates can also be used accurately for other speech applications. Figure 4.6(a) shows the resulting spectrogram when TE-based SS is applied for $SNR_{(4)}$. As expected, TE cannot cope with the remaining background noise in all frequency bands. On the other hand, the proposed HRLE+TE based SS method can suppress the noise effectively as shown in Figure 4.6(b). The similarity between Figure 4.6(b) and Figure 4.5(b) justifies the importance of the proposed configuration in a typical audition system of a mobile robot and that it achieves a similar suppression performance even if the environment changes.

Figure 4.6: Refined spectrograms ($SNR = 3.2dB$)
4.4. Evaluation

4.4.3 Evaluation of Incremental Learning

Experimental Settings

We assess the estimation and suppression capabilities of (1) conventional TE method whose templates represent both BGN and EN, and (2) proposed method in Section 4.2 whose templates represent only motor noise by applying them to the noise signals consisting of ego noise and environmental background noise. One set of noise data (200 seconds long) for training and three sets of noise data (100 seconds long) for testing are collected during a continuous head motion of 2 Degree of Freedom (DoF) and arm motion of 4 DoF, which generates 18 features. The recording environment is a room with the dimensions of $4.0 \times 7.0 \times 3.0$ m with a reverberation time ($RT_{20}$) of 0.2 sec. Because the accuracy of the VAD would also influence the performance of the learning system, we intentionally by-pass (exclude) it in these experiments by recording only ego noise and not human speech. The performance of all methods are compared under 4 different SNR conditions for the same signal segments as in Figure 3.8. Condition (1)-(2): Noise energy is fixed, speech signals are amplified to yield $SNR_{(1)} = 3$ dB and $SNR_{(2)} = -3$ dB; Condition (3)-(4): Gaussian white noise is added to (2) to represent changing conditions of static BGN (e.g. entering into a new room or turning on the air conditioner) with $SNR_{(3)} = -3.1$ dB and $SNR_{(4)} = -3.2$ dB. The parameters of the HRLE are selected in a way that they are optimal for stationary noise estimation [54] and are given in Table 4.1. The optimal value of $\eta = 0.9$ for incremental learning, which pursues stability rather than adaptivity is found empirically. For a database size such as in our experiments, real-time conditions are properly provided with $\tau = 5$ frames. A minor spectral floor $\beta = 0.1$ is used in the SS stage.

Results

Learning performance: To assess the learning capability of our system with respect to threshold $T$, we evaluated the estimation error (NNEE) in incremental steps, i.e. after repeating the same motion $N$ times ($1 \leq N \leq 20$). Because we used bounded ([0 1]) features, the values of $T$ are also bounded to $[0 \sqrt{3J}]$. Figure 4.7 demonstrates the tendency of reduced error ($\epsilon$) with respect to the increased repetitions. The error decreases when $T$ is sufficiently low. However, we also observe that there is a negative correlation between the number of templates stored and the value of $T$ as in Figure 4.8. Therefore, the system designer must consider the trade-off between the size of the tem-
plate database and distance threshold. The settings denoted as “$T \to 0$” indicates that there is a continuous insertion of every incoming template into the database like in conventional template estimation methods [121, 136] and “$T \to \infty$” indicates that there is only one single (mean) template updated during all repetitions. They both yield the baseline performance. Because HRLE is a stationary noise estimator, it cannot deal with the non-stationary ego noise and shows also a poor performance. Since “$T = 0.0001$” among others yielded the smallest error in our experiments, it is selected as the optimal $T$ for the incremental learning and we continue to evaluate the final estimation and suppression results based on this value.

Figure 4.7: Estimation error in relation with the number of iterations

Figure 4.8: Number of templates in relation with the number of iterations
4.4. EVALUATION

Time [s] | Frequency [kHz]
---|---
0 | 2 | 4 | 6
0 | 2 | 4 | 6 | 8 | 10
50 | 60 | 70 | 80 | 90 | 100

(a) using TE \((T \rightarrow 0)\)
(b) Using TE \((T = 0.00001)\)

Figure 4.9: Estimated noise spectrograms

Table 4.5: Ego noise estimation performance for all methods

<table>
<thead>
<tr>
<th>SNR [dB]</th>
<th>SNR for given interval</th>
<th>HRLE</th>
<th>TE ((T \rightarrow 0))</th>
<th>TE ((T \rightarrow \infty))</th>
<th>TE ((T = 0.0001))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) 3.0</td>
<td>Stationary noise</td>
<td>-5.81</td>
<td>-5.08</td>
<td>-1.74</td>
<td>-6.77</td>
</tr>
<tr>
<td></td>
<td>Non-stationary + stationary noise</td>
<td>-4.61</td>
<td>-4.89</td>
<td>-4.44</td>
<td>-6.59</td>
</tr>
<tr>
<td></td>
<td>Total noise+speech</td>
<td>-4.92</td>
<td>-5.06</td>
<td>-4.78</td>
<td>-6.84</td>
</tr>
<tr>
<td>(2) -3.0</td>
<td>Stationary noise</td>
<td>-5.81</td>
<td>-5.08</td>
<td>-1.74</td>
<td>-6.77</td>
</tr>
<tr>
<td></td>
<td>Non-stationary + stationary noise</td>
<td>-4.61</td>
<td>-4.89</td>
<td>-4.44</td>
<td>-6.59</td>
</tr>
<tr>
<td></td>
<td>Total noise+speech</td>
<td>-4.84</td>
<td>-5.06</td>
<td>-4.78</td>
<td>-6.84</td>
</tr>
<tr>
<td>(3)</td>
<td>Stationary noise</td>
<td>-7.96</td>
<td>-4.95</td>
<td>-4.00</td>
<td>-6.69</td>
</tr>
<tr>
<td>-3.1</td>
<td>Non-stationary + stationary noise</td>
<td>-6.30</td>
<td>-4.95</td>
<td>-5.28</td>
<td>-6.71</td>
</tr>
<tr>
<td></td>
<td>Total noise+speech</td>
<td>-6.71</td>
<td>-5.11</td>
<td>-5.57</td>
<td>-6.99</td>
</tr>
<tr>
<td>(4)</td>
<td>Stationary noise</td>
<td>-8.87</td>
<td>-4.52</td>
<td>-5.04</td>
<td>-5.96</td>
</tr>
<tr>
<td>-3.2</td>
<td>Non-stationary + stationary noise</td>
<td>-7.1</td>
<td>-4.76</td>
<td>-5.56</td>
<td>-6.41</td>
</tr>
<tr>
<td></td>
<td>Total noise+speech</td>
<td>-7.42</td>
<td>-4.92</td>
<td>-5.83</td>
<td>-6.69</td>
</tr>
</tbody>
</table>

Performance of conventional TE with incremental learning: It is also important to investigate the performance of these methods in the presence of speech. Final estimates after the 20th iteration can be seen in Figure 4.9(a) and Figure 4.9(b) representing TE with \(T \rightarrow 0\) and TE with \(T = 0.0001\), respectively. The smoothness of the spectrum in Figure 4.9(b) reflects the more accurate estimation results as given in Table 4.5. In conditions (1) and (2), TE \((T = 0.0001)\) performed better than other methods for any given time interval, surprisingly even in the case of stationary noise, where HRLE is more suitable to apply. The reason is that the conditions in (1) and (2) are the same with the database. However, unfamiliar conditions such as in (3) and (4) degrade the accuracy of the TE because the portion of the stationary noise, compared to the ego noise, becomes more dominant in the overall noise energy. We can clearly see
the tendency of HRLE outperforming TE \((T = 0.0001)\) gradually in the conditions of slightly increased BGN power. This justifies the usage of the concatenated processing of HRLE and TE (referred as HRLE+TE) as proposed in Figure 4.2 to compensate the deteriorating performance of TE. An alternative way to deal with the changing acoustic conditions is letting the templates adapt to those conditions entirely using the incremental update mechanism, but it may take a long time and the improvements are in rather small steps during the adaptation process with a high \(\eta\), which makes this option less feasible.

Figure 4.10 shows the distribution of the estimation error (NNEE) over time. Apparently, TE \((T = 0.0001)\) demonstrated the least estimation error even in the first two seconds of stationary noise. We also analyze the distribution of NNEE over the frequency bins (see Figure 4.11). One important advantage of TE is that the error is almost evenly distributed to all frequencies. In contrast, the spectral distortion of the noise estimate provided by semi-blind source separation based on internal NAM microphones [137] is rather large at low frequencies (see also Figure 5 in [137]), which are known to heavily contain acoustic features of speech.
4.4. EVALUATION

Table 4.6: Ego noise reduction performance for all methods

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR(1) = 3dB</th>
<th>SNR(2) = -3dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNR</td>
<td>LSD</td>
</tr>
<tr>
<td>NP</td>
<td>3.00</td>
<td>9.7</td>
</tr>
<tr>
<td>HRLE</td>
<td>3.96</td>
<td>8.94</td>
</tr>
<tr>
<td>TE (T → 0)</td>
<td>5.49</td>
<td>8.51</td>
</tr>
<tr>
<td>TE (T → ∞)</td>
<td>4.92</td>
<td>8.20</td>
</tr>
<tr>
<td>TE (T = 0.0001)</td>
<td>5.24</td>
<td>8.03</td>
</tr>
<tr>
<td>HRLE+TE (T → 0)</td>
<td>6.02</td>
<td>8.66</td>
</tr>
<tr>
<td>HRLE+TE (T → ∞)</td>
<td>5.01</td>
<td>8.36</td>
</tr>
<tr>
<td>HRLE+TE (T = 0.0001)</td>
<td>5.46</td>
<td>8.20</td>
</tr>
</tbody>
</table>

Table 4.7: Ego noise reduction performance for all methods

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR(3) = -3.1dB</th>
<th>SNR(4) = -3.2dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNR</td>
<td>LSD</td>
</tr>
<tr>
<td>NP</td>
<td>-3.1</td>
<td>12.6</td>
</tr>
<tr>
<td>HRLE</td>
<td>-0.75</td>
<td>10.0</td>
</tr>
<tr>
<td>TE (T → 0)</td>
<td>1.97</td>
<td>9.75</td>
</tr>
<tr>
<td>TE (T → ∞)</td>
<td>2.42</td>
<td>8.90</td>
</tr>
<tr>
<td>TE (T = 0.0001)</td>
<td>1.97</td>
<td>9.02</td>
</tr>
<tr>
<td>HRLE+TE (T → 0)</td>
<td>2.63</td>
<td>9.03</td>
</tr>
<tr>
<td>HRLE+TE (T → ∞)</td>
<td>2.59</td>
<td>8.69</td>
</tr>
<tr>
<td>HRLE+TE (T = 0.0001)</td>
<td>2.61</td>
<td><strong>8.66</strong></td>
</tr>
</tbody>
</table>

Performance of proposed noise estimation framework with incremental learning: Finally, we evaluate the noise reduction performance by using the system depicted in Figure 4.3. As Table 4.6 demonstrates, TE-based SS with incremental learning achieves the smallest LSD and largest WCRs among all methods for the trained conditions (1) and (2). Besides, HRLE+TE with incremental learning attains the second-best results to TE, which can allow us to make a small compromise between the best performance and adaptivity of the noise estimation system. The results for the simulation of changing ambient noise (conditions (3) and (4)) are shown in Table 4.7. We observe that the higher the portion of the background noise, the more effective HRLE method gets. Under these conditions, HRLE contributes more to cancelling the overall noise by
eliminating the background noise. Hence in its combination with TE, TE deals only with the non-stationary part of the overall noise regarding the ego-motion noise. This kind of configuration increases the robustness of the noise suppression system and makes it independent of any change in the environmental noise condition. In terms of SNR, HRLE+TE with incremental learning is only outperformed by HRLE+TE with $T \to 0$.

4.5 Summary

Since a speech enhancement system solely based on TE cannot cope with the changing environmental noise or an single-channel noise reduction system is unable to eliminate non-stationary noise, we proposed to concatenate the two stages to obtain a unified preprocessing framework for a robot audition system. In the first part of this chapter we assessed the performance of MCRA/HRLE-based noise reduction methods in the presence of background noise and ego noise, and presented a method exhibiting high robustness even against changing conditions of the environment. Another contribution was that it provides the underlying basis and configuration of further research advancement in online learning of ego-motion noise templates.

In the second part, we proposed an online and adaptive learning mechanism for an adaptive ego noise estimation framework consisting of stationary noise estimation (HRLE) and non-stationary noise estimation (TE) in series. We designed a learning mechanism provided with a measure of performance, which enables the learning process to continue over the entire lifespan of a robot without human intervention (also termed lifelong learning [3]). This Incremental Learning (IL) algorithm is a natural extension of the parameterized template generation method and allows the robot to learn its ego noise not only in isolated training conditions, but also in daily environments (possibly even in the presence of humans). Moreover, it allows us to tackle the curse of dimensionality problem caused by the large number of DoF of a robot. We assessed the learning, estimation and suppression performance of this template-based single-channel noise reduction method in the presence of background noise and ego-motion noise. We showed that the online learning contributes to precise estimation of overall noise and a high ASR accuracy under various SNR conditions.
Chapter 5

Whole Body Motion Noise Reduction of a Robot

5.1 Introduction

The problem with using two consecutive nonlinear noise reduction operations (stationary noise reduction + template subtraction) is that they produce even more musical noise, eventually damaging the acoustic features and reducing the recognition performance of ASR. To compensate for this effect, we extend the single microphone-based template subtraction method to a hybrid system: (1) a multi-channel noise reduction block consisting of sound source localization (SSL), sound source separation (SSS), and speech enhancement (SE); and (2) the previously mentioned template subtraction block. While spectral enhancement techniques are the most suitable way to deal with diffuse noise, source separation improves the signal-to-noise ratio (SNR) of the noisy signal by removing directional noise components of ego noise from speech. In this respect, we want to integrate all the above-mentioned speech processing methods into a single framework to perform ego noise cancellation. Furthermore, we propose an original strategy to solve the whole body motion noise problem of a robot.

5.2 Issues on Whole Body Motion Noise

Our goal was to develop a robot audition system that cancels whole body motion noise of a robot, thus improving the recognition performance. To be able to develop this system, we had to deal with two issues:
5.2. ISSUES ON WHOLE BODY MOTION NOISE

1. Automatic speech recognition in the presence of ego-motion noise:

Conventional ASR systems assume clean inputs of speech signals. Therefore, a speech signal mixed with ego-motion noise must be refined due to the distorting effects of the noise. If the ego-motion noise is not suppressed, there would be a mismatch between noisy signals and the trained acoustic models resulting in degraded speech recognition performance.

2. Applicability of ego-motion noise reduction to the whole body motion of a robot:

The robot is expected to use different parts of its body to accomplish certain tasks, including loco-motion, object manipulation, and object tracking. Based on the complexity of each motion or behavior, the robot may perform several tasks at one time or a task may involve motions of several body parts of the robot, thus involving several joints at one time. We had to confirm whether ego-motion noise could be suppressed, regardless of the body part generating the noise. Furthermore, it was necessary to suppress the whole body motion noise of a robot using a single framework.

We dealt with the above-mentioned issues using the following approaches:

1. Ego-motion noise suppression: We integrated two different methods of ego-motion noise suppression: (1) a template-based ego-motion noise reduction method with spectral enhancement parameters, and (2) a multi-channel noise reduction chain with SSL, SSS and SE stages. The former is suitable in dealing with ego-motion noise especially. The latter method, however, is effective in suppressing directional sound sources. Since the noise originates from the motors that move relative to the positions of the microphones, the noise can be considered directional. Both techniques cannot only be applied individually, but can be merged into a hybrid system. The output of the noise suppression was used as the input of consequent audio processing stages for various purposes. In this study, we were especially interested in one particular application: Automatic speech recognition.

2. Motion type-based selective ASR module: Instead of tackling the whole body motion noise problem holistically, we utilized a synthesis-by-analysis approach. We divided the whole body motion noise problem mainly into three ego-motion noise categories: arm, leg and head motion noise, depending on the spatial locations of each relative to the microphones and their intensity levels. Figure 5.1
illustrates three spectrograms of corresponding limb motions, showing that they clearly differed in loudness recorded with the microphone array located on top of the robot head (see Section 5.4 for details about the system). We also observed an uneven energy distribution, with higher noise energy at lower frequencies.

![Spectrograms of three types of motor noise](image)

Figure 5.1: Typical spectrograms of three types of motor noise

These domains are distinctive in being the three main body parts of a robot with end effectors. To relocate one end effector, it is usually necessary to relocate a series of joints connected to the end effector. Therefore, the accumulated noise of the joints can be considered as originating from a certain area. In addition, some tasks can be performed without needing to use all body joints at the same time. Tasks such as the coordination of body dynamics, sensor processing and perceptual understanding of the environment can become complicated and difficult. Since the immobility of distinct body parts unrelated to the task is always desired to improve the reliability of accomplishing the task, we provide individual solutions for all three types of noise. We also assessed the performance of ASR for all three ego-motion noises and their combinations. After presenting the results of our analyses, we propose a final architecture that can deal with all types of ego-motions and their combinations (whole body motion), using a motion type-based selective ASR module with a switching mechanism for ASR inputs.

### 5.3 Hybrid Framework

In this section, we describe the basic building blocks of the proposed system architecture and illustrate, how single-channel template subtraction can be extended into a hybrid framework to suppress ego noise by incorporating existing multi-channel noise reduction
techniques. Please note that the modules briefly described in Section 5.3.1 can be replaced by other multi-channel solutions capable of separating directional sound sources, thus they are rather generic in that sense. In Section 5.3.2, we explain the key module of our proposed architecture that results in the hybrid system.

5.3.1 Multi-channel Noise Reduction System

To estimate the Directions of Arrival (DoA) of each sound source, we used a popular adaptive beamforming algorithm called MUltiple Signal Classification (MUSIC) as described in Section 2.3.4. This algorithm detects each DoA by performing eigenvalue decomposition on the correlation matrix of the noisy signal, by separating subspaces of undesired interfering sources and sound sources of interest, and finally by identifying the peaks occurring in the spatial spectrum.

Geometric Source Separation (see Section 2.3.3) makes explicit use of source locations to separate different sound sources. Specifically, GSS has three distinct advantages for the ego-noise cancellation problem:

1. The introduction of the concept of geometric constraints, which involves calculations of current transfer functions based on the known locations of the microphones and the positions of the sound sources obtained from SSL. This relaxes the limitations of BSS, such as permutation and scaling problems, and can therefore run in real-time.

2. Sound separation of moving sources is possible. This is especially important since the part of the robot on which the microphones are mounted (e.g. the head) can also move. Relative to a moving microphone array, even stationary sound sources are regarded as moving objects.

3. Generally, an embodied robot has loud ego noises, such as the stationary operational noise of hardware and fan noise, which are located close to each other. If the positions of these high noise emission sources are known, their directions can be specified, because our GSS module has a function that suppresses stationary ego noise as a fixed noise source.

The separation process is followed by a multi-channel post filtering operation, enabling the sounds to be enhanced further (see Section 2.3.4). It can cope with stationary
background noise, as well as the stationary hardware and fan noise of the robot. The configuration of the modules is such as illustrated in Figure 2.10.

5.3.2 Switching Module for ASR Input Selection

After initially analyzing the performances of (1) multi-channel and (2) single-channel noise reduction methods (see Section 5.5), in the synthesis stage, we suggested processing the speech feature outputs of both pathways in a single ASR module, and to use them interchangeably in a motion-dependent fashion, as in Figure 5.2. This switching module is triggered by the output of the motion detector. Because this system gathers information about all joints at every moment of time, the switching module is able to discriminate among joints that are and are not actively involved in the motion by checking their velocities and can therefore determine the motion being performed at that moment. The module then switches between the outputs of single and multi-channel noise reduction based speech features. Because the multi-channel approach works better than single-channel approach for the leg and arm noises (Section 5.5.2), the switch feeds the acoustic features of this branch to the ASR module whenever a leg and/or an arm motion is detected. It also utilizes the acoustic model trained for the multi-channel approach. For a head motion, however, features generated after template subtraction are more suitable (Section 5.5.3). We observed similar recognition performance during simultaneous head and arm motion noises (Section 5.5.3), indicating that head motion contributes to whole body motion noise more than any other motion from other domains. In summary, as long as head motion is present, we can suppress whole body motion noise using \( \text{➀ the single-channel noise reduction methods} \); otherwise \( \text{➁ the multi-channel noise reduction method} \) is used. We implemented the following rule-based routing in the switch:

\[
Decision(k) = \begin{cases} 
\text{Acoustic features of ➀, if any } |\dot{\theta}_{\text{HeadJoint}}(k)| > \epsilon, \\
\text{Acoustic features of ➁, otherwise}
\end{cases}
\]

(5.1)

where \( |\dot{\theta}_{\text{HeadJoint}}(k)| \) denotes the absolute velocity of the pan or tilt motion of the head and \( \epsilon \) is a speed threshold. We proposed to use \( \epsilon \), instead of zero, to prevent the activation of the switch during the tail motion of the head; it is used as a countermeasure to situations during which motion has stopped, but the joint sensors continue to send very small differences in position.
5.4 System Architecture and Implementation

The overall architecture of the proposed noise reduction system consisting of three blocks is shown in Figure 5.2. The multi-channel noise reduction block and the single-channel template subtraction run in parallel. Together, both blocks are responsible for producing spectrograms for the extraction of audio features in the last block, which performs automatic speech recognition. The whole system (including joint status acquisition, sound recording and sound processing stages) runs synchronously based on a single clock. Data flow is realized mainly by means of fixed-length audio frames.

This system was specifically designed for single-speaker speech recognition tasks. This means that we have assumed that there is no directional sound source other than the speaker, whose speech would be recognized. Therefore, external noise is considered only as background noise, which is a diffuse noise by its nature. Both branches of the hybrid system can deal with diffuse noise by utilizing either background noise reduction (single-channel template subtraction) or MCRA inside SE (multi-channel noise reduction). To tackle the motor noise, on the other hand, we use either template subtraction (single-channel noise reduction for head motion) or SSS (multi-channel noise reduction for arm and leg motion). Thus, both branches can deal with internal and external noise in their own ways.

![Diagram of proposed noise cancellation system]

Figure 5.2: Proposed noise cancellation system
CHAPTER 5. WHOLE BODY MOTION NOISE REDUCTION OF A ROBOT

5.5 Evaluation

To evaluate the performance of the proposed techniques, we used a humanoid robot developed by Honda. This robot is equipped with an 8-channel microphone array on top of its head. We used 2 motors for head motion, 5 motors to move each leg, and 4 motors to move each arm, resulting in a total of 20 degrees of freedom. Relative to the microphone array configuration, the neck motors are the closest sound sources, making them the most problematic, because the intensity of a sound wave depends on its distance from its source:

\[
\text{SoundIntensity} = \frac{\text{SoundPower}}{(4\pi R^2)},
\]

where \( R \) denotes the distance. In addition, since all limbs of the robot operate in different, non-overlapping coverage areas, which also help in differentiating noise types by their spatial locations, we decided to handle the noise problem in different domains, each covering a set of joints required for a certain type of interaction with the robot’s environment. We therefore recorded motions performed by a given set of limbs, which could be classified into 3 distinct categories, arm motion, leg motion and head motion, in order of increasing noise intensity.

5.5.1 Experimental Settings

We recorded 1) random, whole-arm pointing motion within the reaching space of the robot body as arm motion, 2) stamping behavior and short distance walking as leg motion, and 3) random head rotation (elevation=[-30° 30°], azimuth=[-90° 90°]) as head motion. The average noise energies of leg and head motion were 5.1dB and 8.4dB higher, respectively, compared with those of arm motions. For the second part of the experiments involving template subtraction, we recorded two additional sets of random motions (performed by the head only and by the head and arm together) and stored a training database of 30 minutes and a test database 10 minutes long. (Due to software constraints the joint positions of the legs cannot be acquired, therefore we could not apply the template subtraction method to leg motion noise.) Sensors determine the angle of the joints every 5 msec, with each audio frame being 10 msec in length. We used constant values for \( \alpha=1 \) and \( \beta=0.5 \) as template subtraction parameters, because we previously observed that, compared with \( \beta=0 \), increased \( \beta \) improves ASR accuracy considerably. (For detailed evaluations regarding \( \alpha \) and \( \beta \), their effects on ASR accuracy,
signal quality and noise suppression rates, see [136].)

Because the noise recordings are longer than the utterances used in isolated word recognition, we selected those segments in which all joints contributed to noise. To generate precise amounts of noise and speech energy for various SNR conditions before mixing them, we amplified clean speech based on its segmental SNR, segSNR. The segSNR estimates the SNR-level within each segment and averages it over the entire signal providing a better representation of energy distribution for speech and noise within the relevant time interval under consideration.

\[
\text{segSNR} = \frac{1}{J} \sum_{j=1}^{J} 10 \log_{10} \left( \frac{\sum_{n} s_j^2(n)}{\sum_{n} d_j^2(n)} \right),
\]

where J is the number of segments with speech activity, and \(s(n)\) and \(d(n)\) are the \(n\)-th discrete speech and noise samples, respectively. The noise signal, consisting of ego noise (including ego-motion noise) and environmental background noise, was mixed with clean speech utterances used in a typical human-robot interactive dialog. This Japanese word dataset includes 236 words for 4 female and 4 male speakers. Acoustic models were trained with the Japanese Newspaper Article Sentences (JNAS) corpus, 60 hours of speech data spoken by 306 male and female speakers, making speech recognition a word-open test. The results for template subtraction (TS) were evaluated using an acoustic model trained with MCRA-applied speech data. In contrast, we used a matched acoustic model for multi-channel noise reduction (GSS+PF) methods. Both of these models were trained with data processed at motor noise conditions of SNR levels ranging from -10 to 5 dB. We used 13 static Mel-Scale Log Spectrum (MSLS) [138] features, 13 delta MSLS features and 1 delta power feature. Speech recognition results are given as average word correct rates (WCR) of instances from the noisy test set. The position of the speaker was kept fixed at 0° throughout the experiments. The recording environment consisted of a 4.0 m × 7.0 m × 3.0 m room with a reverberation time (RT_{20}) of 0.2sec. The implementation was run on HARK, an open-sourced software for robot audition [132].

### 5.5.2 Speech Recognition with Arm and Leg Motion Noise

In this experimental setting, the microphone array and the head were kept stationary, allowing us to fix the direction of the ego noise (fan noise) originating from the backpack of the robot at -180°. Providing a fixed ego-noise direction did not pose any hard constraints on robot audition scenarios or applications, because the robot was already
equipped with sensors to transmit the positions of the joints. Depending on the posture of the body, we were able to determine exactly the source of the ego noise and transmit the direction automatically to our source separation algorithm as input. We present the results for GSS and GSS+PF where the position of the speaker was detected using our implementation. As an additional test, we also determined “GSS+PF with known source location” results, a condition where we assumed that the location of the sound source was estimated precisely.

Figure 5.3: Recognition performance during arm and leg motions

Figure 5.3 shows speech recognition accuracies for arm, leg and arm & leg motions at the same time. Single-channel results without processing were used as baseline. Template subtraction resulted in good ASR accuracy, but its performance was inferior to that of GSS+PF (TS evaluation of leg motion was not possible). Under all three conditions, multi-channel noise reduction system resulted in an up to 40 points improvement compared with single-microphone based recognition. In general, these results indicate that the directional effects of arm and leg motions noise can be treated with GSS and that residual noise (as diffuse components) can be partially handled by PF. Because the arms operate mostly on the right- and left-hand sides of our humanoid robot, their noises can be separated well due to the spatial (angular for GSS) distance between the arms and the target speaker standing in front of the robot. In addition, the leg noise came from below the waist of the robot, making its distance from the microphone array large enough for separating it from the speaker. As long as the direction of the ego-motion noise is not the same as that of the target speaker, this method works well to suppress all ego noise, both ego-motion noise and fan noise. Furthermore, the recognition result curves in Figure 5.3(c) show very similar patterns to the curves in Figure 5.3(a) and
5.5. EVALUATION

Figure 5.3(b). This very promising result indicates that GSS+PF is effective even when it is used against the combination of arm and leg motion noise.

It is also noteworthy to mention that SSL fails in low SNRs due to its fixed threshold operation. SSL estimates additional non-existing “ghost” sources, decreasing the performance of GSS and PF. In contrast, GSS+PF with known source location demonstrates the upper performance limit of our proposed method.

5.5.3 Speech Recognition with Head Motion Noise

One consequence of head motion is the relative motion of sound sources with respect to the microphones. Whenever the head moves, the microphone array also moves. Since we tested only isolated word recognition, we hypothesized that the effects of the moving sound sources on separation and speech enhancement performance were rather small, but in fact not negligible. Nevertheless, to inspect the capabilities of our proposed noise reduction system based on SSS, we did not provide the ego-noise direction of the robot in advance; rather, the SSL system predicted it automatically.

![Graphs showing recognition performance during head motion](image)

Figure 5.4: Recognition performance during head motion

Figure 5.4(a) illustrates the ASR accuracy for head motion noise. The multi-channel approach provided poorer performance than the single-channel template subtraction technique, because short range reverberation effects and multipath propagation inside/outside the head are properties of head-motion noise that are very hard to overcome.
with the current GSS+PF algorithm limits and settings. The neck motors are located inside the head cover, where the microphones are also installed. Because head motor noise propagates inside the head in a highly reverberant way in close proximity to the microphones, the directional noise assumption is violated. Strong noise sources in the very near field of the microphone array have highly complicated propagation patterns. As a consequence, it worsens the separation quality. Thus, the noise model used in the post filtering is not applicable under these conditions. TS resulted in better improvement, because it does not model the noise depending on its directivity-diffuseness nature, but rather instantaneously predicts the current noise template from a database, depending on the position and velocity of the joints. In addition to being prone to modeling errors, it also suffers from musical noise components caused by subtraction in the spectral domain. This distorts the spectrum and degrades features, making the WCR improvement rather limited, but still better than GSS+PF.

Table 5.1: Recognition accuracy (%) for different ego-motions achieved by single-channel template subtraction and multi-channel noise reduction ($SNR=-5dB$)

<table>
<thead>
<tr>
<th></th>
<th>Arm</th>
<th>Leg</th>
<th>Arm+Leg</th>
<th>Head</th>
<th>Head+Arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Channel</td>
<td>73</td>
<td>50</td>
<td>58</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>Template Subtraction</td>
<td>80</td>
<td>-</td>
<td>-</td>
<td>69</td>
<td>69</td>
</tr>
<tr>
<td>Multi-ch. Noise Reduction</td>
<td>96</td>
<td>90</td>
<td>90</td>
<td>53</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 5.1 shows that the proposed hybrid system highly improves the elimination of all motor noise types and their combinations within the limits of our implementation conditions. Although this table presents results for only one moderate SNR level of $-5dB$, the trend of improvement was true for all SNRs used in these experiments. Our switching module always selects the noise suppression method that yields the best performance for the undertaken action.

Finally, to assess the suppression capabilities of both methods, we also assessed SNR improvement by calculating the differences in SNR before and after the application of noise cancellation methods. Mean SNR improvements are shown in Figure 5.5. For head motion noise the template subtraction method resulted in higher suppression performance than GSS and GSS+PF for head motion noise, whereas, GSS+PF reduces the largest portion of arm motion noise. Although the trends of SNR improvements were consistent with the WCR curves, high suppression rates do not necessarily indicate
higher recognition accuracy for TS as long as noise templates are not correctly estimated.

![Graphs showing SNR improvements of compared methods](image)

(a) during arm motion noise  
(b) during head motion noise

Figure 5.5: SNR improvements of compared methods

5.5.4 Discussion and Future Work

The optimal system structure for ego noise suppression depends solely on the characteristics of the ego noise, as long as the ego noise is not picked up by additional sensors and must be estimated. We found that single-channel noise reduction was far superior to multi-channel noise reduction in suppressing motor noise recorded by closely located microphones. Because of the complex characteristics of motor noise propagation inside the robot cover, where the microphones are mounted, blind source separation and speech enhancement perform very poorly in these conditions (i.e. head motion noise). In contrast, when the motors are located further from the speaker and microphones, both in distance and separation angle, the multi-channel noise reduction was more effective (i.e. arm and leg motion noise). Therefore, the proposed parallel system architecture can be considered optimal for any robotic system containing only robot-embedded microphones and with the switch trigger design based on the locations of the microphones and motors.

We also investigated alternative combinations, including a cascaded version of SSS+PF+TS, instead of our hybrid architecture. In practice, however, it was not possible to create a template database for ego noise after the SSS+PF stages. One reason is that the recording must be performed only when there is no external directional sound source, but performing an SSS without any sound sources is not possible. In addition,
the template database must contain ego noise artifacts after SSS+PF for all directions in which a candidate target sound source may be present. It is almost impossible to create such a huge template database. We also evaluated this combination in reverse order, Pre-filtering+TS+SSS, but it yielded far worse results, because the spectral subtraction prior to SSS damaged the spectra of the microphone signals, resulting in poorer performance of the SSS.

By changing the posture of the robot, so that its body is aimed directly at the target speaker, the robot can avoid the interference due to the ego-motion noise of the arms with the target speaker’s utterances by maximizing their spatial distance. Our system remains open for improvements. To apply template subtraction to leg noise, we plan to make changes that will allow us to gather angular information from the legs. One weakness of the current architecture is the fixed threshold operation used in the sound source localization procedure, which determines if a source is present at that location. As motor noise increases, the system becomes more susceptible to the threshold value. Since no optimal threshold is effective for every kind of motor noise, we plan to make it adaptive. Our multi-channel system in its current form can also deal with 4 speakers. The next step is to design a system in real time and in a real situation involving speech recognition of several speakers simultaneously while the robot is performing some motion.

5.6 Summary

We have described methods for eliminating whole-body motion noise from speech signals. Since ego-motion noises arising from the motors of a robot are created in the near field of the microphone array, these noises have both diffuse and directional characteristics. We used a synthesis-by-analysis approach to suppress this noise. We divided whole body motion noise into three domains, depending on their spatial location and intensity levels: arm, leg and head motion noise. The system we proposed extracts information about the motion type performed at that moment and decides on the best choice of processing method for speech recognition by a selective ASR module. We adopted two methods, multi-channel noise reduction and single-channel template subtraction, which are switched depending on the detection of the head motion. If no head motion is detected, the first method is selected because it is effective for arm, leg, and arm&leg motion noises. This method utilizes sound source localization incorporating the MUSIC algorithm and sound source separation using the GSS algorithm, finalized
by a speech enhancement stage that suppresses both background noise and interference/leakage noise. Source separation is particularly effective against noises from the arms and legs, because the limbs are located away from the microphone array and are separated from a speaker standing directly in front of the robot. On the other hand, if head motion is detected, the second method is selected. It is more appropriate for canceling head motion noise (or the combination of the head motion noise with arm and/or leg motion noise), because template subtraction makes no assumptions about the nature of the noise and uses previously recorded noises. We validated the applicability of our approach by evaluating its performance on three different motor noise types and their combinations. Our method demonstrated good performance in suppressing arm, leg and head motion noise and their combinations, as shown by ASR accuracy.
Chapter 6

MFT-based Integration of Ego Noise Reduction and ASR

6.1 Introduction

Strategies, which make use of a confidence-based weighting of the time-frequency representation of audio signals, can enhance the quality of speech. We basically apply a mask to decrease the contribution of unreliable parts of distorted speech [102]. Retaining the reliable parameters that are essential for speech recognition results in a substantial increase in recognition accuracy. In this section, we will discuss the basic steps of such an ASR system and how this approach can be adapted to the ego noise problem by presenting a robust mask design method for estimating the reliability of speech based on current motor noise. We aim to estimate the reliability of the features of speech, which are subject to residuals of motor noise after template subtraction. To improve the performance of ASR, we propose to use MFT with a reliability estimation model based on ego-motion noise predictions. A filtering operation applied to missing or damaged acoustic features help us to solve the ego noise problem of a robot at a higher level for an ASR application. The proposed solutions are twofold based on the problem domain: single talker ASR and multi talker ASR.

6.2 MFT-ASR System for a Single Talker

Ego noise estimation can contribute to ASR in the level of robust feature extraction as well as unreliable feature masking. This section focuses on (1) further improvement
6.2. MFT-ASR SYSTEM FOR A SINGLE TALKER

6.2.1 System Overview

The overall architecture of the proposed noise reduction system is shown in Figure 6.1. The first block of the processing chain, composed of elements for performing SSL, SSS and SE constitute the \textit{multi-channel noise reduction} block [139], while a second block performs \textit{template subtraction}. These two modules together are responsible for the \textit{audio features} of speech recognition and \textit{spectrograms} that will be further processed in the MFM generation stage. Finally, a new third block, \textit{MFT-based speech recognition}, designed to achieve a more robust ASR, uses both the features and spectrograms created in the pre-processing stages to extract the most suitable features.

6.2.2 Selective White Noise Superimposition

Because it is impractical to create matched models for each ego-motion noise, we add white noise with a fixed amplitude value as a known noise source during the training phase. The second advantage of using white noise is that it blurs the \textit{musical noise} distortions caused by the spectral subtraction of the PF. Because the artifacts of the louder motor noise are more harmful compared to the artifacts of less noisier motors, we propose a switching mechanism for white noise level adjustment inside the noise superimposition module. The mechanism performs a decision between two white noise levels \{\(C_1, C_2\}\}, which is triggered by the motion predictor.

This method is scalable according to the physical conditions regarding microphones, motors, their distances and properties. We propose to implement the following rule-
based routing in the switch:

\[
\rho(l) = \begin{cases} 
C_1, & \text{if any } |\hat{\theta}_{\text{LoudJoints}}(l)| > \epsilon \\
C_2, & \text{otherwise}
\end{cases},
\]

(6.1)

where \( \rho \) [dB] represents white noise magnitude relative to clean speech magnitude, \( |\hat{\theta}_{\text{LoudJoints}}(l)| \) denotes absolute velocity of the related joint and \( \epsilon \) is a certain speed value. \( \epsilon \), instead of zero, is used to prevent the activation of the switch when the motion has stopped, but the joint sensors still send very small position differences. Motion detection is compromised by a high \( \epsilon \) value. Please note that the additive white noise will be cancelled out in the spectral mean normalization module of ASR.

### 6.2.3 Missing Feature Mask Generation

We compute the reliability of features for each frame and for each mel-frequency band. Masks composed of continuous values between 0 and 1 are called soft masks, whereas masks composed of only discrete values, either 0 or 1, are called hard masks.

GSS lacks the ability to identify and suppress motor noise originating from the same direction as the speaker, because it regards the noise as part of the speech. Moreover, if the position of the noise source is not detected precisely, GSS cannot separate the sound in the spatial domain. Thus, small amounts of motor noise can spread to the separated sound sources. However, multi-channel noise suppression systems work very well for weaker motion noise, such as arm or leg motions, when compared with head motion noise, as we demonstrated in the previous chapter. In addition, the multi-channel system was optimally designed for “simultaneous multiple speakers” scenarios with background noise and demonstrated a very good performance when no motor noise was present. In contrast, template subtraction makes no assumptions about the directivity or diffuseness of the sound source and can match a pre-recorded template of the motor noise at any moment. The drawback of this approach, however, is that due to its not being stationary, the characteristics of predicted and actual noise may differ to some extent.

As described, the two approaches have distinct strengths and weaknesses and thus may be used in a complementary fashion. A speech feature is considered unreliable if the difference between the energies of refined speech signals generated by multi-channel and single-channel noise reduction systems is above a threshold \( T \). The masks are computed for each frame, \( l \), and for each mel-frequency band, \( f \). First, a continuous mask is
calculated as:

\[ m(f,l) = \frac{||\hat{X}_m(f,l)||^2 - ||\hat{X}_s(f,l)||^2}{||\hat{X}_m(f,l)||^2 + ||\hat{X}_s(f,l)||^2} \]

(6.2)

where \( ||\hat{X}_m(f,l)||^2 \) and \( ||\hat{X}_s(f,l)||^2 \) are the estimated energy of the refined speech signals \( \hat{X}(k,l) \), following multi-channel noise reduction and single-channel template subtraction, respectively. Both signals are computed using a mel-scale filterbank. The numerator represents the deviation of the two outputs, a measure of their uncertainty or unreliability. The denominator, however, is a scaling constant and is the average of the two estimated signals. (To simplify the equation, we removed the scalar values from the denominator, so that \( m(f,l) \) can take on values between 0 and 1). Thereby, the reliability can be defined by \( \frac{1}{m(f,l)} \), which means the smaller the \( m(f,l) \), the higher the reliability and vice versa. Depending on the type of mask (hard or soft) used in the MFT-ASR, Equation (6.3) or Equation (6.4) is selected.

1. For hard (binary) masks:

\[
M(f,l) = \begin{cases} 
1, & \text{if } m(f,l) < T \\
0, & \text{if } m(f,l) \geq T
\end{cases}
\]

(6.3)

2. For soft masks [113]:

\[
M(f,l) = \begin{cases} 
\frac{1}{1 + \exp(-\sigma(m(f,l) - T))}, & \text{if } m(f,l) < T \\
0, & \text{if } m(f,l) \geq T
\end{cases}
\]

(6.4)

where \( \sigma \) is the tilt value of a sigmoid weighting function.

### 6.3 MFT-ASR System for Multiple Talkers

In Section 6.2, we proposed simultaneous usage of (1) single-channel template subtraction and (2) multi-channel noise suppression to improve the ASR accuracy by exerting MFT and designing a mask based on the similarity of the refined signals at the output of the two noise suppression methods ((1) and (2)). However, this mask was vulnerable to the distortions caused by the residuals of motor noise due to incorrect/overestimated
CHAPTER 6. MFT-BASED INTEGRATION OF EGO NOISE REDUCTION AND ASR

noise estimations. Besides, the system was only able to deal with one speaker at a time. In this section, we aim to eliminate the artifacts of musical noise by disposing template subtraction from our system and to extend the speech recognition system so that it can recognize multiple speakers at the same time while the robot is moving.

In the following sections, we introduce the architecture of the overall system (Section 6.3.1), discuss two reliability estimation techniques, one designed especially against speaker separation artifacts (Section 6.3.2) and one for the purpose of eliminating ego-motion noise (Section 6.3.3). They are followed by mask generation algorithms (Section 6.3.4) and proposed method for the integration of the two masks (Section 6.3.5). In general, the masking operation can be considered as a confidence-based weighting of the time-frequency representation of audio signals, therefore the masks and acoustic features must be provided to MFT-ASR simultaneously.

6.3.1 System Overview

![Diagram of Proposed Multi-talker Speech Recognition System]

Figure 6.2: Proposed multi-talker speech recognition system

The overall architecture of the proposed system is shown in Figure 6.2. The system is equivalent to the one used in Figure 6.1 in terms of multi-channel noise reduction elements. We exclude, however, the ego noise suppression block and utilize only the ego noise prediction block. The details of the MFT-based Automatic Speech Recognition Block will be discussed in the following sections in detail. Note that by using a switching arrow the number of inputs of ASR sub-block is reduced to one for the acoustic features and MFM, which indicates that any talker’s utterance can be recognized. The talker
6.3. MFT-ASR SYSTEM FOR MULTIPLE TALKERS

selection can be triggered by a selective attention system that is out of scope of this thesis.

6.3.2 Reliability Estimation for Multiple Talkers

As mentioned in Section 2.3.4 the noise estimates in the post-filter can be decomposed into stationary (background noise) and transient (leakage energies of interfering sources) components for each source of interest. In order to predict the amount of noise present at a certain time in a certain frequency, Yamamoto et al. proposed a computation method for measuring the reliability as given like following [109]:

\[ m_m(f, l) = \frac{\left| \hat{X}(f, l) \right|^2 + \left| \hat{N}_s(f, l) \right|^2}{\left| \hat{X}_{in}(f, l) \right|^2} \]  

(6.5)

where \( |\hat{X}_{in}|^2 \) and \( |\hat{X}_{out}|^2 \) are respectively the post-filter input and output energy estimates for frame \( l \) and Mel-frequency band \( f \). \( \hat{N}_s(f, l) \) denotes the background noise estimate and \( m_m(f, l) \) gives a measure for the reliability based on multi-talker (\( m \)) effects.

6.3.3 Reliability Estimation for Ego Noise

There are several problems associated with GSS and PF based ego-motion noise reduction. First of all, GSS lacks the ability to catch motor noise originating from the same direction of the speaker and separate it, because the noise is considered as part of the speech in that case. Moreover, when the position of the noise source is not detected precisely, GSS cannot separate the sound in the spatial domain. As a consequence, motor noise can be spread to the separated sound sources in small portions. Apart from the directional portion, motor noise also has a diffuse portion that is caused by the highly reverberant propagation patterns of motor noise waves inside the body covers of the robot. Although diffuse noises are tackled by the post filter, the non-stationarity of the motor noise makes PF ineffective against ego-motion noise. Based on those drawbacks, we claim that it is impossible to remove the motor noise completely just by applying source separation or speech enhancement. However this multi-channel noise reduction chain (GSS+PF) is optimally designed for "simultaneous multiple speakers" scenarios with background noise and demonstrates a very good performance when no motor noise is present. Therefore, we plan to support our ASR with a probabilistic framework de-
signed for reliability estimation, MFT-ASR.

We base our reliability measurements for ego noise on the retrieved templates during the motion. In contrary to speech separation and enhancement block (cf. Figure 6.2), ego noise prediction block does not make any assumption about the directivity or diffuseness of the sound source and can correlate the incoming noisy signal to the retrieved template in a frame-by-frame basis. For now, we make the simplifying assumption that the additive motor noise is distributed uniformly among the existing sound sources. Therefore, we divide the noise energy by the number of sources. The ratio of the template (estimated noise) energy and the noisy signal energy of interest yields a reliability measurement about whether the corresponding frequency bin is strongly contaminated by ego-motion noise or not. A continuous measurement for the reliability based on ego noise ($e$) effects, $m_e(f,k)$, is calculated like following:

$$m_e(f,l) = 1 - \min\left(1, \frac{|\hat{N}^n(f,l)|^2}{S \cdot |\hat{X}_{out}(f,l)|^2}\right),$$  

(6.6)

where $|\hat{N}^n(f,l)|^2$ is the noise energy of the template and $S$ represents the number of speakers. To make $m_e(f,l)$ and $m_m(f,l)$ value ranges consistent, the possible values that it can take are limited between 0 and 1. This formula suggests that if high motor noise ($|\hat{N}^n(f,l)|^2$) is estimated, the reliability is zero, whereas low motor noise sets $m_e(f,l)$ close to 1.

### 6.3.4 MFM Generation

The reliability of features is computed for each frame and for each mel-frequency band. We adopted two mask generation mechanisms that 1) meet our needs and 2) are generic in the sense that they generate masks for both ego noise ($M_e$) and for multi-talker ($M_m$) by substituting $x$ with $e$ or $m$ in Equation (6.7) and (6.8).

1. Hard (binary) mask:

$$M_x(f,l) = \begin{cases} 
1, & \text{if } m_x(f,l) \geq T_x \\
0, & \text{if } m_x(f,l) < T_x
\end{cases}.$$  

(6.7)

2. Soft mask:
6.3. MFT-ASR SYSTEM FOR MULTIPLE TALKERS

\[
M_x(f,l) = \begin{cases} 
  1, & \text{if } m_x(f,l) \geq T_x \\
  \frac{1}{1 + \exp(-\sigma_x(m_x(f,l) - T_x))}, & \text{if } m_x(f,l) < T_x \\
  0, & \text{if } \left| \hat{X}_{\text{out}}(f,l) \right|^2 < T_{\text{mec}} \end{cases}
\]  
\hspace{1cm} (6.8)

where \( \sigma_x \) is the tilt value of a sigmoid weighting function and \( T_x \) is a predefined threshold. A speech feature is considered unreliable, if the reliability measure is below \( T_x \). In this thesis, we used both masks to assess their performance.

Additionally, we introduced a new heuristics-based concept called minimum energy criterion (mec) as in Equation (6.9). It is used to override the decision of the above formulae, if the energy of the noisy signal is smaller than a given threshold, \( T_{\text{mec}} \). It is used to avoid wrong estimations caused by computations performed with very low-energy signals, e.g. during pauses or silent moments.

\[
M_x(f,l)=0, \quad \text{if } \left| \hat{X}_{\text{out}}(f,l) \right|^2 < T_{\text{mec}}. 
\]  
\hspace{1cm} (6.9)

6.3.5 MFM Integration

Even if some portion of the motor noise is removed by the pre-processing stages (SSS and SE), the real ego-motion noise suppression is performed in the masking stage. As we have stated, the two masks in Section 6.3.2 and Section 6.3.3 serve two different purposes. Nevertheless, they can be used in a complementary fashion within the context of multi-talker speech recognition under ego-motion noise.

\[
M_{\text{tot}}(f,l) = w_m M_m(f,l) + w_e M_e(f,l), 
\]  
\hspace{1cm} (6.10)

where \( M_{\text{tot}}(f,l) \) is the total mask and \( w_x \) is the weight of the corresponding mask. This is a generic framework that allows the system designer 1) to weight each single mask individually and 2) to perform either an addition (OR) or a multiplication (AND) operation with soft (hard) masks for integration.
CHAPTER 6. MFT-BASED INTEGRATION OF EGO NOISE REDUCTION AND ASR

6.4 Evaluation

6.4.1 Evaluation of MFT-ASR for Single Talker

Experimental Settings

To evaluate the performance of the proposed techniques, we used a humanoid robot developed by Honda. This robot, which has many degrees of freedom, was equipped with an 8-channel microphone array on top of its head. We used 2 motors for head motion, and 4 motors for the motion of each arm, resulting in 10 degrees of freedom. Sensors recorded the angle of each joint every 5 ms and the length of each audio frames was 10 ms. We used a constant $\alpha=1$ and varying $\beta$ values as template subtraction parameters, because we had observed that higher values of $\alpha$ damage speech and an increase in $\beta$ improved ASR accuracy considerably compared with $\beta=0$. (For detailed evaluations of the parameters $\alpha$ and $\beta$ parameters, their effects on ASR accuracy, signal quality and noise suppression rates, see [136].)

We recorded random motions performed by the given set of limbs by storing a training database of 30 minutes’ duration. In a separate session, we recorded a test database 10 minutes long for evaluation. Because the noise recordings were longer than the utterances used in isolated word recognition, we selected those segments, in which all joints contributed to the noise. To generate precise amounts of noise and speech energy for various SNR conditions before mixing them, we amplified clean speech based on its segmental SNR, $SNR_{seg}$, which estimates the SNR-level within each segment and averages them over the whole signal, providing a better representation of the energy distribution of speech and noise within the time interval of interest.

$$SNR_{seg} = \frac{1}{J} \sum_{j=1}^{J} 10 \log_{10} \left( \frac{\sum_{t} x_{j}^2(t)}{\sum_{t} d_{j}^2(t)} \right),$$

(6.11)

where $J$ is the number of segments with speech activity, and $x(t)$ and $n(t)$ are the $t$-th discrete speech and noise sample respectively. The noise signal, consisting of whole ego noise and environmental background noise, was mixed with clean speech utterances used in a typical human-robot interaction dialog. This Japanese word dataset contains 236 words spoken by 4 female and 4 male speakers. Acoustic models are trained with Japanese Newspaper Article Sentences (JNAS) corpus, 60-hours of speech data spoken by 306 people (153 males and 153 females), making speech recognition a speaker-open
and word-open test. We used 13 static MSLS, 13 delta MSLS and 1 delta power for an acoustic feature. Speech recognition results were reported as average Word Correct Rates (WCR, defined as the number of correctly recognized words from the test set divided by the number of all instances in the test set). WCR improvement was calculated by subtracting two WCRs obtained from the experiments with two different parameter sets or two different methods and represented as “points”. The position of the speaker was fixed at 0° throughout the experiments. The recording environment consisted of a room 4.0 m x 7.0 m x 3.0 m in size with a reverberation time ($RT_{20}$) of 0.2 sec.

Although the position of the original sound source was provided in advance to avoid mis-recognition due to localization errors, we did not fix the ego noise direction of the robot. In this experiment, the SSL module predicted it automatically.

Spectrograms and Masks

Figure 6.3 shows a general overview of the effect of each processing stage until the masks are generated. Figure 6.3c) represents a dense mixture of speech (Figure 6.3a)) and motor noise (Figure 6.3b)) with an SNR of -5 dB. GSS+PF in Figure 6.3g) reduced only a minor part of the motor noise while sustaining the speech. In contrast, template subtraction (Figure 6.3h)) reduced the motor noise aggressively while distorting some parts of the speech. The hard mask (Figure 6.3i)) provides a filter that eliminates unreliable and still noisy parts of the speech ($T=0.5$). The soft mask (Figure 6.3j)) provides more detailed information about the degree of reliability of each feature so that the noise-free features are weighted more than the noise-containing parts of the MFT-ASR ($\{T, \sigma\} = \{0.5, 5\}$). Furthermore, we found that features between times 0.10-0.42 sec. and 1.07-1.27 sec., which were composed basically of motor noise alone, were given zero weights in the masks, except for a few mis-detections. The dotted yellow lines in the panels of Figure 6.3 indicate the borders of these regions, with speech features located between the 0.42-1.07 sec. Within this time interval, these masks were able to detect even those speech features that were contaminated by motor noise residuals and set either zero or low weights.

ASR Performance using Missing Feature Masks

We compared our MFT-based noise elimination approach with the single-channel noise suppression (TS) and multi-channel (GSS+PF) noise suppression techniques. The results were evaluated using an acoustic model trained with MCRA-applied speech data,
CHAPTER 6. MFT-BASED INTEGRATION OF EGO NOISE REDUCTION AND ASR

Figure 6.3: Spectra of speech signal (utterance: "Nan desu ka?" (What is this?)), noisy speech signals, refined speech signals and corresponding masks. In a)-h), the y-axis represents 256 frequency bins between 0 and 8kHz and in i)-j) the y-axis represents 13 static MSLS features. The x-axis in all panels is the index of frames.
except that, for the GSS+PF method, we used a matched acoustic model of this particular condition. In preliminary tests, we found that the feature set derived at the output of template subtraction achieved a greater accuracy by 10-20 points in WCR, compared with the features after multi-channel noise reduction. We therefore concluded that the former feature type is more suitable for an MFT-ASR. Single-channel results were used as a baseline for comparing all ego-motion noise reduction methods. Figure 6.4 illustrates the ASR accuracies for all methods under consideration. MFT-ASR outperformed both single (TS) and multi-channel (GSS+PF) noise reduction methods. We also evaluated MFMs for three heuristically selected threshold parameters, $T = \{0.25, 0.5, 0.75\}$, with the outcomes presented in Figure 6.5. If $T<0.5$, ASR was not improved because essential features belonging to speech were discarded, resulting in a deterioration of WCRs. In contrast, higher thresholds improved the outcomes significantly.

In our second set of experiments, we compared the results of hard masking with an optimal threshold ($T=0.75$) obtained during our first set of experiments, with the results of soft masking for the parameter set $\sigma = \{5, 10, 50\}$. All three examples with these parameters yielded similar WCR improvements. Outside this range, however, the results become sensitive to $\sigma$ and eventually deteriorated. Therefore, we will present only the results for $\sigma=5$. We also assessed the effect of decreasing the aggressiveness level of the template subtraction, by leaving an artificial floor on the bottom of the spectra. Thus, in our first set of experiments, the parameter called spectral floor ($\beta$, where $0 \leq \beta \leq 1$) [136] was set to zero. We assessed the results for $\beta = \{0, 0.2, 0.5\}$ in the
framework of soft-hard mask comparison in Figure 7.3 by determining the improvement in WCR relative to that obtained for the hard mask at $\beta=0$ and $T=0.75$. Increasing $\beta$ resulted in considerable improvements in the WCRs, indicating that a tradeoff between “noise reduction level” and “signal distortion” contributed substantially to the quality of the mask. We found that soft masks reduced the WCRs even further by up to 8 points compared with hard masks. This reduction was due to the improved probabilistic representation of the reliability of each feature. Optimal results were obtained when we used a soft mask with the parameter set: $\{T, \sigma, \beta\} = \{0.75, 5, 0.5\}$.

Table 6.1: Recognition accuracies (% WCR) of all methods utilized in this study with realistic SNRs during a robot-human interaction ($SNR=\{-5, 0\} dB$)

<table>
<thead>
<tr>
<th>SNR</th>
<th>Sing. ch.</th>
<th>GSS+PF</th>
<th>TS</th>
<th>TS+MFM($T=0.75$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hard</td>
</tr>
<tr>
<td>-5dB</td>
<td>29</td>
<td>49</td>
<td>69</td>
<td>78 0 (0)</td>
</tr>
<tr>
<td>0dB</td>
<td>38</td>
<td>60</td>
<td>72</td>
<td>80 0.5 (85)</td>
</tr>
</tbody>
</table>

While the masks eliminated unreliable speech features contaminated with motor noise, they also compensated for the erroneous effects of voice activity detection due to additive motor noise containing a large portion of energy. These masks prevented the false identification of motor noise as speech, when speech had not yet started, or had been completed. Table 6.1 shows the average WCRs extracted from the results in
6.4. EVALUATION

Figure 6.6: Speech recognition performance of soft mask - hard mask comparisons for given parameters.

Figures 6.4 and 7.3. This table helps in visualizing the simulated results in a real-world scenario with a robot, where SNRs usually vary by $[-5-0]dB$ for head and arm motion noises depending on the optimal distance and loudness of the speaker. The gain achieved by applying soft masking was 15 points greater than that of the single-channel template subtraction method.

**ASR Performance using Selective White Noise Superimposition**

In this experiment, we used two types of noise caused by (1) random whole-arm pointing behavior as *arm motion* and (2) random head rotation (elevation=$[-30^\circ, 30^\circ]$, azimuth=$[-90^\circ, 90^\circ]$) as *head motion*. MFM parameters are selected empirically: $T=0.75$ and $\sigma=10$. We created a matched acoustic model for multi-channel noise reduction (GSS+PF) methods by adding a white noise of $-40dB$.

We superimpose white noise of various SNR’s and evaluate WCRs with and without MFM. Fig. 6.7 shows the ASR accuracies for all methods under consideration. Single-channel results obtained with clean and noise matched acoustic models and without any processing are used as a baseline. In case of arm motion, which is considered as a relatively weaker noise, white noise of the same intensity level used in acoustic model training has shown the best performance. On the other hand, the best ASR accuracy during a head motion with high noise intensity is achieved with an additive white noise of $-20dB$. Based on the results with our robot, where head motion (pan and tilt) noise
was louder than background, arm-motion and leg motion noise, we suggest finally that $C_1$ and $C_2$ in Eq. 6.1 should be set to $-20dB$ and $-40dB$.

![Graph](a) Under arm motion noise  
![Graph](b) Under head motion noise

Figure 6.7: Recognition performance for different types of ego-motion noise

We also observe that the MFT-ASR outperforms the standard ASR without MFM. Although there is little gain of using MFM for the $-20dB$ white noise (See Fig 6.4(a) 6.4(b)), the masks improved the WCR’s for all other SNR’s during the experiments. While the masks eliminate unreliable speech features contaminated with motor noise, they can also compensate the erroneous effects of voice activity detection due to additive motor noise that contains a large portion of energy. They prevent mis-detection of motor noise as speech, when the speech has not started yet, or is already over.

The reader should also note that in a real-time, real-world scenario with a robot, where the SNR is between 0-5dB for the arm motion and -5-0dB for the head motion noise depending on the distance and loudness of the speaker, 15% and 18% average WCR improvement is attained compared to the WCRs obtained by ego-motion noise matched single-channel speech recognition.

### 6.4.2 Evaluation of MFT-ASR for Multiple Talkers

**Experimental Settings**

To evaluate the performance of the proposed techniques, we use a humanoid robot developed by Honda. The robot is equipped with an 8-channel microphone array on top
of its head. Of the robots many degrees of freedom, we use only a vertical head motion (tilt), and 4 motors for the motion of each arm with altogether 9 degrees of freedom. We recorded random motions performed by the given set of limbs by storing a training database of 30 minutes and a test database 10 minutes long. Because the noise recordings are comparatively longer than the utterances used in the isolated word recognition, we selected those segments, in which all joints of the corresponding limb contribute to the noise. We recorded clean speech utterances and converted them to 8 ch. data by convoluting with a transfer function of the microphone array. This Japanese word dataset includes 236 words for 1 female and 2 male speakers that are used in a typical human-robot interaction dialog. After normalizing the energies of the utterances to yield an SNR of \(-6\text{dB}\) (noise:two other interfering speakers), the noise signal consisting of ego noise (incl. ego-motion noise and fan noise) and environmental background noise is mixed with clean speech utterances. Acoustic models are trained with Japanese Newspaper Article Sentences (JNAS) corpus, 60-hour of speech data spoken by 306 male and female speakers, hence the speech recognition is a word-open test. We used 13 static MSLS, 13 delta MSLS and 1 delta power as acoustic features. Speech recognition results are given as average Word Correct Rates (WCR). The position of the speakers are kept fixed at three position configurations throughout the experiments: \([-80^\circ, 0^\circ, 80^\circ]\), \([-20^\circ, 0^\circ, 20^\circ]\).

To avoid the mis-recognition due to localization errors and evaluate the performance of the proposed method, we set the locations manually by by-passing SSL module. The recording environment is a room with the dimensions of \(4.0\text{m}\times 7.0\text{m}\times 3.0\text{m}\) with a reverberation time \((RT_{20})\) of 0.2s. We evaluated MFM with the following heuristically selected parameters \(T_e=0\), \(T_m=0.2\), \(\sigma_e=2\), \(\sigma_m=0.003\) (energy interval: [0 1]).

**Spectrograms and Masks**

Figure 6.8 gives a general overview about the effect of each processing stage until the masks are generated. In Figure 6.8f), we see a tightly overlapped speech (Figure 6.8a)+b)+c), \(SNR_m=-6dB\) and motor noise (Figure 6.8d), \(SNR_e=-5dB\) mixture. GSS+PF as in Figure 6.8g)–l) reduces only a minor part of the motor noise while sustaining the speech.

The masks presented in Figure 6.9 are all applied to the refined signals in Figure 6.8j)–l). The ego noise masks in Figure 6.9a)–c), show that ego noise located between the 20-45th frames can be detected and the features can be suppressed. Besides, as in Figure 6.9c), the residuals of ego noise between 80-100th frames can be masked, as
Figure 6.8: Spectrograms for preprocessing. y-axis represents frequency bins between 0 and 8kHz. x-axis represents the index of frames.

well. The multi-talker masks, on the other hand, are active especially where the source separation leaves residual energies that belong to interfering speakers. In the integrated masks (in Figure 6.9g–i)), we see the contribution of ego noise masks is more dominant compared to multi-talker masks, as they seem to extend the contours of the ego noise masks. The soft masks in Figure 6.9j)–l), in addition, provide more detailed information about the reliability degree of each feature so that the noise-free features are weighted more than the noise-containing parts in the MFT-ASR.

**ASR Accuracy using MFMs**

Figure 6.10 and 6.11 illustrate the ASR accuracies for a speaker setting with wide, resp. narrow, separation intervals and for all methods under consideration. The results are evaluated using an acoustic model trained with motor noise data. Single channel recognition is between 0-2 percent for all SNRs. Because the task is a multi-talker recognition, GSS+PF is considered as the baseline. There is only little improvement gained from minimum energy criterion (mec), as the results of the comparison for the hard masks presents in both figures. In overall it contributes only up to 1-3% to WCR.
6.4. EVALUATION

Figure 6.9: Masks generated with a mild threshold value ($T_e=0.1$). y-axis represents 13 static mel-features. x-axis represents the index of frames.

Figure 6.10: Recognition performance for the speaker located at $[-80^\circ, 0^\circ, 80^\circ]$
We see three general trends:

1. Soft masks outperform hard masks for almost every condition. This improvement is attained due to the improved probabilistic representation of the reliability of each feature.

2. We observe that the ego noise masks perform well for low SNRs, however WCRs deteriorate for high SNRs. The reason resides in the fact that faulty predictions of ego-motion noise degrade the quality of the mask, thus ASR accuracy, of clean speech more compared to noisy speech. On the other hands, in high SNRs (inferring no robotic motion or very loud speech) multi-talker masks improve the outcomes significantly, but their contribution suffers in lower SNRs instead.

3. As the separation interval gets narrower, the WCRs tend to reduce drastically. The presented WCRs in Figure 6.10 and 6.11 are consistent with a multi-talker recognition study of [140]. We further observe a slight increase in the accuracy provided by $M_m$ compared to $M_e$ in -5dB for narrow separation angles (Figure 6.11), because the artifacts caused by SSS for very close talkers become very dominant.

We evaluated both integration techniques for hard masks: AND and OR-based integration of $M_e$ and $M_m$. However, the WCRs were all worse compared to the individual recognition performances of single masks. We applied a simple binary weighting based on the assessment of trend 1) and 2) above. We set the weights according the following conditional statements and apply an addition operation as in Equation 6.10:

$$\{w_e, w_m\} = \begin{cases} 
1, \ 0 & \text{if } SNR < 0 \\
0, \ 1 & \text{if } SNR \geq 0 
\end{cases}$$
Corresponding results are displayed in the final bin of each SNR segment, which indicate that fused masks work best for this problem. As shown in Figure 6.12 and 6.13, our method demonstrated significant WCR improvement with soft masking up to 10% compared to GSS+PF for the sound sources that are $|80^\circ|$ from each other and up to 15% for the sources with $|20^\circ|$ separation angle.

Figure 6.12: Average WCRs for all speakers located at $[-80^\circ, 0^\circ, 80^\circ]$  

Figure 6.13: Average WCRs for all speakers located at $[-20^\circ, 0^\circ, 20^\circ]$
CHAPTER 6. MFT-BASED INTEGRATION OF EGO NOISE REDUCTION AND ASR

Future Work

In future, we intend to determine optimized parameter sets for template subtraction for this specific task of MFT-ASR in a wider range. We plan to weight the masks not binary, but in a continuous way for a large SNR value interval. Besides, equal distribution of total ego-motion noise to all talkers is not a good representation, e.g. even if the ego-motion noise comes only from right arm, current ego noise masks do not distinguish the direction of the noise and the mask of each talker is affected by the same amount of noise. So, we plan to calculate the noise energy contributions based on the direction of the motors in relation with the directions of speech.

6.5 Summary

The main contribution of our work will be the incorporation of an original Missing Feature Mask (MFM) generation method based on the signals generated by two blocks (template subtraction & multi-channel noise reduction) that run in parallel. The mask relies on a measure of a frequency bin’s quality calculated from the similarity of two totally different- yet complementary- approaches. We first suggest a binary mask to estimate the reliability of each acoustic feature. This method was enhanced by using a soft mask, which yields more detailed information about reliability. We validated the applicability of our method by evaluating its performance at different settings for hard and soft MFMs. Our method demonstrated significant WCR improvements with hard masking (49 points relative to single-channel recognition) and soft masking (up to 55 points).

In the second part of this chapter, we incorporated Missing Feature Theory (MFT) to solve the ego-motion noise problem of a robot within the context of multiple speakers talking simultaneously. We designed a missing feature mask (MFM) generation method based on the measure of a frequency bin’s Signal-to-Noise Ratio (SNR), which is computed from the ratio of speech and estimated motor noise energies and called ego-motion noise MFM. Furthermore, we focus on various integration methods for fusing the ego-motion noise MFM with the multi-talker MFM (originally introduced by Valin et al. [26]). The SNR-weighted integration mechanism of MFMs prevents the robot from applying unnecessary suppression to speech features by exerting the ego-motion masks while the robot is moving slowly or comes to rest. Thus, it optimally balances the contribution of the two masks on the noise masking. We showed that our integration
method achieves a high ASR accuracy for any arbitrary separation angle between the talkers and any SNR value.
Chapter 7

Application to Sound Source Localization

7.1 Introduction

Sound source localization suffers from the contamination of sound signals with ego noise. Especially because the motors are located in the near field of the microphones and are covered with body shells, they emit sounds having both diffuse and directional characteristics, which influences the estimation of sound directions. On the other hand, the noise emitted from the fan of the robot is the main reason of mis-recognition of the sound sources because of its high power. Additionally, when the robot moves its limbs on which the microphones are mounted, the direction of ego noise alters rapidly, therefore the effects created by the moving microphones must be also taken care of.

7.2 Design of Ego Noise Robust Sound Localization

We propose to extend the single-channel based noise estimation method in a way that it can estimate also multi-channel data. For this purpose, we store/retrieve data recorded by multiple microphones such as described in Figure 7.1. They are used in combination with an SSL system, i.e. GEVD-MUSIC as explained in Section 2.3.4 to decorrelate the ego noise and also cope with instantaneous head rotation effects, where the microphones are mounted.
7.3 Evaluation

7.3.1 Experimental Settings

We compare three SSL techniques: (1) SEVD-MUSIC, (2) GEVD-MUSIC with fixed noise Correlation Matrix (CM) (averaged over 2,000 frames) and (3) proposed method, called GEVD-MUSIC with instantaneously estimated noise CMs. The real-world experiments are conducted for two conditions:

E1) The robot is in still-stand (fan noise only)
E2) The robot moves its arms randomly (fan noise + arm motion noise)
E3) The robot moves its arms and head randomly (fan noise + arm & head motion noise + head rotation effect)

The resolution of the steering vectors is 5°. The sound source is located 1 meter away at 0° relative to the body of the robot for all experiments. Two types of signals with varying SNR values ranging from -5 to 10dB are played from a loudspeaker for one minute each: a sinusoidal signal with a fundamental frequency of 600Hz and a white noise signal.

Our evaluation criteria are Mean Localization Error (MLE) [°] and the Peak Accuracy (PA [%]) for different threshold values:

\[
\text{Peak Accuracy} = 100 \frac{\#\text{Frames} - \#\text{Ins} - \#\text{Del} - \#\text{Subst}}{\#\text{Frames}}. \tag{7.1}
\]
7.3.2 Mean Localization Error

Table 7.1 shows that GEVD with estimated noise templates shows almost the same performance like GEVD-fixed in a stationary robot (E1), whereas it achieves far superior performance in terms of MLE compared to the other methods in E2 and E3 (see Tables 7.2 and 7.3). Generally, SEVD-MUSIC is unable to detect the peak of the desired signal due to the loud fan noise. GEVD-MUSIC with fixed noise CM performs well for fan noise only in E1, and fairly for E2, in which the orientation of the fan noise does not change. The trained CM is still able to suppress the fan noise at a fixed position, however the arm motion noise degrades the performance. Through Figure 7.2 to Figure 7.4 we can see the improvement in the spatial spectrum. The spectrum of GEVD-MUSIC is much flatter compared to the previous methods due to the suppression of disturbing noises. The 3-dimensional representations (Figure 7.5-Figure 7.7) gives a better view for the results of the same experiment E2. In E3, on the other hand, the proposed method is the only method that can eliminate the dynamic noise changes in the spatial spectrum of MUSIC (see Figures 7.8-7.10).

7.3.3 Peak Accuracy

We also assess the methods in terms of PA. Figure 7.11 illustrates the performance of each method for two different cases: \( \text{thr} \) shows the results obtained with an optimum threshold value, whereas \( \text{max} \) only takes the largest peak into account, thus the deletion and insertion errors in Equation 7.1 are automatically omitted. The proposed method outperforms the others in case the maximum peak is selected as the estimated position of the sound source. When a threshold value is used, the performance drops significantly due to the increased insertion errors such as in Figure 7.11(a).

7.3.4 Discussion

In SSL systems, the number of sound sources \( M \) and threshold values are the most crucial key points for performance. When the number of sound sources is unknown, a strategy based on a fixed threshold is practical such as in SEVD and GEVD-fixed methods. However, a fixed threshold value for GEVD with estimated noise CM is difficult to determine, because the power of the MUSIC temporal spectrum fluctuates due to the incorrect template predictions, thus its performance is not stable. One way to make the temporal-directional plane of MUSIC smoother is to estimate the CM \( K(\omega, \phi) \).
### 7.3. EVALUATION

Table 7.1: Mean localization error (MLE [°]) results for E1

<table>
<thead>
<tr>
<th>Signal type</th>
<th>SNR</th>
<th>SEVD</th>
<th>GEVD fixed</th>
<th>GEVD est</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinusoidal signal with ( f_f = 600Hz. )</td>
<td>-5</td>
<td>122.5</td>
<td>69.9</td>
<td>59.1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>137.3</td>
<td>5.1</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>34.3</td>
<td>5.0</td>
<td>1.9</td>
</tr>
<tr>
<td>White noise</td>
<td>10</td>
<td>5.0</td>
<td>5.0</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 7.2: Mean localization error (MLE [°]) results for E2

<table>
<thead>
<tr>
<th>Signal type</th>
<th>SNR</th>
<th>SEVD</th>
<th>GEVD fixed</th>
<th>GEVD est</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinusoidal signal with ( f_f = 600Hz. )</td>
<td>-5</td>
<td>142.2</td>
<td>93.62</td>
<td>15.8</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>164.8</td>
<td>69.79</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>105.4</td>
<td>15.0</td>
<td>7.8</td>
</tr>
<tr>
<td>White noise</td>
<td>10</td>
<td>20.4</td>
<td>5.9</td>
<td>5.9</td>
</tr>
</tbody>
</table>

Table 7.3: Mean localization error (MLE [°]) results for E3

<table>
<thead>
<tr>
<th>Signal type</th>
<th>SNR</th>
<th>SEVD</th>
<th>GEVD fixed</th>
<th>GEVD est</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinusoidal signal with ( f_f = 600Hz. )</td>
<td>-5</td>
<td>151.5</td>
<td>82.0</td>
<td>35.2</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>148.3</td>
<td>56.5</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>31.4</td>
<td>20.1</td>
<td>7.1</td>
</tr>
<tr>
<td>White noise</td>
<td>10</td>
<td>26.3</td>
<td>11.6</td>
<td>4.5</td>
</tr>
</tbody>
</table>

124
CHAPTER 7. APPLICATION TO SOUND SOURCE LOCALIZATION

Figure 7.2: MUSIC spectrum obtained with SEVD during fan noise and arm motion (E2).

Figure 7.3: MUSIC spectrum obtained with GEVD (fixed) during fan noise and arm motion (E2).

Figure 7.4: MUSIC spectrum obtained with GEVD (estimated) during fan noise and arm motion (E2).
Figure 7.5: 3-dimensional representation of Figure 7.2

Figure 7.6: 3-dimensional representation of Figure 7.3

Figure 7.7: 3-dimensional representation of Figure 7.4
CHAPTER 7. APPLICATION TO SOUND SOURCE LOCALIZATION

Figure 7.8: Prediction of positions based on the highest peak of MUSIC spectrum in each frame during random arm and head motion (E3).

Figure 7.9: Prediction of positions based on the highest peak of MUSIC spectrum in each frame during random arm and head motion (E3).

Figure 7.10: Prediction of positions based on the highest peak of MUSIC spectrum in each frame during random arm and head motion (E3).
Figure 7.11: Peak Accuracy curves for all three methods

in a longer time window, but it also degrades the noise reduction and SSL performance. Besides, a consequent tracking operation would have improved the final localization accuracies.

In this work, we were mainly interested in our method’s capability of suppressing the MUSIC spectrum of noise and dominant noise peaks. We mostly focused on extracting desired sound’s peak, therefore we used the strategy of selecting the $M$ largest peaks by assuming that $M$ is given in advance or detected by another process. However, the details about this detection process and the exact correspondence between the sound sources and peaks are still open questions.

### 7.4 Summary

We inspected the applicability of our proposed method, estimating ego noise as a sequence of discrete templates, to a task of robust SSL. We created correlation matrices from the estimated templates to whiten the non-stationary and thus to eliminate the disturbing effects of ego noise in the spatial spectrum. The validity of the ego noise estimation technique was confirmed by quantitative assessments. Using our method drastic improvements by up to 70 points in peak accuracy and reduction by up to $60^\circ$ in mean localization error compared to a standard sound source localization method (SEVD-MUSIC) are achieved.
Chapter 8

Discussion

8.1 Observations

We tackled the problem of estimating and removing ego noise from recorded signals. This problem is general and important in human-robot interaction because microphones capture the disturbing noise and the system cannot differentiate if the total signal is a clean speech or contaminated speech. In the framework of computational auditory scene analysis and robot audition, we developed a method for effectively overcoming the ego noise problem.

8.1.1 Template Based Ego Noise Estimation

We modeled the ego noise using templates which are well-suited to capture the dynamic nature of the motion data represented by a sequence of observations. Based on these observations, it was possible to associate a discrete time series data representing the motion (i.e., the angular status of each joint of the robot) with another series of discrete time data representing the total ego noise spectrum. This approach relied on seamless synchronization between data on joint status (i.e., angular position, velocity and acceleration) and audio data. Throughout all our experiments, we used a humanoid robot developed by Honda. The high estimation quality achieved by this approach allows us to suppress noise accurately by applying a template-based spectral subtraction. The method is evaluated using several objective criteria such as normalized noise estimation error, segmental SNR, log-spectral distortion and word correct rates obtained by an ASR. The results demonstrated the effectiveness of proposed method in the sense that
high suppression rates are achieved while keeping the speech as untouched as possible.

8.1.2 Online System for Ego Noise Estimation

Since a speech enhancement system solely based on template estimation cannot cope with the changing environmental noise or an single-channel noise reduction system is unable to eliminate non-stationary noise, we proposed to concatenate the two stages to obtain a unified preprocessing framework for a robot audition system. We assessed the performance of MCRA/HRLE-based noise reduction methods in the presence of background noise and ego noise, and presented a method exhibiting high robustness even against changing conditions of the environment. On top of this underlying basis we proposed an online and adaptive learning mechanism for an adaptive ego noise estimation framework. We designed a learning mechanism provided with a measure of performance, which enables the learning process to continue over the entire lifespan of a robot without human intervention. This incremental learning algorithm is a natural extension of the parameterized template generation method and allows the robot to learn its ego noise not only in isolated training conditions, but also in daily environments (possibly even in the presence of humans). Moreover, it allows us to tackle the curse of dimensionality problem caused by the large number of DoF of a robot. We assessed the learning, estimation and suppression performance of this template-based single-channel noise reduction method in the presence of background noise and ego-motion noise. We showed that the online learning contributes to precise estimation of overall noise and a high ASR accuracy under various SNR conditions. We achieved, thanks to the update mechanism, reduction in the estimation error up to -1.7dB (50% improvement). Furthermore, the size of the database is reduced 10-fold after 20 repetitions.

8.1.3 General Applicability of Ego Noise Estimation

To evaluate the proposed ego noise estimation method, we created frameworks for various applications in robot audition using the ego noise estimation as the core concept:

Whole Body Motion Noise Reduction of a Robot

We have described methods for eliminating whole-body motion noise by using a synthesis-by-analysis approach to suppress this noise. We divided whole body motion noise into three domains, depending on their spatial location and intensity levels: arm,
leg and head motion noise. The system we proposed extracts information about the motion type performed at that moment and decides on the best choice of processing method for speech recognition by a selective ASR module. We adopted two methods, multi-channel noise reduction and single-channel template subtraction, which are selected or not depending on the detection of head motion. If no head motion is detected, the first method is selected because it is effective for arm, leg, and arm&leg motion noises. Source separation is particularly effective against noises from the arms and legs, because the limbs are located away from the microphone array and are separated from a speaker standing directly in front of the robot. On the other hand, if head motion is detected, the second method is selected. It is more appropriate for canceling head motion noise (or the combination of the head motion noise with arm and/or leg motion noise), because template subtraction makes no assumptions about the nature of the noise and uses previously recorded noises. We validated the applicability of our approach by evaluating its performance on three different motor noise types and their combinations. Our method demonstrated good performance in suppressing arm, leg and head motion noises and their combinations. As a consequence, improvements up to 23 pts, 40 pts and 39 pts are achieved in the presence of arm, leg and head motion noises respectively.

**MFT-based Integration of Ego Noise Reduction and ASR**

We incorporated an original MFM generation method based on the signals generated by two blocks (template subtraction & multi-channel noise reduction) that run in parallel. The mask relies on a measure of a frequency bin’s quality calculated from the similarity of two totally different- yet complementary- approaches. We validated the applicability of our method by evaluating its performance at different settings for hard and soft MFMs. Our method demonstrated significant WCR improvements with hard masking (49 points relative to single-channel recognition) and soft masking (up to 55 points).

To solve the ego noise problem of a robot within the context of multiple speakers talking simultaneously, we designed a missing feature mask (MFM) generation method based on the measure of a frequency bin’s Signal-to-Noise Ratio (SNR). It is computed from the ratio of speech and estimated motor noise energies and called ego-motion noise MFM. Furthermore, we focused on various integration methods for fusing the *ego-motion noise MFM* with the *multi-talker MFM*. The SNR-weighted integration mechanism of MFMs prevents the robot from applying unnecessary suppression to speech features by exerting the ego-motion masks while the robot is moving slowly or comes to rest. Thus,
it optimally balances the contribution of the two masks on the noise masking. We showed that our integration method achieves a high ASR accuracy for any arbitrary separation angle between the talkers and any SNR value.

**Ego Noise Robust Sound Source Localization**

We inspected the applicability of our proposed method, estimating ego noise as a sequence of discrete templates, to a task of SSL. The validity of the ego noise estimation technique was confirmed by quantitative assessments in which drastic improvement by up to 70 points in peak accuracy and reduction by up to $60^\circ$ in mean localization error are achieved.

**8.2 Contributions**

We achieved a real-time ego noise estimation system with online learning in this thesis. Our method represents ego noise of a robot as templates and estimates it from a template database. It also overcomes the necessity of offline training sessions for template learning so that the ego noise estimates can be used in various robot audition modules such as, noise reduction, robust feature extraction, automatic speech recognition and sound source localization.

In the robot audition literature, numerous examples of speech enhancement methods do not work properly under the circumstances posed by ego noise. Our study, representing ego noise with parameterized templates and estimating it from a template database recorded in advance, is original. In contrast to attaching additional sensors our method has several advantages, including its ability to be easily implemented on any mobile robot regardless of the physical constraints about the external shielding. By exploiting only existing microphones, it is also cost effective and applicable without any hardware modifications. Some studies using single-channel templates pursued to represent these templates as whole blocks of noise data, as from the onset to the offset time of the motion, and some researches aimed at using neural networks to learn and estimate the templates, but their performance was not high and their system was unable to cope with all possible templates covering the whole motion space of a robot with a high number of DoF. Furthermore, approximate search strategies for selecting the appropriate templates made our method more suitable for online learning. In addition, we enhanced the accuracy of the templates further by incorporating more information related to the
joints, such as angular acceleration.

In our study, we updated the template database in an online manner, which enables the robot to learn its own noise without human intervention and on the fly. Being the first of its kind, this strategy of template learning makes use of previously learned knowledge about the templates to speed up the learning. It also makes the noise estimation module more robust because errors in the training set can be corrected during operation and it enables the system to adapt to partially-known or dynamic environments.

Automatic speech recognition has been chosen as the primary application environment for design and evaluation of ego noise reduction techniques in our thesis. To demonstrate the general applicability of the proposed approach, we extended it also to other applications such as sound source localization, and MFT-based integration of noise reduction and ASR.

8.3 Remaining Issues and Future Work

Three problems remain to be solved regarding ego noise estimation for robot audition:

8.3.1 Synchronization of Motion and Audio Data Streams

The current version of ego noise estimation relies on seamless synchronization between data on joint status (i.e., angular position, velocity and acceleration) and audio data. This scheme is, however, sensitive to the delays/latencies that can come up during the acquisition, processing and transmission stages before the data can be observed in the end point. As shown in Figure 8.1, audio and motion data experience delays of \( t' \) and \( t'' \), respectively. In our approach, we regarded one frame, the unit of audio processing at a single time, as a reference and assumed that \( \max[t', t''] \) is shorter than a frame.

However, not all robotic systems are as precise as ours. In case the acquisition frequency of audio and motor encoders are out of proportion, we suggest to use interpolation to generate missing data. In case the data are out of synchronization with a constant delay, the best strategy would be to utilize time alignment through matched filters. Computing the cross-correlation of the input signals with a pre-selected pilot signal can help to detect the delay between two data streams and delaying the earlier stream can bring the system to synchronization. We also plan to conduct experiments on a robotic system, which provides data streams with changing delay, to evaluate the limitations of our method in this most difficult synchronization problem.
8.3. REMAINING ISSUES AND FUTURE WORK

8.3.2 Higher-level Representation for Templates

In the current system, the data is labeled for each state of ego noise, i.e., each frame (see Figure 3.2 and 3.4(b)). The label in its current form consists of only meta-information regarding instantaneous motion data. However, the sequence of these motion data can bear useful information, as long as the motions are repeated often and with fixed trajectories. Therefore, additional features that utilize cues about time series expansion of consecutive motion elements and incorporate information on motion primitives would improve the reliability and performance of the predictions. This is, for example, possible by using higher-level transcriptions for the templates that represent the consecutive states of distinct data streams. The template can be extended to represent motion primitives with the following 3 components:

1. Transcription: Description (or ID) of the motion primitive,
2. Label: Sequence of consecutive joint state of the robot (i.e., feature vectors),
3. Data: Sequence of consecutive ego noise spectra during the motion primitive (i.e., spectral vectors).

Hidden Markov Models [141] or Gaussian Mixture Models [142] are suitable techniques used to represent the dynamic motion primitives in the robot motion and behavior generation community. Repertoires of motion primitives can be updated in terms of their sequential data and labels. It is also possible to compute the similarity of several motion primitives by calculating their Kullback Leibler distances. In case the
recently acquired template does not exist in the template, it can be easily inserted into the database as a new motion primitive. One major challenge of this approach is, however, the requirement of a robust and automated primitive derivation method for motion classification.

### 8.3.3 Performance Improvement in Database Storage and Retrieval System

As the number of observed templates becomes huge, the robot must have an effective way of storing the acquired database for easier retrieval and organization. This is only possible if either 1) the retrieved information is reduced (dimension reduction) or 2) the retrieved set is reduced (different indexing methods).

Our initial experiments with *Principal Component Analysis (PCA)* failed to reduce the number of features involved in the classification because each feature is important to characterize the representative motion and the number of features are optimal in that sense.

In our implementation we used a hierarchical scheme called KD-trees, however the selectivity of these schemes degrades rapidly as dimensionality increases. There are more complex and advanced database storage technologies than KD-trees which utilize sequential access schemes [143]. They outperform hierarchical schemes as dimensionality increases and they are a lot easier to integrate within a database engine. They are designed to reduce the I/O cost associated with external memory. These methods all support database expansions during active periods because the size of incoming data cannot be known in advance at the construction. They also work well in dynamic settings, such as our incremental learning method, which requires efficient insertion and update operations.

For example, methods called *Vector Approximation (VA)-File* [143] and *VA+File* [144] proposed efficient indexing structures and algorithms for various models of K-NN queries on multiple data streams. Index structures are built by dynamically quantizing the incoming dimensions. This is an important advantage for us because our joint status space is not uniformly distributed, therefore bit assignment to the feature vectors can be done in a non-uniform manner. This will enable the query time to improve significantly.
Chapter 9

Conclusion

The ultimate goal of our research is to develop a robot audition system that allows the robot to understand the surrounding auditory world, such as human voices, music, and other environmental sounds. Robots equipped with microphones hear their own noise and sounds of interest simultaneously. Therefore, the robots have to recognize a target sound from the captured audio signal even if it is a contaminated sound. To eliminate the disturbing effects of ego noise, the robots need to have a proper ego noise estimation method.

The following points are the requirements for optimal ego noise estimation:

- **Requirements for ego noise estimation**

  1. Low hardware and computational cost,
  2. Online and real-time operation,
  3. Connectivity and compatibility with other useful techniques designed for robot audition.

To solve the problem of ego noise, while satisfying the requirements, we developed the following techniques:
Developed techniques

1. Parameterized template-based ego noise estimation,
2. Online system, which incorporates a unified framework for stationary and non-stationary noise estimation, and incremental learning of templates,
3. Frameworks for ego noise reduction, noise robust feature extraction, ASR and SSL.

Finally, we achieved online, real-time ego noise estimation for robot audition. We evaluated the estimation and suppression performance of our method. The experiments showed that our method was effective in terms of estimation error, log-spectral distortion, signal-to-noise ratio, word correct rates for ASR, localization error and peak accuracies for SSL.
Relevant Publications

Chapter 3

Template Based Ego Noise Estimation


Chapter 4

Online System for Ego Noise Estimation


RELEVANT PUBLICATIONS

Chapter 5
Whole-body Motion Noise Reduction of a Robot


Chapter 6
MFT-based Integration of Ego Noise Reduction and ASR


Chapter 7

Application to Sound Source Localization

References


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


155
REFERENCES


REFERENCES


