

© 2010 Lae-Hoon Kim

Amazon v. SoundClear
US Patent 9,031,259
Amazon EX-1006

STATISTICAL MODEL BASED MULTI-MICROPHONE SPEECH
PROCESSING: TOWARD OVERCOMING MISMATCH PROBLEM

BY

LAE-HOON KIM

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Associate Professor Mark A. Hasegawa-Johnson, Chair
Professor Stephen E. Levinson
Associate Professor Minh N. Do
Research Associate Professor Margaret M. Fleck

UMI Number: 3452187

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3452187

Copyright 2011 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

In this thesis, a joint optimal method for clean speech estimation and ASR in a mismatched condition will be described with a unified speech model under a generalized expectation maximization (GEM) scheme. From this perspective, multi-microphone optimal speech estimation can be interpreted as pre-processing to increase reliability of feature components before the actual speech recognition or model based speech estimation is performed. Also, ideal binary mask (IBM) estimation from the context of the statistical model for ASR can be regarded as an initialization step to exclude the unreliable portion for ASR and to increase the estimation accuracy based only on the reliable components and trained speech process model.

Optimal multi-microphone speech processing is performed in the short-time Fourier transform (STFT) domain, since the atomic speech information can be meaningfully represented with a series of 10 to 30 ms short frames. Convolution in the time domain is formulated as filtering via a feed-forward network in the STFT domain, and is shown to be an appropriate representation under the overlap-add framework. With this structure in mind, sufficient statistics for estimating target speech from the multi-microphone measurements are formulated, and realistic relaxations for them are discussed since we need to estimate not only the target speech information but also the room impulse responses (RIRs), which have unavoidable uncertainty due to the movement of speakers.

Firstly, reverberant speech mixture separation with typical background noise is tackled. Standard adaptive independent component analysis (ICA) implemented with the natural gradient method is extended into the STFT domain with regularized feed-forward ICA (RFFICA) and post-processing based on direction-per-frequency. This method showed up to almost an order of magnitude performance improvement (29 dB in C-weighting) compared with the state of the art methods.

Secondly, we try to update the filters fast enough, with a smaller amount of measured data sharing the same directional information about target and interference location. Expectation maximization beamforming (EMB) followed by minimum mean squared error (MMSE) post-filtering is proposed to reduce the number of filter taps to update. Because we can obtain generative model based information about the target speech presence probability per each frequency bin and per each frame with enhanced robust DOA estimation capability, EMB can also be used to replace the direction-per-frequency based post-processing, which has been applied independently after RFFICA.

Thirdly, the DOA only based beamforming is extended to early response based beamforming. We estimate the RIRs from target and interference speech given the robust estimation on DOAs and construct linearly constrained minimum variance (LCMV) beamforming, which can be easily extended with the EMB framework. Because we perform a two-step approach, estimating RIR first and applying a demixing filter, without introducing more taps in the frame for adaptation purposes, we can have good demixing or dereverberation results.

Finally, IBM estimation and ASR are jointly formulated under a GEM framework. Even with the optimal front-end pre-processing, there always exists a mismatched portion with the statistical speech process model which is going to be used for ASR. Therefore, identifying the corrupted portions and removing them in ASR from the perspective of ASR itself is a necessary procedure. The cepstral domain ASR models are transformed into the spectral domain without loss of information through the global tying process. The proposed algorithm achieved much higher absolute ASR accuracy, ranging from 14.69% at 0 dB signal-to-noise ratio (SNR) to 40.10% at 15 dB SNR, than a normal ASR method with an optimal front-end processing in a highly non-stationary mismatch environment.

To my Suna

ACKNOWLEDGMENTS

“It was God, not the task, he had in view.”

First of all, I would like to thank my advisor Professor Mark Hasegawa-Johnson. His continuous encouragement and guidance toward innovation keep inspiring me to pursue new perspectives with sound theory on the problems known to be hard to solve.

I would like to thank Dr. Ivan Tashev. I could actually convince myself that a wonderful innovation really leads to great ground breaking results.

I would like to thank my thesis committee, Professor Stephen Levinson, Professor Minh Do, and Professor Margaret Fleck, who provided me valuable suggestions and comments for a better-quality thesis.

There are many friends I would like to thank: my colleagues in the statistical speech technology group, the Korean Church of Champaign-Urbana, special thanks to the choir Soli Deo, members of the Korean ECE Association. I would like to acknowledge Han Yoon for his generous help in making the self-made eight-channel pre-amp for my thesis experiments.

Finally, I would like to thank my wife Suna, and two sons David and Edward for their endless support and understanding.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER 1 INTRODUCTION	1
1.1 Graphical model for joint optimization of distortionless speech estimation and ASR	3
1.2 Multi-microphone based speech estimation in an adverse environment	6
1.3 Decoder oriented IBM estimation	10
1.4 Thesis organization	12
CHAPTER 2 BACKGROUND	14
2.1 Multi-microphone sound capturing and processing	14
2.2 Classical MMSE estimation and sufficient statistic	26
2.3 HMM/GMM speech model based approach	30
2.4 Missing feature approach	38
CHAPTER 3 REVERBERANT SPEECH SEPARATION USING RFFICA AND IDOA BASED POST-PROCESSING	43
3.1 Introduction	43
3.2 Problem formulation and background	46
3.3 Regularized feed-forward ICA (RFFICA) with IDOA based post-filtering	51
3.4 Experimental results	57
3.5 Conclusions and future work	63
CHAPTER 4 JOINT ESTIMATION OF DOA AND SPEECH BASED ON EM BEAMFORMING	70
4.1 Introduction	70
4.2 Multi-microphone speech enhancement	73
4.3 Sufficient statistic: EM beamforming	75
4.4 Experiments	79
4.5 Conclusions and future work	81

CHAPTER 5	EARLY ROOM IMPULSE RESPONSE ESTIMATION BEYOND DOA ESTIMATION	87
5.1	Multi-path GSC	88
5.2	Iterative blind estimation of RIR based on multi-path GSC	99
5.3	Sequential MVDR beamforming and postfiltering	108
5.4	Probabilistic RIR sufficient statistic approach	114
CHAPTER 6	DECODER ORIENTED IDEAL BINARY MASK ESTIMATION FOR ROBUST ASR	123
6.1	Introduction	123
6.2	Proposed method	125
6.3	Experiment	130
6.4	Conclusions and future work	131
CHAPTER 7	CONCLUSIONS AND FUTURE DIRECTION	134
APPENDIX A	DIFFERENT FORMULATIONS FOR JOINT OPTIMIZATION OF SPEECH ESTIMATION AND ASR	136
A.1	MLE of both model and missing features	136
A.2	MLE of model and missing features as latent variables	137
A.3	Both model and missing features as latent variables	137
APPENDIX B	SUFFICIENT STATISTIC CONSIDERING REVERBERATION PROCESS	138
APPENDIX C	PROOF OF EQUATION (1.8)	140
APPENDIX D	PRE-AMP CIRCUIT DIAGRAM	143
APPENDIX E	CORPUS GENERATION FOR SOURCE SEPARATION EXPERIMENT	145
E.1	Room impulse response measurement	145
APPENDIX F	ROBUST SPEECH RECOGNITION BASED ON MONTE CARLO UNCERTAINTY DECODING	151
F.1	Problem formulation	151
F.2	Monte Carlo method	153
F.3	Experiment	154
F.4	Conclusion	155
REFERENCES		158

LIST OF TABLES

5.1	Digit recognition accuracy(%) using half of the AVICAR corpus or enhanced AVICAR corpus for training	112
5.2	Digit recognition accuracy(%) using TIDIGITS training corpus for training	112
6.1	Recognition accuracy(%)	131

LIST OF FIGURES

1.1	Simplified diagram for computer assisted speech communication.	2
1.2	Graphical model representation for ASR and ideal binary mask estimation: q_t is a state in a given model, $\vec{O}(t)$ is an observation vector or feature vector, and $\vec{O}_L(t)$ is a binary mask vector of reliability for each component of $\vec{O}_L(t)$	4
1.3	Block diagram for decoder oriented IBM estimation procedure.	11
2.1	Propagation of an impulse sound in the time domain across a linear array of eight electret condenser microphones with incident angle of 60°	15
2.2	Impulse response measurement setup: mouth simulator (B&K Type 4227), a linear microphone array with a penny in the center, and eight channel self-made pre-amplifier (circuit diagram and interior are shown in Appendix D.) . . .	16
2.3	Illustration of double main lobes at -90° and 90° with f_c : inter-microphone delay $\frac{T}{2}$ corresponds to -90° and 90° incident angles, which has the same phase and can be summed to be constructive. Note that -90° and 90° incident angles produces the largest inter-microphone delay.	18
2.4	Direction of arrival estimation result (target speech spoken at 32° up to 300 frames, interference speech was at -32° during all frames) (a) sinc function applied GCC-PHAT based DOA estimation with pair of microphones at distance 0.076 m, (b) MUSIC based DOA estimation (8-microphones linear array with 0.03795 m inter-distance). . . .	21
2.5	Conceptual diagram of blind source separation via ICA.	23
2.6	Histogram of the speech signal in a subband domain: clean speech (the sharpest, blue curve), mixed speeches (the second sharpest, green curve), mixed speeches with background noise (the third sharpest, red curve), and noise-only (cyan, typical Gaussian curve).	24

2.7	Wavelet packet decomposition to simulated 24 band mel-filter bank, where a gray code method was used to determine frequency order . If a node is dominated by 0, its children are 0 to left and 1 to right; if a node is dominated by 1, its children are 1 to left and 0 to right. Gray code reads from bottom to top. Bold face bottom digits represent frequency order 1 to 24.	26
2.8	Subband output of WPD (top figure: x axis represents sample and y axis represents amplitude) and delay estimation result (bottom figure: x axis represents frame and y axis represents estimated delay in samples) using cross-correlation from the first lowest subband (a) to the last highest subband (x).	27
3.1	Block diagram of separation for two concurrent speech streams in the modulated complex lapped transform (MCLT) domain.	47
3.2	Microphone array magnitude response (dB) from -180° to 180° : (a) MVDR beamforming directivity pattern, (b) MVDR beamforming with nullforming directivity pattern, (c) adaptive MPDR beamforming without nullforming directivity pattern, and (d) adaptive MPDR beamforming with nullforming directivity pattern.	49
3.3	Conceptual diagram for the proposed method compared with individual beamforming and ICA.	51
3.4	Proposed regularized feed-forward ICA (RFICA) with IDOA based post-processing: doubling represents and single line represent vector and scalar, respectively.	52
3.5	Multi-microphone room impulse response measurement setup: total number of measured source positions is 24.	59
3.6	SIR measurement accuracy test.	60
3.7	(a) Contour plot of the improvement in SIR (dBC) as a function of the distance and separation angle, and (b) PESQ MOS improvement 2-D contour plot.	61
3.8	SIR improvement comparison for the setup of 1.22 m distance and 55° angle: (1) beamforming plus nullforming, (2) Saruwatari et. al., (3) beamforming plus nullforming plus IDOA based post-processing, and (4) RFFICA plus IDOA based post-processing.	62
3.9	Time domain speech signals: (a) one of the mixed channels, (b) and (c) original clean target speech and separated speech by the proposed method, (d) and (e) another original clean target speech and separated speech by the proposed method.	64

3.10	Spectrogram of speech signals (from 5 to 27.6 seconds and up to 4 kHz for purpose of clear presentation): (a) one of the mixed channels, (b) and (c) original clean target speech and separated speech by the proposed method, (d) and (e) another original clean target speech and separated speech by the proposed method.	65
3.11	(a) SIR versus iteration numbers, (b) SIR versus tap numbers.	66
3.12	Time domain speech signals (spontaneous real speech scenario, 1.22 m, 55°): (a) one of the mixed channels, (b) and (c) separated speech streams by the proposed method.	67
3.13	Spectrogram of speech signals (spontaneous real speech scenario, 1.22 m, 55°): (a) one of the mixed channel, (b) and (c) separated speech streams by the proposed method.	68
4.1	Conceptual diagram of the proposed joint optimization of DOA and speech based on EM beamforming.	71
4.2	Conceptual diagram of the generative model based determination of $M(k)$: example of frequency bin $N = 6$ and target speech dominated frequency bin $K = 3$ at $k = 2, 4, 6$	78
4.3	RIR measurement setup.	79
4.4	DOA estimation result	82
4.5	EMB result	83
4.6	DOA estimation result	84
4.7	Spectrogram	85
4.8	Time domain results: unprocessed channels (top), EM beamforming with one iteration (middle), EM beamforming + MMSE-logSA (bottom).	86
5.1	Block diagram of GSC beamformer. y_1 through y_N are measured input for beamforming, w_0 is a fixed beamformer shown in (5.7), C_n is the blocking matrix shown in (5.9), and $\vec{w}_n$ is the adaptable beamforming coefficient vector.	90
5.2	Results of applying multi-path GSC for the task of deconvolving the channel response from an impulse input: (a) original impulse input, (b) measured signal at microphone #1, (c) measured signal at microphone #2, (d) beamformed output (channel response is entirely removed and noise is ameliorated).	95
5.3	Time response of utterance “three”: (a) original speech signal taken from the TIDIGITS corpus, (b) measured speech at microphone #1, (c) beamformed output, (d) result of the two-step, beamform + estimate procedure, using an MMSE-logSA estimator, (e) result of applying the MMSE-logSA to measured speech at microphone #1.	96

5.4	Spectrogram of utterance “three”: (a) original speech signal taken from the TIDIGITS corpus, (b) measured speech at microphone #1, (c) beamformed output, (d) result of the two-step, beamform + estimate procedure, using an MMSE-logSA estimator, (e) result of applying the MMSE-logSA to measured speech at microphone #1.	97
5.5	(a) FBF output: blue dotted line is DSB output, black dotted lines are updated FBF output, red line is final FBF output after 20 iterations. Updated FBF output produces more impulse-like output by eliminating the effect of the designated reflections, in other words, more deconvolved output. (b) Estimated channel h_1 . (c) Estimated channel h_2 : red dots show the converged channel response after 20 iterations and the blue dotted lines are updated responses. The black line is for original RIR. The designated channel responses are almost perfectly identified.	104
5.6	(a) Estimated channel h_1 . (b) Estimated channel h_2 : red dots show the converged channel response after 20 iteration and the black line is for original RIR. The designated channel responses above the predefined threshold are almost correctly identified.	106
5.7	(a) Estimated channel h_1 . (b) Estimated channel h_2	107
5.8	Block diagram of optimal multi-microphone speech enhancement.	110
5.9	Time response of utterance “three” in the 55D condition: original (upper), after MVDR beamforming (middle), and after MVDR beamforming and MMSE-logSA postfiltering (lower).	113
5.10	Spectrogram of utterance “three” in the 55D condition: original (upper), after MVDR beamforming (middle), and after MVDR beamforming and MMSE-logSA postfiltering (lower).	113
5.11	EM-ML estimate result: (top) source spectrum (an utterance of the word “three”) only contaminated by additive noise: SNR is about 12 dB, (middle) one of the measurements: source spectrum contaminated by RIR as well as additive noise, (bottom) EM-ML estimated spectrum: sufficient statistic S_{ML}	119
6.1	Conceptual diagram for the marginalization: the fifth feature has been marginalized by deleting the fifth mean and the fifth column and row of covariance.	127
6.2	Recognition accuracy(%) vs SNR (dB) with white noise as background noise.	132

6.3	Recognition accuracy(%) vs SNR (dB) with dance music in the movie “Dream Girls” as background noise.	133
D.1	Block diagram of one unit for the audio pre-amp.	144
D.2	Inside the actual self-made 8-in 8-out pre-amp.	144
E.1	RIR measurement configuration.	146
E.2	Logarithmic sweep signal: time domain signal.	147
E.3	Conceptual diagram for deconvolution filter	148
E.4	Sample room impulse response measured with logarithmic sweep signal: (a) left smaller repetitions of RIR correspond to harmonic distortion. (b) RIR by excluding harmonic distortion.	149
E.5	Difference between generated and measured signal.	150
F.1	Graphical representation of HMM based speech recognition in a noisy environment.	152
F.2	(a) Log-likelihood given each model in the almost noise-free situation: blue is for model 1, red is for model 2. (b) Log- marginalized likelihood given by (F.6): blue is for model 1, red is for model 2.	156
F.3	(a) Log-likelihood given each model in the high-level noise situation: blue is for model 1, red is for model 2. (b) Log- marginalized likelihood given by (F.6): blue is for model 1, red is for model 2.	157

CHAPTER 1

INTRODUCTION

Speech communication is easily degraded by *mismatch* between a favorable and a realistic acoustic quality above certain allowable thresholds. The mismatch may come from environmental conditions such as the existence of interference speech, reverberation, and background noise. It may also come from non-environmental sources such as speaker dependent accents, especially in the case of people using a target language, which is not their mother tongue. Note that the source of the mismatch is necessarily neither stationary nor expected. Figure 1.1 shows a simplified diagram of the possible mismatch circumstances.

Although humans are naturally trained to defeat the mismatched circumstances, there surely exist practical limits on achieving reasonable decoding accuracy. In the case of hearing impaired people, the dependency on the mismatch is much more critical. It is also true that better-matched speech causes less fatigue, regardless of impairment. The dependence of automatic speech recognition (ASR) on the degree of mismatch is also critical. Therefore, speech processing to mitigate or even eliminate the mismatch has been a very active research topic not only for human-to-human interaction but also for ASR.

Methods to solve the mismatch problem have progressed in parallel depending on the application purpose. For human-to-human communication, the problem has been cast into an optimal estimation problem where the aim is to recover the most probable target speech signal with fewer unnatural side effects such as musical noise. Depending on the nature of the mismatch condition, differently named techniques have been proposed: source separation [1–4], de-reverberation [5–7], and de-noising or speech enhancement [8–12]. In fact, in a real acoustic environment, those methods should be combined optimally.

For robust ASR, we can simply use the optimally estimated speech to ob-

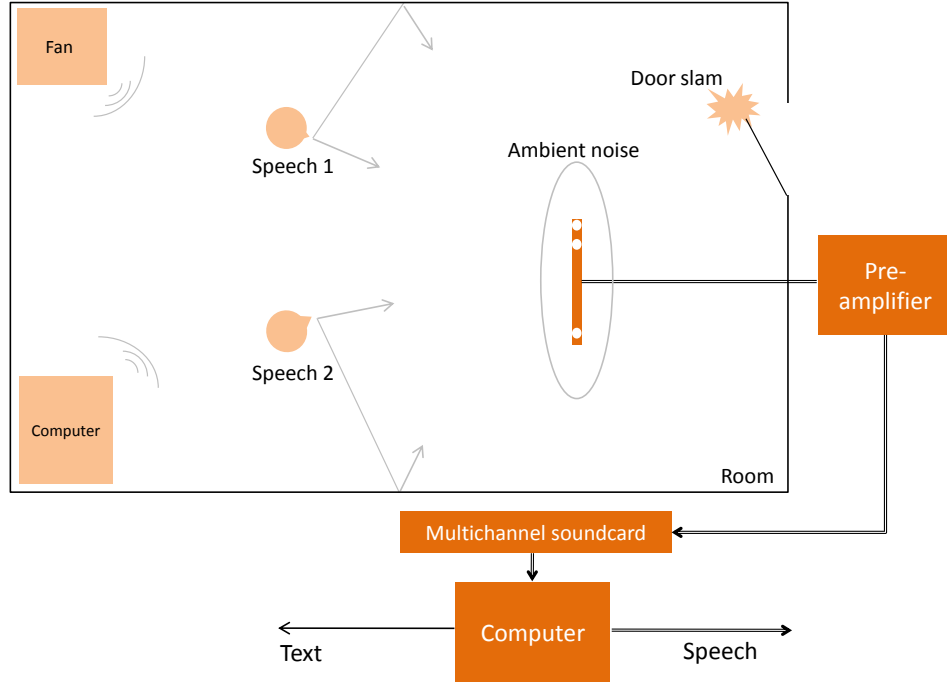


Figure 1.1: Simplified diagram for computer assisted speech communication.

tain feature vectors such as the mel-frequency cepstral coefficient (MFCC) [13] or perceptual linear prediction (PLP) [14]. We are also able to adapt or modify HMMs to explain the feature contamination or model mismatch with a small number of parameters, under the assumption that the mismatch condition is stationary enough so that the parameters can be trained [15, 16]. Recently, missing data theory has been applied to both approaches: the optimal estimation is rephrased as “imputation” [17], and model adaptation is supplemented with marginalization over missing feature components [18].

The clean speech estimation problem and robust speech recognition problem have been combined, usually in a straightforward way:

1. Simplified statistical modeling of the speech process.
2. Clean speech estimation based on 1), which is called “front-end processing” if ASR is the final destination.
3. Speech recognition using the clean speech estimation, or only the reliable portion of observed speech based on more complicated speech process modeling such as the hidden Markov model (HMM).

The key idea of this thesis work started from this simple discrepancy between

the statistical models of speech for speech estimation and speech recognition. In this thesis, a joint optimal method for clean speech estimation and robust ASR in a mismatched condition will be derived based on a generalized expectation maximization (GEM) scheme [19]. The following questions will be addressed.

- How can we estimate the speech signal as optimally as possible in the adversely mismatched environment?
- How can we use the optimally estimated speech in ASR with the same statistical model of the speech process?
- How can we use a more complicated statistical speech model, typical of ASR, for speech estimation purposes?
- How can we actually jointly optimize the performance in both contexts, speech estimation and ASR?

1.1 Graphical model for joint optimization of distortionless speech estimation and ASR

Figure 1.2 shows a graphical model representation for ASR and ideal binary mask (IBM) estimation. Inside the dotted box corresponds to the typical HMM based ASR [20]. To take the mismatch into consideration, the IBM part has been added. With perfect estimation of IBM [21], we can perform ASR with reliable feature components only [18]. Similarly, with perfect knowledge of the utterance, we can estimate the missing part in support of the reliable components [17]. To fulfill those two different objectives in a jointly optimal manner, a GEM algorithm is proposed as follows.

The missing speech estimation problem is cast into a maximum likelihood estimation given the reliable speech parts.

$$\hat{\vec{O}}_m(1:T) = \arg \max_{\vec{O}_m(1:T)} \log p \left(\vec{O}_r(1:T) | \vec{O}_m(1:T) \right), \quad (1.1)$$

where $\hat{\vec{O}}_m(t)$ is an estimated missing part, $\vec{O}_r(t)$ is the reliable portion, and $t = 1, 2, \dots, T$ where T is total number of frames. Estimation of $\hat{\vec{O}}_m$ can

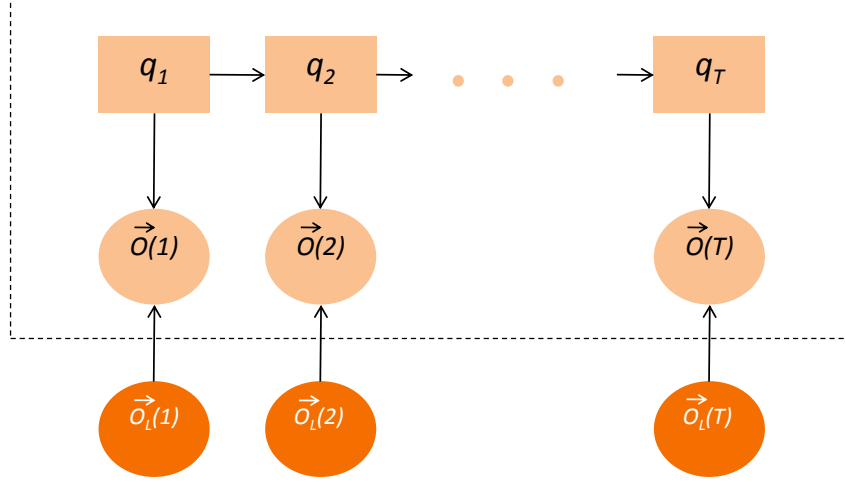


Figure 1.2: Graphical model representation for ASR and ideal binary mask estimation: q_t is a state in a given model, $\vec{O}(t)$ is an observation vector or feature vector, and $\vec{O}_L(t)$ is a binary mask vector of reliability for each component of $\vec{O}_L(t)$.

be reformulated into a GEM problem, considering the model for ASR as a latent variable.

$$\hat{\vec{O}}_m(1:T) = \arg \max_{\vec{O}_m(1:T)} Q(\vec{O}_m(1:T), \vec{O}_m(1:T)^0), \quad (1.2)$$

where $\vec{O}_m(1:T)^0$ corresponds to the previous estimate and

$$\begin{aligned} & Q(\vec{O}_m(1:T), \vec{O}_m(1:T)^0) \\ &= E_{p(C|\vec{O}_r(1:T), \vec{O}_m(1:T)^0)} \left[\log p(\vec{O}_r(1:T), C | \vec{O}_m(1:T)) \right] \\ &= \sum_{C \in \mathbf{C}} p(C | \vec{O}_r(1:T), \vec{O}_m(1:T)^0) \\ & \quad \cdot \log p(\vec{O}_r(1:T) | C, \vec{O}_m(1:T)) p(C | \vec{O}_m(1:T)) \\ &\sim \sum_{C \in \mathbf{C}} p(C | \vec{O}_r(1:T), \vec{O}_m(1:T)^0) \\ & \quad \cdot \log p(\vec{O}_r(1:T) | C, \vec{O}_m(1:T)), \end{aligned} \quad (1.3)$$

where $C \in \mathbf{C}$ represents an utterance model in the space of all possible models, e.g. a single digit model among all possible digit models.

In the maximization step, we need to solve the following equation to esti-

mate $\hat{\vec{O}}_m(t)$.

$$\begin{aligned}
0 &= \frac{\partial}{\partial \vec{O}_m(t)} \sum_{C \in \mathbf{C}} p(C | \vec{O}_r(1:T), \vec{O}_m(1:T)^0) \\
&\cdot \sum_{q \in \mathbf{q}} \left[\sum_{q' \in \mathbf{q}} \alpha_{t-1}(q' | C, \vec{O}_r(1:t-1), \vec{O}_m(1:t-1)^0) p(q | q') \right] \\
&\cdot \beta_t(q | C, \vec{O}_r(t+1:T), \vec{O}_m(t+1:T)^0) \log p(\vec{O}_r(t) | \vec{O}_m(t), C, q),
\end{aligned} \tag{1.4}$$

where $\alpha_t(q)$ and $\beta_t(q)$ are the conventional forward and backward variables at time t and state q [22]. Under the Gaussian mixture model (GMM) modeling of the likelihood, the following is the solution of (1.4).

$$\begin{aligned}
\hat{\vec{O}}_m(t) &\approx \sum_{C \in \mathbf{C}} p(C | \vec{O}_r(1:T), \vec{O}_m(1:T)^0) \\
&\cdot \sum_{q \in \mathbf{q}} \left[\sum_{q' \in \mathbf{q}} \alpha_{t-1}(q' | C, \vec{O}_r(1:t-1), \vec{O}_m(1:t-1)^0) p(q | q') \right] \\
&\cdot \beta_t(q | C, \vec{O}_r(t+1:T), \vec{O}_m(t+1:T)^0) E[\vec{O}_m(t) | \vec{O}_r(t), C, q].
\end{aligned} \tag{1.5}$$

After the iteration, we do not need to adhere to the initial IBM labeling about the reliability, because we can re-estimate each feature component given the estimate of the remaining portion. After the convergence, we can use the estimated gain to adjust the level of the pre-processed speech: this is more like a mel-filter bank graphical equalization using an utterance specific HMM. Note that GEM guarantees a local maximum; therefore, a good initial estimate of the IBM is very important. Also, it is obvious that more reliable features are better. Therefore, various optimal multi-microphone based speech processing methods are proposed to provide better opportunities to get more reliable components. After obtaining optimally estimated speech, the remaining part is to have optimal estimation of IBM to minimize the negative influence due to unavoidable mismatched components even after the optimal pre-processing. A speech decoder oriented IBM estimation method is proposed to optimally estimate the mask from the perspective of ASR. We can employ the trained speech process model using HMM structure and the location of speech source information to support the estimation as well. Given the best speech and IBM estimate, we can expect (1.5) will produce the best solution near the global optimum.

1.2 Multi-microphone based speech estimation in an adverse environment

1.2.1 Problem Formulation

The observed signal of $y_i(t)$ and $i = 1, 2, \dots, N_{mic}$, where N_{mic} is the total number of microphones, is given by

$$y_i(t) = \sum_{j=1}^{N_S} h_{i,j}(t) * s_j(t) + n_i(t), \quad (1.6)$$

where $h_{i,j}$ are the RIR from the j^{th} source speaker to the i^{th} microphone, $s_j(t)$ is the j^{th} source speech signal, $*$ denotes convolution, and n_i is the noise signal measured by the i^{th} microphone. The frequency domain representation of (1.6) is given by

$$Y_i(e^{j\omega}) = \sum_{j=1}^{N_S} H_{i,j}(e^{j\omega}) S_j(e^{j\omega}) + N_i(e^{j\omega}), \quad (1.7)$$

where $Y_i(e^{j\omega})$, $H_{i,j}(e^{j\omega})$, $S_j(e^{j\omega})$, and $N_i(e^{j\omega})$ denote the spectral component of $y_i(t)$, $h_{i,j}(t)$, $s_j(t)$, and $n_i(t)$, respectively, and N_S represents the number of speech sources. However, with the fact that speech processing is normally performed in the short-time Fourier transform (STFT) domain in mind [23], note that in fact this representation only holds properly when the window length is much larger than the whole room reverberation time. If it is not sufficiently long, $Y_i(e^{j\omega})$ is not a function of only the current frame $S_j(e^{j\omega})$ but needs to include the effects of the previous frames. Therefore, we modify (1.7) into a STFT domain representation:

$$Y_i(k) = \sum_{j=1}^{N_S} \sum_{l=0}^{N_H-1} H_{i,j}^l(k) S_j^{-l}(k) + N_i(k), \quad (1.8)$$

where $H_{i,j}^l(k)$ represents the k^{th} bin of the l^{th} frame STFT response of the original $h_{i,j}(t)$, $S_j^{-l}(k)$ represents the k^{th} bin of l frames before the current frame $S_j(k)$, and N_H is the number of frames to cover the whole impulse response. The proof, that (1.8) is a precise STFT representation with longer RIR than the frame length, is given in Appendix C in the context of overlap

addition [23]. Equation (1.8) can be written conveniently as

$$\vec{Y} = \sum_{l=0}^{N_H-1} \mathbf{H}^l \vec{S}^{-l} + \vec{N}, \quad (1.9)$$

where $\vec{Y} = [Y_1 \cdots Y_{N_{mic}}]^T$, $\vec{S}^{-l} = [S_1^{-l} \cdots S_{N_S}^{-l}]^T$, $\vec{N} = [N_1 \cdots N_{N_{mic}}]^T$, and

$$\begin{aligned} \mathbf{H}^l &= \begin{bmatrix} H_{1,1}^l & H_{1,2}^l & \cdots & H_{1,N_S}^l \\ H_{2,1}^l & H_{2,2}^l & \cdots & H_{2,N_S}^l \\ \vdots & \vdots & \ddots & \vdots \\ H_{N_{mic},1}^l & H_{N_{mic},2}^l & \cdots & H_{N_{mic},N_S}^l \end{bmatrix} \\ &= \begin{bmatrix} \vec{H}_1^l & \vec{H}_2^l & \cdots & \vec{H}_{N_S}^l \end{bmatrix}. \end{aligned}$$

Since we can assume that $N_i(k)$ is a zero mean circularly symmetric complex Gaussian random variable independent of $S(k')$ and $N_i(k')$ for $k' \neq k$, “k” is omitted for readability. Note that (1.9) holds for every specific frequency bin. We can also assume that $\vec{N}$ is a zero mean circularly symmetric complex Gaussian random vector [24] with the spectral covariance matrix $\mathbf{R}_n = E[\vec{N}\vec{N}^H]$. These assumptions about the source and noise signals give the posterior probability density function (PDF).

$$\begin{aligned} &p(\vec{Y} | \vec{S}, \vec{S}^{-1}, \dots, \vec{S}^{-(N_H-1)}, \mathbf{H}^0, \mathbf{H}^1, \dots, \mathbf{H}^{N_H-1}) \\ &= \frac{1}{\pi^N \det(\mathbf{R}_n)} e^{-(\vec{Y} - \sum_{l=0}^{N_H-1} \mathbf{H}^l \vec{S}^{-l})^H \mathbf{R}_n^{-1} (\vec{Y} - \sum_{l=0}^{N_H-1} \mathbf{H}^l \vec{S}^{-l})}, \quad (1.10) \end{aligned}$$

where superscript H means conjugate transpose. Note a couple of facts:

- To model a noise vector following circularly symmetric complex Gaussian, we assume the relation matrix $C = E[\vec{N}\vec{N}^T]$ [24], which is a common assumption [9, 25], but we need to be aware that in fact this is a rather strong assumption.
- $\mathbf{H}^l$ is not easily obtainable; therefore, we will estimate it in a blind and adaptive manner.
- The number of sources N_S is not necessarily the same as the number of microphones N_{mic} .

Given the problem formulation and statistical modeling as in (1.10), our main objective is to estimate the current frame source speech vector $\vec{S}$, or some of them in an optimal manner with some reasonable manipulation of the ambiguous room impulse response, which will be addressed in the subsequent chapters.

1.2.2 Sufficient Statistic for Multi-Microphone Speech

Balan and Rosca [25] introduced a sufficient statistic approach to get a one-microphone summary of $\vec{Y}$ using the Fisher-Neyman factorization theorem [26] under the assumption of only one speech source and the frequency representation of (1.7). Obtaining a sufficient statistic is important in the sense that multi-dimensional manipulation is not needed after having the one-microphone summary. Also, this leads to the minimum variance unbiased (or minimum mean squared error (MMSE)) estimator of the target speech according to the Rao-Blackwell theorem [26]. In this thesis, the sufficient statistic is generalized to deal with multiple target sources in a STFT domain representation as in (1.8).

Because the Gaussian PDF is a member of an exponential family, (1.10) can be factored as

$$\begin{aligned}
& p(\vec{Y} | \vec{S}, \dots, \vec{S}^{-(N_H-1)}, \mathbf{H}^0, \dots, \mathbf{H}^{N_H-1}) \\
&= C(\vec{S}, \dots, \vec{S}^{-(N_H-1)}) \\
& \cdot \exp \left(\sum_{j=1}^{N_S} \sum_{l=0}^{N_H-1} S_j^{-l*} \vec{H}_j^l H \mathbf{R}_n^{-1} \vec{Y} + (S_j^{-l*} \vec{H}_j^l H \mathbf{R}_n^{-1} \vec{Y})^* \right) h(\vec{Y}) \\
&= g_{S_1}(T_{\vec{H}_1^0}(S_1)) \cdots g_{S_1}^{-(N_H-1)}(T_{\vec{H}_1^{N_H-1}}(S_1^{-(N_H-1)})) \\
& \cdot g_{S_2}(T_{\vec{H}_2^0}(S_2)) \cdots g_{S_2}^{-(N_H-1)}(T_{\vec{H}_2^{N_H-1}}(S_2^{-(N_H-1)})) \\
& \cdots \\
& \cdot g_{S_{N_S}}(T_{\vec{H}_{N_S}^0}(S_{N_S})) \cdots g_{S_{N_S}}^{-(N_H-1)}(T_{\vec{H}_{N_S}^{N_H-1}}(S_{N_S}^{-(N_H-1)})) h(\vec{Y}), \quad (1.11)
\end{aligned}$$

where C , g , and h are arbitrary functions, and

$$T_{\vec{H}_j^l}(S_j^{-l}) = \frac{\vec{H}_j^l H \mathbf{R}_n^{-1} \vec{Y}}{\vec{H}_j^l H \mathbf{R}_n^{-1} \vec{H}_j^l}. \quad (1.12)$$

Note that $T_{\vec{H}_j^l}(S_j^{-l})$ is the maximum likelihood estimate of S_j^{-l} , obtained by taking the derivative of the log likelihood function (1.10) with respect to S_j^{-l} and setting it to zero.

$$\begin{aligned} 0 &= \frac{\partial \log p(\vec{Y}|\vec{S}, \dots, \vec{S}^{-(N_H-1)})}{\partial S_j^{-l}} \\ &= \vec{H}_j^l{}^H \mathbf{R}_n^{-1} \vec{Y} - \vec{H}_j^l{}^H \mathbf{R}_n^{-1} \vec{H}_j^l S + (\vec{H}_j^l{}^H \mathbf{R}_n^{-1} \vec{Y} - \vec{H}_j^l{}^H \mathbf{R}_n^{-1} \vec{H}_j^l S)^H. \end{aligned} \quad (1.13)$$

The Fisher-Neyman factorization theorem observes that, given (1.11),

$$p(S_j^{-l}|\vec{Y}, \vec{H}_j^l) = p(S_j^{-l}|T_{\vec{H}_j^l}(\vec{Y})). \quad (1.14)$$

Therefore, for any function $z(S)$

$$p(z(S_j^{-l})|\vec{Y}, \vec{H}_j^l) = p(z(S_j^{-l})|T_{\vec{H}_j^l}(\vec{Y})), \quad (1.15)$$

and

$$E[z(S_j^{-l})|\vec{Y}, \vec{H}_j^l] = E[z(S_j^{-l})|T_{\vec{H}_j^l}(\vec{Y})]. \quad (1.16)$$

Note a couple of important points:

- Again, (1.15) assumes that we know $\mathbf{H}^l$. How can we relax this assumption in such a way that we minimize the loss of optimality?
- We have overlapped observation vectors to estimate a specific frame of speech, because the current observation vector is composed not only of the current speech frame but also the previous frames via the reverberation process as in (1.9). If we introduce a sufficient statistic for the observation vector m frame before the current frame $\vec{Y}^{\cdot-m}$ as $T_{\vec{H}_j^l}^{\cdot-m}(S_j^{-l})$, then

$$T_{\vec{H}_j^l}^{\cdot-m}(S_j^{-l}) \approx T_{\vec{H}_j^{(l+m)}}(S_j^{-(l+m)}), \quad (1.17)$$

where $m \leq N_H - 1 - l$. Therefore, we can take the previous observation frames into consideration along with modeling delay to estimate the speech source as in the case of time domain deconvolution [27]. The Appendix B derives a sufficient statistic by taking the series of observation frames into account.

In this thesis, various multi-microphone based MMSE optimal speech estimation methods that depend on relaxation direction are discussed.

- Regularized feed-forward independent component analysis (ICA) with post-processing based on direction information is introduced in Chapter 3. In this method, instead of relying on an indirect method, estimation of $\mathbf{H}^l$ followed by estimation of S_j^{-l} , we directly estimate S_j^{-l} through learning the demixing matrix $\mathbf{W}^l$, which is assumed to be the re-framed STFT representation of the inverse of RIR.
- Joint estimation of direction of arrival (DOA) and speech based on expectation maximization (EM) beamforming is introduced in Chapter 4. In this method, DOA is assumed to be the only significant RIR so that it can be tracked with much less observation data that share the same DOAs for the target and interference speech. Note that we limit the number of frames N_H as 1, and the sufficient statistic, which produces a maximum likelihood estimate of the target speech, is implemented under EM formulation so that a generative model based speech estimation can improve the DOA estimation as well.
- Multipath generalized sidelobe canceller (GSC) based iterative blind estimation of distinct early room impulse response (RIR) and sequential MVDR beamforming is introduced in Chapter 5. In this method, the DOA estimation is extended to include dominant reflection paths from the target and interference speech source positions so that the sufficient statistic includes not only the direct path with DOA but also significant reflection paths.
- EM formulation based on modeling of the remaining RIR as a random variable is introduced in Chapter 5. In this method, the sufficient statistic is updated in the process of expectation by regarding the RIR as a latent variable.

1.3 Decoder oriented IBM estimation

With the multi-microphone optimal processing introduced in this thesis, a realizable sufficient statistic can be obtained, which means we have an equiv-

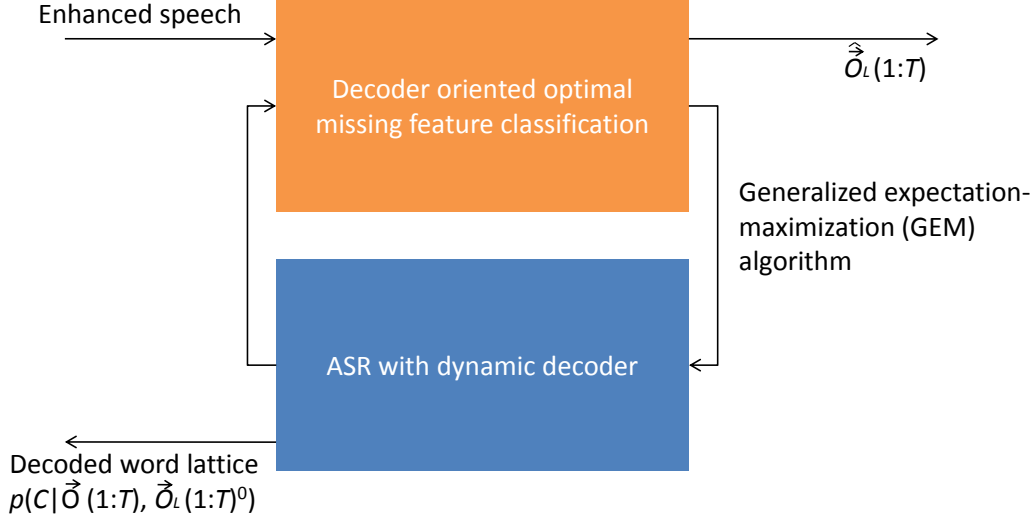


Figure 1.3: Block diagram for decoder oriented IBM estimation procedure.

alent one-channel summary for not only speech estimation but also ASR. Decoder oriented ideal binary mask (IBM) estimation is introduced in Chapter 6. In this method, according to (1.15), multi-microphone IBM estimation can be formulated with single-channel IBM estimation with the sufficient statistic. However, there may still exist unresolved mismatch. This portion of speech will be regarded as “missing” through the decoder oriented IBM estimation process. Since the IBM estimation is performed with the goal of increasing ASR performance, the newly proposed method does not need any assumptions about the mismatch statistic and can deal with even unseen, unexpected types of mismatch. Also, to solve the joint optimization problem (1.2), it is crucial to have the good IBM estimation, so that in the first iteration we can only rely on the reliable portion to estimate the missing portion.

Figure 1.3 shows a conceptual block diagram for the proposed method. Features are labeled as “missing” if doing so increases ASR performance; at the same time, ASR is performed aiming to increase the accuracy of the missingness label. The IBM estimation produces an estimated binary mask vector sequence (EBMVS),

$$\hat{\vec{O}}_L(1:T) = \arg \max_{\vec{O}_L(1:T)} \log p \left(\vec{O}(1:T) | \vec{O}_L(1:T) \right), \quad (1.18)$$

where $\vec{O}_L(t, f) \in \{0, 1\}$ according to 1 is for reliable feature, 0 is for unreliable feature, the frame index $t = 1, 2, \dots, T$ where T is the total number of frames, the feature index $f = 1, 2, \dots, F$ where F is the total number of features. Estimation of $\hat{\vec{O}}_L(1 : T)$ can be reformulated into a GEM problem, considering the EBMVS as a parameter to estimate and the model for ASR as a latent variable.

$$\hat{\vec{O}}_L(1 : T) = \arg \max_{\vec{O}_L(1:T)} Q(\vec{O}_L(1 : T), \vec{O}_L(1 : T)^0), \quad (1.19)$$

where

$$\begin{aligned} & Q(\vec{O}_L(1 : T), \vec{O}_L(1 : T)^0) \\ &= E_{p(C|\vec{O}(1:T), \vec{O}_L(1:T)^0)} \left[\log p(\vec{O}(1 : T), C | \vec{O}_L(1 : T)) \right]. \end{aligned} \quad (1.20)$$

In the maximization step, we need to solve the following equation to estimate $\hat{\vec{O}}_L(t, f)$.

$$\begin{aligned} & \hat{O}_L(t, f) \\ &= \arg \max_{O_L(t,f) \in \{0,1\}} \sum_{i=1}^{N_C} p(C_i | \vec{O}(1 : T), \vec{O}_L(1 : T)^0) \\ & \cdot \sum_q \left[\sum_{q'} \alpha_{t-1}(q' | C_i, \vec{O}(1 : t-1), \vec{O}_L(1 : t-1)^0) p(q | q') \right] \\ & \cdot \beta_t(q | C_i, \vec{O}(t+1 : T), \vec{O}_L(t+1 : T)^0) \\ & \cdot \log p(\vec{O}(t) | O_L(t, f), \vec{O}_L(t, \neg f) = \vec{O}_L(t, \neg f)^0, C_i, q), \end{aligned} \quad (1.21)$$

where C_i represents an utterance, and N_C is the total number of utterance models, $\vec{O}_L(t, \neg f)$ represents $\vec{O}_L(t)$ excluding $\vec{O}_L(t, f)$, $\vec{O}_L(1 : T)^0$ is the estimate of EBMVS at previous iteration, and $\alpha_t(q)$ and $\beta_t(q)$ are the conventional forward and backward variables at time t and state q [22].

1.4 Thesis organization

This thesis derives and tests the methods that have been introduced in Chapter 1, in the same order as they were introduced, except that the joint optimization of ASR and speech estimation will be dealt with last, because all

the preceding chapters can be regarded as pre-processing or initialization for that problem. In Chapter 2, background will be provided, which includes important concepts of sound capture using microphone arrays, array processing, source separation, de-reverberation, de-noising, and robust speech recognition. Chapter 3 presents a reverberant source separation method, where the reported signal-to-interference ratio (SIR) is almost an order of magnitude higher than the state of the art in a conventional setup (wide angle and short distance between speaker and microphones). Chapter 4 describes a joint estimation of speech and DOA based on an expectation maximization algorithm newly named “EM beamforming” with a probabilistic representation of the DOA per each frequency bin. In Chapter 5, the DOA consideration is expanded to the distinct early RIR estimation problem, if any, via an iterative method of system (RIR) identification and speech source estimation in a multi-path GSC setup. The remaining indistinct response is modeled as a random variable with the justification of more modal overlapping in the higher frequency range; a multi-microphone EM algorithm is derived to compensate for this statistical response part. Chapter 6 presents a hidden Markov model (HMM) based decoder oriented IBM estimation method. After convergence, since the IBM is estimated in the direction of increasing ASR performance, we can notice that it provides much better word recognition accuracy, especially in a very non-stationary background noise (e.g. dance music) situation. The proposed algorithm achieved word error rate reductions of 14.69% at 0 dB SNR and 40.10% at 15 dB SNR compared to a normal ASR method with optimal front-end processing. Chapter 7 concludes the thesis with a summary of the proposed methods, and the original joint optimization problem will be re-envisioned with a discussion about the direction of future work.

CHAPTER 2

BACKGROUND

In this chapter, important concepts and methods closely related to the main objectives of this thesis are addressed and reinterpreted.

2.1 Multi-microphone sound capturing and processing

Microphone array sound capturing provides another dimension (the “spatial domain”) compared with single microphone sound capturing, which means we can localize sources by focusing on the desired direction. Figure 2.1 shows the propagation of an impulse sound in the time domain across a linear array of eight electret condenser microphones with a three-dimensional plot. The diameter of each microphone was 6 mm. The actual array can be seen in Figure 2.2. We can visually confirm the same amount of time delay s_d in samples between microphones, which is actually expected to be calculated according to the following equation:

$$s_d = \frac{d \cdot \sin \theta}{c} \cdot f_s, \quad (2.1)$$

where, e.g. if $d = 37.95$ mm, $\theta = 60^\circ$, and $c = 340$ m/s, then $s_d = 4.64$ sample with sampling rate of $f_s = 48$ kHz.

Given the direction information on the target source, we can adjust the inter-delay among the microphone channels to produce a constructive or destructive summation. In the time domain multi-microphone representations, the angle resolution for beamforming is limited by the sampling rate f_s because the sample delay is bounded by the following equation:

$$|s_d| \leq \frac{d}{c} \cdot f_s, \quad (2.2)$$

which also determines the angle resolutions. For example, with $f_s = 16$

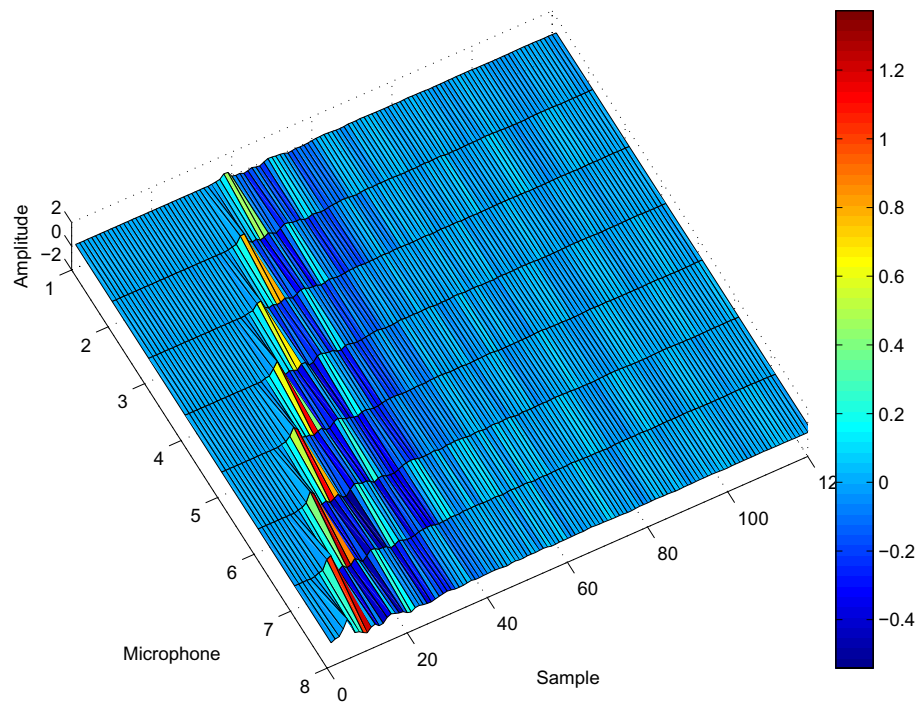


Figure 2.1: Propagation of an impulse sound in the time domain across a linear array of eight electret condenser microphones with incident angle of 60° .



Figure 2.2: Impulse response measurement setup: mouth simulator (B&K Type 4227), a linear microphone array with a penny in the center, and eight channel self-made pre-amplifier (circuit diagram and interior are shown in Appendix D.)

kHz, $s_d \leq 1.79$; therefore, in this case only three different angles (-34.05° , 0° , and 34.05°) can be evaluated in an 180° angle span. However, with a frequency domain approach, this restriction is relaxed, because non-integer delay can be evaluated by applying $\exp(-j2\pi \frac{k}{N} \cdot s_d)$ in the frequency domain, where N is the total number of samples for discrete Fourier transform (DFT) and $k = \lfloor -\frac{N}{2} \rfloor, \dots, \lceil \frac{N}{2} \rceil - 1$, or $\text{sinc}(\cdot - s_d)$ in the time domain with the consideration on circular shift due to the circular convolution of length N . Note that we need to sample in the frequency domain as well as in the time domain to apply DFT; therefore, the accuracy of non-integer delay is limited not only by the interpolation function, e.g. sinc function [28], but also the truncation and circular convolution due to DFT.

The distance between microphones is related to the lowest frequency range and the highest frequency range as well. To have better spatial resolution, longer distances compared to wavelength can be preferred. Without proper manipulation to reduce inter-channel correlation due to the background noise and reverberation, the array gain becomes lower in the lower frequency range. However, note also that the greater distance between microphones limits the high-frequency range because of spatial aliasing, called grating lobes, which is determined by the following equation [29]:

$$f_c = \frac{c}{2d}. \quad (2.3)$$

At frequency f_c , the wavelength is twice the distance between microphone. If the search range is from $\theta_1 = -90^\circ$ to $\theta_2 = 90^\circ$, then with delay of $T/2$, where the period $T = \frac{1}{f_c}$, the wave coming from left or right can be summed to be constructive; therefore, two main lobes are visible in the 180° angle span. Above f_c , two or more main lobes are produced. Figure 2.3 shows a conceptual explanation of this mechanism with -90° to 90° search range. Note that by decreasing the search angle $[\theta_1 \ \theta_2]$ as in (2.4), the spatial aliasing effect can be avoided at higher frequency for the purpose of DOA estimation [30].

$$f_c = \frac{c}{2d \cdot (\sin(\theta_1) - \sin(\theta_2))}. \quad (2.4)$$

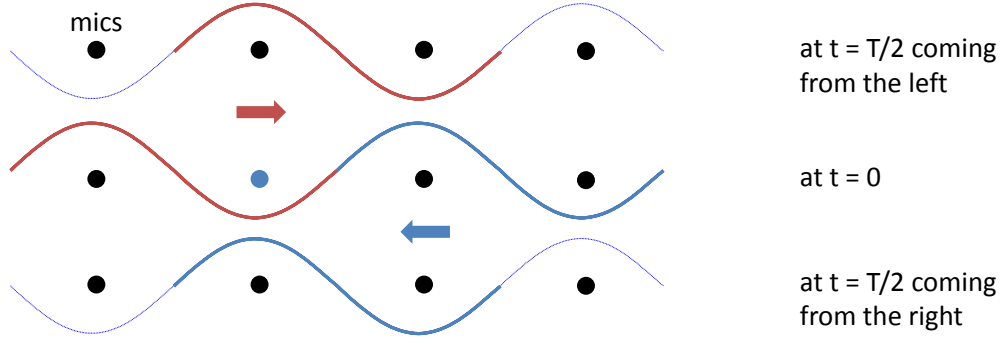


Figure 2.3: Illustration of double main lobes at -90° and 90° with f_c : inter-microphone delay $\frac{T}{2}$ corresponds to -90° and 90° incident angles, which has the same phase and can be summed to be constructive. Note that -90° and 90° incident angles produces the largest inter-microphone delay.

2.1.1 Beamforming

Without background noise and reverberation, all we need to beamform using a perfectly matched microphone array is “delay and sum” with proper information on the direction of arrival (DOA). However, because of the background noise, we need to compensate for the effect of noise, and this can be formulated as minimum variance distortionless response (MVDR) beamforming:

$$\arg \min_{\vec{W}} \vec{W}^H \mathbf{R}_n \vec{W} \text{ subject to } \vec{W}^H \vec{D} = 1, \quad (2.5)$$

where $\vec{D}$ is a steering vector toward the target direction, $\mathbf{R}_n$ is a noise covariance matrix, and $\vec{W}^H$ is a beamforming weight vector for multi-microphone input vector $\vec{Y}$. Often the noise-only covariance $\mathbf{R}_n$ is replaced by $\mathbf{R}$, which is the covariance matrix of the input vector (signal plus noise). The beamformer with $\mathbf{R}$ is known as the minimum power distortionless response (MPDR). In fact, $\mathbf{R}$ should also be estimated and the sample covariance matrix can normally be used [31]. To prevent instability due to the DOA mismatch, a regularization term can be added to the sample covariance matrix, which is called “diagonal loading” [31]. The actual MPDR beamformer in the frequency domain can be described as

$$\vec{W}^H = \frac{\vec{D}^H [\mathbf{R} + \lambda \mathbf{I}]^{-1}}{\vec{D}^H [\mathbf{R} + \lambda \mathbf{I}]^{-1} \vec{D}}, \quad (2.6)$$

where λ is the regularization term for diagonal loading. In fact, the diagonal loading is more important in the low-frequency range, where coherence becomes close to one such that $\mathbf{R}$ becomes more ill-conditioned. Coherence of diffuse noise is known to follow the equation given below [32]:

$$\Gamma(kd) = \text{sinc}\left(\frac{2\pi fd}{c}\right), \quad (2.7)$$

where as frequency f becomes close to 0, Γ becomes close to 1. Due to the uncorrelated self-noise produced by microphone or pre-amplifier, without the regularization, unnecessary self-noise boosting may happen.

2.1.2 DOA estimation

Robust estimation of DOA is one of the important parts of array processing. Estimation of DOA can be the pre-processing for beamforming as in (2.5), and it can be a main objective in the area of source tracking, e.g. navigation systems, and speaker tracking can also be a useful tool for smart room applications [33].

There are several well-established methods in the STFT domain. An overall review of them is given in [34]. In the frequency domain, we can estimate DOA with finer angle resolution with a relatively lower sampling rate. However, if the computation power is affordable, a higher sampling rate is better, since we cannot avoid problems from interpolation between integer samples. The simplest method, which is also fast, of estimating DOA by estimating the inter-delay t_d between the microphone pair is based on the weighted cross-correlation, called GCC-PHAT [35]:

$$t_d = \arg \min_{t \in (t_l, t_r)} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lceil \frac{N}{2} \rceil - 1} \frac{Y_1(k)Y_2(k)^*}{|Y_1(k)Y_2(k)^*|} \cdot e^{j2\pi \frac{k}{N}t}. \quad (2.8)$$

Note that t_d will be obtained considering all frequency bins with the weighting function $\frac{1}{|Y_1(k)Y_2(k)^*|}$, which actually normalizes the magnitude responses so that the cross-correlation of all the frequency bins is obtained only by considering phase difference with the same weighting. Therefore, we need to be careful when we use this method in a high-noise or reverberant environment.

We may use a different way of implementing GCC-PHAT:

$$t_d = \arg \min_{t \in (t_l, t_r)} \sum_{k=0}^{N-1} \frac{Y_1(k)Y_2(k)^*}{|Y_1(k)Y_2(k)^*|} \cdot \text{Kernel}_t(k), \quad (2.9)$$

where

$$\text{Kernel}_t(k) = \sum_{n=0}^{N-1} \text{sinc}(n + t_{[0 \ N-1]}) e^{-j2\pi \frac{k}{N} n}, \quad (2.10)$$

where

$$t_{[0 \ N-1]} = \begin{cases} t & t \leq 0 \\ t - N & t > 0 \end{cases}. \quad (2.11)$$

Figure 2.4(a) based on (2.9) shows the non-integer delay estimation result with the angle span of 20° of target DOA at 32° . Note the broad pattern on the cross-correlation around the target angle span with the GCC-PHAT based method.

Another category of DOA estimation is based on beamforming by searching for a maximum power producing angle. Among these methods, a direction finding based method called “multiple signal classification” (MUSIC) [36] per each frequency bin k and per each DOA θ can be formulated as follows:

$$\begin{aligned} P(k, \theta) &= \frac{1}{||\vec{U}_N^H \vec{d}_{k,\theta}||^2} \\ &= \frac{1}{\vec{d}_{k,\theta}^H (\vec{U}_N \vec{U}_N^H) \vec{d}_{k,\theta}} \\ &= \frac{1}{\vec{d}_{k,\theta}^H (\mathbf{I} - \vec{U}_S \vec{U}_S^H)}, \end{aligned} \quad (2.12)$$

where

$$\vec{d}_{k,\theta} = e^{-j2\pi \frac{k}{N} \vec{\Delta}}, \quad (2.13)$$

and where $\vec{\Delta} = \frac{[\dots -3d/2 \ -d/2 \ d/2 \ 3d/2 \ \dots] \sin \theta}{c}$ in the case of a linear array with distance d between microphones. The vectors $\vec{U}_S$ and $\vec{U}_N$ are the signal and noise subspace obtained from the sample covariance matrix per each frequency bin [31, 37]:

$$\begin{aligned} \hat{\mathbf{R}}_k &= \frac{1}{M} \sum_{m=1}^M \vec{Y}_k \vec{Y}_k^H \\ &= [\vec{U}_S | \vec{U}_N] \Lambda [\vec{U}_S | \vec{U}_N]^H, \end{aligned} \quad (2.14)$$

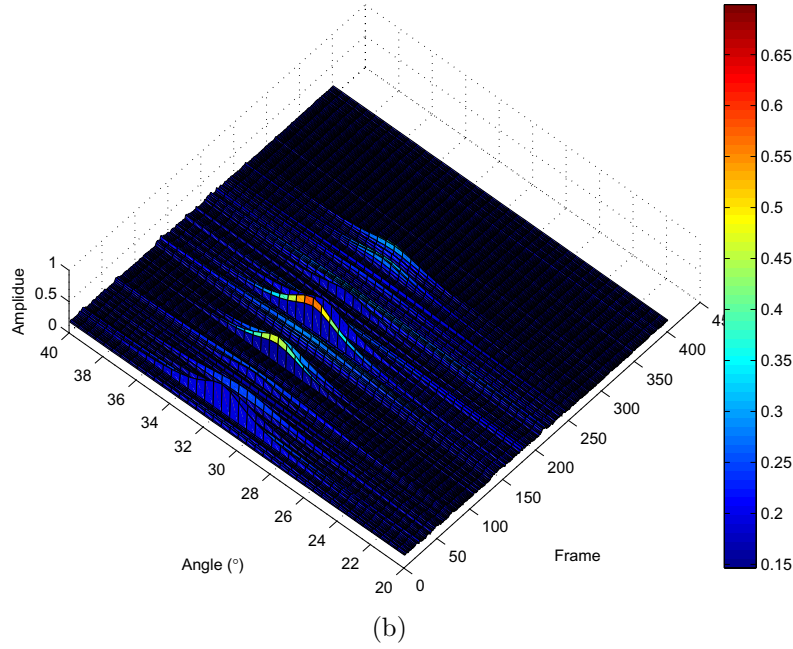
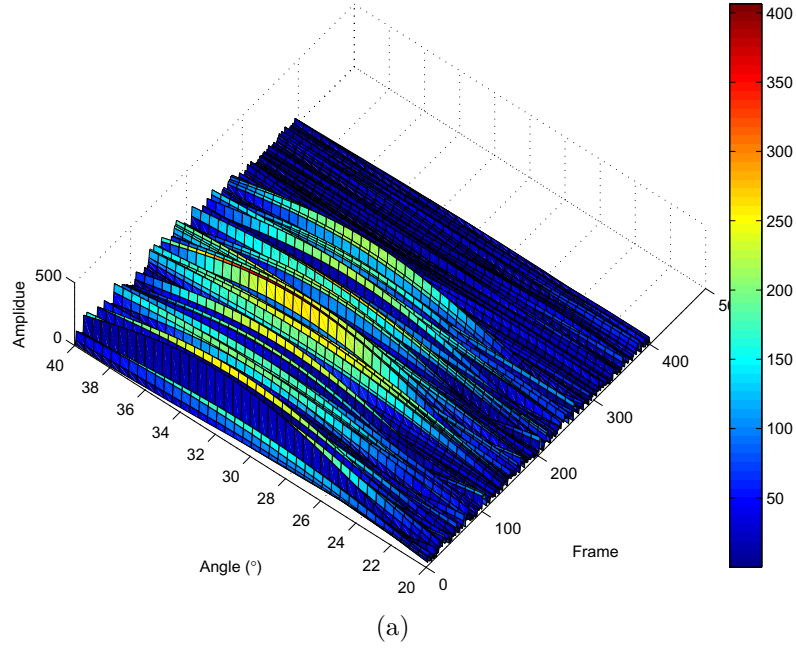


Figure 2.4: Direction of arrival estimation result (target speech spoken at 32° up to 300 frames, interference speech was at -32° during all frames) (a) sinc function applied GCC-PHAT based DOA estimation with pair of microphones at distance 0.076 m, (b) MUSIC based DOA estimation (8-microphones linear array with 0.03795 m inter-distance).

where Λ is a diagonal matrix with real value due to the fact that $\hat{R}_k$ is Hermitian by spectral theorem [37]. Figure 2.4(b) shows the DOA estimation result using an eight-microphone linear array, the maximum of which corresponds to the designated DOA as in Figure 2.4(a). However, note that due to the existence of a strong interference speaker (male) possibly dominating certain frequency bins more than the target speaker (female), neither MUSIC nor GCC-PHAT can produce the same DOA estimate during all the target speech active frames, even with the reduced search angle around the target angle. This issue will be explored in Chapter 4.

2.1.3 Independent component analysis (ICA)

Independent component analysis (ICA) has been a popular tool for blind source separation based on the higher-order independence assumption among the original sources before the mixing process [2, 3]. From the perspective of the original information-maximization (INFOMAX) approach [38], the goal is to recover the source signal at each time instance by maximizing the mutual independence of the nonlinear mapping from the demixed outputs. Maximizing the mutual independence is equivalent to maximization of entropy $H(\vec{U})$ and is also shown to be equivalent to minimization of the Kullback-Leibler divergence between the joint probability of $\vec{U}$ and the product of probability of each component of $\vec{U}$ [2]

$$D(p(\vec{U})||p(U_1)p(U_2)\cdots p(U_{N_{mic}})). \quad (2.15)$$

Figure 2.5 shows a conceptual diagram of blind source separation via ICA. The nonlinear mapping has the form of the cumulative density function (CDF) of the source signal so that after convergence (the output of nonlinear mapping follows the uniform probability density function (PDF)) the PDF of the source signal follows the derivative of the nonlinear mapping. With a proper choice of the nonlinear mapping function based on the speech signal statistic [2, 39], maximizing of the mutual independence results in maximization the super-Gaussianity of the separated speech signals.

To speed up the convergence, we can rescale the learning rule $\Delta \mathbf{W}$ of the separation filters by post-multiplying $\mathbf{W}^H \mathbf{W}$, which makes the metric in $\mathbf{W}$ independent of the particular point in $\mathbf{W}$. This method, called the natural

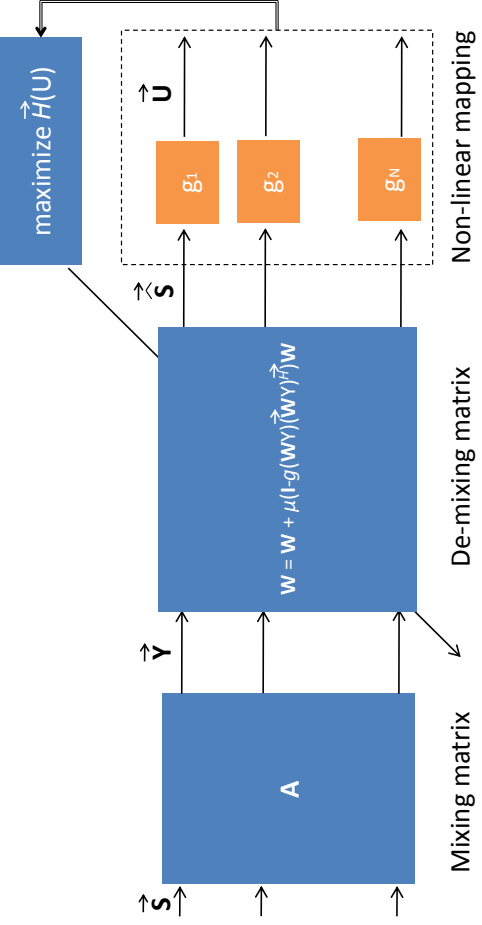


Figure 2.5: Conceptual diagram of blind source separation via ICA.

gradient method [40] is as follows.

$$\mathbf{W} = \mathbf{W} + \mu \left(\mathbf{I} - g(\mathbf{W}\vec{Y})(\mathbf{W}\vec{Y})^H \right) \mathbf{W}, \quad (2.16)$$

where μ is a step-size and g is a nonlinear function to explain the CDF of the original sources. To follow the convention of notation for beamforming and ICA, $\vec{W}^H$ is used for beamforming and $\mathbf{W}$ is used for ICA, both of which are actually the same except that beamforming is expressed using vector notation for a specific target source, but ICA is expressed using matrix notation for all sources, with one target source per column. Note that ICA is based on mutual information, and this objective function can be augmented with other constraints, if necessary.

The original ICA for the instantaneous mixing case can be extended to the time domain convolutive mixing such as mixing of multiple simultaneous speech signals in a reverberant acoustic space. The conventional approach is performed in the frequency domain as represented by (1.7), which will be extended to feed-forward network based ICA in Chapter 3. Figure 2.6 shows the average PDF of the real and imaginary values in all subbands. The clean speech signals show a typical super-Gaussian shape, but the sharpness of the peak is reduced when we mix two speeches and more greatly reduced when the background noise is also added. Note that the noise statistics show the Gaussian PDF. Therefore, maximizing mutual independence between separated speeches with a super-Gaussian prior in each subband results in

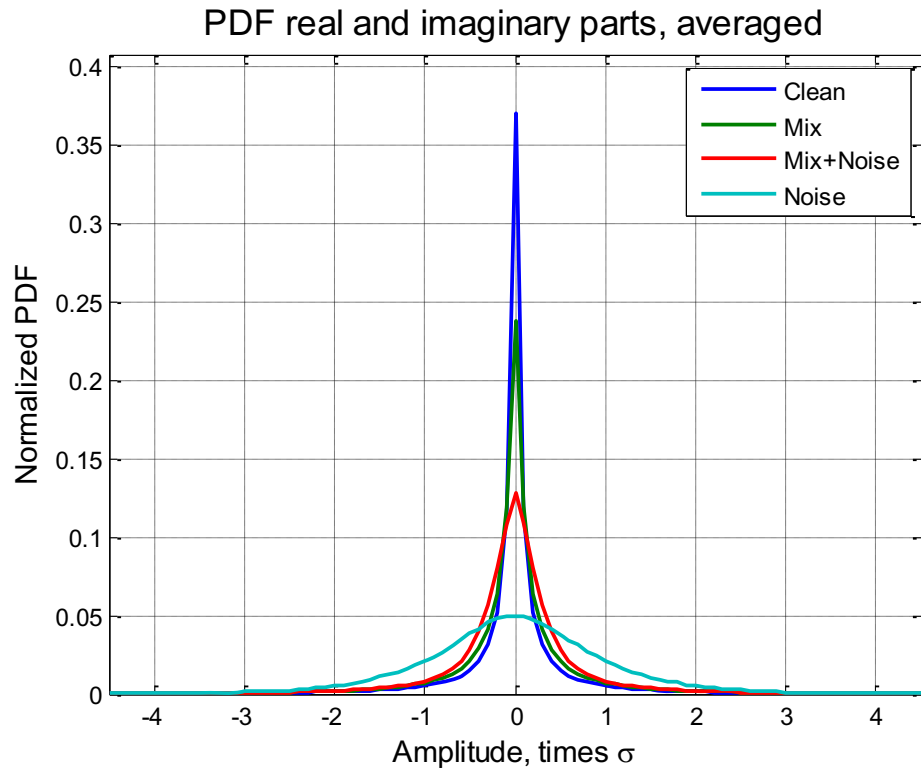


Figure 2.6: Histogram of the speech signal in a subband domain: clean speech (the sharpest, blue curve), mixed speeches (the second sharpest, green curve), mixed speeches with background noise (the third sharpest, red curve), and noise-only (cyan, typical Gaussian curve).

the separation of speeches.

2.1.4 Subband microphone array processing for ideal binary mask estimation

Conventionally, multi-microphone processing has been performed to increase SNR based on estimated DOA [41]. However, we have noticed that the SNR improvement is actually different depending on the frequency bands. If we can assume that noisy speech observation follows the MIXMAX condition [42] shown in (2.17), the frequency bands which exceed some threshold larger than 0 dB SNR can be simply considered as reliable.

$$\log |Y(\omega)| \approx \max(\log |S(\omega)|, \log |N(\omega)|). \quad (2.17)$$

Also, because we are now interested in subband-by-subband results rather than having a single whole-band result, we do the processing independently on each subband. In fact, the DOA estimation per each subband produces important cues for deciding whether the subband is reliable or not before taking the subband beamforming. Most DOA estimation algorithms produce one best direction, such that the RIR has the highest energy in that direction [35, 43]. Therefore, if there exists one speaker, the DOA will be the same regardless of the subbands once the SNR is high enough in those bands. The noise dominant subband might be able to be enhanced using beamforming toward the source direction, once the source direction has been confirmed by other high SNR subbands. We will discuss this issue in more detail in Chapter 4.

A simple simulation has been performed to validate whether the DOA estimation for each subband is actually useful as a reliability criterion. Two different utterances with some amount of delay (four samples) were generated using one clean utterance (“three” from the TIDIGITS corpus [44]). Uncorrelated white Gaussian noise was added to each utterance. A wavelet packet decomposition (WPD) with 24 subbands was generated using Haar wavelets [45]. The DOA was estimated before downsampling the subband, in order to maximize the angle resolution of the DOA. Figure 2.7 shows the WPD which was used for simulating a mel-filter bank [46], where, to determine frequency order after decomposition, a gray code method was used [47].

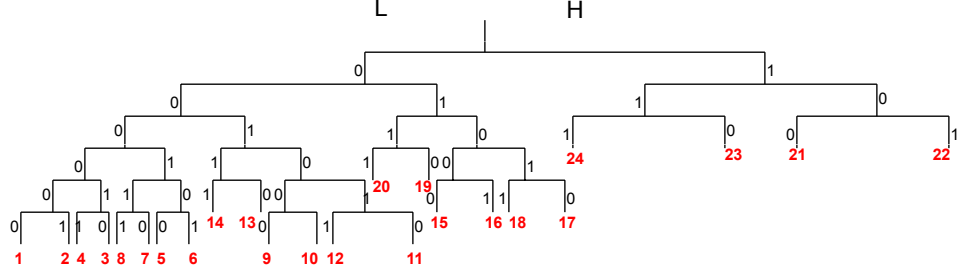


Figure 2.7: Wavelet packet decomposition to simulated 24 band mel-filter bank, where a gray code method was used to determine frequency order . If a node is dominated by 0, its children are 0 to left and 1 to right; if a node is dominated by 1, its children are 1 to left and 0 to right. Gray code reads from bottom to top. Bold face bottom digits represent frequency order 1 to 24.

Gray code reads from bottom to top. Bold face digits on leaf nodes represent the subband filters from first to twenty-fourth. Figure 2.8 shows the DOA estimation result, actually the inter-microphone delay estimation result, in signal dominant subbands and noise dominant subbands. In the signal dominant subbands, we can obtain robust delay estimation by using just a simple cross-correlation method.

2.2 Classical MMSE estimation and sufficient statistic

In this section, we first briefly review the classical optimal minimum mean squared error (MMSE) estimators [8, 9, 48] in conjunction with the known RIR for speech enhancement. After that, we will see how the sufficient statistic can link the multi-microphone measurement into the classical MMSE estimation.

2.2.1 Classical one-channel MMSE estimation

Based on the assumption that the spectral coefficient of the speech and the noise can be modeled as a complex Gaussian, MMSE spectral estimation can

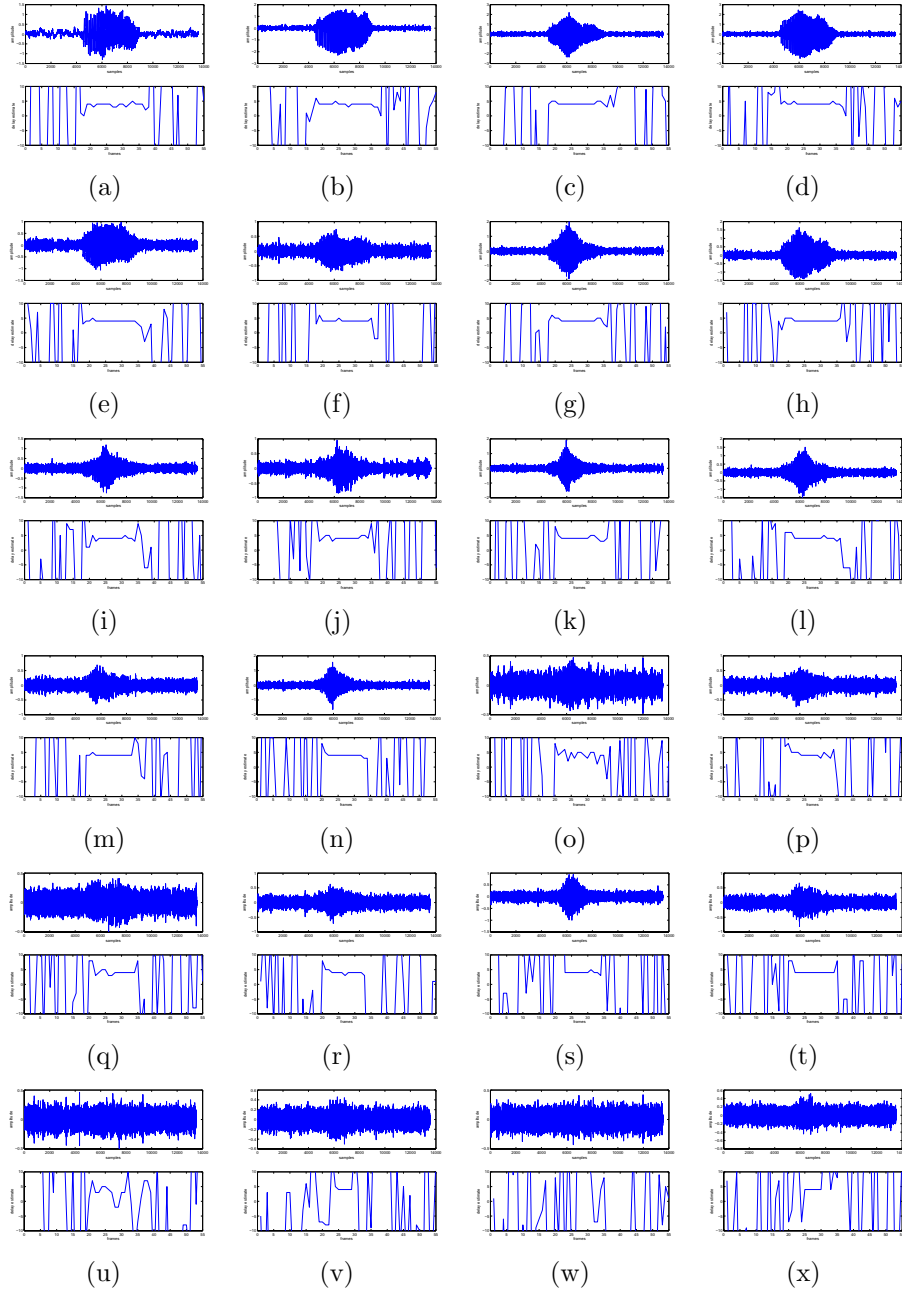


Figure 2.8: Subband output of WPD (top figure: x axis represents sample and y axis represents amplitude) and delay estimation result (bottom figure: x axis represents frame and y axis represents estimated delay in samples) using cross-correlation from the first lowest subband (a) to the last highest subband (x).

be defined as following:

$$E[S|Y_1, H_1] = E[S] + \frac{\text{Cov}(S, Y_1)}{\text{Cov}(Y_1)}(Y_1 - E[Y_1]). \quad (2.18)$$

If we define $\text{Cov}(S) = \sigma_S^2$ and $\text{Cov}(N_1) = \sigma_{N_1}^2$, then (2.18) can be interpreted as a Wiener filtering:

$$E[S|Y_1, H_1] = \frac{H_1^H \sigma_S^2}{|H_1|^2 \sigma_S^2 + \sigma_{N_1}^2} Y_1. \quad (2.19)$$

The MMSE spectral amplitude estimation can also be made under the same assumption. First, Y_1 can be represented in a form a little different from (1.7) with the number of sources $N_S = 1$ and the number of microphones $N_{mic} = 1$:

$$Y_1 = a|H_1|e^{j\alpha} + N_1, \quad (2.20)$$

where $a = |S|$ is a Rayleigh random variable, and uniformly distributed random phase α is the sum of the phase of S and H . Then Y_1 follows a complex Gaussian PDF with mean $a|H|e^{j\alpha}$:

$$p(Y_1|a, \alpha, H_1) = \frac{1}{\pi \sigma_{N_1}^2} \exp\left(-\frac{1}{\sigma_{N_1}^2} |Y_1 - a|H_1|e^{j\alpha}|^2\right). \quad (2.21)$$

Therefore, the MMSE estimator of $|S|$ is [49]

$$\begin{aligned} & \widehat{|S|}_{MMSE-STSA} \\ &= E[|S||Y_1, H_1] \\ &= \frac{\int_0^\infty a^2 \exp\left(-\left(\frac{a^2}{\sigma_S^2} + \frac{a^2|H_1|^2}{\sigma_{N_1}^2}\right)\right) \frac{1}{2\pi} \int_0^{2\pi} \exp(2a|H_1|\frac{Y_1}{\sigma_{N_1}^2} \cos\beta) d\beta da}{\int_0^\infty a \exp\left(-\left(\frac{a^2}{\sigma_S^2} + \frac{a^2|H_1|^2}{\sigma_{N_1}^2}\right)\right) \frac{1}{2\pi} \int_0^{2\pi} \exp(2a|H_1|\frac{Y_1}{\sigma_{N_1}^2} \cos\beta) d\beta da} \\ &= \frac{1}{|H_1|} \frac{\int_0^\infty a^2 |H_1|^2 \exp\left(-\left(\frac{a^2|H_1|^2}{\sigma_S^2|H_1|^2} + \frac{a^2|H_1|^2}{\sigma_{N_1}^2}\right)\right) \frac{1}{2\pi} \int_0^{2\pi} \exp(2a|H_1|\frac{Y_1}{\sigma_{N_1}^2} \cos\beta) d\beta da}{\int_0^\infty a |H_1| \exp\left(-\left(\frac{a^2|H_1|^2}{\sigma_S^2|H_1|^2} + \frac{a^2|H_1|^2}{\sigma_{N_1}^2}\right)\right) \frac{1}{2\pi} \int_0^{2\pi} \exp(2a|H_1|\frac{Y_1}{\sigma_{N_1}^2} \cos\beta) d\beta da} \\ &= \frac{1}{|H_1|} \widehat{H_1 S}_{MMSE-STSA}. \end{aligned} \quad (2.22)$$

where $|\widehat{H_1 S}|_{MMSE-STSA}$ is the Ephraim-Malah MMSE-STSA estimator [9].

$$|\widehat{H_1 S}|_{MMSE-STSA} = \Gamma(1.5) \frac{\sqrt{v}}{\gamma} \exp\left(-\frac{v}{2}\right) \left[(1+v) I_0\left(\frac{v}{2}\right) + v I_1\left(\frac{v}{2}\right) \right] R. \quad (2.23)$$

The gamma function is denoted by Γ ; I_0 and I_1 denote the modified Bessel functions of zero and first order.

$$v \triangleq \frac{\xi}{1+\xi} \gamma. \quad (2.24)$$

$$\xi \triangleq \frac{|H_1|^2 \sigma_S^2}{\sigma_{N_1}^2}. \quad (2.25)$$

$$\gamma \triangleq \frac{R^2}{\sigma_{N_1}^2}, \quad (2.26)$$

where $R \triangleq |Y_1|$. Note that the phase response α , which is the sum of the channel and source phase, can be marginalized; therefore, we can have the intuitive and simple way of getting MMSE estimation as in (2.22). Similarly the MMSE-logSA estimator can be obtained as [49]

$$\begin{aligned} \widehat{\log |S|}_{MMSE-logSA} &= E[\log |S| | Y_1, H_1] \\ &= \widehat{\log |H_1 S|}_{MMSE-logSA} - \log |H_1|, \end{aligned} \quad (2.27)$$

where $\widehat{\log |H_1 S|}_{MMSE-logSA}$ is the Ephraim-Malah MMSE-logSA estimator [48].

$$\widehat{\log |H_1 S|} = \frac{1}{2} \log \lambda + \frac{1}{2} \left(\log v' + \int_{v'}^{\infty} \frac{e^{-t}}{t} dt \right), \quad (2.28)$$

where λ satisfies

$$\frac{1}{\lambda} = \frac{1}{|H_1|^2 \sigma_S^2} + \frac{1}{\sigma_{N_1}^2}, \quad (2.29)$$

and

$$v' \triangleq \frac{\xi}{1+\xi} \lambda. \quad (2.30)$$

2.2.2 Multi-microphone MMSE estimation with sufficient statistic

Under the Gaussian model of the speech in a subband domain [9, 48], multi-channel MMSE estimation can be described as

$$E[z(S_j^{-l})|\vec{Y}, \vec{H}_j^l] = E[z(S_j^{-l})|T_{\vec{H}_j^l}(\vec{Y})], \quad (2.31)$$

where we have used the sufficient statistic given in (1.12). Therefore, once we have the sufficient statistic, we can perform the MMSE optimal estimation in the sense of the one-channel MMSE estimation. In other words, to obtain MMSE estimation, we need to take two steps, first obtaining a sufficient statistic and then applying conventional one-channel MMSE estimation.

2.3 HMM/GMM speech model based approach

In this section, we address how we can use some advanced statistical speech models for speech enhancement purpose. Conventionally, we assume that the speech follows the following statistical model [8, 9, 48]:

$$\begin{aligned} p(\vec{S}) &\sim N(\vec{0}, \text{diag}(\vec{\lambda}_S)) \cdot N(\vec{0}, \text{diag}(\vec{\lambda}_S)) \\ &\triangleq N^2(\vec{0}, \text{diag}(\vec{\lambda}_S)), \end{aligned} \quad (2.32)$$

where $\vec{S}$ is a vector with K equally spaced frequency bins and $\text{diag}(\vec{\lambda}_S)$ is a $K \times K$ diagonal covariance matrix with $\vec{\lambda}_S = [\lambda_{S,1} \ \lambda_{S,2} \cdots \ \lambda_{S,K}]^T$ as the diagonal entry. Note that S is a complex variable and the real and imaginary part are assumed to be independent with the same covariance matrix.

Before starting this section, we should note that the sufficient statistic is not dependent on the statistical model of speech, but on the noise. Hereafter, we assume the statistical model of the noise spectrum is as in (2.33); the sufficient statistic has the same form (1.12) regardless of the speech model:

$$p(\vec{N}) \sim N^2(\vec{0}, \text{diag}(\vec{\lambda}_N)). \quad (2.33)$$

In the speech recognition area, we usually use HMM/GMM as a speech model [20]. We can easily find a large gap between the speech models of

the classical speech enhancement literature and the speech recognition literature. There have been many attempts to close the gap [10, 17, 18, 50–57], but each such approach makes assumptions that can be improved. In speech recognition, we usually model the speech spectrum with mel-frequency dependent smoothing. This process is a crude dimensional reduction computed as

$$\tilde{S}(i) = \sum_{k \in K} S(k) M_i(k), \quad (2.34)$$

where $\tilde{S}(i)$ is the mel-spectral coefficients, $M_i(k)$ is the mel-filter, and $i = 1, \dots, N_m$ where N_m is the number of mel-spectral coefficients.

In this section, we do not consider the RIR as in the classical speech enhancement approach in Section 2.2, because the sufficient statistic for the multi-microphone followed by this type of single-channel MMSE estimator produces the optimal estimation in the sense of the single-channel estimation.

$$\vec{Y} = \vec{S} + \vec{N}, \quad (2.35)$$

where $\vec{Y}, \vec{S}, \vec{N}$ are vectors with K equally spaced frequency bins of the measured signal, speech signal, and noise signal.

2.3.1 Uncorrelated spectral coefficient model

In speech recognition, we are generally interested in $\log |\tilde{S}_i|$. The MMSE estimation of $\log |\tilde{S}_i|$ can be expressed as following:

$$\begin{aligned} \widehat{\log |\tilde{S}_i|} &= E\{\log |\tilde{S}_i| | y(t), 0 \leq t \leq T\} \\ &= E\{\log |\tilde{S}_i| | Y(1), Y(2), \dots, Y(K)\}. \end{aligned} \quad (2.36)$$

Let us first follow the statistical model of (2.32):

$$\widehat{\log |\tilde{S}_i|} = E\{\log |\tilde{S}_i| | \sum_{k \in K} Y(k) M_i(k)\}, \quad (2.37)$$

where

$$\begin{aligned}\sum_{k \in K} Y(k)M_i(k) &= \sum_{k \in K} S(k)M_i(k) + \sum_{k \in K} N(k)M_i(k) \\ \tilde{Y}(i) &= \tilde{S}(i) + \tilde{N}(i),\end{aligned}\tag{2.38}$$

and

$$\tilde{S}(i) \sim N^2(0, \tilde{\lambda}_{S,i}),\tag{2.39}$$

where

$$\tilde{\lambda}_{S,i} = \sum_{k \in K} \lambda_{S,k} M_i(k)^2,\tag{2.40}$$

and

$$\tilde{N}(i) \sim N^2(0, \tilde{\lambda}_{N,i}),\tag{2.41}$$

where

$$\tilde{\lambda}_{N,i} = \sum_{k \in K} \lambda_{N,k} M_i(k)^2.\tag{2.42}$$

Therefore, (2.38) is essentially the same as the classical setting without mel-filtering [8, 9, 48]; in other words, we can obtain MMSE optimal estimation of the mel-filter coefficients under the statistical model of (2.32). The MMSE log-mel-coefficient amplitude (MMSE-logMCA) estimation can also be derived as

$$\widehat{\log |\tilde{S}(i)|} = \frac{1}{2} \log \lambda_i + \frac{1}{2} \left(\log v_i + \int_{v_i}^{\infty} \frac{e^{-t}}{t} dt \right),\tag{2.43}$$

where λ_i satisfies

$$\frac{1}{\lambda_i} = \frac{1}{2\tilde{\lambda}_{S,i}} + \frac{1}{2\tilde{\lambda}_{N,i}}\tag{2.44}$$

and

$$v_i \triangleq \frac{\xi_i}{1 + \xi_i} \lambda_i\tag{2.45}$$

$$\xi_i \triangleq \frac{\tilde{\lambda}_{S,i}}{\tilde{\lambda}_{N,i}}.\tag{2.46}$$

2.3.2 Correlated spectral coefficient model

Classical speech enhancement methods are based on the assumption of statistical independence of the spectral coefficients in order to simplify the algorithms [9, 48]. If the analysis frame length approaches infinity, then this

assumption might be true asymptotically. However, in fact, the frame length is quite short (typically 10 to 30 ms) to extract the quasi-stationary non-stationary speech [23]. Therefore, there surely exist correlations among the coefficients, and this explains why we might be more successful when we use the cepstral domain feature with the independence assumption in speech recognition [17, 56]. In fact, most of dependence with this short-time frame based representation comes from speech production mechanism, which determines the frequency response of the frame so that the frequency components can be correlated each other, e.g. a vowel with some distinct pitch normally is presented with a typical harmonic structure. Our goal in this section is to derive the MMSE estimators without the independence assumption.

MMSE spectral estimation under the dependent spectral coefficient assumption is newly derived. According to the linear innovations sequences [58] under the assumption that $S(k), Y(1), Y(2), \dots, Y(K)$ are jointly Gaussian, the Wiener estimator can be given as follows:

$$\begin{aligned} E[S(k)|Y(1), Y(2), \dots, Y(K)] &= E[S(k)|\check{Y}(1), \check{Y}(2), \dots, \check{Y}(K)] \\ &= \sum_{i=1}^K E[S(k)|\check{Y}(i)] \\ &= \sum_{i=1}^K \frac{\text{Cov}(S(k), \check{Y}(i))}{\text{Cov}(\check{Y}(i))} \check{Y}(i), \end{aligned} \quad (2.47)$$

where

$$\check{Y}(k) = Y(k) - \sum_{i=1}^{k-1} E[Y(k)|\check{Y}(i)], \quad k \geq 2. \quad (2.48)$$

In (2.47), the $\check{Y}(k)$ s are all Gaussian and the $K \times K$ correlation matrix of $\vec{S}$ is assumed to be given; therefore, the single frequency bin Wiener filtering has been extended to whole-FFT estimation. Note that correlation could be sufficiently low, for the coefficients that are far apart, to cause the correlation matrix to be sparse.

As in the previous case, what if we do the mel-frequency smoothing? The only difference with the full spectrum case is just in the reduced size of the

correlation matrix from $K \times K$ to $N_m \times N_m$. For example,

$$\begin{aligned} E[\tilde{S}(i)\tilde{S}(j)] &= E[(\sum_{k \in K} S(k)M_i(k))(\sum_{k \in K} S(k)M_j(k))] \\ &= \sum_{k \in K} \sum_{l \in K} M_i(k)M_j(l)E[S(k)S(l)]. \end{aligned} \quad (2.49)$$

Note that the $N_m \times N_m$ correlation matrix might also be empirically obtained after smoothing in the training phase. Given the $N_m \times N_m$ covariance matrix,

$$\begin{aligned} E[\tilde{S}(i)|\tilde{Y}(1), \tilde{Y}(2), \dots, \tilde{Y}(N_m)] &= E[\tilde{S}(i)|\check{\tilde{Y}}(1), \check{\tilde{Y}}(2), \dots, \check{\tilde{Y}}(N_m)] \\ &= \sum_{j=1}^{N_m} E[\tilde{S}(i)|\check{\tilde{Y}}(j)], \end{aligned} \quad (2.50)$$

where

$$\check{\tilde{Y}}(i) = \tilde{Y}(i) - \sum_{j=1}^{i-1} E[\tilde{Y}(i)|\check{\tilde{Y}}(j)], \quad j \geq 2. \quad (2.51)$$

Without independence assumptions, MMSE-logSA or MMSE-logMCA do not seem to be easily obtainable, because $\log |\tilde{S}|$ cannot be modeled to be a Gaussian PDF if we model S to follow Gaussian. Therefore, the linear innovation sequences as in (2.47) cannot be directly constructed. However, note that in speech recognition we usually use the approximation that $\log |\tilde{S}|$ follows a Gaussian PDF. Once we assume that this approximation is good enough, then the same argument can be made as in (2.47). Note that in this case we assume that we know the correlation between $\log |\tilde{S}(i)|$ and $\tilde{S}(j)$ rather than $\tilde{S}(i)$ and $\tilde{S}(j)$. Therefore, MMSE-logMCA can be given as

$$\begin{aligned} &E[\log |\tilde{S}(i)||\tilde{Y}(1), \tilde{Y}(2), \dots, \tilde{Y}(N_m)] \\ &\simeq E[\log |\tilde{S}(i)||\check{\tilde{Y}}(1), \check{\tilde{Y}}(2), \dots, \check{\tilde{Y}}(N_m)] \\ &= \sum_{j=1}^{N_m} E[\log |\tilde{S}(i)|\check{\tilde{Y}}(j)]. \end{aligned} \quad (2.52)$$

Recently, $S(k)$ has been proposed to be modeled as a super-Gaussian distribution, such as Laplacian, due to the short length of the analysis frame [59, 60]. In this section, we try to use the standard Gaussian modeling with full-covariance matrix, but the non-Gaussian approaches do not address the correlation among the DFT coefficients. Note that by using GMM we may

have similar PDF with the non-Gaussian PDFs. Even for the mel-smoothed spectrum, the central limit theorem plays a greater role in causing them to be nearer to the Gaussian PDF. In fact, because of the non-Gaussianity, Chen et al. also used an approximation [60], and it might be interesting to compare the GMM approximation of the mel-spectra with Chen et al.'s approximation with Laplacian. In Appendix F, a Monte Carlo method will be exploited to replace the integration over the non-Gaussian source PDF.

2.3.3 Uncorrelated spectral coefficient Gaussian mixture model (GMM)

In this section, we investigate how we can derive MMSE estimation with the uncorrelated spectral coefficient Gaussian mixture model (GMM). The case considering correlation will be covered in the following section. The main difficulty in the GMM case lies in the non-Gaussianity of the mixture model. However, usually when we deal with this mixture model, we assume there exists a latent variable which specifies the membership of the mixture. Let us denote this as C . The GMM can be defined as follows:

$$p(\vec{S}) \sim \sum_{i=1}^{N_c} p(C = i) N^2(\vec{0}, \text{diag}(\vec{\lambda}_{N_i})). \quad (2.53)$$

MMSE spectral estimation can be given as

$$\begin{aligned} E[S(k)|Y(1), Y(2), \dots, Y(K)] &= E[S(k)|Y(k)] \\ &= \int_S S p(S|Y) dS \\ &= \int_S \left(\sum_{C=1}^{N_c} S p(S, C|Y) \right) dS \\ &= \sum_{C=1}^{N_c} \left(\int_S S p(S|Y, C) dS \right) p(C|Y) \\ &= \sum_{C=1}^{N_c} E[S|Y, C] p(C|Y) \\ &= \sum_{C=1}^{N_c} \left(\frac{\text{Cov}(S, Y|C)}{\text{Cov}(Y|C)} Y \right) p(C|Y), \end{aligned} \quad (2.54)$$

where we omit k from the second line for readability,

$$p(C|Y) \sim p(Y|C)p(C), \quad (2.55)$$

and we assume that we know $p(S|C)$ and $p(C)$ in the training phase. Note that if C is given, $E[S|Y, C]$ is an ordinary MMSE spectral estimation process. Therefore, Gaussian mixture MMSE spectral estimation is summarized as a weighted sum of mixture components. Even with mel-frequency smoothing, the MMSE spectral estimation under GMM is straightforward. The MMSE-logMCA can be also easily derived as follows:

$$E[\log |\tilde{S}(k)| | Y(1), Y(2), \dots, Y(K)] = \sum_{C=1}^{N_c} E[\log |\tilde{S}(k)| | \tilde{Y}(k), C] p(C | \tilde{Y}(k)) \quad (2.56)$$

and $E[\log |\tilde{S}(k)| | Y(k), C]$ is almost same with (2.43) except that we do the estimation for each mixture and average them according to the weights. Note that (2.54) and (2.56) are the extension of Huang et al.'s approach given in (10.146) of [51].

2.3.4 Correlated spectral coefficient Gaussian mixture model (GMM)

Now, MMSE estimation can be extended into the correlated spectral coefficient Gaussian mixture model (GMM) case. We might expect this extension to be straightforward. However, as we see now, in fact this extension requires much more computational complexity.

$$\begin{aligned} & E[S(k) | Y(1), \dots, Y(K)] \\ &= \int_{S(k)} S(k) p(S(k) | Y(1), \dots, Y(K)) dS(k) \\ &= \int_{S(k)} \left(\sum_{C_1=1}^{N_c} \dots \sum_{C_K=1}^{N_c} S(k) p(S(k), C_1, \dots, C_K | Y(1), \dots, Y(K)) \right) dS(k) \\ &= \sum_{C_1=1}^{N_c} \dots \sum_{C_K=1}^{N_c} E[S(k) | Y(1), C_1, \dots, Y(K), C_K] \\ & \cdot p(C_1, \dots, C_K | Y(1), \dots, Y(K)), \end{aligned} \quad (2.57)$$

where $E[S(k)|Y(1), C_1, \dots, Y(K), C_K]$ can be obtained as in (2.47). Similarly, the MMSE-logMCA can be derived.

$$\begin{aligned}
& E[\log |\tilde{S}(i)| | \tilde{Y}(1), \dots, \tilde{Y}(N_m)] \\
& \simeq \sum_{C_1=1}^{N_c} \cdots \sum_{C_K=1}^{N_c} E[\log |\tilde{S}(i)| | \tilde{Y}(1), C_1, \dots, \tilde{Y}(N_m), C_K] \\
& \cdot p(C_1, \dots, C_K | \tilde{Y}(1), \dots, \tilde{Y}(N_m)), \tag{2.58}
\end{aligned}$$

where $E[\log |\tilde{S}(i)| | \tilde{Y}(1), C_1, \dots, \tilde{Y}(N_m), C_K]$ can be obtained as in (2.52).

2.3.5 Hidden Markov model with GMM

In speech recognition, we generally use hidden Markov models (HMMs) to model the dynamics of the speech process [20, 22]. Hidden Markov models for speech recognition can be represented using a probabilistic relationship among the hidden states, which is responsible for segregating frames sharing similar acoustical characteristics, and using observation vectors such as MFCCs and PLPs to describe the speech process with sufficient dimensional reduction, which are shown to efficiently capture distinctive features, such as the spectral envelope. To model the dynamics of the speech process, the inter-state relationship is defined using transition probability. Different types of topology to model the relationship, oftentimes a left-to-right model allowing the self-looping or transition only from the lower indexed state, are used for speech modeling, but in some cases an ergodic model can also be used [61]. Given the state index, the observation probability can be modeled as a parametric continuous PDF.

The GMM is a popular model for capturing not only the multi-modality but also the super-Gaussianity of speech observation PDF. With the cepstral representation, the inter-correlation among the feature components is often assumed to be zero such that the covariance matrix of each Gaussian is modeled as a diagonal matrix. With the trained model, speech recognition is performed on an incoming speech feature stream by searching for a model producing maximum likelihood with marginalization on possible hidden state streams. For computational reasons, the likelihood may be calculated with

the best state sequence by the Viterbi algorithm [20].

The standard HMM can be more generalized to reflect a top-down hierarchy such as a language model or to deal with different types of modality in feature vectors, e.g. a video stream with lip motion detection or an articulatory feature stream [62], using a dynamic Bayesian network [63] with a graphical model based dependency structure representation [64]. Once we set the structure, inference and learning of the probabilistic relationships among the network parameters can be performed using message passing algorithms such as the sum-product algorithm [65], belief propagation [66], the forward-backward algorithm or the Baum-Welch algorithm [22].

With a pre-trained HMM for the speech process, MMSE estimation based on this model could easily be thought possible. However, note that in this case we have another important assumption on the given model because different HMMs are trained for each different target utterance. Therefore, we should know what utterances have been spoken to obtain MMSE estimation of the speech spectral coefficients. This actually contradicts the objective of speech enhancement as a method for improving speech recognition, and it is not always possible to have the transcription with the noisy speech for other applications. In order to avoid this contradiction, we defer hard decisions about the speech signal until the actual decision stage of speech recognition. In other words, rather than having an enhanced signal followed by a standard recognizer. In Chapter 6, we maximally utilize the probabilistic information of the corrupted speech signal to improve recognition accuracy. Similar ideas have been already introduced in various recent studies [18, 52–54, 56, 67]. However, all these approaches again start the discussion based on the HMM with GMM of $\log(|\tilde{S}|)$. In the next section, we will discuss the model-based robust speech recognition method where we will see how we can realize similar ideas based on the HMM with GMM of S or $\tilde{S}$. After having N-best lattice of recognized utterances, we might be able to use the HMM with the information about the transcription for optimal speech enhancement.

2.4 Missing feature approach

In speech recognition with hidden Markov models (HMMs) with Gaussian mixture models (GMMs) trained to observe cepstral domain features, a sin-

gle contaminated spectral component affects all the cepstral components. In the classical estimation approach given in Section 2.2, our objective was to estimate the source spectrum given the measured noisy spectrum. The estimated source spectrum can produce better recognition accuracy than when we use the noisy spectrum without any front-end pre-processing. There has been another approach, the so-called missing feature approach [18, 56], where we simply admit that there exist frequency components which cannot be reliably measured. To exclude those components, marginalization over those components is performed, and the marginalized likelihood function $p(\vec{S}_r|M)$ replaces the original one $p(\vec{S}_r, \vec{S}_u|M)$, where $\vec{S}_r$ and $\vec{S}_u$ are the reliable and unreliable components, respectively, and M is a model or simply an HMM state.

$$p(\vec{S}_r|M) = \int p(\vec{S}_r, \vec{S}_u|M) d\vec{S}_u. \quad (2.59)$$

To be successful with this approach, the process of classifying the reliable and unreliable features is critical. There have been several approaches for solving this classification problem [68–71]; we will extend the classification scheme toward increasing speech recognition performance in Chapter 6.

Marginalization over cepstral features does not seem to be done easily, because a small portion of the spectrum can affect all the cepstral features. To mitigate this problem, spectral features with a diagonal covariance matrix have been used [17, 18, 56]. However, spectral components are not conditionally independent given knowledge of the phoneme; therefore, diagonal-covariance modeling of the spectrum is ineffective. If we try to use full-covariance with spectral features to reflect the inter-correlation among spectral coefficients, then we should have far more training data; this is not promising when the training data are limited. We will also discuss how we can apply the missing feature approach reflecting the inter-correlation in Chapter 6.

2.4.1 Bounded marginalization and weighted likelihood

In the missing feature scheme, it is known that bounded marginalization works better than complete marginalization. When we classify unreliable features, sometimes we might not need to discard the observation entirely. For example, when we know that the observation is dominated by noise, we

might say that the speech level is at least below the noise level. This information is actually valid under the MIXMAX approximation in (2.17) [42]. For example, if we have partial information sufficient to specify that $B_{low} < \log |S(\omega)| < B_{high}$ given $Y(\omega)$ or $\log |Y(\omega)|$, then we can estimate

$$p(\log |Y(\omega)| | model) = \int w(x) p(x | model) dx, \quad (2.60)$$

where $x = \log |S(\omega)|$, and

$$w(x) = \frac{1}{B_{high} - B_{low}} U[B_{low} : B_{high}](x), \quad (2.61)$$

and

$$U[a : b](u) = \begin{cases} 1 & a \leq u \leq b \\ 0 & otherwise \end{cases}. \quad (2.62)$$

Since $\frac{1}{B_{high} - B_{low}}$ is the same regardless of the different *models*, the weighted marginalization scheme with $w(x)$ (2.60) becomes the bounded marginalization of $p(\log |S(\omega)| | model)$.

Note that the bounded marginalization scheme for the missing feature approach is regarded as a problem of finding the weighting function, and the weight function has a hard decision boundary. Once classified as unreliable, we take the uniform smoothing for those bounds, and otherwise we just take a delta function on the observation. We study how this problem can be handled. Multi-microphone processing for reliability decision and SNR improvement will be introduced, so that the decision can be more accurate, and more reliable features can be obtained. This approach still potentially has an uncertainty problem, but it has the merit of low complexity in terms of performing the integration. However, note that even with this scheme, there remains a computational issue, because the integration of the multivariate Gaussian with a full covariance matrix is still demanding [72].

2.4.2 Using $p(z(\vec{Y}) | z(S), \vec{H})$ as weighting function for the weighted likelihood

In this section, we try to utilize the distribution $p(z(\vec{Y}) | z(S), \vec{H})$ as the weighting function rather than using $E[z(S) | \vec{Y}, \vec{H}]$ to enhance the unreli-

able speech. In [53], $p(\log |\tilde{Y}_1| |\log |\tilde{S}|)$ has been introduced, and the idea of using it in the spectral domain under the independence assumption among the coefficients is currently the state of the art in terms of speech recognition accuracy [73]. This section is an extension of the approach into the multi-microphone case. However, we want to avoid the unnecessary error in the following approximation, which is a common approximation in the previous approach, by dropping $2N^H HS$ [53, 73, 74]:

$$\begin{aligned} |Y|^2 &= |H|^2 |S|^2 + |N|^2 + 2N^H HS \\ &\approx |H|^2 |S|^2 + |N|^2. \end{aligned} \quad (2.63)$$

We start our inference from the spectral coefficients themselves by assuming S or $\tilde{S}$ to be a mixture Gaussian, not by assuming $\log |\tilde{S}|$ to be a mixture Gaussian. The weighted likelihood is given as

$$\begin{aligned} p(z(\vec{Y})|\vec{H}, M) &= \int_{z(S)} p(z(\vec{Y})|z(S), \vec{H}) p(z(S)|M) dz(S) \\ &= \int_{z(S)} \frac{p(z(S)|\vec{Y}, \vec{H}) p(z(\vec{Y})|\vec{H})}{p(z(S))} p(z(S)|M) dz(S) \quad (2.64) \\ &\simeq \int_{z(S)} \frac{p(z(S)|T_{\vec{H}}(\vec{Y})) p(z(T_{\vec{H}}(\vec{Y})))}{p(z(S))} p(z(S)|M) dz(S), \end{aligned}$$

where M is a given model and in the third line we approximate $p(z(\vec{Y})|\vec{H}) = p(z(T_{\vec{H}}(\vec{Y})))$. This weighting scheme is in fact shown to be a generalization of bounded marginalization in the missing feature approach.

The weighted likelihood has some interesting characteristics. For illustration, let us define $z(S)$ as S , which means that the HMMs are assumed to be trained using spectral coefficients or mel-smoothed spectral coefficients. If $T_{\vec{H}}(\vec{Y})$ is high enough compared with the noise variance, then $p(\vec{Y}|z(S), \vec{H})$ becomes nearly a delta function centered around $T_{\vec{H}}(\vec{Y})$. In other words, this case is similar to the case of obtaining likelihood using the most likely estimate of S . However, if the noise variance is too large, the weighting function approximates a wide flat function, which becomes very similar to the missing feature weighting function as in (2.61).

Unlike the previous case of $z(S)$ as S , if we model $z(S)$ as $\log |S|$ or $\log |\tilde{S}|$, then $p(z(S)|T(\vec{Y}))$ is no longer Gaussian. Therefore, we should consider realistic approximation methods to perform this integration. The Laplace ap-

proximation and expectation propagation methods [75] can be used to tackle this approximation problem. Laplace approximation is a method used to approximate the posterior as a Gaussian around the maximum of the posterior using second-order Taylor expansion, and it has been used successfully in many applications [53, 73, 74]. The expectation propagation (EP) algorithm is also an approximation method with many applications. In Chapter 6, the weighting function is simplified as two extreme functions with IBM estimation in order to avoid multiple integration in a full covariance structure: Dirac-delta centered on the clean speech S_i or uniform regardless of Y_i , where $i = 1, 2, \dots, N_{feature}$, corresponding more to (2.17).

CHAPTER 3

REVERBERANT SPEECH SEPARATION USING RFFICA AND IDOA BASED POST-PROCESSING

In this chapter, independent component analysis (ICA) in a subband domain has been extended into a regularized feed-forward network with instantaneous direction of arrival (IDOA) based post-processing. The feed-forward network maximizes mutual independence of separated current frames using information from both the current and previous multi-channel frames of speech signals captured by a microphone array. Feed-forward realization in the subband domain is shown to be equivalent to the time domain feed-forward realization. To ensure a proper separation preventing permutation and arbitrary scaling, we not only rely on the steered response for the first tap of the demixing filter but also penalize on the direction, thus drastically increasing the mean squared error with the spatial filtered output in the first stage. After convergence, by applying IDOA based post-processing, we can additionally suppress the leakage of the interference as well as the remaining reverberated target signal. The signal-to-interference ratio (SIR) is improved more than 20 dB in C-weighting (dBC) for distances up to 2.7 m and angle differences down to 26° . We can achieve up to 29 dBC SIR improvement and 0.6 MOS score increase and more than 11.3 dBC SIR and 0.13 MOS in the most adverse condition.

3.1 Introduction

Speech separation has been an active research topic for various interesting applications. One of the compelling applications is simultaneous voice control of multimedia equipment supplemented by robust speech recognition. In this scenario, we should be able to successfully handle various realistic acoustic environments (background noise and reverberation conditions) to achieve a robust separation of the simultaneous spoken speech streams.

The speech separation problem has traditionally been tackled from two distinct disciplines: beamforming and blind source separation such as independent component analysis (ICA). The optimal beamforming approach has been used with the robust estimation of direction of arrival (DOA) for a target source source to suppress noise or interference speech [31]. The independent component analysis (ICA) approach has also been widely attempted with the assumptions that the original speech sources are independent and we approximately know the probability distribution function (PDF) of them [2].

Recently, combinations of beamforming and independent component analysis have been proposed [4, 76]. Considering the fact that these two schemes are based on different optimality criteria (minimizing power for non-look direction signal [31] versus maximizing non-Gaussianity or mutual independence [2]), we might be able to expect that combining two heterogeneous technologies would provide better results than each individual approach alone. However, in most current approaches, ICA demixing filters converge near to a nullformer on interference sources [4, 77, 78]. Although this observation is appealing because by nulling the interferences we can increase super-Gaussianity in such a way that we can suppress the unwanted speech signals, this also means that the ICA does not contribute much to the conventional beamforming combined with nullforming on the direction of interferences. In fact, the beamforming plus nullforming scheme seems to be the best we can achieve using only the information from the current frame, and this observation will be verified experimentally in the Section 3.4.3.

Frequency domain ICA was originally proposed to solve convolutive mixing with separated instantaneous demixing in each individual frequency bin [2]. Although the fact that convolution in the time domain can be represented as multiplication in the frequency domain has been a reasonable justification for the benefits of the frequency domain approach, this is true only when the frame length is long enough. In a typical frame length (10 to 30 ms), the reverberated target and interference cannot be compensated properly, because reverberation time (typically 100 to 300 ms) far exceeds the frame length [77]. Low et al. experimentally showed the effectiveness of using temporal processing in the context of the adaptive noise cancellation (ANC) [79] with the conventional ICA as a pre-processor, where they assume that they can obtain a target-signal-free reference channel based on the Kurtosis measure after the conventional ICA for ANC. However, with the aforementioned

limitation in the conventional ICA, the leakage of speech is unavoidable in a reverberant environment, and it may cause well-known problem of the target signal cancellation. Note that if we replace the conventional ICA part with a fixed-beamformer in a generalized sidelobe canceller (GSC) framework, then Low et. al.'s approach [79] becomes equivalent to the subband domain GSC [80].

In this chapter, we overcome the fundamental limitation of the subband domain ICA in a reverberant acoustic environment by taking previous multi-channel frames into consideration as well as the current frame in the context of ICA. In other words, we introduce a feed-forward network in the subband domain, rather than just using current frames for instantaneous demixing. A feed-forward demixing filter structure with several taps in the subband domain is accommodated with natural gradient update rules [1] to speed up convergence. A feed-forward network followed by a proper nonlinear filter is known to be a general way of representing a neural network, which has the so-called universal approximation property according to Kolmogorov's theorem [81]. To prevent permutation and arbitrary scaling and to guide the separated speech sources into the designated channel outputs, we not only use the estimated spatial information on the target and interference, but also add a regularization term on the update equation, thus minimizing the mean squared error between separated output signals and the outputs of spatial filters in the first stage of the proposed method. The regularization can also be interpreted as a priori information in a Bayesian sense as in [82]. After convergence of the regularized feed-forward demixing filter, we observe better separation of the speech signals, with a little late reverberation remaining for both desired and interference speech signals. These reverberation tails can be further suppressed by using spatial filtering based on instantaneous direction of arrival (IDOA), which gives us the probability for each frequency bin to be in the original source direction [34]. This post-processing also suppresses the remaining leakage of the interference speech coming from non-look directions.

The proposed method is evaluated using two criteria. Physical separation is measured by the signal-to-interference ratio (SIR) in dB with C-weighting (dBC), and separated speech quality is measured by the perceptual evaluation of speech quality (PESQ) algorithm [83]. Experiments are performed in a relatively adverse acoustic environment (T60 of 375 ms, SNR of 15 dBC, where dBC stands for dB in C-weighting [34]), with a distance from speak-

ers to microphone array of 1.5 to 4.3 m, and with an angle between the two speakers ranging from 6° to 70° . The proposed algorithm achieves an improvement of 29 dBC in SIR and 0.6 in mean opinion score (MOS) by PESQ for the best case. These improvements remain above 10 dBC and 0.084 PESQ points in most adverse conditions. For distances up to 2.7 m and separation angles down to 26° , the SIR stays above 20 dBC, above which the interference speech was barely audible.

The remaining parts of the chapter are presented as follows. In Section 3.2, the speech separation problem is formulated in the subband domain, and background for beamforming and conventional ICA are briefly introduced with clear connection with the proposed method. In Section 3.3, we introduce the proposed regularized feed-forward ICA (RFFICA) with IDOA based post-processing. Firstly, to validate the appropriateness of the feed-forward network in a subband domain, we explain the equivalence between the subband domain feed-forward network implementation and the time domain feed-forward network. Secondly, a detailed implementation of the proposed method is introduced in a step-by-step manner. In Section 3.4, experimental results are presented and discussed from the several different perspectives. In Section 3.5, we conclude our chapter and mention some directions for future work.

3.2 Problem formulation and background

Figure 3.1 shows a block diagram for separation of two independent speeches in a subband domain. Time domain signals captured using multiple microphones are converted to a subband domain using a modulated complex lapped transform (MCLT) that can produce better separation between frequency bands in an efficient manner [84]. The source separation can be performed using a demixing filter in each individual frequency bin $k = 1, 2, \dots, K$ where K is the number of the frequency bins. Then the resulting signal can be converted back to the time domain using inverse MCLT. Conventionally, subband domain source separation per each bin can be formulated as the following:

$$\vec{S}_k = \mathbf{W}_k \vec{Y}_k, \quad (3.1)$$

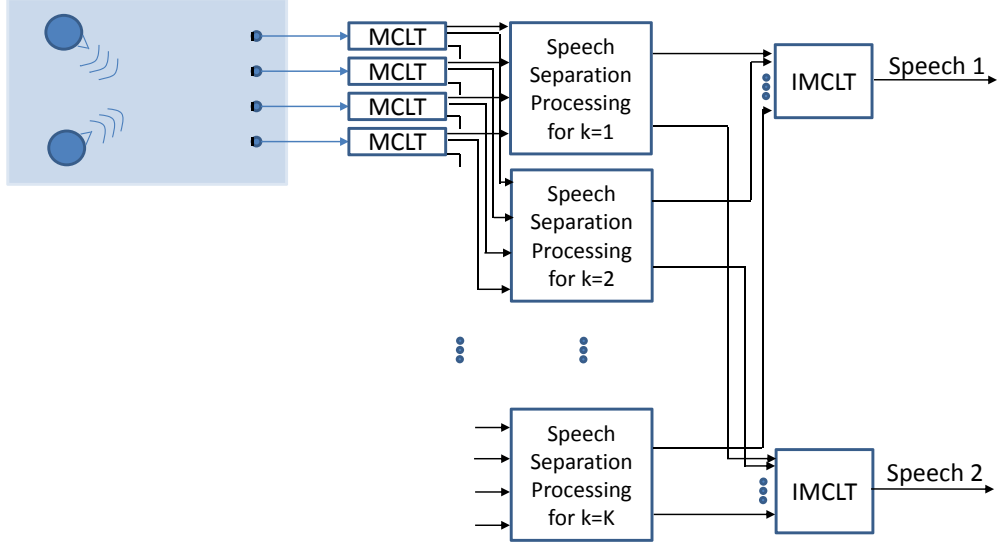


Figure 3.1: Block diagram of separation for two concurrent speech streams in the modulated complex lapped transform (MCLT) domain.

where $\vec{S}_k$ is the separated speech vector, $\mathbf{W}_k$ is the demixing matrix, and $\vec{Y}_k$ is the measured speech vector in a reverberant and noisy environment. We omit bin indexes k for clarity of presentation hereafter.

In this section, we explore two classical source separation methods given the conventional problem formulation in (3.1) and discuss how we can improve the performance by extending the problem formulation itself. Firstly, a linearly constrained minimum variance beamformer is reviewed with the spatial information on target and interference speech at hand. Secondly, ICA in the frequency domain is reviewed with the conventional assumption that the multiplication operation in the frequency domain is equivalent to convolution operation in the time domain. Finally, the combination of these two methods is briefly addressed and extended to the proposed method to overcome its limitations.

3.2.1 Beamforming with nullforming

Beamforming is spatial filtering to suppress the unwanted interference and noise given the information on DOA for the incoming signals. Usually the optimal beamforming is constructed on a narrow band signal as in the communication system, but we extend it into the wide band signal such as speech

by subband decomposition. One of the most commonly used beamformers is the minimum variance distortionless response (MVDR) beamformer, which in the frequency domain can be described as

$$\vec{W}^H = \frac{\vec{D}^H \mathbf{R}_n^{-1}}{\vec{D}^H \mathbf{R}_n^{-1} \vec{D}}. \quad (3.2)$$

Here $\vec{D}$ is a steering vector, $\mathbf{R}_n^{-1}$ is a noise covariance matrix, and $\vec{W}^H$ is a beamforming weight vector for multi-microphone input vector $\vec{Y}$. Often the noise-only covariance $\mathbf{R}_n$ is replaced by $\mathbf{R}$, which is the covariance matrix of the input vector (signal plus noise). The beamformer with $\mathbf{R}$ is known as minimum power distortionless response (MPDR). In fact, $\mathbf{R}$ should also be estimated and the sample covariance matrix is normally used [31] to estimate $\mathbf{R}$. To prevent instability due to the direction of arrival mismatch, a regularization term can be added to the sample covariance matrix, which is called “diagonal loading” [31]. In our case, we also add an additional null constraint in the direction of the interference, which makes sense because the ICA filter has been reported to converge toward the nullformer in the direction of the interference source [4, 78]. The beamformer with extra nullforming constraint can be formulated as

$$\vec{W}^H = [1 \ 0] \left([\vec{D}_t \ \vec{D}_i]^H [\mathbf{R} + \lambda \mathbf{I}]^{-1} [\vec{D}_t \ \vec{D}_i] \right)^{-1} [\vec{D}_t \ \vec{D}_i]^H [\mathbf{R} + \lambda \mathbf{I}]^{-1}, \quad (3.3)$$

which is obtained by solving the following constrained optimization problem:

$$\arg \min_{\vec{W}} E[|\vec{W}^H \vec{Y}|^2] \text{ subject to } \vec{W}^H [\vec{D}_t \ \vec{D}_i] = [1 \ 0], \quad (3.4)$$

where $\vec{D}_t$ and $\vec{D}_i$ are steering vectors toward the target and interference direction, respectively, and λ is the regularization term for diagonal loading.

The separation result will be reported in Section 3.4.3. Figures 3.2(a) and 3.2(b) show the directivity patterns of MPDR on DOA 19° and MVDR on DOA 19° + nullforming on DOA −9°, respectively. We can confirm the direction of target and null clearly. The solution of (3.4) can also be obtained in an adaptive manner. Figure 3.2(c) shows the directivity pattern of an adaptive MPDR after 1000 frames with DOA 19°. We can see the adaptively constructed null around the interference direction of −9°. Figure

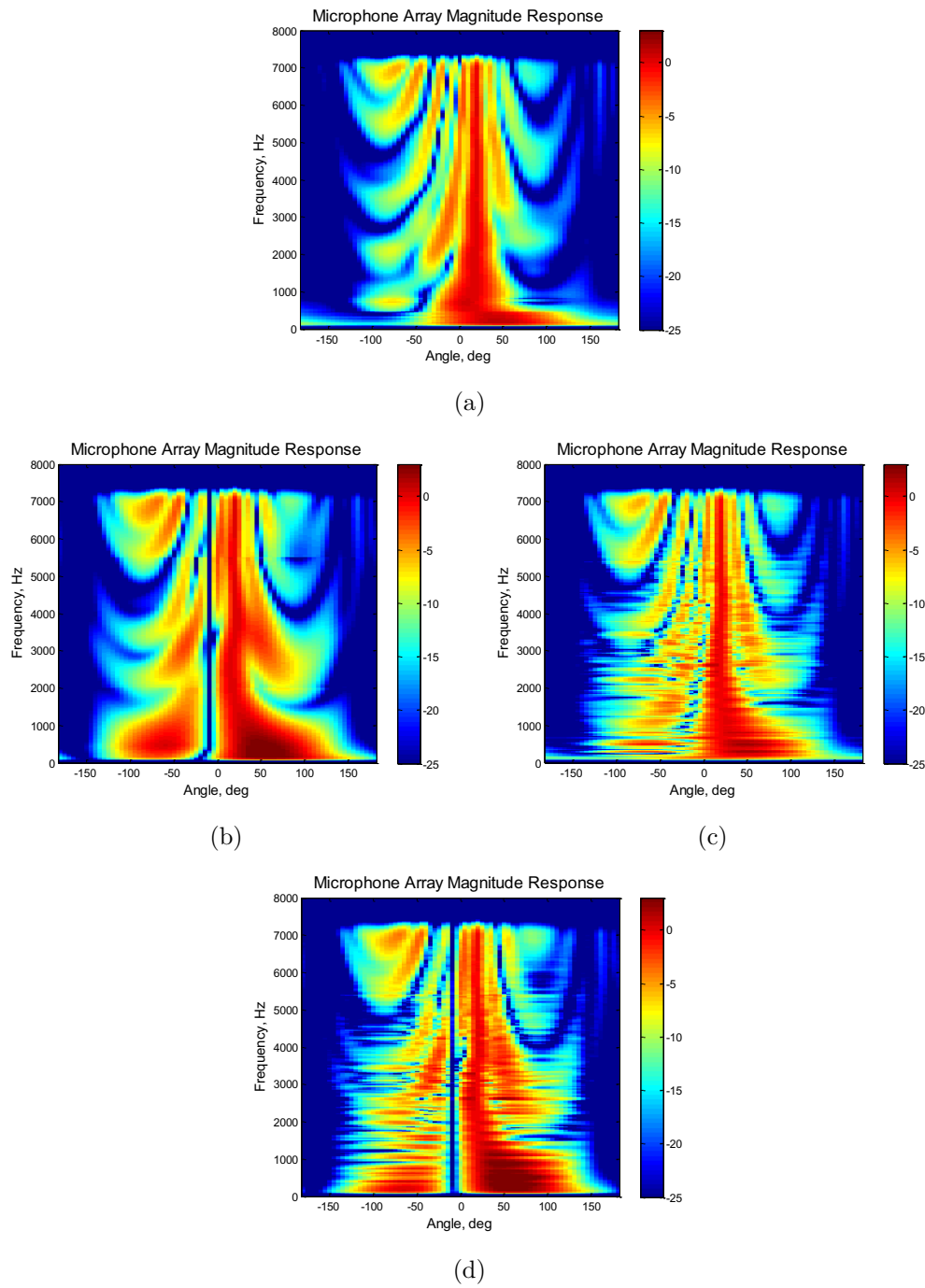


Figure 3.2: Microphone array magnitude response (dB) from -180° to 180° : (a) MVDR beamforming directivity pattern, (b) MVDR beamforming with nullforming directivity pattern, (c) adaptive MPDR beamforming without nullforming directivity pattern, and (d) adaptive MPDR beamforming with nullforming directivity pattern.

3.2(d) shows the directivity pattern of the adaptive MPDR with additional null constraint; the expected null is seen much more clearly. With this clear beam and null on the target and interference directions, respectively, we can initialize the first tap of the feed-forward ICA filter for the appropriate channel assignment. MATLAB implementation of the beamforming was modified from the scripts provided in [34].

3.2.2 Combination of beamforming and ICA

The observation that the ICA filter converges toward the nullformer in the direction of the interference source has been used in the ICA filter update [4] to prevent misleading the convergence and to speed up the convergence itself. Also, the steered response of the converged ICA filter per each frequency bin has been utilized as a cue for solving the permutation and scaling problem. We report the separation results on this approach [4] as another baseline result.

As will be noticed in the results section, the performance of this combined method does not produce impressive results, compared with the individual conventional ICA and beamforming plus nullforming. From beamforming, we rely on the spatial information on sources such that we can reliably solve the permutation and arbitrary scaling problem. Furthermore, we can assign the output channel for some designated target sources by initializing the separation filter with the spatial information and regularizing the permutation using a first-stage separation result. From ICA, we do have identifiability of each source; therefore, with the correct identification and alignment, we should be able to have better results than with ICA alone if the spatial information gives us extra resources. However, not much improvement can be expected using the conventional method.

The missing part is that both the conventional beamforming and ICA in the subband domain do not have the identifiability for reverberated sources exceeding the current frame time length. For alignment purposes, the current spatial information only on the direct paths from speakers to multi-microphones would be enough; therefore, in the next section we propose a regularized feed-forward ICA (RFFICA) in the subband domain with a proper alignment procedure. Figure 3.3 shows the conceptual diagram.

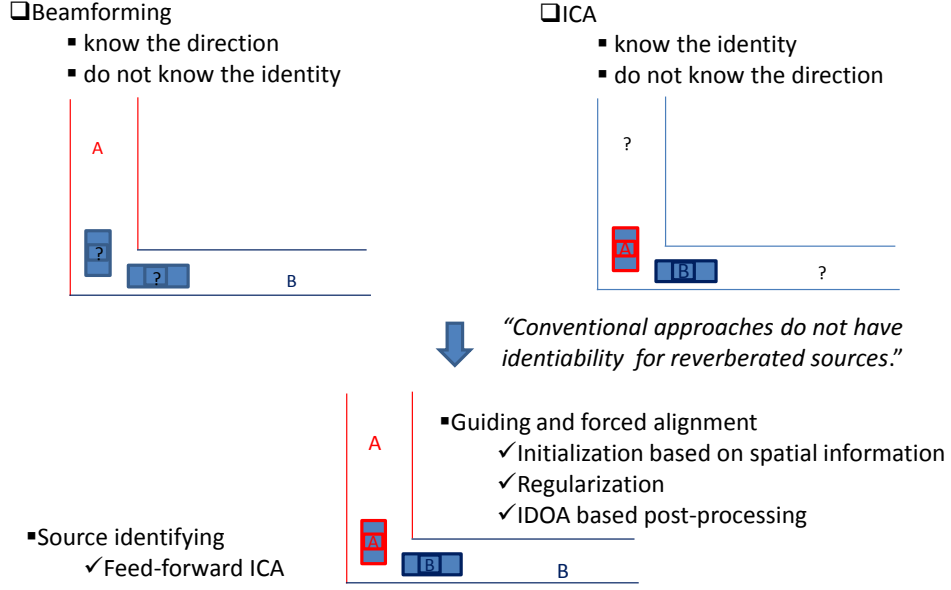


Figure 3.3: Conceptual diagram for the proposed method compared with individual beamforming and ICA.

3.3 Regularized feed-forward ICA (RFFICA) with IDOA based post-filtering

Figure 3.4 shows a block diagram of the proposed two-step method for a specific subband. The first step is the beamforming plus nullforming, followed by an IDOA based spatial filter which produces additional suppression of the interference signals. The second step is the regularized feed-forward ICA with post-processing, which uses the output of the first step for regularization. Here we maximize the mutual independence of the separated speech streams by using both current and previous multi-channel frames. A secondary spatial filter is applied at the end of the second step.

3.3.1 Realization of sufficient statistic with unknown RIR

Our goal in this chapter is to identify $\vec{S}^{-(N_H-1)}$ with the time series of a multichannel measurement vector $\vec{Y}, \dots, \vec{Y}^{-(N_H-1)}$ in (1.9). For this purpose, as we discussed in Section 1.2.2, we need to have sufficient statistics of $T_{\vec{H}_1^0}^{-(N_H-1)}(S_1), \dots, T_{\vec{H}_{N_S}^{N_H-1}}(S_j^{-(N_H-1)})$. Although we can assume to know or reasonably estimate $\mathbf{R}_n^{-1}$ with the stationarity assumption on the background

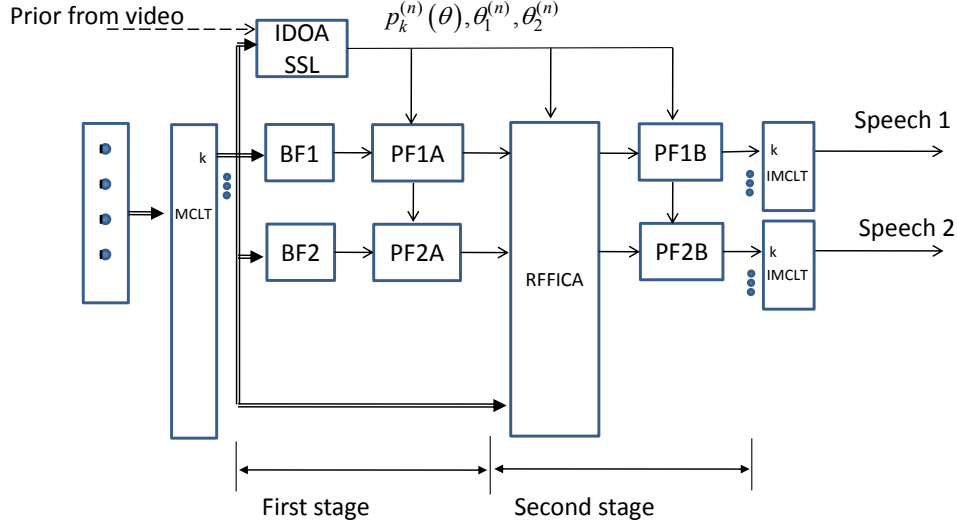


Figure 3.4: Proposed regularized feed-forward ICA (RFICA) with IDOA based post-processing: doubling represents and single line represent vector and scalar, respectively.

noise or update about the noise statistic, we still need to blindly estimate $\vec{H}_j^l$, where $j = 1, \dots, N_S$, and $l = 0, \dots, N_H - 1$.

In this chapter, we turn the RIR identification problem into a direct deconvolution and separation problem as in (3.9) and try to blindly estimate the deconvolution and separation filter network based on RFFICA. Since we do not have the source speech signals before the reverberant mixing, the deconvolution should be implemented in a blind manner. Therefore, the proposed method can be interpreted as an efficient adaptive way of estimating all the necessary sufficient statistics.

3.3.2 Subband domain separation

When we perform a subband domain decomposition, often a separation filter length, should be longer than a frame length as in the case of the current reverberant speech separation, since normally a room impulse response (RIR) is much longer than a frame length. A common mistake is to begin by ignoring this fact and just modeling the separated speech in the subband domain as instantaneous multiplication of the measured speech and the separation filter at each frequency bin and frame as in (3.1). In this section, we first show that a subband domain deconvolution is the correct representation of

the time domain deconvolution in the context of overlap addition; then we extend it into the separation scenario. Therefore, a feed-forward network in a subband domain is shown to be a correct implementation of ICA based reverberant speech separation. Proofs are given in Appendix C using separation filter W instead of convolutional filter H .

Proposition 1. *Deconvolution in the time domain by a filter with longer length than a frame length used for subband decomposition is equivalent to a subband domain convolution with delay of the frame shift.*

$$S(e^{j\omega_k}) = \sum_{l=0}^{N_W-1} W^l(e^{j\omega_k}) Y^{\cdot-l}(e^{j\omega_k}), \quad (3.5)$$

where ω_k is the angular frequency at frequency index k , superscript $\cdot-l$ stands for the frame l frames before the current frame, and N_W stands for the total number of frames to include the deconvolution filter length sufficiently and the l^{th} tap subband domain convolution filter

$$W^l(e^{j\omega_k}) = \sum_{t=-\infty}^{\infty} w^l[t] e^{-j\omega_k t}, \quad (3.6)$$

where

$$w^l[t] = w[t] \text{win}[lR - t], \quad -\infty < t < \infty, \quad (3.7)$$

where $\text{win}[t]$ is a proper window function and from the overlap-add context

$$w[t] = \sum_{l=0}^{N_W-1} w^l[t]. \quad (3.8)$$

Note that the deconvolution filter has been chopped into N_W frames with the same length as the frame shift. The equivalence between the subband domain deconvolution operation and the original time domain deconvolution can be demonstrated by taking the inverse Fourier transform on (3.5) and summing them over all possible frames, which is eventually shown to be equivalent to the time domain deconvolution when we simply implement the deconvolution in the time domain with the original incoming signal and the deconvolution filter before applying subband decomposition. Note that each chopped deconvolution filter and source signal do not need to have zero values outside the frame length such that frequency domain equivalence can

be satisfied without zero padding, because the deconvolution aims not to have a nonzero value outside the frame. This fact allows FFICA to be constructed in the subband domain more easily.

Proposition 2. *Multi-channel convolutive demixing in the time domain with room response longer than a frame length used for subband decomposition is equivalent to a subband domain convolutive demixing with delay of the frame shift.*

$$S_i(e^{j\omega_k}) = \sum_{j=1}^{N_S} \sum_{l=0}^{N_W-1} W_{i,j}^l(e^{j\omega_k}) Y_j^{\cdot-l}(e^{j\omega_k}), \quad (3.9)$$

where the subscript i means the i^{th} convolutive mixing output and j means the channel index. The number of the mixed sources N_S is not necessarily less than N_{mic} , but it is generally assumed so to perform source separation [2].

Note that the superiority of the subband domain approach over the time domain approach is obvious, because we can train each subband individually in a more stable manner with fewer filter coefficients.

3.3.3 First stage: beamforming plus nullforming followed by a spatial filter

To determine the direction of arrival (DOA) of the desired and interference speech signals, we use an IDOA based sound source localizer. Instantaneous direction of arrival (IDOA) space is $N_{mic} - 1$ dimensional with the axes being the phase differences between the non-repetitive pairs [34]. Here N_{mic} is the number of microphones. This space allows estimation of the probability density function $p_k(\theta)$ as a function of the direction θ for each subband. The results from all subbands are aggregated and clustered [34]. At this stage, additional cues (from a video camera, for example) can be applied to improve localization and tracking precision. The sound source localizer provides directions to target θ_1 and interference θ_2 signals. Given the proper estimation of the DOAs for the target and interference speech signals, we apply the constrained beamformer plus nullformer according to (3.3).

The consequent spatial filter is applying a time-varying real gain for each subband, acting as a spatio-temporal filter for suppressing the sounds coming

from non-look directions. The suppression gain is computed as

$$G_k(n) = \frac{\int_{\theta_1 - \Delta\theta}^{\theta_1 + \Delta\theta} p_k(\theta) d\theta}{\int_{-\pi}^{+\pi} p_k(\theta) d\theta}, \quad (3.10)$$

where $\Delta\theta$ is the range around the desired direction θ_1 from which we want to capture the sound.

3.3.4 Regularized feed-forward ICA followed by IDOA based post-processing

In this chapter, we utilize the virtue of the time domain source separation approach [1] in the subband domain case by allowing multiple taps in the demixing filter structure in each subband:

$$\vec{S} = \sum_{l=0}^{N_W-1} \mathbf{W}^l \vec{Y}^{\cdot -l}, \quad (3.11)$$

where $\vec{S} = [S_1 \ S_2 \ \dots \ S_{N_S}]^T$, $\vec{Y}^{\cdot -l} = [Y_1^{\cdot -l} \ Y_2^{\cdot -l} \ \dots \ Y_{N_{mic}}^{\cdot -l}]^T$ and

$$\mathbf{W}^l = \begin{bmatrix} W_{1,1}^l & W_{1,2}^l & \dots & W_{1,N_S}^l \\ W_{2,1}^l & W_{2,2}^l & \dots & W_{2,N_S}^l \\ \vdots & \vdots & \ddots & \vdots \\ W_{N_{mic},1}^l & W_{N_{mic},2}^l & \dots & W_{N_{mic},N_S}^l \end{bmatrix}.$$

Note that (3.11) is a vector representation of (3.9).

The proposed update rule for the regularized feed-forward ICA (RFFICA) is given below:

$$\mathbf{W}^l = \mathbf{W}^l + \mu \left((1 - \alpha) \cdot \Delta_{\text{ICA}}^l - \alpha \cdot \Delta_{\text{First stage}}^l \right), \quad (3.12)$$

where $l = 0, 1, \dots, N_W - 1$, and N_W is the number of taps. The terms Δ_{ICA}^l and $\Delta_{\text{First stage}}^l$ represent the portion of the ICA update and the regularized portion on the first stage output.

$$\Delta_{\text{ICA}}^l = \mathbf{W}^l - \left\langle g \left(\vec{S}^{\cdot -(N_W-1)} \right) \vec{Y}^{\cdot -lH} \right\rangle_t, \quad (3.13)$$

$$\vec{Y} = \sum_{l=0}^{N_W-1} \mathbf{W}^{N_W-1-lH} \vec{S}^{-l}, \quad (3.14)$$

$$\Delta_{\text{First stage}}^l = \left\langle \vec{Y}^{-l}|_{\text{Ref}} \left(\vec{S}|_{\text{Ref}} - \vec{S}_{\text{First stage}} \right)^H \right\rangle_t, \quad (3.15)$$

where $\langle \cdot \rangle_t$ represents time averaging, $(\cdot - l)$ represents l sample delay, $\vec{S}|_{\text{Ref}}$ is the first stage output vector for regularization, and $|_{\text{Ref}}$ represents the reference channels. The penalty term has been only applied to the channel where the references are assigned; the other entries for the mixing matrix are set to zero so that the penalty term vanishes on those channel updates. For initialization of the subsequent filters, we modeled the dereverberation process as exponential attenuation:

$$\mathbf{W}_{\text{ini}}^l = \exp(-\beta l) \cdot \mathbf{I}, \quad (3.16)$$

where $\mathbf{I}$ is an identity matrix, β is selected to model the average reverberation time, and l is the tap index. Note that we initialized the first tap of RFFICA for the reference channels as a pseudo-inversion of the steering vector stack for the current experiment so that we can assign 1 to the target direction and null to the interference direction:

$$\mathbf{W}_{\text{ini}}|_{\text{Ref}} = \left([\vec{D}_t \ \vec{D}_i]^H [\vec{D}_t \ \vec{D}_i] \right)^{-1} [\vec{D}_t \ \vec{D}_i]^H, \quad (3.17)$$

where $\vec{D}_t$ and $\vec{D}_i$ are steering vectors toward the target and interference direction, respectively. Because we update the initialized filter using ICA, a slight mismatch with actual DOA can be adjusted in the updating procedure. For the current experiment, we set α as 0.5 just to penalize the larger deviation from the first stage output. As a nonlinear function $g(\cdot)$, we used a polar-coordinate based tangent hyperbolic function, suitable to the super-Gaussian sources with a good convergence property [39]:

$$g(\vec{X}) = \tanh(|\vec{X}|) \exp(j\angle \vec{X}), \quad (3.18)$$

where $\angle \vec{X}$ represents the phase response of the complex value $\vec{X}$. To deal with the permutation and scaling, we also used the steered response of the

converged first tap demixing filter as following:

$$S^i = \frac{S_i}{F_i} \cdot \left(\frac{|F_i|}{\max(|\vec{F}|)} \right)^\gamma, \quad (3.19)$$

where $i = 1, 2, \dots, N_{mic}$ is the designated channel number, F_i is the steered response for the channel output to the target DOA [4], and $\vec{F}$ is the steered response vector to the candidate DOAs. To penalize the non-look direction in the scaling process we added the nonlinear attenuation with the normalization using the steered response. For our current experiment, we set γ as 1. The spatial filter also penalizes the non-look directional sources in each frequency bin.

3.4 Experimental results

In this section, we report the experimental results of the proposed method. Firstly, for the qualitative evaluation purpose, we generated a simulated corpus for the reverberated speech mixture captured by i^{th} microphone, where $i = 1, 2, \dots, N_{mic}$, by summing after convolving the clean speech streams $s_j(t)$ with the measured room impulse responses from the source positions $j = 1, 2, \dots, N_S$ to the designated i^{th} microphone:

$$y_i(t) = \sum_{j=1}^{N_S} h_{i,j}(t) * s_j(t) + n_i(t), \quad (3.20)$$

where $h_{i,j}(t)$ is a RIR from j^{th} source to i^{th} microphone and $n_i(t)$ is measured noise by the i^{th} microphone. Secondly, to validate the proposed method, we applied our algorithm to the spontaneous real speech scenario, where two speakers have spoken for about 30 seconds in the same space where we measured the RIRs and background noise.

3.4.1 Experimental setup for corpus generation

For room impulse response measuring, we used a four-element microphone array with unidirectional microphones of a length 225 mm. In a large office room (T_{60} of 375 ms), we measured the room impulse responses (RIRs) be-

tween each point of the rectangular grid and each of the microphones. This was done by using a mouth simulator B&K Type 4227 playing a wideband chirp signal from the loudspeaker located at each point of the grid. Positions of the mouth simulator represent the possible positions of the simulated speaker, and the diagram for the positions is given in Figure 3.5. Using the same microphone array, we recorded ambient noise in the same space (air conditioning plus five computers).

Using a clean speech corpus convolved with the corresponding RIRs and adding the natural noise, we generated 18 different evaluation cases with two speech sources. The TIMIT corpus has been used as the clean speech source. We used concatenated 10 utterances; interestingly, all elongated sentences started with the same two utterances but the remaining 8 utterances are different. We assigned speech streams randomly and added randomly segmented measured noise with the same duration as the speech mixture among the original 30 minute recordings of the noise. The distance from the microphone array to the speech source was varied from 1.22 to 4.23 m.

The distance between the two speech sources varied from 0.6 to 1.8 m, resulting in a angle distance between 6° and 70° from the microphone array point of view. Comprehensive detail on the corpus generation is given in the Appendix E.

3.4.2 Measures to evaluate performance

Due to the fact that we have the clean speech signals for evaluation in our experimental setup, we used two different measures, signal-to-interference ratio (SIR) and PESQ mean opinion score (MOS) [83]. We define SIR as

$$\text{SIR} = 10 \log_{10} \frac{\text{Target portion energy}}{\text{Interference portion energy}} \text{ dBC}, \quad (3.21)$$

where target and interference portions can be estimated precisely using the clean speech signals. Note that we calculated the decibel with C-weighting so that we cover all the frequency range almost flat at a 16 kHz sampling rate. After subtracting each portion in the original mixture, we simply get the noise plus a negligible level of the leaking speech streams. The accuracy of this measure has been tested by manually changing the SIR of the reverberated target and interference with additive measured real noise, and we can obtain

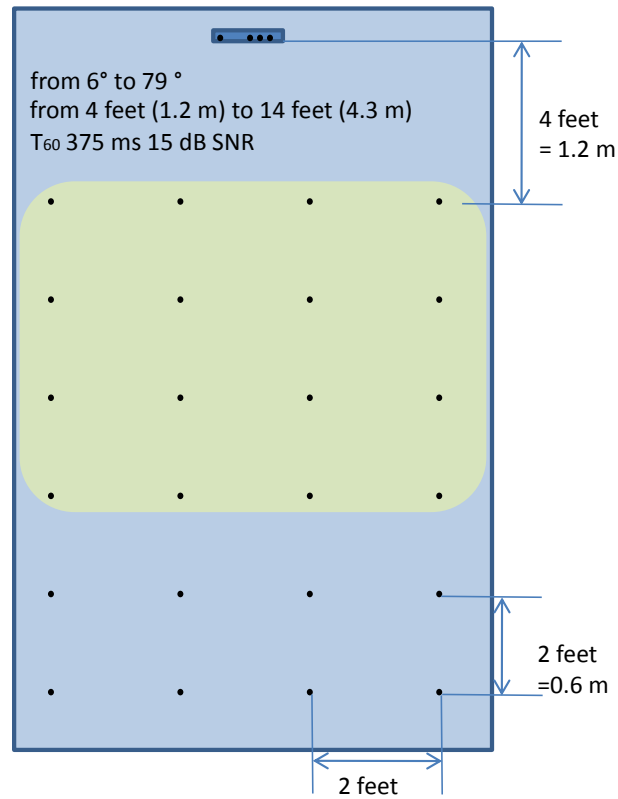


Figure 3.5: Multi-microphone room impulse response measurement setup: total number of measured source positions is 24.

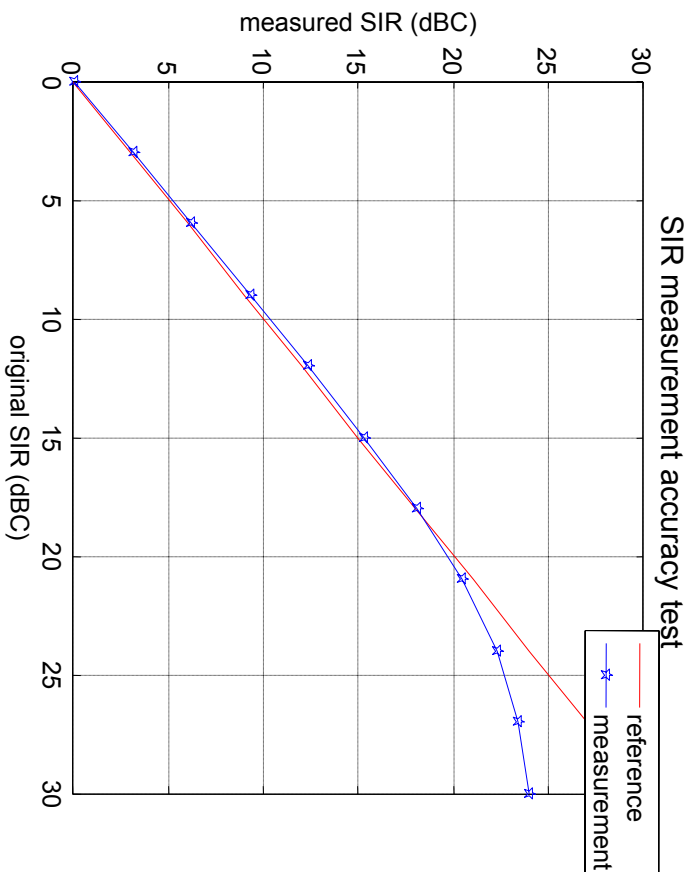


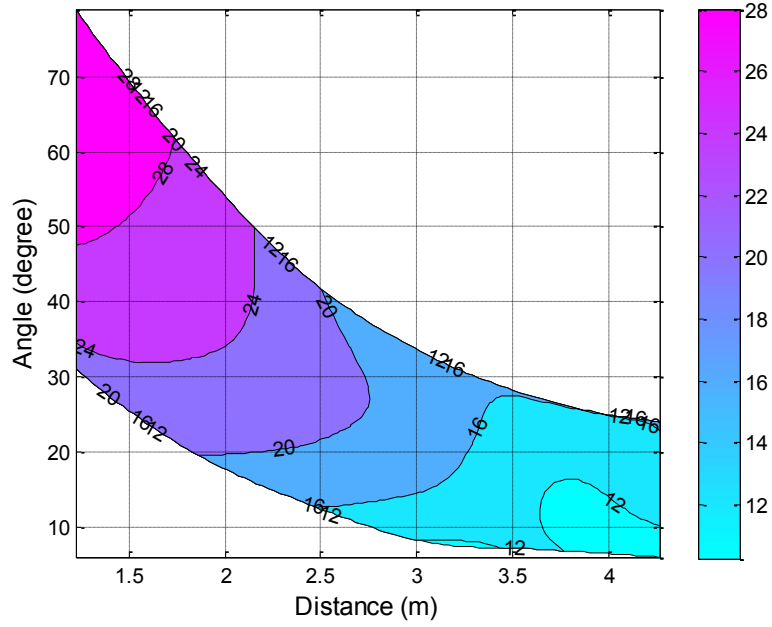
Figure 3.6: SIR measurement accuracy test.

an accurate measurement up to 20 dBC as in Figure 3.6. If we exceed 20 dBC, it underestimates the actual SIR. This is a measure of the physical improvement of the separation quality.

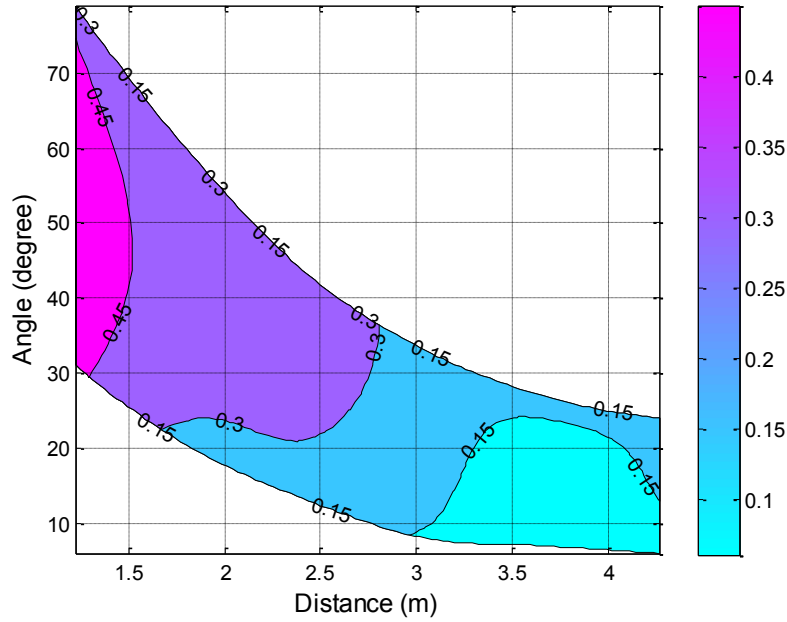
For the perceptual measure, PESQ MOS has been used. The PESQ score is rated from -0.5 to 4.5, and a higher score represents better perceived quality of speech [83]. Although PESQ is not a typical measure for source separation purposes, it allows us to track the distortions in the desired speech signal. We actually saw the general trend that a better SIR produces a better MOS score.

3.4.3 Results and discussion

We have evaluated all 18 two-speech reverberant mixing cases. With a conservative setting of 1000 iterations and 20-tap filters for each subband, the proposed approach improves SIR in the range of 10 to 29 dBC and the PESQ MOS score in the range of 0.1 to 0.6 points. Figures 3.7(a) and 3.7(b) represent contour plots of the improvement in SIR and PESQ, respectively. Note that the contour plots have been generated by interpolating the results of



(a)



(b)

Figure 3.7: (a) Contour plot of the improvement in SIR (dBC) as a function of the distance and separation angle, and (b) PESQ MOS improvement 2-D contour plot.

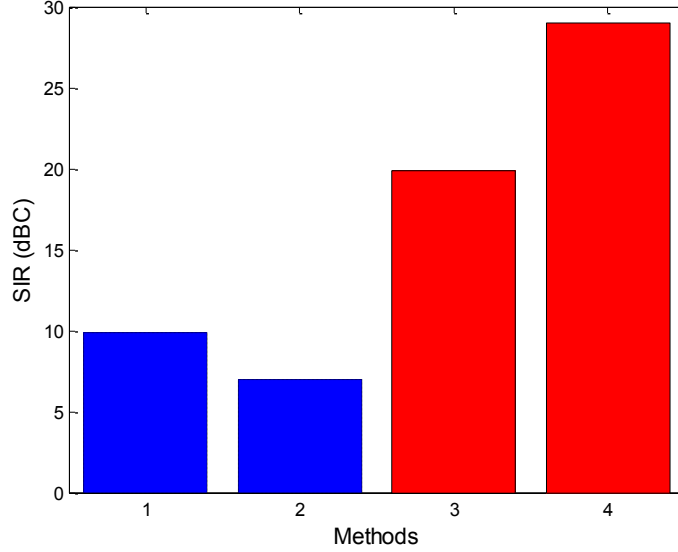


Figure 3.8: SIR improvement comparison for the setup of 1.22 m distance and 55° angle: (1) beamforming plus nullforming, (2) Saruwatari et. al., (3) beamforming plus nullforming plus IDOA based post-processing, and (4) RFFICA plus IDOA based post-processing.

18 actual measurements. Assuming 20 dBC separation as good enough for practical purposes, we can say that the proposed algorithm is good enough for distances up to 2.7 m for two speakers at 26° , i.e. standing shoulder to shoulder. In the most difficult condition of 6° between speakers at 4.23 m distance, we can still maintain a 10 dBC SIR and 0.1 improvement in PESQ points.

Figure 3.8 provides a comparison of the results for several methods for a distance of 1.22 m, 55° angle between speakers, and T60 of 375 ms. These conditions are close to those published in [4, 78], but are a more difficult configuration in general. Just the conventional beamformer plus nullformer provides nearly a 10 dBC improvement, which is close to the reported results in those papers. Adding the spatial filter (the first stage of the proposed algorithm) increases the suppression close to 20 dBC, which is practically the maximum a spatial separation can achieve. Adding the second stage increases the SIR to 29 dBC while keeping an acceptable quality of the separated speech signal. Note that the performance of the beamforming plus nullforming is already comparable with the result of [4], as we expected. Figures 3.9 and 3.10 show the time domain speech streams and corresponding spectrograms (1.22 m, 55°): (a) one of the mixed channels, (b) and (c)

original clean target speech and separated speech by the proposed method, (d) and (e) another original clean target speech and separated speech by the proposed method. We can clearly see the similarity between the original signals and the outputs. Note also that the interference dominant portion did not smear into the target separated speech, which is the reason why we cannot hear the interference speech at all.¹ In fact, in this case SIR was around 24 dBC. Figures 3.11(a) and 3.11(b) show the results of the dependency on iteration number and tap number on the middle configuration (2.44 m and 26°) among all 18 configurations. Actually, with 200 iterations and 5 taps, the SIR score can already reach the satisfactory level.

3.4.4 Results on a real scenario

We applied our algorithm to the spontaneous real speech scenario, where two speakers have spoken for about 30 seconds in the same space (1.22 m, 55°) where we measured the RIRs and background noise. They were asked not to move their positions while they were speaking, but natural head movement was allowed. Figures 3.12 and 3.13 show the time domain representations and spectrograms of the measured reverberant speech mixture together with the separated two-target speech streams, respectively. As in the simulated results, we could barely perceive the interference speech.¹

3.5 Conclusions and future work

In this chapter, we proposed a subband domain based, two-stage speech separation method that utilizes spatial filtering and regularized feed-forward ICA with multi-taps demixing filter. This approach produces substantial improvement in terms of SIR, while maintaining sound quality, measured with the PESQ algorithm. The proposed approach can be interpreted from both points of view independently. From the beamforming point of view, with proper prior knowledge of DOAs for the target and interferences, we can construct a subband domain filter structure augmented with higher-order independence maximization criterion for better suppression of the unwanted

¹The wave files of this figure are linked in the following web site: <http://www.isle.uiuc.edu/~lkim9/Demo>.

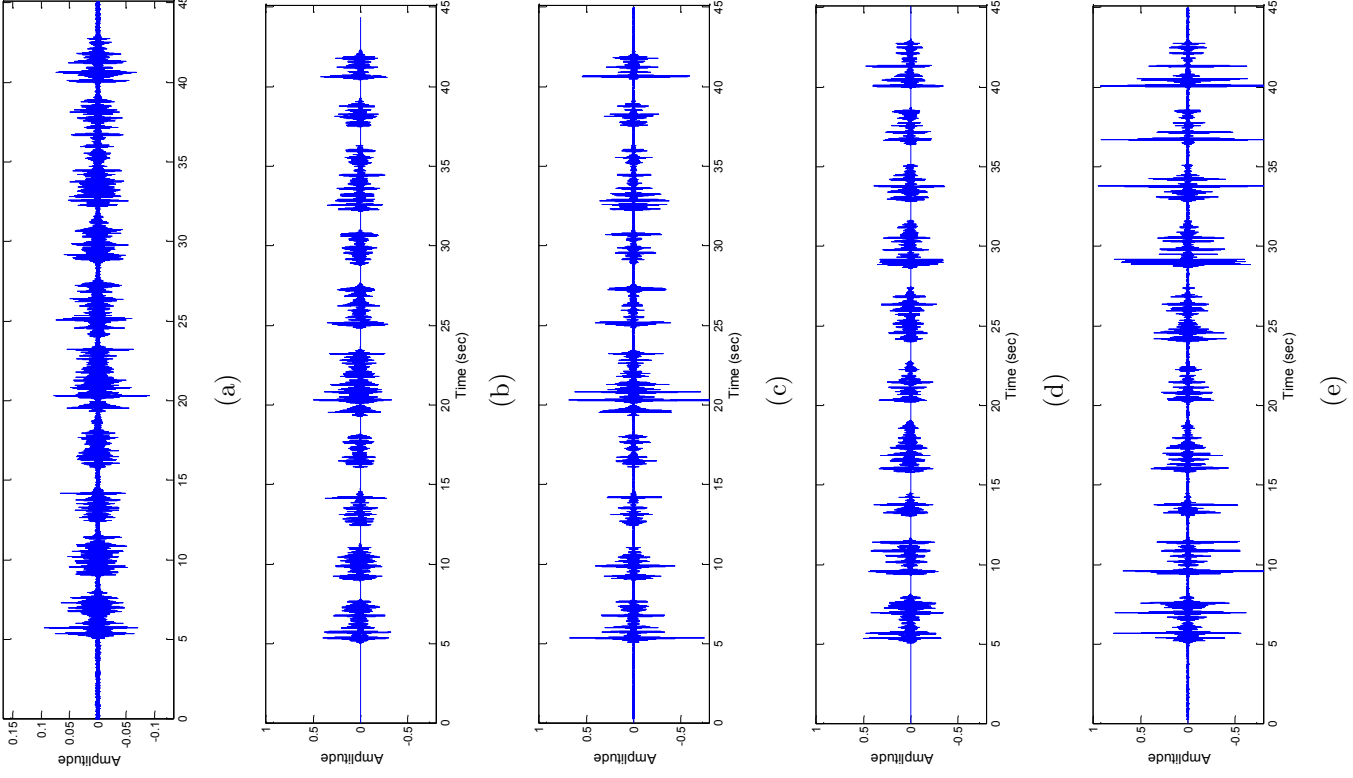


Figure 3.9: Time domain speech signals: (a) one of the mixed channels, (b) and (c) original clean target speech and separated speech by the proposed method, (d) and (e) another original clean target speech and separated speech by the proposed method.

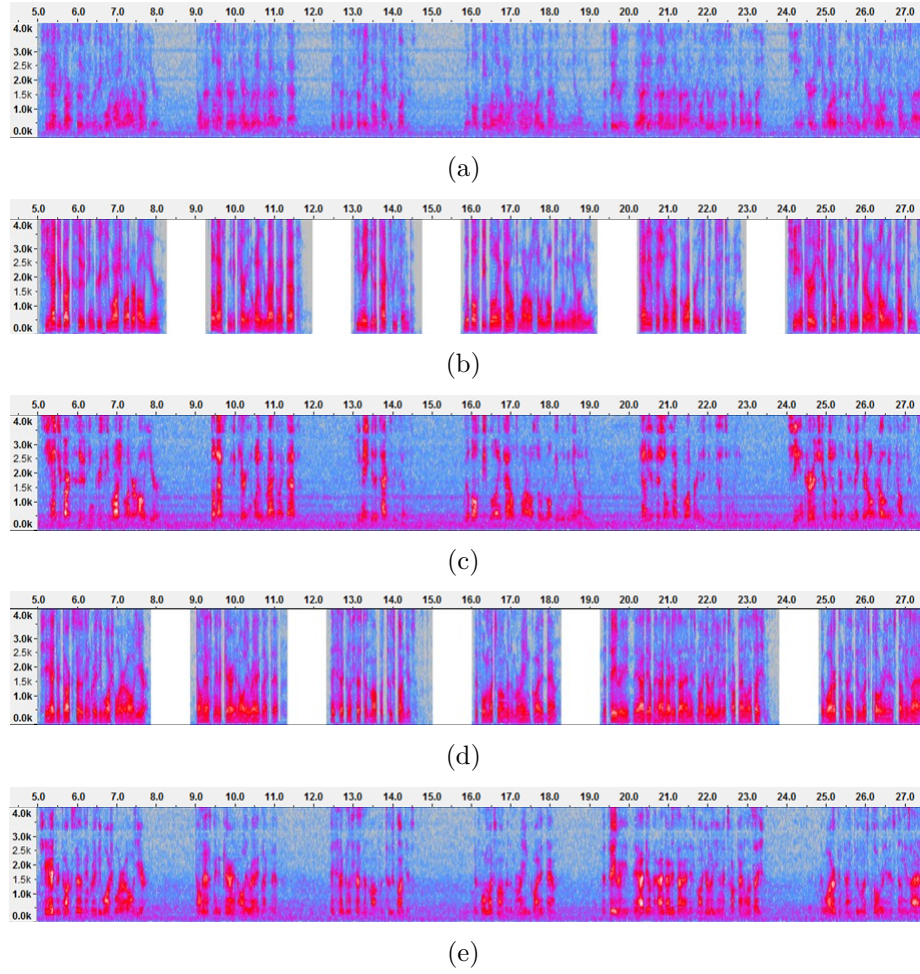
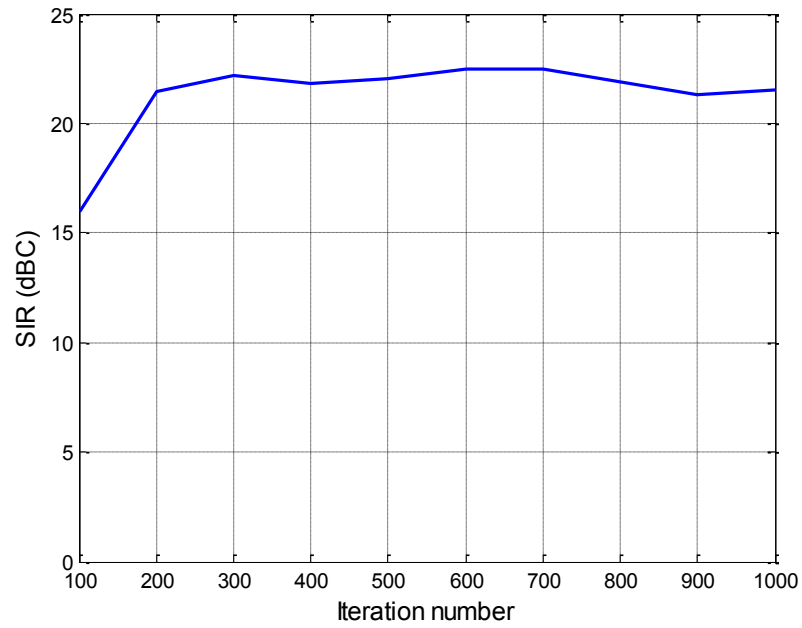
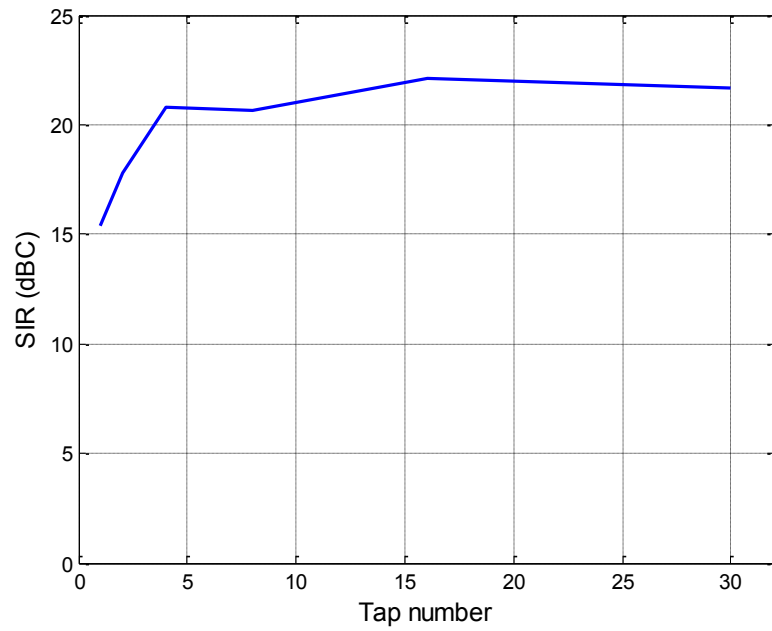


Figure 3.10: Spectrogram of speech signals (from 5 to 27.6 seconds and up to 4 kHz for purpose of clear presentation): (a) one of the mixed channels, (b) and (c) original clean target speech and separated speech by the proposed method, (d) and (e) another original clean target speech and separated speech by the proposed method.



(a)



(b)

Figure 3.11: (a) SIR versus iteration numbers, (b) SIR versus tap numbers.

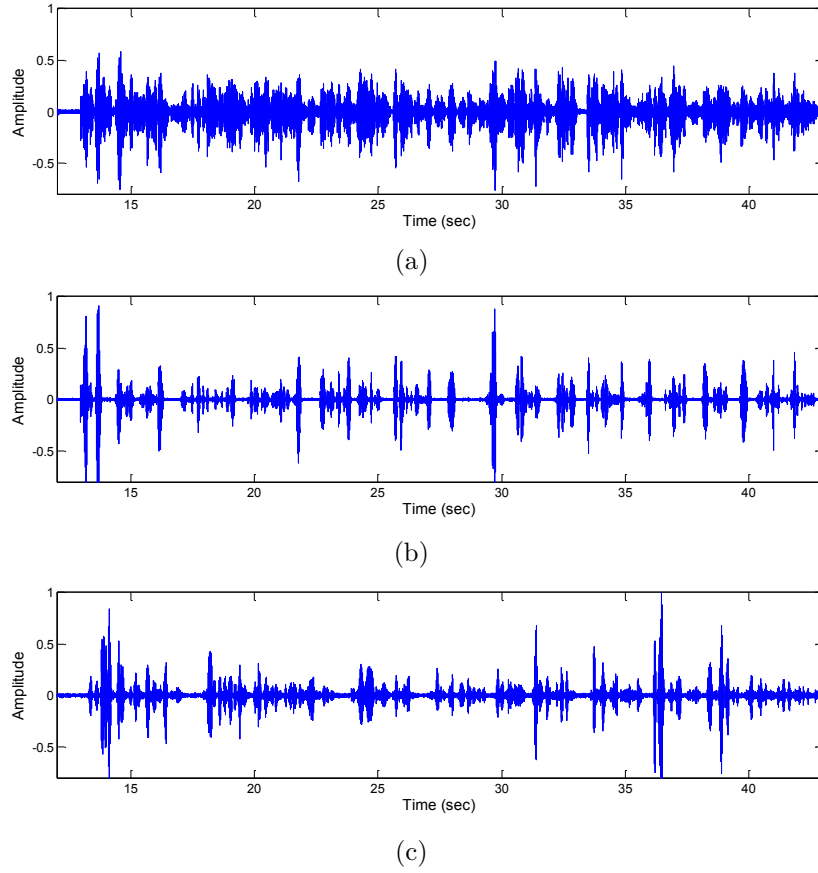


Figure 3.12: Time domain speech signals (spontaneous real speech scenario, 1.22 m, 55°): (a) one of the mixed channels, (b) and (c) separated speech streams by the proposed method.

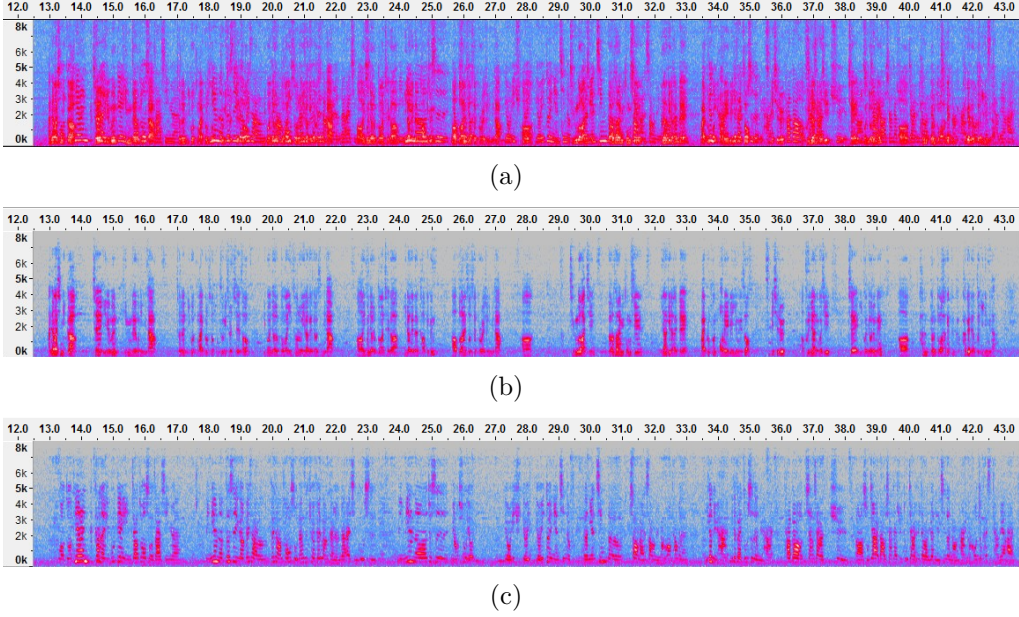


Figure 3.13: Spectrogram of speech signals (spontaneous real speech scenario, 1.22 m, 55°): (a) one of the mixed channel, (b) and (c) separated speech streams by the proposed method.

interference and noise. Note that we are relying on the same constraint of “unity response with zero phase on look direction” but with a different cost function of maximizing independence of separated or beamformed output signals. With the power minimization criterion, the source signal will be canceled out unless we have perfect knowledge of room impulse responses to the microphones, which is not practically possible, or we confine the beamforming to be a scalar multiplication process as in the conventional methods even in a subband domain. Therefore, the proposed method can be interpreted as a blind beamforming method overcoming the source cancellation problem with a simple unity response constraint on look direction.

From the ICA point of view, with a prior knowledge on the DOAs for the target and interference and a regularization term preventing large deviation from the beamforming results in the first stage, we can expand the conventional instantaneous demixing in the subband domain into the feed-forward network, which turned out to increase the mutual independence without additional processing to solve the permutation problems. The initialization and regularization based on the beamforming method is useful in solving the permutation problem from the beginning of the iteration. The unity response

constraint of beamforming is useful in solving the scaling problem. Many different ways of solving the permutation problem can be combined without hurting current schemes. In our case, we only used the IDOA based sound source localization information.

There have been other approaches to accomplish speech separation based on computational auditory scene analysis. Speech model based approaches have not been combined with ICA yet in this chapter, but combination may assist current schemes [56]. In this chapter, we have exploited the statistical aspect of speech with just a simple assumption of super-Gaussianity. A more complicated dynamic speech model is widely accommodated in the domain of automatic speech recognition, and this can also be explored to enhance the separated speech quality.

CHAPTER 4

JOINT ESTIMATION OF DOA AND SPEECH BASED ON EM BEAMFORMING

In this chapter, we propose a multi-microphone joint optimal estimation of the direction of arrival (DOA) and the source speech signal through newly introduced EM beamforming. This produces a posterior PDF for the DOA, based only on the reliably estimated speech spectrum. By maximizing over the posterior PDF of the DOA, we achieve maximum *a posteriori* DOA estimation. After convergence, the estimated source spectrum through a weighted sum in the Bayesian sense is a maximum likelihood estimate (MLE). This is a sufficient statistic for minimum mean squared error (MMSE) optimal estimation using a subsequent single channel MMSE filter.

4.1 Introduction

Beamforming has been a versatile technology in various fields of application including radar, sonar, and speech enhancement. However, due to the fact that everyday speech is a non-stationary broad-band signal contaminated with reverberation and noise, the application of beamforming for speech enhancement has been a unique issue. Recently, minimum mean squared error (MMSE) multi-microphone optimal speech enhancement has been theoretically shown to be achieved using minimum variance distortionless response (MVDR) beamforming, which is a maximum likelihood estimation (MLE) of the source speech [85], followed by an MMSE post-filter, under the assumption of perfect knowledge of channel responses from source position to the positions of multiple microphones [25, 85]. However, the underlying assumption is not realistic, because perfect blind channel response identification is infeasible.

Standard beamforming follows a two-step procedure: (i) direction of arrival (DOA) estimation; and (ii) beamforming based on the best estimation

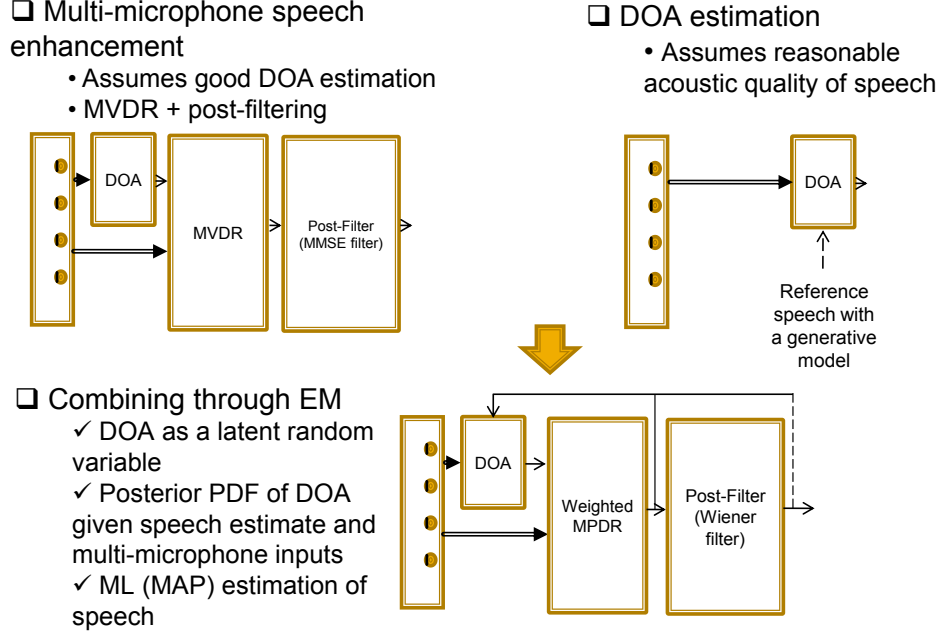


Figure 4.1: Conceptual diagram of the proposed joint optimization of DOA and speech based on EM beamforming.

of DOA [41]. However, mismatch between the “real DOA” and “estimated DOA” causes challenging robustness problems, which results in poor enhancement of the target speech. On the other hand, to estimate DOA accurately, we need to have better-enhanced speech. In the extreme case, if we know the original speech before convolving with room impulse response from the speaker to multi-microphones and before being deteriorated by the environmental background noise, we can much more reliably estimate the DOA than when we do not have it, because we can adaptively estimate the RIR including DOA information much more accurately using the source signal [27]. Note that those two different but closely connected problems, speech enhancement via beamforming and the estimation of DOA, have been solved independently based on the assumption of the good quality of the preceding part as illustrated in Figure 4.1. In this chapter, we propose to solve those two problems in a joint optimal manner by a newly introduced expectation-maximization (EM) algorithm, where we consider the DOA as a latent random variable with a *a posteriori* PDF given the multi-microphone observations and the maximum likelihood estimate of the target speech in the previous EM step. Figure 4.1 shows the conceptual diagram of the proposed EM beamforming.

Due to the fact that the speech signal is sparse in the short-time spectral domain, a large percentage of the frequency bins are corrupted by the coexisting noise and reverberation. In the process of EM beamforming, we only use the speech-dominant reliable frequency bins to update the DOA posterior. The reliability of each frequency bin is estimated based on the generative likelihood model, which will be discussed in detail. Because we avoid using the contaminated bins to estimate the DOA posterior PDF, the proposed approach provides more robust DOA estimation than the conventional methods based on the whole band, and it results in a better optimal speech estimation as well. In fact, after convergence, the estimated source spectrum is the maximum likelihood estimate (MLE) if we follow the solid loop in Figure 4.1. This is a sufficient statistic for an MMSE estimation using the subsequent single-channel MMSE filter such as the Wiener filter. This separate process, ML estimation of the sufficient statistic and MMSE post-filtering, can be done in the iteration as well if we follow the dotted loop. Then, after convergence, the estimated source spectrum is now the maximum *a posteriori* (MAP) estimate.

The ML estimate of the target speech in the proposed EM beamforming turns out to have a form similar to the Bayesian beamformer [86], but takes only the reliably estimated frequency bins into account to update the posterior PDF. Note that through iterative updates of both estimates of source DOA and signal, we introduce an explicit link between the source signal estimate and the DOA estimate.

To validate the proposed method, we performed a simple experiment using real recordings inside the IBM smart room from four microphones arranged into a T-shape array configuration. In this setting, the distance between the microphones is 0.26 m, which means that spatial aliasing is unavoidable above 654 Hz. However, unlike a traditional application such as radar, for speech inside smart rooms, we often have video as a supplemental modality that can be used for DOA estimation with limited time and spatial resolution. For example, [87] reports 3D tracking of the speaker’s head position every second. Therefore, we may have an acoustic viewfinder with narrow beam angles as candidate DOAs, and need only refine them to accommodate higher temporal resolution in the audio side (e.g. 16 ms frame rate). We obtained maximum *a posteriori* estimation of DOA and compared it with two baselines (GCC-PHAT and MUSIC based) by plotting DOA estimation per frame with

the ground truth at every second. The proposed method produces smoothly changing maximum *a posteriori* estimation of the DOA around the ground truth, which yields better spectral enhancement at the high frequency range.

4.2 Multi-microphone speech enhancement

4.2.1 Problem formulation

Multi-microphone speech measurements in a reverberant acoustic space with background noise can be modeled in the time domain by (1.6). In this chapter, we interpret this problem more from the perspective of a single target speech estimation. Note that we may do the original multiple target estimation by performing a sequence of target source estimations. We also further consider tracking of the speaker movement, which means we need to quickly update the estimation process with small amounts of data to estimate the source speech. To satisfy this criterion, we reformulate (1.6) as follows:

$$y_i(t) = \sum_{j=1}^{N_S} \left(h_{i,j}^{d,\theta}(t) * s_j(t) + h_{i,j}^r(t) * s_j(t) \right) + n_i(t), \quad (4.1)$$

where $y_i(t)$ is the measured signal, $h_{i,j}^{d,\theta}(t)$ and $h_{i,j}^r(t)$ denote the early room impulse response (RIR) and late RIR from the j^{th} speech source to the i^{th} microphone, $s_j(t)$ is the j^{th} source signal, $n_i(t)$ is the noise, and i varies from 1 to N_{mic} , where N_{mic} represents the number of microphones. Note that only $h_{i,j}^{d,\theta}(t)$ is assumed to be dependent on the DOA parameter θ .

In this chapter, we use only the direct path response as the *early* RIR, as in previous research where the statistical reverberation model has been successfully employed [88, 89]. The direct path can be modeled as

$$h_{i,j}^{d,\theta}(t) = \frac{1}{d_{i,j}^\theta} \delta(t - t_{i,j}^\theta), \quad (4.2)$$

where $d_{i,j}^\theta$ and $t_{i,j}^\theta$ denote the distance and time delay between the j^{th} source and the i^{th} microphone, respectively. Note that if the distance among the array microphones is relatively much smaller than the distance to the target speaker, $d_{i,j}^\theta$ can be assumed to be the same without loss of generality.

The time domain representation of (4.1) yields a subband domain representation as

$$\begin{aligned}\vec{Y} &= \mathbf{D}\vec{S} + \sum_{l=1}^{N_H-1} \left(\mathbf{R}^l \vec{S}^{-l} + \vec{N} \right) \\ &= \mathbf{D}\vec{S} + \vec{N}',\end{aligned}\tag{4.3}$$

where $\vec{Y} = [Y_1 \cdots Y_{N_{mic}}]^T$, $\vec{S}^{-l} = [S_1^{-l} \cdots S_{N_S}^{-l}]^T$, $\vec{N} = [N_1 \cdots N_{N_{mic}}]^T$, as in (1.9),

$$\begin{aligned}\mathbf{D} &= \begin{bmatrix} D_{1,1} & D_{1,2} & \cdots & D_{1,N_S} \\ D_{2,1} & D_{2,2} & \cdots & D_{2,N_S} \\ \vdots & \vdots & \ddots & \vdots \\ D_{N_{mic},1} & D_{N_{mic},2} & \cdots & D_{N_{mic},N_S} \end{bmatrix}, \\ &= \begin{bmatrix} \vec{D}_1 & \vec{D}_2 & \cdots & \vec{D}_{N_S} \end{bmatrix}.\end{aligned}$$

where $D_{i,j}$ corresponds to short-time Fourier transform of $h_{i,j}^{d,\theta}(t)$, and

$$\begin{aligned}\mathbf{R}^l &= \begin{bmatrix} R_{1,1}^l & R_{1,2}^l & \cdots & R_{1,N_S}^l \\ R_{2,1}^l & R_{2,2}^l & \cdots & R_{2,N_S}^l \\ \vdots & \vdots & \ddots & \vdots \\ R_{N_{mic},1}^l & R_{N_{mic},2}^l & \cdots & R_{N_{mic},N_S}^l \end{bmatrix}, \\ &= \begin{bmatrix} \vec{R}_1^l & \vec{R}_2^l & \cdots & \vec{R}_{N_S}^l \end{bmatrix}.\end{aligned}$$

where $R_{i,j}^l$ corresponds to short-time Fourier transform of $h_{i,j}^r(t)$. If we assume we know the DOA response (4.4), the sufficient statistic in (1.12) can be reformulated as MVDR beamforming:

$$T_{\vec{D}_j}(S_j) = \frac{\vec{D}_j^H \mathbf{R}_n'^{-1} \vec{Y}}{\vec{D}_j^H \mathbf{R}_n'^{-1} \vec{D}_j}.\tag{4.4}$$

Since we can assume that $N_i(k)$ is a zero mean complex Gaussian random variable independent of $S(k')$ and $N_i(k')$ for $k' \neq k$, “k” is omitted for readability.

Our objective is to jointly estimate the source speech and the DOA given the multi-microphone measurements together with the statistical assump-

tions about the source speech, late room impulse response, and the additive noise. We make the following statistical assumptions:

- The spectrum of the late room impulse response follows a normal distribution with zero mean, i.e.,
 $\vec{R}_j^l \sim N(\vec{0}, \Sigma_{\vec{R}_j^l})$ [88].
- The spectrum of the additive noise follows a normal distribution with zero mean, i.e.,
 $\vec{N} \sim N(\vec{0}, \Sigma_{\vec{N}})$ [48].

Σ is an N -by- N covariance matrix.

Note that in (4.4), we redefined a noise vector $\vec{N}'$, and this can also be modeled as a zero mean Gaussian, i.e., $\vec{N}' \sim N(\vec{0}, \Sigma_{\vec{N}'})$. Therefore, in this chapter, instead of estimating the covariance matrix of the reverberation and noise individually, we try to estimate the combined noise statistic. Furthermore, we reformulate the MVDR beamforming into the minimum power distortionless response (MPDR) beamforming, as in Section 3.2.1, which can be shown to be the same as MVDR under the assumption of a known DOA [31].

4.3 Sufficient statistic: EM beamforming

Our goal is to iteratively estimate S_j given the complete observation of $\vec{Y}(k)$ and the incomplete observation of θ with the updated PDF

$$p(\theta | \vec{Y}(1), \dots, \vec{Y}(K), S_j). \quad (4.5)$$

This situation fits well into the EM setup by introducing the functional

$$Q(S_j(k), S_j^0(k)) = E_{p(\theta | \vec{Y}(1), \dots, \vec{Y}(K), S_j^0(k))} [\log p(\vec{Y}(k), \theta | S_j(k))], \quad (4.6)$$

where $S_j^0(k)$ and $S_j(k)$ denote the k^{th} frequency bin estimate at the previous and current step, respectively. Hereafter, without the subscript k , all the terms are assumed to be in the same frequency bin. We decompose $p(\vec{Y}, \theta | S_j)$ as a product of a likelihood PDF given θ and *a priori* PDF of θ .

$$p(\vec{Y}, \theta | S_j) = p(\vec{Y} | \theta, S_j) p(\theta | S_j), \quad (4.7)$$

where the likelihood PDF is a Gaussian PDF,

$$p(\vec{Y}|\theta, S_j) \sim \frac{1}{\pi^N |\Sigma_{\vec{Y}}|} \exp^{-[\vec{Y} - \vec{D}_j S_j]^H \Sigma_{\vec{Y}}^{-1} [\vec{Y} - \vec{D}_j S_j]}. \quad (4.8)$$

Note that we exploited the assumption that the residual term

$$\sum_{j'=1, \neq j}^{N_S} \vec{D}_{j'} S_{j'} + \sum_{l=1}^{N_H-1} \left(\mathbf{R}^l \vec{S}^{-l} + \vec{N} \right) \sim N(\vec{0}, \Sigma_{\vec{N}'}) \quad (4.9)$$

can be modeled to follow a zero mean Gaussian PDF. If we simply assume that S_j , $j = 1, 2, \dots, N_S$ are independently modeled to follow a zero mean Gaussian [9, 48], the assumption of Gaussianity for the residual perfectly holds. However, even with a non-Gaussian model for speech sources, the assumption of Gaussianity for the residual is still very reasonable, because it is composed of not only the speech component but also late reverberation and background noise, which follow a Gaussian PDF. Therefore, the MMSE speech estimation can be performed differently in the post-filtering part by using a different modeling of speech, e.g. super-Gaussian such as Laplacian, which is known to be more precise [59, 60, 90]. A super-Gaussian prior was successfully employed in speech source separation in Chapter 3.

In the maximization step, we estimate $S_j(k)$ of the current step:

$$\begin{aligned} \hat{S}_j(k) &= \operatorname{argmax}_{S_j(k)} Q(S_j(k), S_j(k)^0) \\ &= \int_{\Theta} p(\theta | \vec{Y}(1), \dots, \vec{Y}(K), S_j(k)^0) \\ &\quad \cdot \frac{\vec{D}_j(k)^H \Sigma_{\vec{Y}(k)}^{-1} \vec{Y}(k)}{\vec{D}_j(k)^H \Sigma_{\vec{Y}(k)}^{-1} \vec{D}_j(k)} d\theta. \end{aligned} \quad (4.10)$$

Equation (4.10) is a weighted average of MPDR beamformers, weighted by the posterior PDF of θ rather than a single specific MPDR beamformer with single DOA estimation, where the weights are given by the *a posteriori* PDF of θ . Therefore, (4.10) constitutes a more general formulation of the conventional MPDR beamformer. Note that if we can assume no inter-channel correlation for the noise covariance matrix, which is not achievable especially in the low-frequency region with a single DOA estimate, it becomes the well-known delay-and-sum beamformer. The proposed approach is similar to the

case of the previous Bayesian beamformer [86], except that we are taking the current estimate of the source spectrum into consideration, and we are dealing with a sparse broad-band signal. Statistical reverberation $\mathbf{R}^l$ plays the role of increasing the covariance from noise-only to noise-plus-reverberation, and it also affects the *a posteriori* PDF of θ .

Equation (4.10) is a maximum likelihood estimate of $S_j(k)$, and by applying an MMSE post-filter such as a Wiener filter after the convergence of EM iteration, we can optimally estimate $S_j(k)$ in the MMSE sense as depicted in Figure 4.1 using the solid loop. On the other hand, we can put the MMSE post-filtering process into the iteration loop as depicted in Figure 4.1 using the dotted loop, by assuming that the speech source $S_j(k)$ follows a Gaussian, with the following slight modification of (4.6):

$$Q(S_j(k), S_j^0(k)) = E_{p(\theta|\vec{Y}(1), \dots, \vec{Y}(K), S_j^0(k))}[S_j(k) | \log p(\vec{Y}(k), \theta)]. \quad (4.11)$$

Because maximum *a posteriori* (MAP) estimation of $S_j(k)$ is the same as MMSE estimation with the Gaussian PDF, by performing the two-step, EM beamforming followed by MMSE post-filtering in each iteration, we can achieve MAP estimation of $S_j(k)$ instead of ML estimation in (4.10).

The *a posteriori* PDF of θ follows Bayes' rule:

$$p(\theta|\vec{Y}(1), \dots, \vec{Y}(K), S_j(k)^0) \sim p(\vec{Y}(1), \dots, \vec{Y}(K)|\theta, S_j(k)^0)p(\theta|S_j(k)^0), \quad (4.12)$$

where $p(\theta|S_j(k)^0)$ is the *a priori* PDF of θ , reflecting our prior knowledge of the possible source DOAs.

In the expectation step, we calculate the likelihood

$$p(\vec{Y}(1), \dots, \vec{Y}_K|\theta, S_j(k)^0) \sim \exp^{\sum_{k=1}^K M(k) \log N(\vec{D}_j(k) S_j(k)^0; \Sigma_{\vec{Y}(k)})}, \quad (4.13)$$

where $M(k) = 1$ if $S_j(k)$ is dominated by the target speech S_j , if not, $M(k) = 0$. The determination of $M(K)$ is based on the likelihood similarity, because across all frequency bins the target source should have the same DOA. Figure 4.2 shows a simple conceptual example, where the second, fourth, and sixth bins have a similar likelihood shape within a possible DOA range set by *a priori* PDF of θ . If K is below a certain threshold, which means most frequency bins are dominated by noise or interference, we consider the entire frame as noise, and this process is interpreted as a gen-

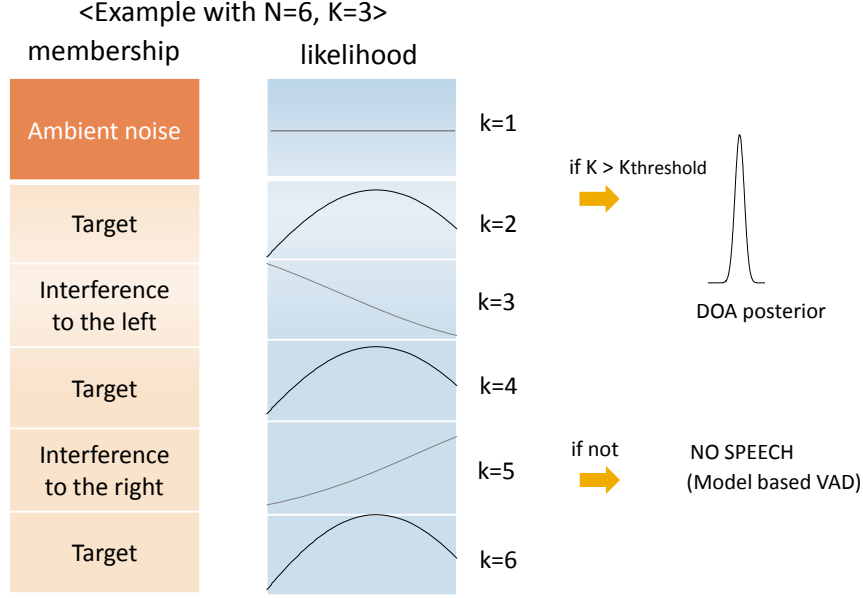


Figure 4.2: Conceptual diagram of the generative model based determination of $M(k)$: example of frequency bin $N = 6$ and target speech dominated frequency bin $K = 3$ at $k = 2, 4, 6$.

erative model based voice activity detection (VAD). The introduction of the membership classification to update the *a posteriori* PDF is very important, because it allows us to have much more robust DOA estimation due to the fact that we only use the target dominated frequency bins. Given the fact that speech does only occupy a relatively small portion of the frequency bin, this procedure is thought to be critical.

For prior $p(\theta|S_j^0)$ we might be able to simply use real prior information about θ based on the application scenario, e.g. from video tracking results we may have a possible search range for the target speech region. The hard binary thresholding of $M(k)$ can be replaced with a soft thresholding as in (3.19), which can be interpreted as target speech existence probability in that frequency bin.

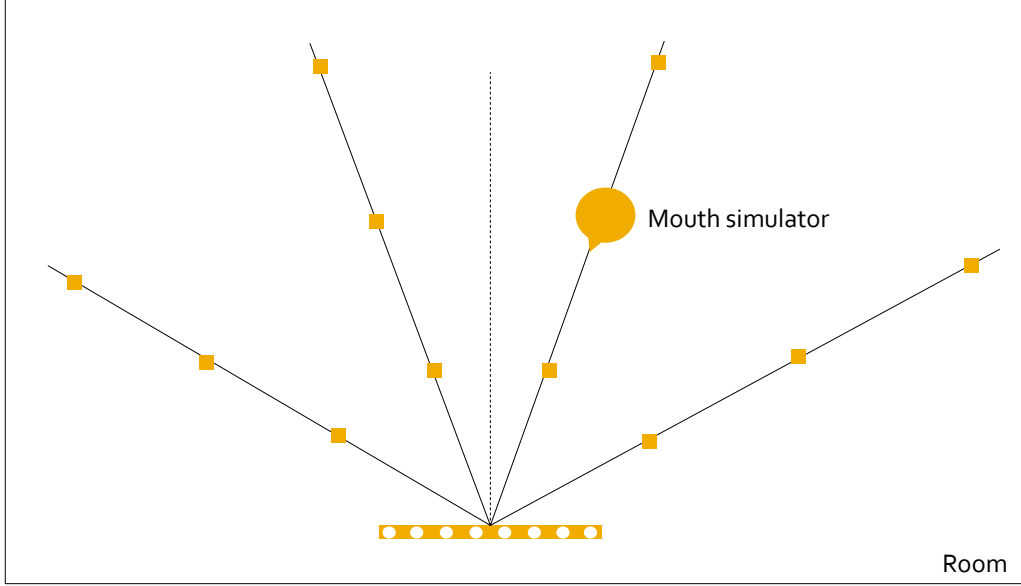


Figure 4.3: RIR measurement setup.

4.4 Experiments

4.4.1 Experiment with a generated corpus

In this section, we tested the proposed method with a generated corpus by measuring the room impulse responses in an office environment. Generating the corpus is documented in Appendix E. Natural ambient noise was also measured in the same place. Twelve different configurations shown in Figure 4.3 are generated. Among them we tested the most difficult setting of a two-speaker case, in which angle distance was about 32° and distance from speakers to array was 2.1 m. Note that as the distance from the speaker position to the microphone array gets longer, the direct-to-reverberation ratio (DRR) can become less in the same angle direction.

For the DOA accuracy evaluation, in average 17 dB SNR by adding the recorded noise, we first compared the results of our proposed method with the results of the conventional methods such as GCC-PHAT [35] and MUSIC [36]. As shown in Figure 4.4, we can verify that in this scenario the one best angle for all the methods closely corresponds to the designated angle 32° . Note that since we do not employ inter-frame smoothing such as Kalman filtering to utilize the information of the previous frames, the estimation results look noisier around the designated DOA. For MUSIC, the angles corresponding

to the maximum of mean values for all frequency bins per each search angle have been assigned [34].

With the capability of finding a generative model based likelihood per frequency bin, we can separate the sources by selecting the frequency bins sharing similar DOA. To validate this idea, we simply estimate the ideal binary mask for target source location based on EMB, and confirm that EMB based post-processing can make further separation as shown in Figure 4.5, where we can compare the EMB result with the original clean speech.¹

4.4.2 Real scenario with a more challenging setup

In this real experiment, we used three horizontal channels, which come from one of the four T-shape microphone arrays of the IBM smart room [87]. five-second long speech is arbitrarily selected with the one-second rate of head tracking results from video data taken as the ground truth. We set the search DOA angle from 31° to 71° with a 0.5° increase based on the head tracking result. Initial estimation of the speech spectrum has been performed using an average MPDR based on uniform distribution among the candidate DOAs. The sampling rate was 16 kHz and the room geometry was roughly $7 \times 6 \times 3$ m. The detailed configuration can be found in [87].

Figure 4.6 shows DOA tracking results every 16 ms with the ground truth marked as a red star every 1 second. Figure 4.7 shows the spectrograms of one of the three channel measurements, the output of the EM beamforming based on the result of Figure 4.6, and the MMSE-logSA [48] estimate of the EM beamforming output. Finally, Figure 4.8 depicts the corresponding time-domain signals.

We first observe that the DOA tracking result of the proposed method produces a reasonably smooth trajectory connecting the ground truth angles, compared to conventional methods. Instantaneous update of DOA based on MUSIC produces severe perturbations due to the unreliable covariance matrix estimation at noise-dominant frequency bins. To perform GCC-PHAT, the nearer two channels have been used. Again we observe the more severe jittering due to the same reason, namely that we do not have the ability to exclude the unreliable frequency portion in the short-time spectral represen-

¹The wave files of this figure is linked in the following web site: <http://www.isle.uiuc.edu/lkim9/Demo>.

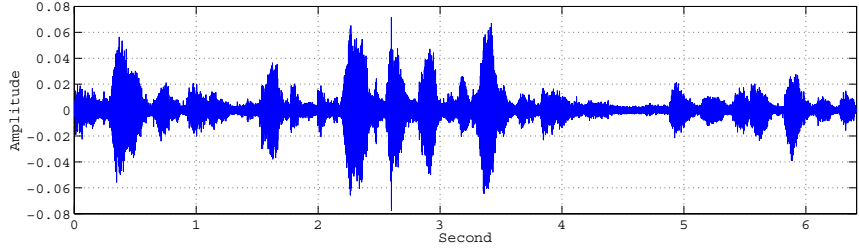
tation. We do not introduce a conventional trajectory smoothing method such as Kalman filtering, but the proposed method can be easily combined with it because we can obtain the DOA observation likelihood by default.

The DOA estimate performance can be evaluated not only by smoothness of the trajectory but also by the enhanced spectrogram in the high-frequency range. As we see in the spectrogram, the estimated DOA actually boosts the speech spectrum (darker yellow) up to the relatively high frequency range around 4 kHz and attenuates the region of silence (darker blue) effectively, which is ultimately further suppressed by the MMSE postfiltering. Note that the wavelength of 4 kHz is around 8.5 cm; therefore, enhancement in the high-frequency range can only be achieved with more accurate DOA estimates. The speech spectrum has been boosted, and these boosted spectra are still prominent after the MMSE process. Note also that during non-speech regions, the DOAs have been reported as -10, which is an arbitrarily assigned number when there exist no reliable frequency bins at a specific frame, without relying on conventional voice activity detection (VAD) algorithms.

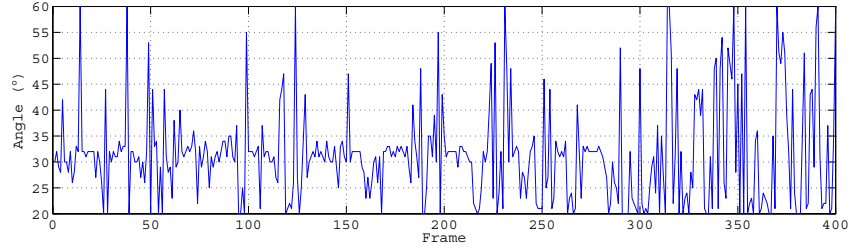
Note that because we are implementing the proposed algorithm in the short-time Fourier transform domain, we use the overlap-addition method to synthesize back to a time domain signal with a hanning window and 50% overlap.

4.5 Conclusions and future work

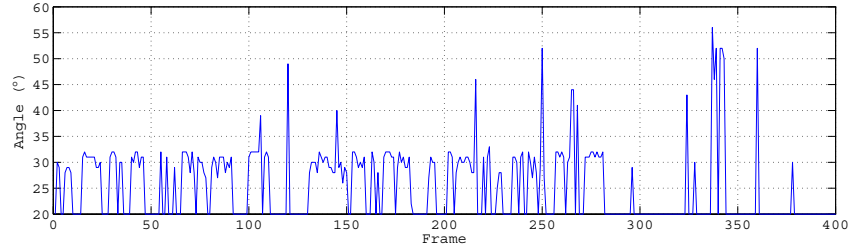
In this chapter, joint optimal estimation of the DOA and the source spectrum is proposed through a newly derived EM beamforming algorithm. The EM beamformed spectrum constitutes a realistic sufficient statistic for MMSE post-filter. The maximum *a posteriori* PDF of DOA is obtained after convergence by taking the maximum of the updated PDF. An experiment with real data shows that the proposed algorithm represents a realistic way to achieve good speech signal estimation with good instantaneous DOA estimation even at the high-frequency range. The proposed method is a general approach which can be accommodated with any microphone array configuration. Combination with trajectory smoothing based on temporal dynamics will be studied in the near future.



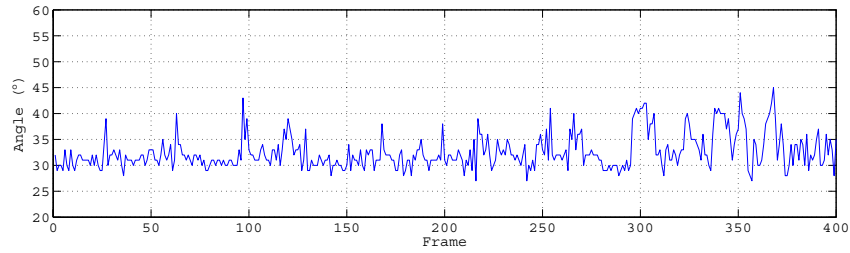
(a)



(b)

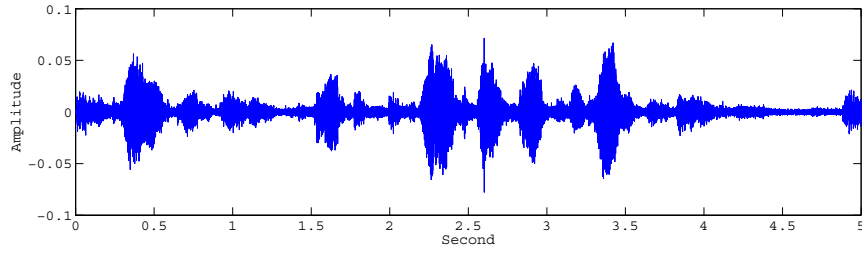


(c)

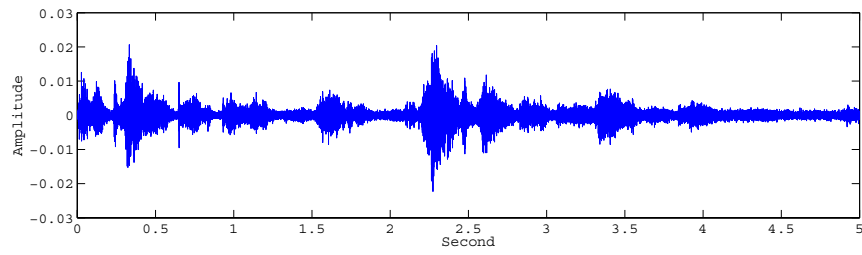


(d)

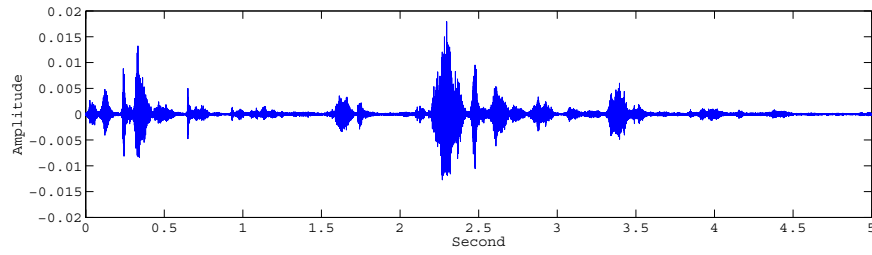
Figure 4.4: DOA estimation results with original reverberant speech mixture (target speech was active up to 280 frames at 32° , interference speech was active during the entire frames at -30°): (a) reverberant speech mixture captured by one of the array microphones, (b) DOA estimation by GCC-PHAT with a pair of microphones at distance 0.076 m, (c) DOA estimation by MUSIC (8-microphone linear array with 0.03795 m inter-distance), and (d) DOA estimation by EMB (8-microphone linear array with 0.03795 m inter-distance).



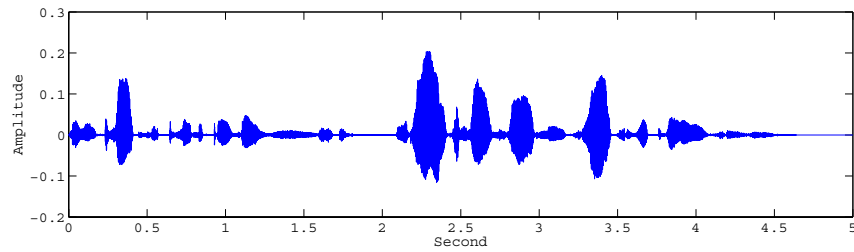
(a)



(b)

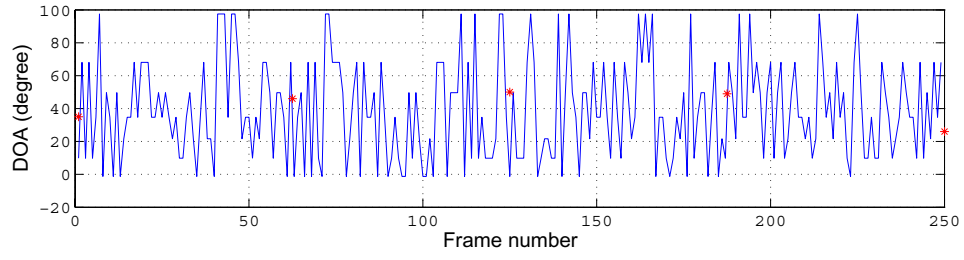


(c)

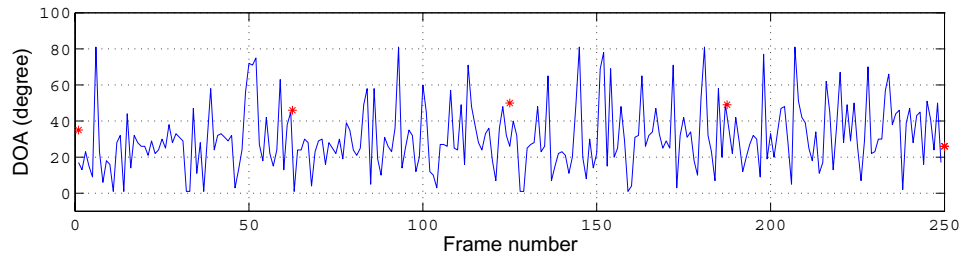


(d)

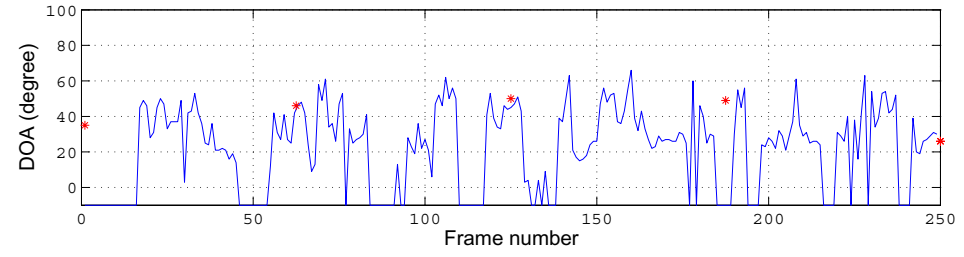
Figure 4.5: EMB result: (a) reverberant mixture, (b) EMB result, (c) EMB with direction-per-frequency based IBM estimation, (d) original clean target speech.



(a)



(b)



(c)

Figure 4.6: DOA estimation with ground truth (every 1 second marked as red star): Proposed method (top), MUSIC (middle), GCC-PHAT (bottom).

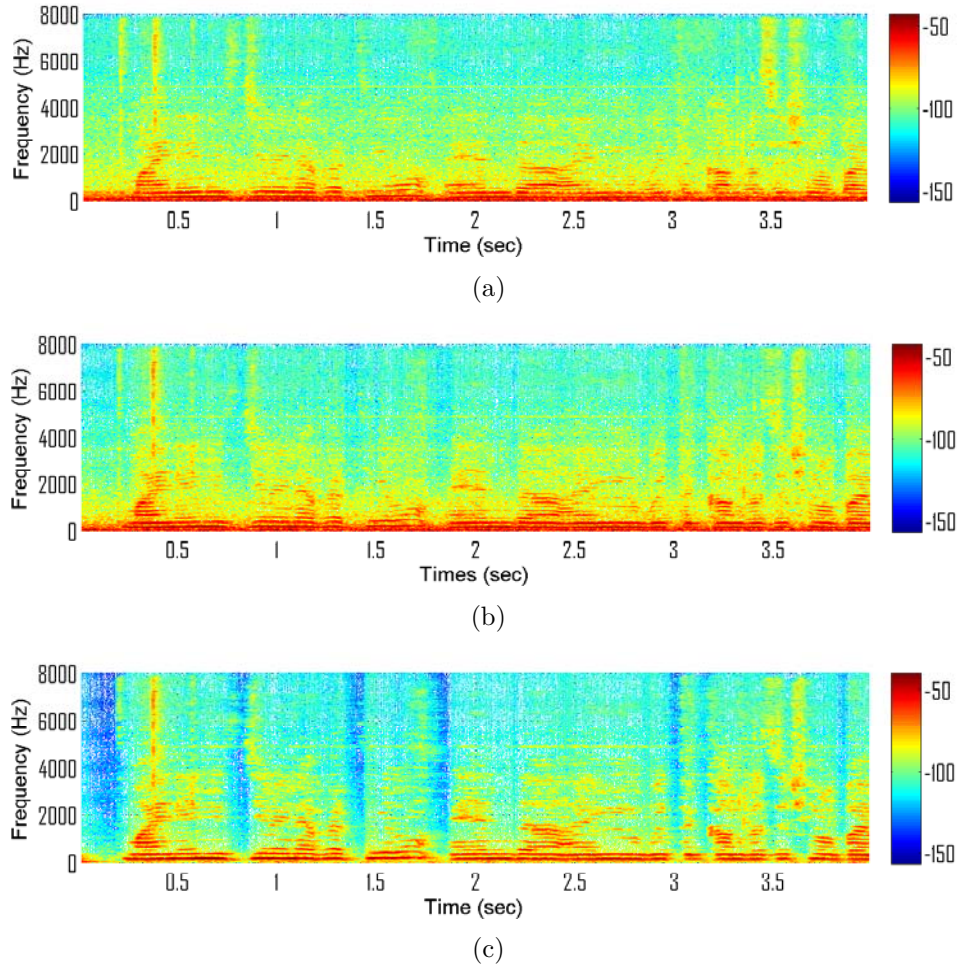


Figure 4.7: Spectrogram: unprocessed channels (top), EM beamforming with one iteration (middle), EM beamforming + MMSE-logSA (bottom).

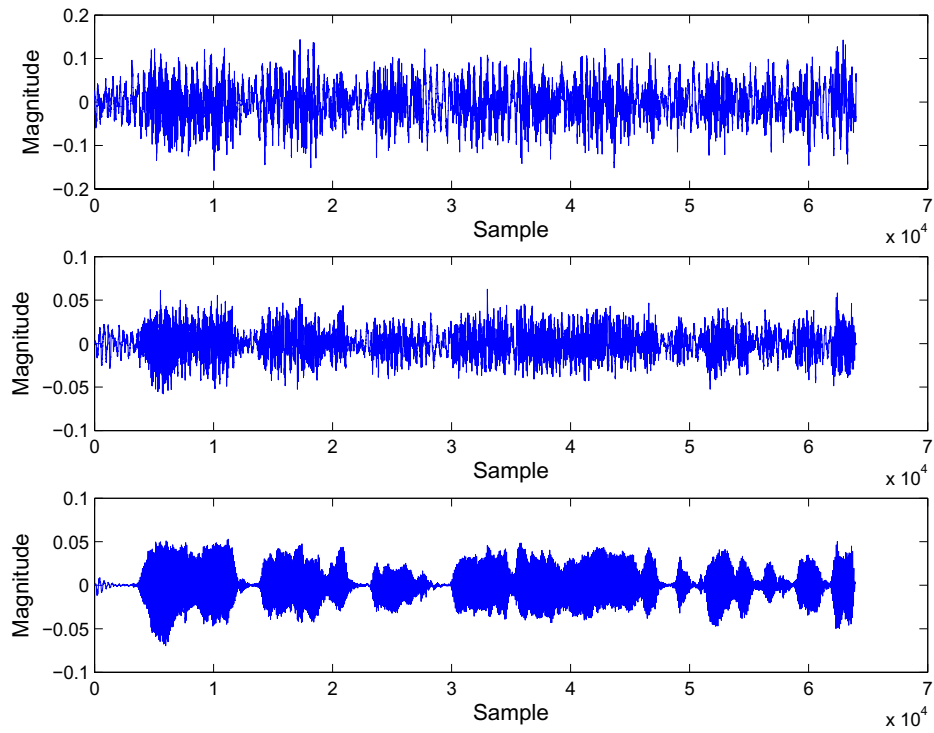


Figure 4.8: Time domain results: unprocessed channels (top), EM beamforming with one iteration (middle), EM beamforming + MMSE-logSA (bottom).

CHAPTER 5

EARLY ROOM IMPULSE RESPONSE ESTIMATION BEYOND DOA ESTIMATION

In this chapter, we give more attention to early RIR so that we can speed up the demixing, dereverberation, and de-noising process to track the moving sources. The joint estimation of DOA and speech in Chapter 4 is extended to the distinct early response estimation, if any, and probabilistic compensation on the remaining part. If we do not compensate the distinct echo paths as in Chapter 4, we may experience unsatisfactory speech estimation results. Also, in the situation where we need to update the speech estimation process fast enough, we noticed that the direct deconvolution and separation method proposed in Chapter 3 may not be suitable, because it requires several sub-band taps even for the case of a few distinct early response paths. This can be easily understood by comparing a simple situation, e.g. estimation of impulse response of $[1 \ 0.5]$ and its inverse, where its inverse should be represented using a longer filter length to ensure sufficient inverse filtering. This requires a much longer observation to estimate all the parameters. If we can achieve similar performance with the method proposed in Chapter 3 with fast updating, this might be the method to achieve separation and tracking at the same time.

If we can assume that we know the early RIRs and treat the late RIR as a random variable as in (4.3), optimal estimation of the speech signal can be done using a simple two-step method: the sufficient statistic compensating for the distinct early response followed by a MMSE estimator. In this chapter, we will first address some realistic implementations of the sufficient statistic considering the distinct response based on a deterministic RIR approach first and then consider the remaining part of RIR based on a probabilistic RIR approach.

5.1 Multi-path GSC

In this section, we derive a multi-path GSC, which is a generalized version of a generalized sidelobe canceller (GSC) [41,91] under a multi-path acoustic environment; we also show that we can get the optimal estimator in a Bayesian sense via multi-path GSC followed by Bayesian estimation by showing that the multi-path GSC has the same formulation as the sufficient statistic. An adaptive beamforming algorithm, GSC is quite common in the source signal enhancement community, but to our knowledge, this is the first time that GSC or multi-path GSC has been shown to be a sufficient statistic for multi-microphone Bayesian optimal spectral estimation. We therefore are showing that multi-path GSC followed by Bayesian single-channel estimation is the optimal multi-microphone source signal estimation in a multi-path environment. The beauty of this approach is in the fixed beamformer (FBF) and the subsequent blocking matrix, which can be constructed as the null space of the multi-channel convolution matrix. The FBF can be simply regarded as a multi-channel deconvolution problem. Therefore, we might be able to apply any kind of multi-channel deconvolution scheme [6, 7, 92, 93] for this fixed part. The blocking matrix can also be constructed by using an echo cancellation scheme [94], because ideally the output of the fixed beamformer is the deconvolved and beamformed speech signal.

Conventional GSC as shown in Figure 5.1 is derived from the linearly constrained adaptive beamforming framework (5.1) [91, 95].

$$\arg \min_{\vec{w}} E \{ \vec{w}^T \vec{y} \vec{y}^T \vec{w} \} \quad \text{subject to } \mathbf{C}^T \vec{w} = \vec{f}, \quad (5.1)$$

where $\hat{s}(n) = \vec{w}^T \vec{y}$ is the time domain estimated source signal, superscript T is transpose, $\vec{w} = [\vec{w}_1^T \vec{w}_2^T \cdots \vec{w}_N^T]^T$ and $\vec{w}_i^T = [w_{(i-1)L+1} \ w_{(i-1)L+2} \cdots w_{(i-1)L+L}]$ where $i = 1, 2, \dots, N$, $\vec{y} = [y_{1,[1:L]} \ y_{2,[1:L]} \cdots y_{N,[1:L]}]^T$ where $y_{i,[1:L]} = [y_i(n - (i-1)n_0) \ y_i(n - (i-1)n_0 - 1) \cdots y_i(n - (i-1)n_0 - (L-1))]^T$, $i = 1, 2, \dots, N$ and n is the current time index, steered to a look-direction of $\theta = \arcsin(n_0 F_s d / c)$ for microphone spacing d and sampling rate F_s , and $\mathbf{C}^T \vec{w} = \vec{f}$ is a linear constraint. To derive the multi-path GSC, we simply need to manipulate the constraint part in (5.1). The constraint part has the following convolutional

form:

$$\begin{bmatrix} \mathbf{C}_{h_1} & \mathbf{C}_{h_2} & \cdots & \mathbf{C}_{h_N} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{NL} \end{bmatrix} = \vec{f} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (5.2)$$

where $l_h + L - 1$ by L matrix $\mathbf{C}_{h_i}$ is constructed from the room response $h_i(n)$:

$$\mathbf{C}_{h_i} = \begin{bmatrix} h_i(0) & 0 & \cdots & 0 \\ h_i(1) & h_i(0) & \ddots & \vdots \\ \vdots & h_i(1) & \ddots & 0 \\ h_i(l_h - 1) & \vdots & \ddots & h_i(0) \\ 0 & h_i(l_h - 1) & \vdots & h_i(1) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_i(l_h - 1) \end{bmatrix}, \quad (5.3)$$

$i = 1, 2, \dots, N$

Equation (5.3) is a typical linear convolution matrix which has a Toeplitz structure. The constraint in (5.2) is therefore equivalent to the requirement that $\vec{W}$ inverts $\vec{H}$, i.e. (5.2) is actually a channel deconvolution in the look-direction [92]. Note that the length of the constraint vector $\vec{f}$ is determined as $l_h + L - 1$ in (5.3). Here is a simple example of an anechoic case where $N = 2$, $L = 2$, and $h_1(n) = h_2(n) = \delta(n)$ so $l_h = 1$; then by (5.3)

$$\mathbf{C}_{h_1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, i = 1, 2 \quad (5.4)$$

and

$$\mathbf{C}^T = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}. \quad (5.5)$$

This gives the original GSC [91] result because we consider only the direction of arrival. Therefore, conventional GSC is a special case of multi-path GSC, in which the room response is a delta function. Similarly, the anechoic case is always a special case of the more general (5.2). Equation (5.2) guarantees deconvolution of the acoustic channel; at the same time, the minimization of total power (5.1) provides noise suppression in the non-look-directions.

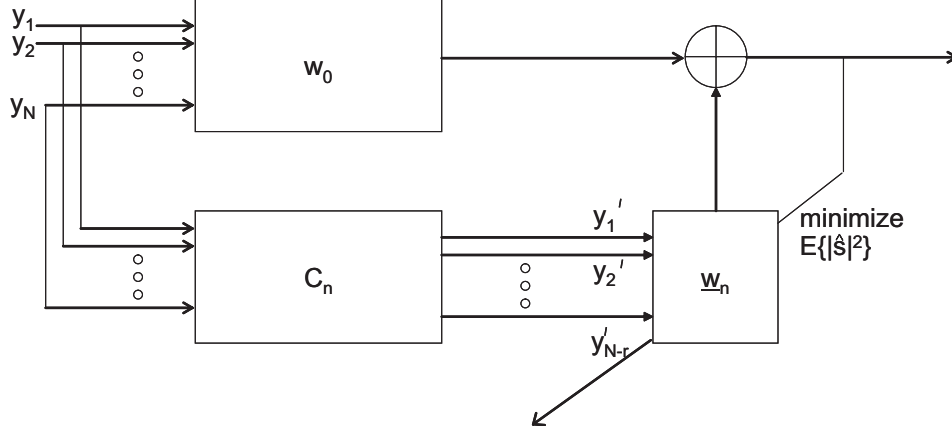


Figure 5.1: Block diagram of GSC beamformer. y_1 through y_N are measured input for beamforming, w_0 is a fixed beamformer shown in (5.7), C_n is the blocking matrix shown in (5.9), and $\vec{w}_n$ is the adaptable beamforming coefficient vector.

In conventional GSC, the coefficient vector $\vec{w}$ is decomposed as $\vec{w} = \vec{w}_0 - \vec{v}$, where $\vec{w}_0$ is obtained by the Moore-Penrose pseudo-inverse $\vec{w}_0 = \mathbf{C}(\mathbf{C}^T \mathbf{C})^{-1} \vec{f}$ which guarantees a minimum norm least square (MNLS) solution. The minimization of total power in (5.1) is computed by adopting the coefficients $\vec{v} = \mathbf{C}_n \vec{w}_n$ as shown in Figure 5.1, where the blocking matrix $\mathbf{C}_n$ can be obtained from $\mathbf{C}$ using orthogonalization procedures such as singular value decomposition (SVD) as shown below. The total length of the array filter, NL , should be larger than the row count of $\mathbf{C}^T$, $l_h + L - 1$, to guarantee the existence of the solution $\vec{w}_0$. In other words, the difference, $NL - (l_h + L - 1)$, is the number of degrees of freedom to be used for suppressing noise. First, the fixed upper part $\vec{w}_0$ can be obtained easily using singular value decomposition (SVD) of the rank r , $(l_h + L - 1) \times (NL)$ constraint matrix $\mathbf{C}^T$.

$$\mathbf{C}^T = \mathbf{U} \Sigma \mathbf{V}^H, \quad (5.6)$$

where $(l_h + L - 1) \times r$ matrix $\mathbf{U}$ and $(NL) \times r$ matrix $\mathbf{V}$ are orthonormal, $\mathbf{U}^H \mathbf{U} = \mathbf{V}^H \mathbf{V} = \mathbf{I}$, and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)$ is the $r \times r$ diagonal matrix whose diagonal entries, called singular values, are positive values sorted in

decreasing order. Then $\vec{w} = \vec{w}_0 + \vec{v}$, where

$$\vec{w}_0 = \mathbf{V} \cdot \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_r} \end{bmatrix} \cdot \mathbf{U}^H \cdot \vec{f} \quad (5.7)$$

lies in the range space of $\mathbf{C}^T$ and $\vec{v}$ lies in the null space of $\mathbf{C}^T$ such that $\mathbf{C}^T \vec{v} = 0$, $\vec{v} = \mathbf{C}_n \cdot \vec{w}_n$ where $\mathbf{C}_n$ is called the blocking matrix and is obtained by extending the SVD of $\mathbf{C}^T$ as shown below.

$$\begin{aligned} \mathbf{C}^T &= [\mathbf{U} \ u_{r+1} \ u_{r+2} \ \cdots \ u_{l_h+L-1}] \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \\ &\cdot [\mathbf{V} \ v_{r+1} \ v_{r+2} \ \cdots \ v_{N \cdot L}]^H \end{aligned} \quad (5.8)$$

$$\mathbf{C}_n = [v_{r+1} \ v_{r+2} \ \cdots \ v_{N \cdot L}]. \quad (5.9)$$

The linearly constrained adaptive beamforming framework as shown in (5.1) can be reformulated in the discrete frequency domain. First, (5.1) is expanded to include a time frame with length l_f as in (5.10).

$$\arg \max_{\vec{w}} E \{ \vec{w}^T \mathbf{y}_m \mathbf{y}_m^T \vec{w} \} \text{ subject to } \mathbf{C}^T \vec{w} = \vec{f}, \quad (5.10)$$

where $\mathbf{y}_m^T = [\mathbf{y}_{1,m} \ \mathbf{y}_{2,m} \ \cdots \ \mathbf{y}_{N,m}]$ and

$$\mathbf{y}_{i,m} = \begin{bmatrix} y_i(0) & 0 & \cdots & 0 \\ y_i(1) & y_i(0) & \ddots & \vdots \\ \vdots & y_i(1) & \ddots & 0 \\ y_i(l_f - 1) & \vdots & \ddots & y_i(0) \\ 0 & y_i(l_h - 1) & \vdots & y_i(1) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & y_i(l_h - 1) \end{bmatrix}, \quad i = 1, 2, \cdots, N, \quad (5.11)$$

and we set the beginning time index of the frame as 0 for convenience. After transforming the convolution matrices in (5.3) and (5.11) into the circulant

matrices as in (5.12) and (5.13),

$$\mathbf{C}'_{h_i} = \begin{bmatrix} h_i(0) & h_i(L-1) & h_i(L-2) & \cdots & h_i(1) \\ h_i(1) & h_i(0) & h_i(L-1) & \cdots & h_i(2) \\ h_i(2) & h_i(1) & h_i(0) & \cdots & h_i(3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_i(L-1) & h_i(L-2) & h_i(L-3) & \cdots & h_i(0) \end{bmatrix}, i = 1, 2, \dots, N, \text{ and} \quad (5.12)$$

$$\mathbf{y}'_{i,m} = \begin{bmatrix} y_i(0) & y_i(L-1) & y_i(L-2) & \cdots & y_i(1) \\ y_i(1) & y_i(0) & y_i(L-1) & \cdots & y_i(2) \\ y_i(2) & y_i(1) & y_i(0) & \cdots & y_i(3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_i(L-1) & y_i(L-2) & y_i(L-3) & \cdots & y_i(0) \end{bmatrix}, i = 1, 2, \dots, N. \quad (5.13)$$

Equation (5.10) can be reformulated as in (5.14).

$$\begin{aligned} \arg \max_{\vec{w}} E \left\{ \sum_{i=1}^N \vec{w}_i^H \mathbf{F} \mathbf{F}^H y'_{i,m} {}^H \mathbf{F} \sum_{i=1}^N \mathbf{F}^H y'_{i,m} \mathbf{F} \mathbf{F}^H \vec{w}_i \right\} \\ \text{subject to } \sum_{i=1}^N \mathbf{F}^H \mathbf{C}'_{h_i} \mathbf{F} \mathbf{F}^H \vec{w}_i = \mathbf{F}^H \vec{f}, \end{aligned} \quad (5.14)$$

where $\mathbf{F}^H$ and $\mathbf{F}$ are discrete Fourier transform (DFT) and inverse DFT matrices, respectively, and $\mathbf{F}^H \mathbf{F} = \mathbf{F} \mathbf{F}^H = \mathbf{I}$. Equation (5.14) is converted into the frequency domain representation of (5.15).

$$\arg \max_{[\vec{W}_1^T \ \vec{W}_2^T \ \cdots \ \vec{W}_N^T]^T} E \left\{ \sum_{i=1}^N \vec{W}_i^H \Lambda_{y,i}^H \sum_{i=1}^N \Lambda_{y,i} \vec{W}_i \right\} \text{ subject to } \sum_{i=1}^N \Lambda_{H,i} \vec{W}_i = \vec{F}, \quad (5.15)$$

where $\Lambda_{y,i} = \mathbf{F}^H y'_{i,m} \mathbf{F}$ and $\Lambda_{H,i} = \mathbf{F}^H \mathbf{C}'_{h_i} \mathbf{F}$, where $i = 1, 2, \dots, N$, are diagonal matrices, whose diagonal entries $\vec{Y}_i$ and $\vec{H}_i$ are frequency samples of the input signal $[y_i(0) \ y_i(1) \ \cdots \ y_i(l_f - 1) \ \vec{0}_y^T]^T$ and the i^{th} room response $[h_i(0) \ h_i(1) \ \cdots \ h_i(l_h - 1) \ \vec{0}_h^T]^T$ computed by applying the DFT, and $\vec{0}_h$ is a zero vector with length of $L - l_h$, and $\vec{W}_i = \mathbf{F}^H \vec{w}_i$, $\vec{F} = \mathbf{F}^H \vec{f}$. In a specific

frequency bin, if we assume $\vec{F} = \vec{1}$, (5.15) is reformulated as in (5.16).

$$\begin{aligned} \arg \max_{\vec{W}} E \left\{ \vec{W}^H \vec{Y} \vec{Y}^H \vec{W} \right\} \\ \text{subject to } \vec{H}^H \vec{W} = 1, \end{aligned} \quad (5.16)$$

where $\vec{W} = [W_1(f) \ W_2(f) \ \cdots \ W_N(f)]^T$, $\vec{Y} = [Y_1(f) \ Y_2(f) \ \cdots \ Y_N(f)]^T$ and $\vec{H} = [H_1(f) \ H_2(f) \ \cdots \ H_N(f)]^T$. Equation (5.16) can be solved by the method of Lagrange multipliers at the specific frequency. The noise covariance matrix is $\mathbf{R}_n$, and

$$E\{\vec{Y}\vec{Y}^H\} = \mathbf{R}_n. \quad (5.17)$$

The Lagrangian functional for (5.16), with Lagrange multiplier λ , is

$$J(\vec{W}) = \frac{1}{2} \vec{W}^H \mathbf{R}_n \vec{W} + \lambda^H (\vec{H}^H \vec{W} - 1). \quad (5.18)$$

We get the optimal $\vec{W}$ by differentiating $J(\vec{W})$ with respect to $\vec{W}$ and setting it to zero.

$$\vec{W} = -\mathbf{R}_n^{-1} \vec{H} \lambda. \quad (5.19)$$

From the definition of the constraint

$$\vec{H}^H \vec{W} = 1 = -\vec{H}^H \mathbf{R}_n^{-1} \vec{H} \lambda. \quad (5.20)$$

From (5.20)

$$\lambda = \frac{-1}{\vec{H}^H \mathbf{R}_n^{-1} \vec{H}}. \quad (5.21)$$

From (5.21) and (5.19)

$$\vec{W} = \frac{\mathbf{R}_n^{-1} \vec{H}}{\vec{H}^H \mathbf{R}_n^{-1} \vec{H}}. \quad (5.22)$$

$$\vec{W}^H = \frac{\vec{H}^H \mathbf{R}_n^{-1}}{\vec{H}^H \mathbf{R}_n^{-1} \vec{H}}. \quad (5.23)$$

If we multiply $\vec{Y}$, then the solution derived from the frequency domain representation of the multi-path GSC has the identical formulation as the sufficient statistic. Therefore, multi-path GSC is the sufficient statistic to obtain an optimal estimate from the multi-microphone data, and we can get the Bayesian

optimal estimate of the source signal using multi-path GSC followed by a single-channel Bayesian estimator.

The fixed part of the multi-path GSC is nothing but a multi-channel deconvolution problem. Although we might be able to apply any kind of multi-channel deconvolution scheme [6, 7, 92, 93] for this fixed part, in the subsequent chapter we propose a blind multi-channel RIR identification algorithm, which in fact exploits the structure of multi-path GSC. Note that if we can estimate the channel in a blind manner, we do not have to measure RIRs.

5.1.1 Results of simple simulation experiment

The multi-microphone Bayesian optimal estimate of any function $z(S)$ of the speech signal S can be obtained by applying multi-path GSC followed by a single-channel Bayesian estimator. In this section, we implement the two-step approach in a simple simulation environment to see how each step works in practice. First, to see the effect of applying beamforming itself, we assumed that there are two microphones, and that the channel responses from the source to the microphones are $h_1 = [1 \ 0 \ 0.2 \ 0]^T$ and $h_2 = [1 \ 0 \ 0.2 \ 0.1]^T$. The source speech signal is replaced by an impulse because we are now concentrating on the effect of the beamforming. Noise signals for each microphone are taken from real measurements inside a car. As a beamformer, we applied the multi-path GSC. As shown in Figure 5.2 (d), the channel effect is completely removed. Noise is somewhat suppressed: the beamformed output usually lies between the two microphone measurements, which usually correspond to noise suppression. With more microphones, noise would be better suppressed.

Second, we applied a real speech signal as the source: in this case, an example of the utterance “three” taken from the TIDIGITS corpus. As a Bayesian estimator, we applied the MMSE-logSA estimator after multi-path beamforming. Figures 5.3 and 5.4 show the enhancement result. After beamforming, we can confirm that we get a slight signal-to-noise ratio (SNR) improvement (c), and after the Bayesian estimator we get much better enhancement of speech (d). In order to demonstrate that a two-channel Bayesian estimator is better than a one-channel Bayesian estimator, (e) shows the result of one-channel MMSE-logSA estimation (applied to the signal re-

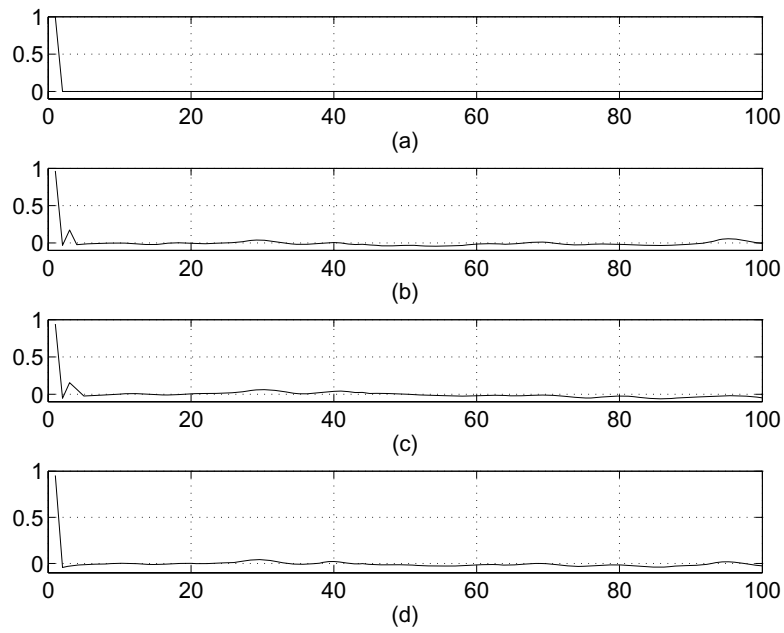


Figure 5.2: Results of applying multi-path GSC for the task of deconvolving the channel response from an impulse input: (a) original impulse input, (b) measured signal at microphone #1, (c) measured signal at microphone #2, (d) beamformed output (channel response is entirely removed and noise is ameliorated).

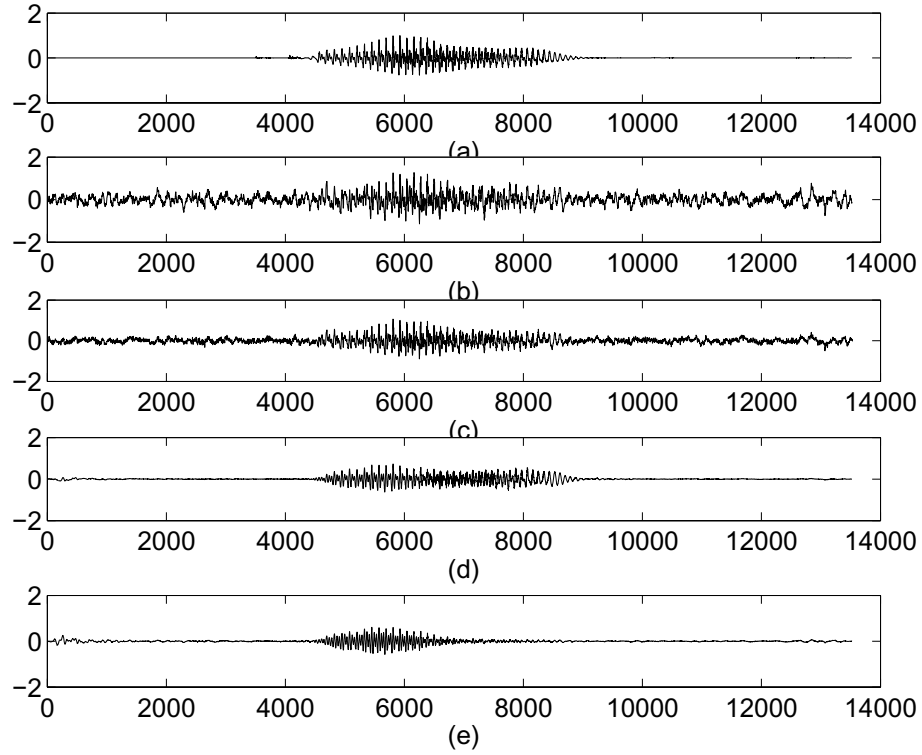


Figure 5.3: Time response of utterance “three”: (a) original speech signal taken from the TIDIGITS corpus, (b) measured speech at microphone #1, (c) beamformed output, (d) result of the two-step, beamform + estimate procedure, using an MMSE-logSA estimator, (e) result of applying the MMSE-logSA to measured speech at microphone #1.

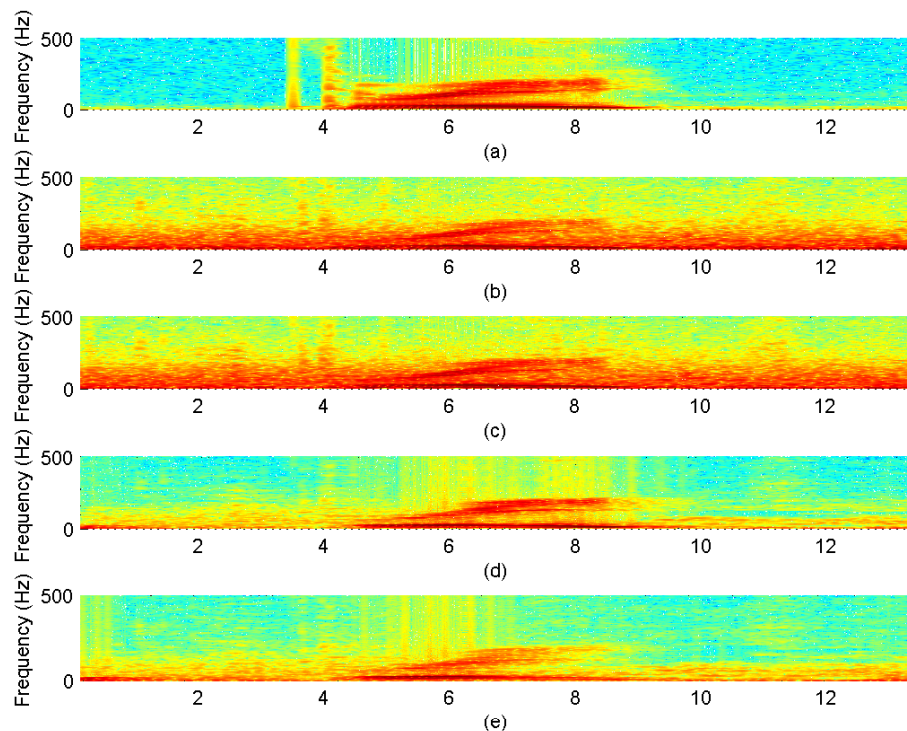


Figure 5.4: Spectrogram of utterance “three”: (a) original speech signal taken from the TIDIGITS corpus, (b) measured speech at microphone #1, (c) beamformed output, (d) result of the two-step, beamform + estimate procedure, using an MMSE-logSA estimator, (e) result of applying the MMSE-logSA to measured speech at microphone #1.

ceived by microphone #1). As we see by comparing (d) and (e), by just using one more microphone we can get a significantly better enhancement result. Note that SNR of the late part of the original utterance “three” (b) is somewhat less than the SNR of the early part. Beamforming improves the SNR very slightly; this slight improvement is magnified by the MMSE-logSA estimator, whose performance is critically dependent on the SNR of measured data. The MMSE-STSA and MMSE-logSA estimators are effective above a threshold SNR and ineffective below that threshold [9, 48]; the beamformer in a multi-microphone optimal spectral estimator serves the function of improving SNR before application of the Bayesian estimator. In this section, we see simulation results that confirm channel response elimination and noise suppression. Both results are the essence of our optimal two-step approach.

5.1.2 Summary and discussion

This section showed how we can get an optimal multi-microphone estimate of any function of the source speech signal via a two-step approach in which a multi-path GSC beamformer is followed by any optimal classical estimator such as MMSE-logSA in the multi-path and additive noise environment. The sufficient statistic $T(\vec{Y})$ can be obtained by maximum likelihood estimation of S given $\vec{Y}$; therefore, beamforming can be interpreted as a maximum likelihood estimation procedure. In other words, in the first stage (beamforming), we get an estimate of S which is as accurate as possible in the maximum likelihood sense. Then in the second stage (Bayesian estimation), we refine the estimate using an *a priori* source and noise statistics.

The method proposed may be compared to previously published “beam-form + estimate” procedures. Since [12, 43, 96, 97] are based on the ideal delay assumption, their algorithms do not give optimal results if the reflections are large and cannot be ignored. Gannot and Cohen’s methods [98–100] are based on a realistic acoustic channel assumption. Therefore, their approach is close to our optimal two-step approach. However, they modified the definition of the acoustic channel: they optimally estimate the speech spectrum at a reference microphone rather than the original speech signal. In other words, the acoustic channel effect is not eliminated by this method.

There exist practical matters that should be considered more carefully be-

fore application of the method proposed in this section. First, we assume that we know the channel response and it does not change, but actually this may not be a realistic assumption. If we can track the channel response changes due to talker movement, we can still use our current two-step approach by updating the constraint matrix according to tracked acoustic channel information. However, note that all of the difficulties that plague the science of dereverberation [5, 7, 89, 101–103] are also practical problems for multi-path GSC. For example, the room acoustic response is usually non-minimum phase, so we should apply a “modeling delay” to $\vec{f}$ [92]. Also, the room response may change; adequate dereverberation of a changing room response is also an on-going research topic. In this section, we have demonstrated a theoretically optimal framework for multi-microphone multi-path Bayesian spectral estimation; the following sections will address the ways to fit practical applications into our proposed framework.

5.2 Iterative blind estimation of RIR based on multi-path GSC

In an acoustic echo cancellation scenario, if we know the source signal, we can adaptively estimate the channel response [94]. Because more correctly beamformed output more closely matches to the source signal, we might be able to use the beamformed output as an input to estimate the channel response from the output signal. Good channel estimation makes the beamformer based on multi-path GSC more accurate, and this again guarantees better channel estimation, where the multi-path GSC is different from the conventional GSC in the sense that it reflects the multi-path effect on the constraint part realized as the fixed beamformer (FBF). Until we can get a satisfactory channel estimation result, in other words, a satisfactory beamforming result (satisfactory deconvolution), we keep iterating this adaptive procedure with some reasonable channel constraint. The iterative procedure can be used for the multi-channel identification as well as the optimal beamforming. Even though we may not get perfect channel identification, this is still a useful scheme in multi-path GSC, because we might use the converged multichannel information as a coefficient vector for the FBF, rather than using a naive delay and sum beamformer as in the conventional GSC; and by leveraging

the converged channel we actually mitigate the inherent signal cancellation problem due to the reverberation.

To visualize the situation more or less in a simple and tractable way, we first show the convergence of a simplified version of the proposed scheme. The preliminary simulation test has been conducted to show how this concept works. The results of the simple preliminary simulation show that this method seems to achieve sufficient blind deconvolution at the output of FBF after enough iterations. We expand the proposed algorithm into the realistic environment in a car, and it converged into a direct-path-only channel response, which is reasonable in a small acoustic space like the inside of a car.

5.2.1 Introduction

The main purpose of this method is to introduce a realistic way of blindly finding unknown early distinct echo paths for the multi-path GSC algorithm given in Section 5.1. The multi-path GSC with the identification of the dominant echo paths can be a realistic sufficient statistic. This method is inspired by Hoshuyama et al.’s robust beamforming approach [104]. They tried to minimize the signal cancellation due to room reverberation by introducing a new adaptive blocking matrix, where the beamformed output of a fixed beamformer (FBF) was used as an input to the acoustic channel, and each delayed channel response was used as an output of the acoustic channel. The acoustic channel is modeled as an FIR filter, and the channel identification is performed. The roughly identified channel response is used for cancellation of the target signal in the lower part of GSC. Note that GSC often suffers from signal cancellation, due to room reverberation inadequately blocked by the blocking matrix [41, 86, 104]. Once we have RIR and construct an FBF W_0 in Figure 5.1, the blocking matrix C_n is constructed as the null space of the space spanned by the RIR. Therefore, it automatically acts like the blocking matrix introduced in Hoshuyama et al.’s approach. However, Hoshuyama et al. used a simple delay and sum beamformer (DSB) as the FBF; therefore, their FBF does not reflect the deconvolution property of the multi-path GSC. In their approach, there cannot be more than one iteration for the RIR identification, because they just used the GSC, in which FBF

should be made with DSB after finding the direction of arrival.

We can extend their approach by using the adaptively estimated channel information from the source to each microphone to make a better FBF. If we can get better channel estimation using this better FBF output, and if we keep iterating this procedure, eventually we can get maximally deconvolved output from the FBF and more complete blocking of the source signal at the output of the blocking. This can be called optimal beamforming, in the sense that the signal cancellation problem due to reverberation is maximally alleviated for the single reverberant speech case. With existing interfering speeches, identification of the dominant echo paths for every concurrent speech source is important as in Chapter 3. Once we have reliable estimate of them, we can deconvolve and separate speech sources with a linearly constrained minimum variance (LCMV) framework. The following is an example of two coexisting reverberant speech sources:

$$\arg \min_{\vec{W}} E[|\vec{W}^H \vec{Y}|^2] \quad \text{subject to} \quad \vec{W}^H [\vec{H}_t \quad \vec{H}_i] = [1 \quad 0], \quad (5.24)$$

where $\vec{H}_t$ and $\vec{H}_i$ are corresponding to $\vec{D}_t$ and $\vec{D}_i$, respectively, in (3.4), but they include the dominant echo paths as well as the direct path.

To make the situation more or less simple and tractable, we first prove the convergence of a simplified version of the proposed scheme. A simple experiment has been conducted to make sure that this concept works. The result of the simple simulation shows that this method achieves sufficient blind deconvolution at the output of FBF after enough iterations.

5.2.2 Problem formulation

The channel response estimation follows the optimization process below.

$$\begin{aligned} \hat{h}_i(t) = & \arg \min_{\hat{h}_i(t)} \|s(t) * (h_1(t) * w_1(t) + \\ & \cdots + h_N(t) * w_N(t)) * \hat{h}_i(t) - s(t) * h_i(t)\|^2, \end{aligned} \quad (5.25)$$

where $\vec{\hat{h}}_i$ is the estimated channel.

$$\vec{\hat{h}}_i = \arg \min_{\vec{\hat{h}}_i} \|\hat{\mathbf{C}} \vec{\hat{h}}_i - \vec{h}_i\|^2 = (\hat{\mathbf{C}}^T \hat{\mathbf{C}})^{-1} \hat{\mathbf{C}}^T \vec{h}_i, \quad (5.26)$$

where $\hat{\mathbf{C}}$ is the convolution matrix obtained with the beamformed output of impulse responses as input for the beamformer. Ideally if $\hat{\mathbf{C}} = \mathbf{I}$, in other words if the FBF with RIR as input produces perfectly deconvolved output, then we can obtain the real channel response. As in (5.26), the estimated channel responses are obtained in the optimal sense of least squares by forcing them to have zero values, except for the RIR value at the estimated time stamps of each dominant reflection. This is similar to the concept of acoustic echo cancellation except for the constraint part.

5.2.3 Algorithm

The proposed algorithm is introduced below step by step. In this section, we just care about the deconvolution, because the noise suppression after the deconvolution is quite straightforward. Based on the assumption that we know the time stamps for the reflections, we can successfully estimate the magnitude of the reflections using the following algorithm, where the number of microphones is N . The way of estimating the time position for the reflections is discussed in the subsequent section.

1. Initialize the magnitude of the time location of reflections with ϵ and 0 otherwise.
2. Perform multi-path GSC to get output $\hat{s}$ and update $\hat{h}_1(r)$ with the solution of (5.25).
3. Set the updated magnitude $\hat{h}_1$ of the time location to be 0 if those are not the position for the designated reflections.
4. Iterate 2 and 3 until there is no more significant change in the magnitude of the reflection.

If you follow the first iteration, you will get the first update of $\hat{h}_1(r) \approx h_1(r) - \frac{1}{N}(h_1(r) - \epsilon + h_2(r) + \dots + h_N(r))$; and if this number is bigger than ϵ , it will be updated until there is no change of $\hat{h}_1(r)$ and this will be

$$\hat{h}_1(r) = h_1(r) - \frac{1}{N-1}(h_2(r) + \dots + h_N(r)). \quad (5.27)$$

In the early part of the RIR, reflections are infrequent; therefore, typically $h_2(r) = \dots = h_N(r) = 0$ or at least $h_2(r), \dots, h_N(r) \ll h_1(r)$ and there-

fore $\hat{h}_1(r) \cong h_1(r)$ in (5.27). Even if there exists noise, because we take a mean of iteration measurements, we can regard it as zero since we can easily assume that the noise process is zero mean. However, in reality, because of low-pass filtering for sampling and other low-pass filtering effects acting on the reflections, the response will not contain perfect impulses, and this imperfection will produce some errors. Therefore, reflections with similar direction of arrival (DOA) will not be estimated correctly using this scheme. This intuitively makes sense; the benefits of using beamforming are reduced when the direction of interference is in the DOA of the source.

Figure 5.5 shows the converged result of a two-channel measurement with a seven-reflection RIR, including one negative component and one merged component in the RIR as in the following:

$$x1 = [1 \ 0 \ 0 \ 0 \ 0.5 \ 0 \ 0 \ 0.4 \ 0 \ 0.05 \ 0.3 \ 0 \ 0 \ -0.1 \ 0.09 \ 0 \ 0 \ 0.04]^T, \quad (5.28)$$

and

$$x2 = [1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0.5 \ 0 \ 0.45 \ 0 \ 0 \ 0.3 \ -0.1 \ 0 \ 0 \ 0.09 \ 0.04 \ 0]^T. \quad (5.29)$$

The first three reflection time stamps are assumed to be known and others are set as zero. We can confirm that by having correct time stamps for some, but not all, early reflections, we can estimate the channel responses up to the given reflection points, and at the same time, the deconvolution can be performed up to the reflection points. These results are very promising because we can track and deconvolve dominant early reflections, which are usually sparse enough and deterministically treatable within a reasonably small time frame where we can assume that the early responses are time invariant.

5.2.4 Algorithm with reflection time stamp estimation

In this section, we propose a heuristic dominant reflection time stamp estimation together with the proposed algorithm. The algorithm is as follows.

1. Initially we choose DSB as a first FBF and perform a normalized least mean squared algorithm to estimate the RIR as FIR coefficients using the output of DSB.

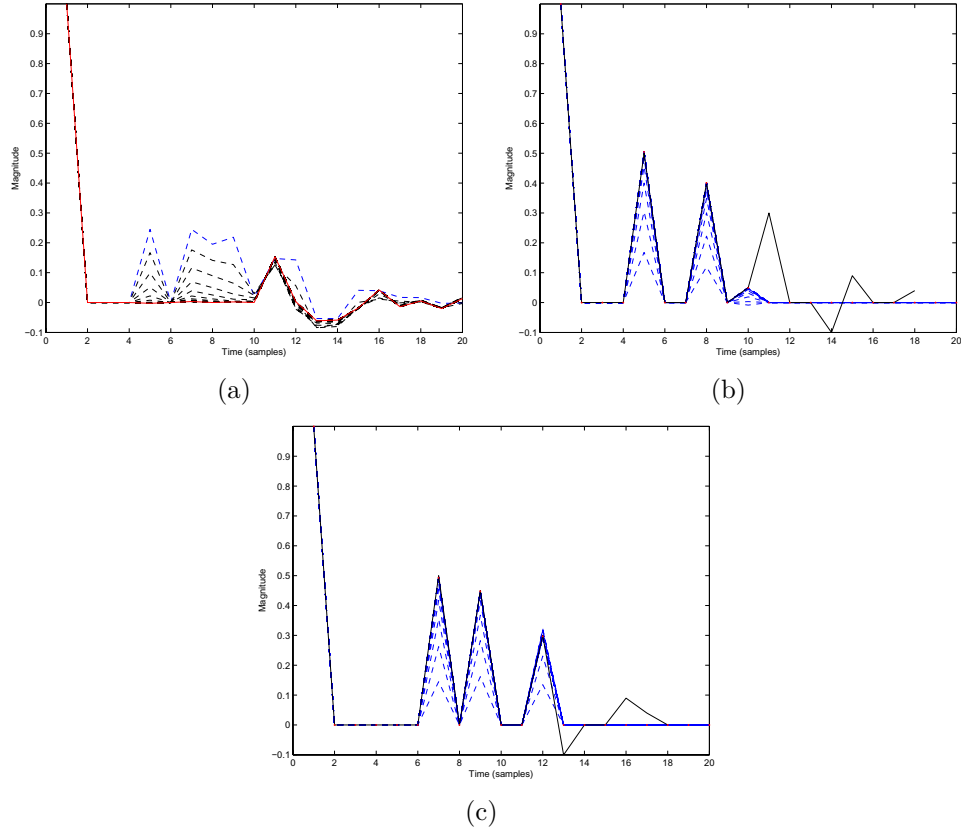


Figure 5.5: (a) FBF output: blue dotted line is DSB output, black dotted lines are updated FBF output, red line is final FBF output after 20 iterations. Updated FBF output produces more impulse-like output by eliminating the effect of the designated reflections, in other words, more deconvolved output. (b) Estimated channel h_1 . (c) Estimated channel h_2 : red dots show the converged channel response after 20 iterations and the blue dotted lines are updated responses. The black line is for original RIR. The designated channel responses are almost perfectly identified.

2. Select the time stamps in which the estimated RIR magnitudes are above a predefined threshold, which determines the significance level of the reflection.
3. Perform the proposed algorithm.
4. Iterate 2 and 3 enough until convergence.

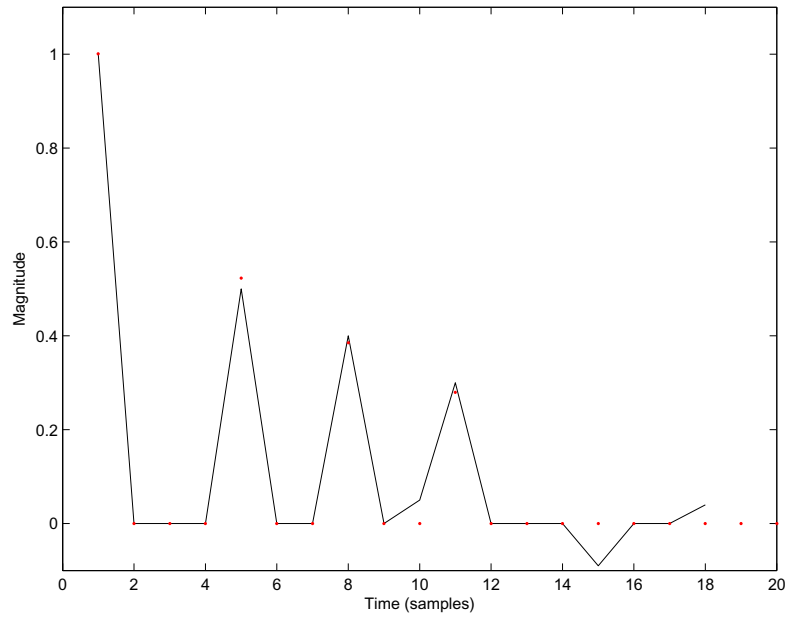
Figure 5.6 shows the converged result, where the simulated output of two channels has been obtained by convolving the channel response with a white Gaussian noise source and the threshold has been set as “0.08”. Note that most of the significant reflection points above the threshold can be estimated almost correctly.

5.2.5 Experiment with real car data

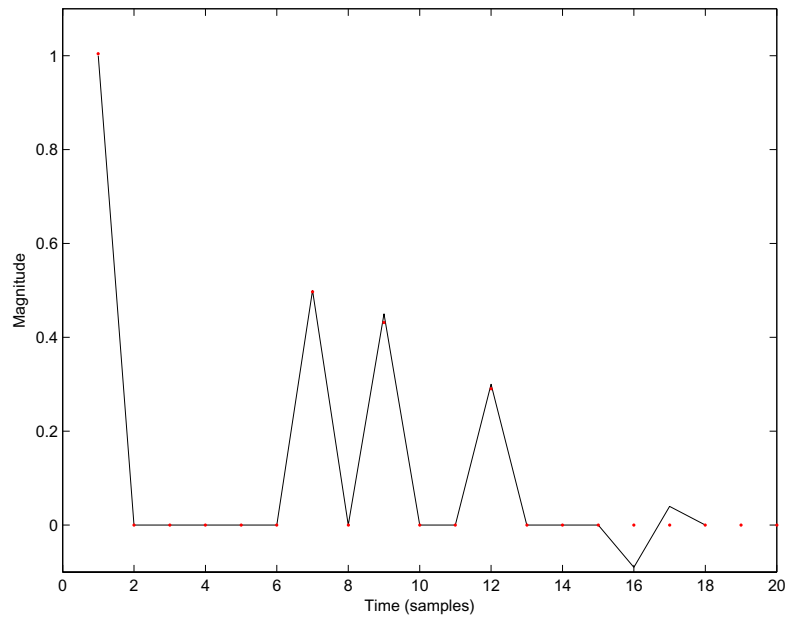
In this section, we test the proposed algorithm using the real multi-channel sources measured in cars. The whole procedure for testing can be summarized as following.

1. Interchannel delay is estimated using the GCC-PHAT method [35] and adjusting the delay to formulate DSB.
2. Perform the proposed algorithm.

Figure 5.7 shows the two-channel identification results with no estimated distinctive reflection except direct path. using one of the single digit utterances in the AVICAR database [105]. A possible explanation for this result is that the space inside a car is too small to have distinctive reflections which could be sparsely separable using the proposed algorithm. However, because this fact also means that there are no significantly correlated reflections in the original signals with the beamformed output using the direct path information (DSB), we can avoid the signal canceling problem when we use the GSC structure with DSB as FBF. Optimal signal enhancement result and isolated digit recognition results with conventional GSC have been reported in [85].

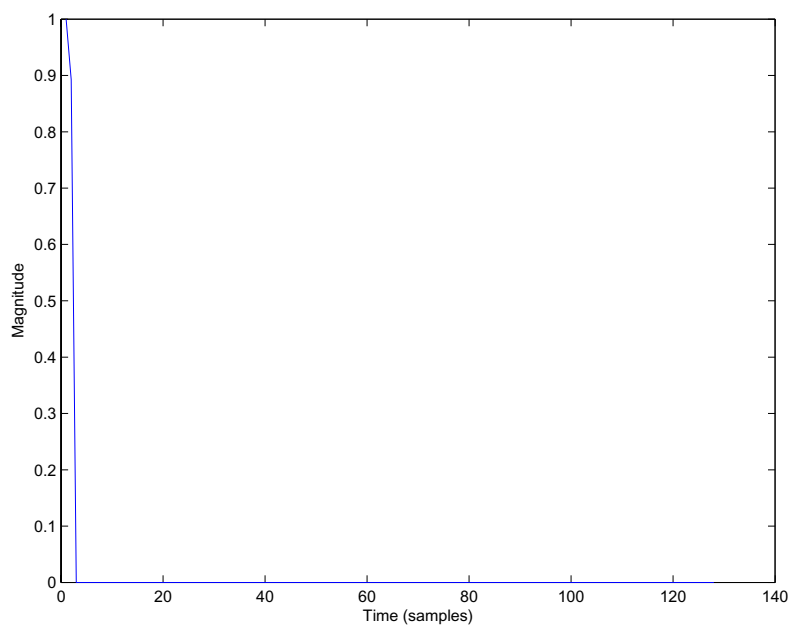


(a)

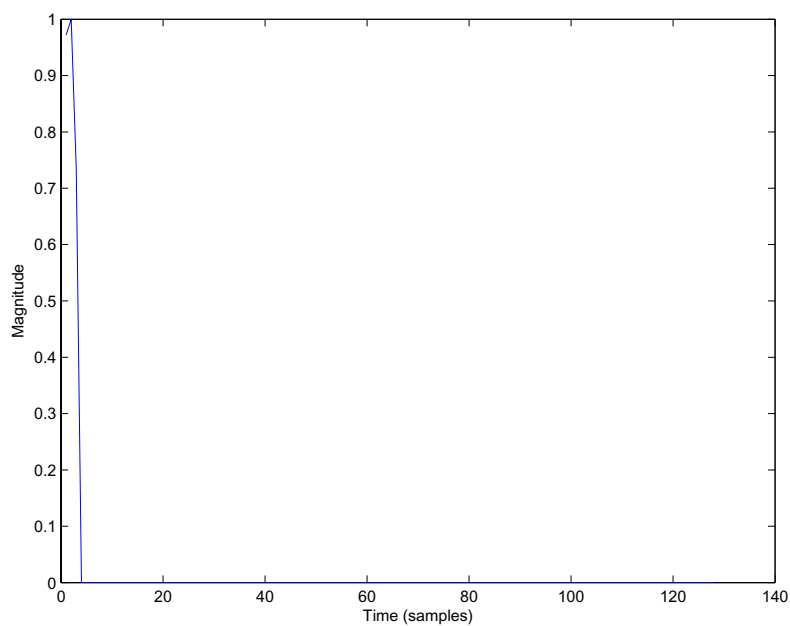


(b)

Figure 5.6: (a) Estimated channel h_1 . (b) Estimated channel h_2 : red dots show the converged channel response after 20 iteration and the black line is for original RIR. The designated channel responses above the predefined threshold are almost correctly identified.



(a)



(b)

Figure 5.7: (a) Estimated channel h_1 . (b) Estimated channel h_2 .

5.2.6 Summary and discussion

In this section, we propose the multi-path GSC based blind channel identification method, which can be plugged in as a realistic replacement for the sufficient statistic for optimal speech enhancement. The simulation with artificially generated sparse channels shows that the proposed algorithm can converge into the original channel responses above a predefined significance threshold. A channel estimation experiment with real data measured in a car results in the fact that there exists no distinctive significant reflection which can contribute to the signal cancellation when we use the GSC structure.

5.3 Sequential MVDR beamforming and postfiltering

Any Bayesian optimal one-channel estimator for speech enhancement can be generalized to the multichannel case as a sequentially constructed minimum variance distortionless response (MVDR) beamformer followed by an optimal one-channel postfilter. We present experimental results using the minimum mean squared error log-spectral amplitude (MMSE-logSA) optimality criterion, applied to a statistical model with a simplified channel but realistic inter-microphone noise coherence. The word error rate in the audio-visual speech in a car (AVICAR) corpus (moving car, windows open) is reduced from 18% to 9%.

5.3.1 Fisher-Neyman factorization with the time structure of channel responses

If we assume that the channel responses have finite impulse responses

$$H = d + r_1 + r_2 \cdots + r_M, \quad (5.30)$$

where d is the direction or arrival vector, $r_i = [r_{i1} \ r_{i2} \ \cdots \ r_{iN}]$ for $i = 1, 2, \dots, M$ is the delay response from i^{th} reflection to microphones, and M

is the number of echoes, then

$$\begin{aligned} T(Y) &= \rho \cdot \frac{H^H \mathbf{R}_n^{-1} Y}{H^H \mathbf{R}_n^{-1} H} \\ &= \rho \cdot \frac{d^H \mathbf{R}_n^{-1} Y + r_1^H \mathbf{R}_n^{-1} Y + \dots + r_M^H \mathbf{R}_n^{-1} Y}{H^H \mathbf{R}_n^{-1} H}, \end{aligned} \quad (5.31)$$

and if $\rho(\cdot) = \frac{H^H \mathbf{R}_n^{-1} H}{d^H \mathbf{R}_n^{-1} d}(\cdot)$ then

$$\begin{aligned} T(Y) &= \frac{H^H \mathbf{R}_n^{-1} Y}{d^H \mathbf{R}_n^{-1} d} \\ &= \frac{d^H \mathbf{R}_n^{-1} Y}{d^H \mathbf{R}_n^{-1} d} + \frac{r_1^H \mathbf{R}_n^{-1} r_1}{d^H \mathbf{R}_n^{-1} d} \cdot \frac{r_1^H \mathbf{R}_n^{-1} Y}{r_1^H \mathbf{R}_n^{-1} r_1} \\ &\quad \dots + \frac{r_M^H \mathbf{R}_n^{-1} r_M}{d^H \mathbf{R}_n^{-1} d} \cdot \frac{r_M^H \mathbf{R}_n^{-1} Y}{r_M^H \mathbf{R}_n^{-1} r_M}. \end{aligned} \quad (5.32)$$

The number of dominant reflections, M , is dependent on how reverberant the room is. In a large room, the acoustic wave travels quite far before any reflection reaches a microphone; therefore, the relative response $\frac{r_i^H \mathbf{R}_n^{-1} r_i}{d^H \mathbf{R}_n^{-1} d}$ is quite small. As can be seen in (5.32), this sufficient statistic $T(Y)$ can be interpreted as a sequentially constructed ML beamformer where each sequential MVDR beamformer is normalized by the direct term. If we assume that the direct and a few dominant reflections can be found, $T(Y)$ may be approximated without knowledge of the whole H . Because we can get the one-channel summary using the sequentially constructed MVDR beamformer without loss of the desired source information present in a multi-channel signal, we can simply apply any classical MMSE speech enhancement technique based on the speech *a priori* PDF. From the previously stated sufficient statistics, we can get a simpler *a priori* PDF $p(S|Y)$ as shown in (5.33).

$$p(S|Y) = p(S|T(Y)) = \frac{p(T(Y)|S)p(S)}{\int p(T(Y)|S)p(S)dS}, \quad (5.33)$$

where $T(Y)$ can be anything with the form of (5.31) including the sequentially constructed MVDR beamformer as shown in (5.32). From this beamformer and postfiltering as shown in Figure 5.8, we can derive any desired multi-channel speech estimator. The MMSE spectral component estimator (the

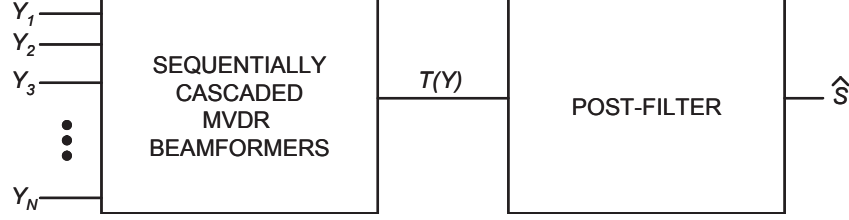


Figure 5.8: Block diagram of optimal multi-microphone speech enhancement.

Wiener filter) based on multi-channel input is given by

$$\begin{aligned}\hat{S}_{MMSE} &= E\{S|Y\} = E\{S|T(Y)\} \\ &= \rho^{-1} \left(\left[\frac{R_s}{R_s + R_{n_{out}}} \right] \cdot T(Y) \right),\end{aligned}\quad (5.34)$$

where $R_{n_{out}} = \frac{\rho^2}{H^H R_n^{-1} H}$. Wiener postfiltering of an MVDR output has been previously derived using the matrix inversion lemma [43].

There has been no theoretical basis for generalized postfiltering after an MVDR beamformer. The MMSE-STSA postfiltering of the output signal of the sequentially cascaded MVDR beamformer provides an optimal MMSE solution for the spectral amplitude enhancement, which can also be formulated as (5.35).

$$\begin{aligned}|\hat{S}|_{MMSE-STSA} &= E\{|S||Y\} = E\{|S||T(Y)\} \\ &= \rho^{-1} \left(\Gamma(1.5) \frac{\sqrt{v}}{\gamma} \exp\left(-\frac{v}{2}\right) \right. \\ &\quad \cdot \left[(1+v) I_0\left(\frac{v}{2}\right) + v I_1\left(\frac{v}{2}\right) \right] \\ &\quad \cdot |T(Y)|, \end{aligned}\quad (5.35)$$

where $\Gamma\{\cdot\}$ denotes the gamma function, with $\Gamma(1.5) = \frac{\sqrt{\pi}}{2}$, I_0 and I_1 denote the modified Bessel functions of zero and first order, respectively. We define v by

$$v = \frac{\xi}{1 + \xi} \gamma, \quad (5.36)$$

where ξ and γ are called the *a priori* and *a posteriori* signal-to-noise ratio (SNR) of $T(Y)$, respectively, as in the Section 2.2. Similarly the multi-

microphone MMSE-logSA estimator can be described as

$$\begin{aligned}
\widehat{|S|}_{MMSE-logSA} &= E\{\log(|S|)|Y\} = E\{\log(|S|)|T(Y)\} \\
&= \rho^{-1}(\frac{\xi}{1+\xi} \exp\{\frac{1}{2} \int_v^\infty \frac{e^{-t}}{t} dt\}) \\
&\quad \cdot |T(Y)|,
\end{aligned} \tag{5.37}$$

As a postfilter for spectral amplitude enhancement, the MAP estimator (5.38), (5.39) of Wolfe and Godsill [106] can be used.

$$\begin{aligned}
\widehat{|S|}_{MAP} &= \arg \max_{|S|} p(|S||Y) \\
&= \arg \max_{|S|} p(T(Y)||S|)p(|S|).
\end{aligned} \tag{5.38}$$

$$\widehat{|S|}_{MAP} = \rho^{-1}(\frac{\xi + \sqrt{\xi^2 + (1+\xi)\frac{\xi}{\gamma}}}{2(1+\xi)} \cdot |T(Y)|). \tag{5.39}$$

5.3.2 Experiment

We applied MVDR beamforming ($M = 0$ in (5.32)) followed by MMSE-logSA postfiltering and conducted a single-digit recognition test using the AVICAR corpus [105]. We chose $M = 0$ because the reflections are assumed to be weak relative to the direct sound in the noisy situation. If we consider more reflections, we may get slightly better performance. To do the digit recognition test, we first trained 11 HMMs (from “zero” to “nine”, “oh”, which is a differently pronounced model for “zero”, and “silence”, which is for the noise-only period) using data from 25 male speakers and conducted speech recognition tests using the other twenty five. We used mel-frequency cepstral coefficients (MFCC), energy, delta coefficients, and acceleration coefficients, for a total feature vector of size 39. We then used the HTK toolkit [20] to build an isolated digit recognition system, and test recognition accuracy. The number of states per word was 8, and the number of Gaussians per mixture was 1 to 14. Secondly, we used 55 talkers from TIDIGITS’ training corpus for training the 11 HMMs, and tested on the AVICAR corpus and enhanced AVICAR corpus.

The AVICAR corpus [105] is data recorded in a real car environment using a multi-sensory array consisting of eight microphones on the sun visor and

Table 5.1: Digit recognition accuracy(%) using half of the AVICAR corpus or enhanced AVICAR corpus for training

Training set	Test set	Number of Gaussians per Mixture							
		1	2	4	6	8	10	12	14
Original	Original	81.09	82.14	82.98	83.47	83.42	83.39	83.69	83.85
Enhanced	Enhanced	89.74	90.98	90.88	90.31	90.79	91.07	90.88	90.79

Table 5.2: Digit recognition accuracy(%) using TIDIGITS training corpus for training

Training set	Test set	Number of Gaussians per Mixture							
		1	2	4	6	8	10	12	14
TIDIGIT	Original	30.49	37.57	33.84	30.16	30.13	29.67	30.84	32.35
TIDIGIT	Enhanced	62.3	58.88	55.94	53.85	54.61	52.14	52.99	53.09

four video cameras on the dashboard. The script for the corpus consists of four categories: isolated digits, isolated letters, phone numbers, and phonetically balanced sentences. Speakers from various language backgrounds are included, 50 male and 50 female. Each script has five different noise conditions: idling (IDL), driving at 35 mph with windows up (35U) and down (35D), and driving at 55 mph with windows up (55U) and down (55D).

We chose the LMS-GSC as an MVDR beamformer with seven filter taps per channel [91]. We chose this adaptive filtering algorithm because it can update the change of noise covariance in real time, and this real time update is thought to be appropriate for in-car recordings like the AVICAR corpus. Before applying GSC, we obtained the delays between microphones to align direction of arrival to the source location using the SRP-PHAT algorithm [43]. The SRP-PHAT is known to be a robust localizing algorithm even in reverberant rooms. As a postfiltering method, we chose the MMSE-logSA algorithm, because the cepstral features used in speech recognition are nearly a linear transform of the log spectrum. Figures 5.9 and 5.10 show the time response and spectrogram of utterance “three”: original (upper), after MVDR beamforming (middle), and MMSE-logSA postfiltering after MVDR (lower). The proposed algorithm shows well-enhanced results even though the noisy speech data was recorded in a real moving car with windows down at 55 mph. Informal listening found a remarkable enhancement in signal quality.

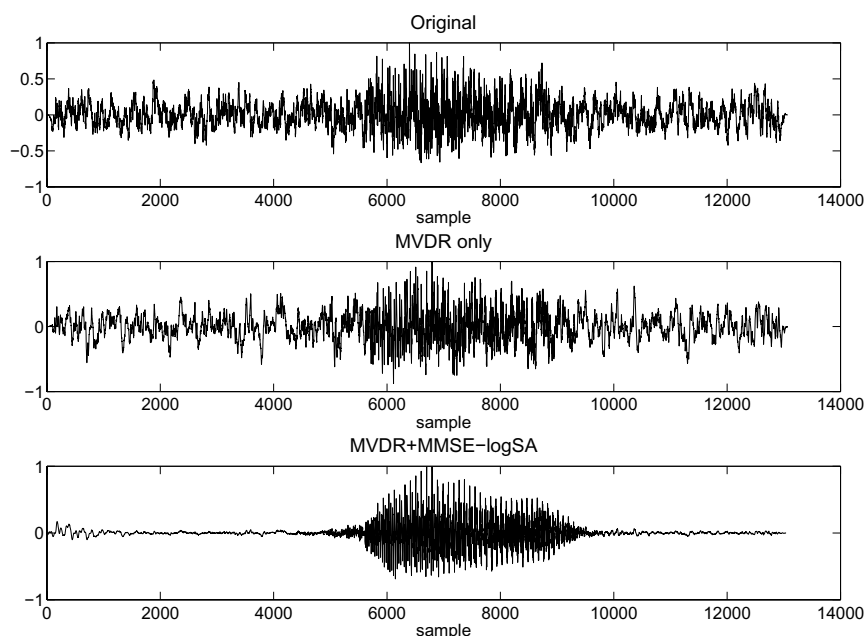


Figure 5.9: Time response of utterance “three” in the 55D condition: original (upper), after MVDR beamforming (middle), and after MVDR beamforming and MMSE-logSA postfiltering (lower).

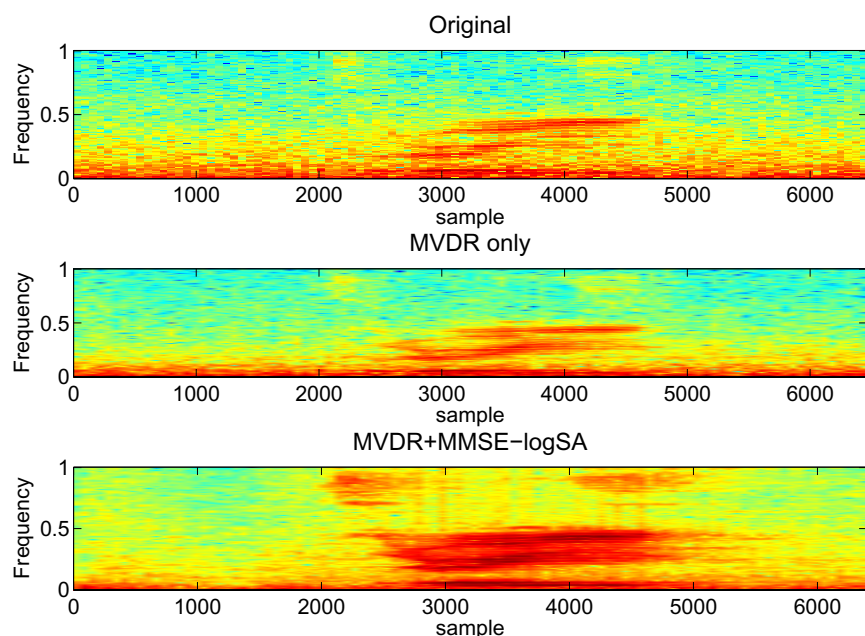


Figure 5.10: Spectrogram of utterance “three” in the 55D condition: original (upper), after MVDR beamforming (middle), and after MVDR beamforming and MMSE-logSA postfiltering (lower).

Table 5.1 contains the results when HMMs are trained using half of the speakers in the AVICAR corpus and tested on the other half. Both training set and test set include data from all five noise conditions in equal proportions. Table 5.2 contains the results when TIDIGITS’ training corpus is used as the training corpus, and the models are tested on half of the AVICAR corpus, or of the enhanced AVICAR corpus, respectively. In all of the experiments, we can confirm the success of our proposed approach. The proposed approach gives better recognition accuracy than when the HMMs were trained using speech from five different noise conditions, and tested even on the same five different noise conditions. This result can be interpreted to mean that this enhancing algorithm makes the acoustic characteristics of the enhanced training and test data more similar.

5.4 Probabilistic RIR sufficient statistic approach

In this section, we model the frequency response of RIR above the Schroeder frequency given in (5.40) as a random variable with Gaussian PDF. We address an expectation-maximization algorithm based on the statistical models of the RIR.

5.4.1 Modeling of RIR

From a classical room acoustic theory for a closed space, we introduce a model of the room impulse response (RIR) from one point to the other in a closed space, e.g. from a driver to the hands-free microphone for telephone conversation with his family in a car. This modeling will be done in the low-frequency range and in the high-frequency range separately, because the statistical characteristics of the room response in the high-frequency range are different from the statistical characteristics of the room response in the low-frequency range. In the high-frequency range, the room response can be assumed to have a Gaussian PDF due to sufficient overlapping of modes; by the central limit theorem, the sum of modal responses with an independent uniformly distributed phase approaches a Gaussian random variable [107]. In the low-frequency range, modes are discrete and easily measurable; therefore, we model the low-frequency range using a deterministic frequency response.

To get the crossover frequency, we use Schroeder's frequency formula. Equation (5.40) specifies the frequency above which the acoustic modes overlap too much in frequency to be discretely modeled:

$$f_s = c\sqrt{\frac{6}{A}} \approx 2000\sqrt{\frac{T}{V}}, \quad (5.40)$$

where c is the speed of the sound, A is the absorption area of the room (m^2), Sabine's reverberation time $T = 0.163V/A$ (sec), and V (m^3) is the volume of the space. In actual computation, we use the measured reverberation time. For example, a typical automobile space with volume of 5 m^3 and reverberation time of 0.2 s will have a crossover frequency of around 400 Hz .

5.4.2 One-channel optimal estimators based on probabilistic room response

Above the Schroeder frequency, we can assume that $\vec{H}$ is a random variable with Gaussian PDF.

$$\vec{H}(\omega) \sim N(\vec{0}, \Sigma_{\vec{H}(\omega)}). \quad (5.41)$$

First of all, the single-channel case is considered; discussion is then extended to the multi-channel case. In this case, optimal estimators should be based on the pdf of H . Because many modes overlap at each frequency, by the central limit theorem, we can assume that H has a complex Gaussian pdf, and therefore $|H|$ is a Rayleigh random variable. The MMSE estimator is

$$\begin{aligned} & |\hat{S}|_{MMSE-STSA,prob} \\ &= E[|S| | Y] = E[E[|S| | Y, H]] \\ &= \int_0^\infty |\hat{S}|_{MMSE-STSA,det} p(|H|) d|H|, \end{aligned} \quad (5.42)$$

where $|\hat{S}|_{MMSE-STSA,det}$ is the spectral amplitude estimator of Ephraim and Malah [9]. Because we have been unable to analytically integrate the equation above, a novel ML estimator followed by the conventional optimal filter is proposed.

ML estimator as a sufficient statistic

The ML estimate of S given H , $S_{ML|H}$, is a sufficient statistic for optimal estimation of any function $f(S)$:

$$E[f(S) | Y, H] = E[f(S) | S_{ML|H}], \quad (5.43)$$

where

$$S_{ML|H} = (H^* H)^{-1} H^* Y. \quad (5.44)$$

In our formulation, H is a random variable; therefore, $S_{ML|H}$ is also a random variable. We propose a two-step procedure in which S , and H are jointly estimated according to

$$(S_{ML}, H_{ML}) = \arg \max p(Y | S, H, \lambda_N) \quad (5.45)$$

and then MMSE estimates of functions $f(S)$ are computed. Note that by including $\lambda_{N,ML}$ in (5.45), the noise variance λ_N can also be jointly estimated as in (5.53), but we can also simply assume that we can obtain λ_N without this ML estimation as in the previous methods [9, 48].

EM algorithm

The expectation-maximization algorithm (EM) is an iterative method to get the ML estimates of parameters [108]. Hidden parameters may be defined to include the expectation and variance of H . Define (Y, H) to have pdf

$$p(Y, H | S, \lambda_N) = p(Y | H, S, \lambda_N) p(H). \quad (5.46)$$

Assume that the noise N has a zero mean Gaussian pdf of variance λ_N , and that H is a zero mean Gaussian with variance λ_H . The objective function has the form below.

$$Q(S, S^{(i-1)}) = E[\ln p(Y, H | S, \lambda_N) | Y, S^{(i-1)}]. \quad (5.47)$$

From (5.46),

$$\begin{aligned}
& Q(S, S^{(i-1)}) \\
&= E \left[\ln p(Y|H, S, \lambda_N) \mid Y, S^{(i-1)} \right] \\
&\sim -\ln \lambda_N - \frac{|Y|^2}{\lambda_N} + \frac{E[H^*|Y, S^{(i-1)}, \lambda_N^{(i-1)}](S)^* Y}{\lambda_N} \\
&+ \frac{E[H|Y, S^{(i-1)}, \lambda_N^{(i-1)}]SY^*}{\lambda_N} - \frac{E[H^*H|Y, S^{(i-1)}, \lambda_N^{(i-1)}]|S|^2}{\lambda_N},
\end{aligned} \tag{5.48}$$

because $p(H)$ has nothing to do with S . The EM algorithm has two steps: the “Expectation step” computes (5.48) and the “Maximization step” finds S and possibly λ_N to maximize it.

<Expectation step>

$$\begin{aligned}
& E[H^*|Y, S^{(i-1)}] \\
&= E[H^*] + \frac{\text{Cov}(H^*, Y^*|S^{(i-1)})}{\text{Cov}(Y^*|S^{(i-1)})}(Y^* - E[Y^*]) \\
&= \frac{\lambda_H S^{(i-1)} Y^*}{|S^{(i-1)}|^2 \lambda_H + \lambda_N},
\end{aligned} \tag{5.49}$$

where this is obtained from $Y = HX^{i-1} + N^{i-1}$.

$$\text{Cov}[H^*|Y^*, S^{(i-1)}] = \frac{\lambda_H \lambda_N}{|S^{(i-1)}|^2 \lambda_H + \lambda_N}, \tag{5.50}$$

and

$$\begin{aligned}
E[H^*H|Y, S^{(i-1)}] &= \text{Cov}[H^*|Y^*, S^{(i-1)}] \\
&+ |E[H^*|Y^*, S^{(i-1)}]|^2.
\end{aligned} \tag{5.51}$$

<Maximization step>

$$S^i = \frac{E[H^*|Y^*, S^{(i-1)}]}{E[H^*H|Y^*, S^{(i-1)}]} Y, \tag{5.52}$$

where this is obtained from $\frac{\partial Q}{\partial S} = 0$, and note that S is complex. The update

formulas for $\lambda_{N,ML}$ may be similarly derived.

$$\begin{aligned}\lambda_N^i &= |Y|^2 - E[H^*|Y, S^{(i-1)}, \lambda_N^{(i-1)}](S^{(i-1)})^*Y - E[H|Y, S^{(i-1)}, \lambda_N^{(i-1)}]S^{i-1}Y^* \\ &+ E[H^*H|Y, S^{(i-1)}, \lambda_N^{(i-1)}]|S^{i-1}|^2.\end{aligned}\quad (5.53)$$

Multiple measurements around a target position may be modeled as independent and identically distributed (IID). In this case, the Q function is

$$Q(S, S^{(i-1)}) = \sum_{j=1}^N E \left[\ln p(Y_j, H_j | S, \lambda_N) \mid Y_j, S^{(i-1)} \right], \quad (5.54)$$

and the maximization step for S^i is (5.55).

$$S^i = \frac{\sum_{j=1}^N E[H_j^*|Y_j^*, S^{(i-1)}]Y_j}{\sum_{j=1}^N E[H_j^*H_j|Y_j^*, S^{(i-1)}]}. \quad (5.55)$$

The multi-channel case without the assumption of IID will be discussed in the following section.

Experimental evaluation

In this experimental evaluation, we focus on the newly derived EM algorithm for the sufficient statistic S_{ML} . To verify the proposed EM algorithm, 50 room impulse responses (RIRs) were simulated with randomly chosen positions for the source, and 50 receivers using the conventionally used room simulation method “image-method” [109, 110]. A shoebox type of acoustic space has been used, the dimension of which are $6.25 \times 3.75 \times 2.5$ m and the volume of which is 58.59 m^3 . The speed of sound, the average absorption coefficient, and the reverberation time are set to 343 m/s, 0.45, and 0.25, respectively. The source was one copy of the speech waveform “three,” extracted from TIDIGITS [44]. Figure 5.11(a) shows the source spectrum (an utterance of the word “three”) contaminated only by additive noise, where the signal-to-noise ratio (SNR) was about 12 dB. Fifty measurements have been simulated not only by convolving the source signal with the 50 RIRs but also by adding noise. Figure 5.11(b) shows one of the measured responses.

At each measurement location, the sufficient statistic S_{ML} can be estimated using EM. In this experiment, we use multiple measurements to ob-

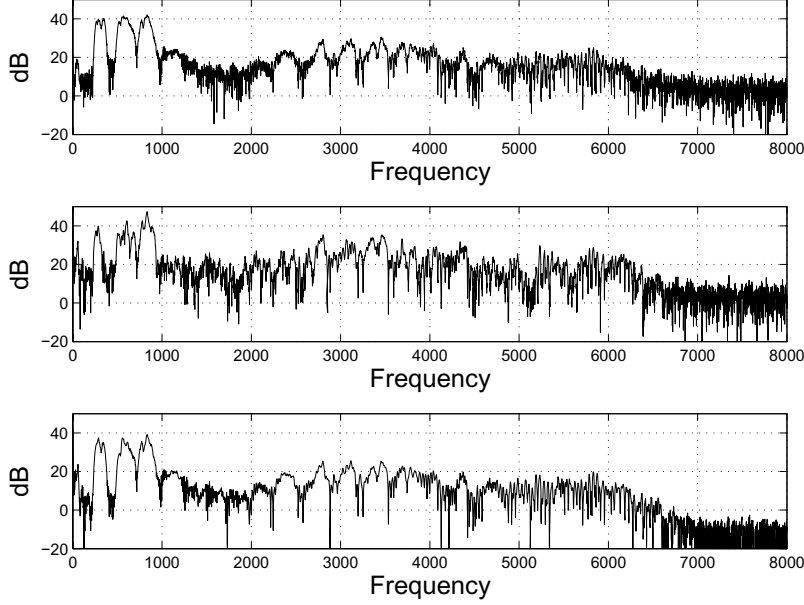


Figure 5.11: EM-ML estimate result: (top) source spectrum (an utterance of the word “three”) only contaminated by additive noise: SNR is about 12 dB, (middle) one of the measurements: source spectrum contaminated by RIR as well as additive noise, (bottom) EM-ML estimated spectrum: sufficient statistic S_{ML} .

tain S_{ML} more accurately. Figure 5.11(c) shows the source spectrum $|\hat{S}|_{ML}$ estimated using the proposed EM-based algorithm. Above about 200 Hz, the estimated source spectrum above the noise floor is almost the same as the source spectrum above the noise floor. In other words, we could successfully eliminate the effect of the room responses in the maximum likelihood sense, and this result is the sufficient statistic for the classical MMSE speech enhancement algorithms. Below 200 Hz, the estimation starts to break down, apparently because the room response is not a Gaussian random variable at lower frequencies. Although the range between 100 and 200 Hz is above the Schroeder frequency, it seems that the modes of the room response do not overlap thickly enough in this frequency range for the response to approach a Gaussian distribution.

Discussion

In this section, MMSE optimal estimators of a signal which has been contaminated by convolutive noise as well as additive noise are defined. In

the high-frequency range, the MMSE spectral estimator is expressed as the conditional expectation of $|S|$ given knowledge of the maximum likelihood estimate, $S_{ML|H}$. Rather than integrating over the PDF of the random variable $S_{ML|H}$, we approximate $p(S_{ML|H})$ using a point distribution centered at the joint maximum likelihood estimates S_{ML} and H_{ML} ; these joint maximum likelihood estimates are computed using EM.

5.4.3 Maximum likelihood estimator using multiple microphones

Now we expand the previous discussion into the original multi-channel situation of (1.9) and (1.10). We have seen that the MLE of S has the same form as the sufficient statistic of (1.12). However, if we don't know the transfer function vector $\vec{H}$ and the covariance matrix $\mathbf{R}_n$, the MLE S_{ML} is itself a random variable. Having the EM algorithm in mind, we define a complete data vector $Z \triangleq (\vec{Y}, \vec{H}, \mathbf{R}_n)$.

$$p(\vec{Y}, \vec{H}, \mathbf{R}_n | S) = p(\vec{Y} | S, \vec{H}, \mathbf{R}_n) \cdot p(\vec{H}, \mathbf{R}_n). \quad (5.56)$$

The objective function for the EM algorithm $Q(S, S^{i-1})$ is defined as

$$Q(S, S^{i-1}) = E[\log p(\vec{Y}, \vec{H} | S, \mathbf{R}_n) | \vec{Y}, S^{i-1}]. \quad (5.57)$$

Our purpose is to get the maximum likelihood estimate of S from the objective function.

$$\begin{aligned} \frac{\partial \log Q}{\partial S} &= E[\vec{H}^H \mathbf{R}_n^{-1} \vec{Y} | \vec{Y}, S^{i-1}] - E[\vec{H}^H \mathbf{R}_n^{-1} \vec{H} S | \vec{Y}, S^{i-1}] \\ &\quad + (E[\vec{H}^H \mathbf{R}_n^{-1} \vec{Y} | \vec{Y}, S^{i-1}] - E[\vec{H}^H \mathbf{R}_n^{-1} \vec{H} S | \vec{Y}, S^{i-1}])^H. \end{aligned} \quad (5.58)$$

In the “Maximizing step,” by setting the RHS of (5.58) to zero, we find the ML estimator $\hat{S}_{ML}$ to be

$$\hat{S}_{ML} = \frac{E[\vec{H}^H \mathbf{R}_n^{-1} \vec{Y} | \vec{Y}, S^{i-1}]}{E[\vec{H}^H \mathbf{R}_n^{-1} \vec{H} | \vec{Y}, S^{i-1}]}. \quad (5.59)$$

We obtain $\hat{\mathbf{R}}_n$ as

$$\begin{aligned}\hat{\mathbf{R}}_n(l, m) &= Y_l \cdot Y_m^* - E[H_m^* | \vec{Y}, S^{i-1}] \cdot Y_l \cdot S^{i-1*} - E[H_l | \vec{Y}, S^{i-1}] \cdot Y_m^* \cdot S^{i-1} \\ &+ E[H_l \cdot H_m^* | \vec{Y}, S^{i-1}] \cdot |S^{i-1}|^2.\end{aligned}\quad (5.60)$$

To get the solution, $E[\vec{H}^H \mathbf{R}_n^{-1} \vec{Y} | \vec{Y}, S^{i-1}]$, $E[\vec{H}^H \mathbf{R}_n^{-1} \vec{H} | \vec{Y}, S^{i-1}]$, $E[H_m^* | \vec{Y}, S^{i-1}]$, $E[H_l | \vec{Y}, S^{i-1}]$, and $E[H_l \cdot H_m^* | \vec{Y}, S^{i-1}]$ should be obtained in the “Expectation step.” $E[H_l | \vec{Y}, S^{i-1}]$, $E[H_m^* | \vec{Y}, S^{i-1}]$, $E[H_l \cdot H_m^* | \vec{Y}, S^{i-1}]$ are obtained as

$$E[H_l | \vec{Y}, S^{i-1}] = \frac{\lambda_{H_l} S^{(i-1)*} Y_l}{|S^{(i-1)}|^2 \lambda_{H_l} + \lambda_{N_l}}. \quad (5.61)$$

$$E[H_m^* | \vec{Y}, S^{i-1}] = \frac{\lambda_{H_m} S^{(i-1)} Y_m}{|S^{(i-1)}|^2 \lambda_{H_m} + \lambda_{N_m}}. \quad (5.62)$$

$$\begin{aligned}\text{Cov}[H_l \cdot H_m^* | \vec{Y}, S^{i-1}] &= E[(H_l - E[H_l | \vec{Y}, S^{i-1}])(H_m^* - E[H_m^* | \vec{Y}, S^{i-1}])] \\ &= \text{Cov}(H_l, H_m) - \frac{\text{Cov}(H_l, Y_l)}{\text{Cov}(Y_l)} \text{Cov}(H_m, Y_l) \\ &\quad - \frac{\text{Cov}(H_m, Y_m)}{\text{Cov}(Y_m)} \text{Cov}(H_l, Y_m) \\ &\quad + \frac{\text{Cov}(H_l, Y_l) \text{Cov}(H_m, Y_m)}{\text{Cov}(Y_l) \text{Cov}(Y_m)} \text{Cov}(H_m, Y_l) \text{Cov}(Y_l, Y_m) \\ &= \text{Cov}(H_l, H_m) - \frac{\lambda_{H_l} S^{(i-1)*} Y_l}{|S^{(i-1)}|^2 \lambda_{H_l} + \lambda_{N_l}} \text{Cov}(H_l, H_m) S^* \\ &\quad - \frac{\lambda_{H_m} S^{(i-1)} Y_m}{|S^{(i-1)}|^2 \lambda_{H_m} + \lambda_{N_m}} \text{Cov}(H_l, H_m) S^* \\ &\quad + \frac{\lambda_{H_l} S^{(i-1)*} Y_l}{|S^{(i-1)}|^2 \lambda_{H_l} + \lambda_{N_l}} \frac{\lambda_{H_m} S^{(i-1)} Y_m}{|S^{(i-1)}|^2 \lambda_{H_m} + \lambda_{N_m}} \\ &\quad \cdot (|S^{(i-1)}|^2 \text{Cov}(H_l, H_m) + \text{Cov}(N_l, N_m)),\end{aligned}\quad (5.63)$$

where $l, m = 1, 2, \dots, N$, and

$$E[H_l \cdot H_m^* | \vec{Y}, S^{i-1}] = \text{Cov}[H_l \cdot H_m^* | \vec{Y}, S^{i-1}] + E[H_l | \vec{Y}, S^{i-1}] \cdot E[H_m^* | \vec{Y}, S^{i-1}]. \quad (5.64)$$

From (5.60), (5.61), (5.62), and (5.63), we can get the scalar values of $E[\vec{H}^H \mathbf{R}_n^{-1} \vec{Y} | \vec{Y}, S^{i-1}]$ and $E[\vec{H}^H \mathbf{R}_n^{-1} \vec{H} | \vec{Y}, S^{i-1}]$ as shown below.

$$E[\vec{H}^H \mathbf{R}_n^{-1} \vec{H} | \vec{Y}, S^{i-1}] = \sum_{l=1}^N \sum_{m=1}^N E[H_l^* H_m | \vec{Y}, S^{i-1}] \cdot (\mathbf{R}_n)^{-1}(l, m). \quad (5.65)$$

$$E[\vec{H}^H \mathbf{R}_n^{-1} \vec{Y} | \vec{Y}, S^{i-1}] = \sum_{l=1}^N \sum_{m=1}^N E[H_l^* | \vec{Y}, S^{i-1}] \cdot (\mathbf{R}_n)^{-1}(l, m) \cdot Y_m. \quad (5.66)$$

Considering

$$p(S|Y) = p(S|T(Y)) = \frac{p(T(Y)|S)p(S)}{\int p(T(Y)|S)p(S)dS}, \quad (5.67)$$

where

$$T(Y) = \frac{H^H \mathbf{R}_n^{-1} Y}{H^H \mathbf{R}_n^{-1} H}, \quad (5.68)$$

and this sufficient statistic is the maximum likelihood estimate of S given $Y, \vec{H}$, and $\mathbf{R}_n$, we can replace (1.12) by (5.59). Furthermore the converged ML estimate of S can be enhanced more by applying a postfilter such as the Wiener filter. A postfiltered sufficient statistic is optimal in the sense of the postfilter, e.g. the Wiener filter is optimal in the MMSE sense. Similarly, MMSE-logSA applied to $T(Y)$ provides the MMSE estimate of $\log |S|$ given Y and H , and any other function of S may be similarly estimated on the basis of $T(Y)$.

CHAPTER 6

DECODER ORIENTED IDEAL BINARY MASK ESTIMATION FOR ROBUST ASR

In this chapter, a joint optimal method for automatic speech recognition (ASR) and ideal binary mask (IBM) estimation in the spectral domain, which is transformed from the cepstral domain, is introduced through a newly derived generalized expectation maximization algorithm. First, cepstral domain missing feature marginalization is established using a linear transformation, after tying the mean and variance of non-existing cepstral coefficients. Second, IBM estimation is formulated using a generalized expectation maximization algorithm directly to optimize the ASR performance. Experimental results show that even in highly non-stationary mismatch conditions (e.g. dance music as background noise), the proposed method achieves much higher absolute ASR accuracy improvement: ranging from 14.69% at 0 dB SNR to 40.10% at 15 dB SNR compared with the conventional noise suppression method.

6.1 Introduction

Speech intelligibility is easily degraded by unexpected mismatch. Therefore, speech processing to mitigate or even eliminate the adverse effect of the mismatch has been a very important research topic not only for human-to-human interaction but also for robust automatic speech recognition (ASR). Note that the source of the mismatch is necessarily neither stationary nor expected. Humans are good at decoding speech even in adverse acoustic environments, e.g. a cocktail party [56, 111]. However, the dependence of ASR on the degree of mismatch is much more critical, and is a very important unsolved problem.

Recently, missing feature theory has been applied to mitigate the negative influence of the mismatch. A missing feature theory classifies the reliable

and unreliable feature components first, and uses the reliability information to enhance ASR accuracy [17, 18, 56, 71, 112]. The optimal estimation is rephrased as “imputation,” and model adaptation is supplemented with marginalization over missing feature components. Missing feature approach based on complicated spectral domain modeling is reported to be inferior to the cepstral domain modeling with optimal imputation [17]. However, even the optimally enhanced feature is not the same as the original clean feature, and this mismatch is transferred to all the cepstral components, such that it deteriorates the recognition accuracy. Therefore, one of the important current issues is how we can import the missing feature approach into the cepstral domain feature without having to impute a full feature vector from the pre-processing.

Many papers since the early 1990s have asked how we can optimally classify the missingness directly for speech recognition performance. Recently, suppressing the missing part via ideal binary mask (IBM) has been demonstrated to be effective for increasing the intelligibility of human-to-human speech communication [21]. Typical methods for automatic IBM estimation use relatively less complicated front-end processing [18, 21, 71], and the recognition is performed based on the results of the separated pre-processing, which is not necessarily optimal for speech recognition based on a hidden Markov model (HMM).

In this chapter, we first introduce how we can use the HMMs trained on the cepstral domain features with a missing feature approach. Our method has the merits of both: optimal imputation for cepstral domain HMM [17] and missing feature marginalization in spectral domain HMM [18]. Secondly, we introduce decoder-oriented IBM estimation such that we can maximally utilize the information encoded in the parameters of a complete ASR framework. Since the missingness is declared by the speech models, any mismatch different enough from the speech will be classified as unreliable. For example, non-stationary abrupt impulsive noise can be detected and removed. Therefore, limitless cases of mismatch conditions can be handled simply if they are different from the speech used for training the HMMs.

6.2 Proposed method

6.2.1 Linear transformation of a Gaussian mixture model (GMM)

Speech is contaminated differently in each frequency band. Missing features (features corrupted by mismatch) can be marginalized out, if the HMM has been trained using spectral features [56]. However, an HMM trained on cepstral domain features produces a lower word error rate [17]. In this section, this gap is bridged by using the simple fact that the discrete cosine transform (DCT) is a linear transformation of the spectrum, and high-order cepstral coefficients are not useful for discriminating among different models. To get the cepstral features, we apply a DCT operation to the spectral coefficients. For example, mel-frequency cepstral coefficients (MFCCs) $\vec{y}$ can be obtained by the following equation.

$$\vec{y} = \mathbf{D}\vec{x}, \quad (6.1)$$

where $\vec{x}$ is mel-frequency spectral coefficients (MFSCs) and $\mathbf{D}$ is a DCT matrix such that

$$\mathbf{D}(i, j) = \sqrt{\frac{2}{N}} \cos \left(\frac{\pi(i-1)}{N}(j-0.5) \right), \quad (6.2)$$

where $i, j = 1, 2, \dots, F$ and F is the number of feature components in a frame. The first row is divided by $\sqrt{2}$ to make the DCT matrix orthonormal. If we train an HMM with a Gaussian mixture model (GMM)

$$\vec{y}|q, C \sim \sum_{i=1}^M \omega_{i|q, C} N(\vec{\mu}_{i|q, C}, \text{diag}(\vec{\sigma}_{i|q, C}^2)), \quad (6.3)$$

where C is the target utterance model, q means a specific state, and usually we use a diagonal covariance matrix, $\text{diag}(\vec{\sigma}_{i|q,C}^2)$. Then,

$$\begin{aligned}
\vec{x}|q, C &\sim \sum_{i=1}^M \omega_{i|q,C} \\
&\cdot N\left(\mathbf{D}^T \begin{bmatrix} \vec{\mu}_{i|q,C} \\ \vec{\mu}_g \end{bmatrix}, \mathbf{D}^T \text{diag} \begin{bmatrix} \vec{\sigma}_{i|q,C}^2 \\ \vec{\sigma}_g^2 \end{bmatrix} \mathbf{D}\right) \\
&\sim \sum_{i=1}^M \omega_{i|q,C} N(\vec{\mu}'_{i|q,C}, \Sigma'_{i|q,C}),
\end{aligned} \tag{6.4}$$

i.e. the spectral feature distribution is given by a trained HMM with full covariance GMM. Note that it is possible to fill the missing mean and variance for high-order coefficients by using a global mean vector $\vec{\mu}_g$ and variance vector $\vec{\sigma}_g^2$, independent of the utterance models, states, and mixture components. For example, our experiments use 13 MFCCs transformed from 26 MFSCs.

This dimensionality reduction is justified because the remaining coefficients do not play an important role in differentiating the trained models. The replacement with a global mean and variance may be interpreted as a mean and variance tying process across all the different models [20]. Only the mean and variance for the first half of the coefficient vector, multiplied by the number of mixture components, need to be trained using model dependent utterances. Compared with the case when we have to train the full-covariance GMM using spectral features, this is really a huge parameter reduction; at the same time, we are now able to use the missing feature marginalization scheme, since we have equivalent spectral domain HMMs. Furthermore, we may increase the spectral domain feature dimension, e.g. from 26 to 64 as in [18] if necessary, with relatively small increase of the corresponding cepstral domain.

Marginalization of a GMM is performed by simply deleting unreliable components from the mean vector, and deleting the row and column of the unreliable components for the covariance matrix as shown conceptually in Figure 6.1. Therefore, the missing parts will be recognized as “missing” and will not contribute any false information to the acoustic score produced by each model.

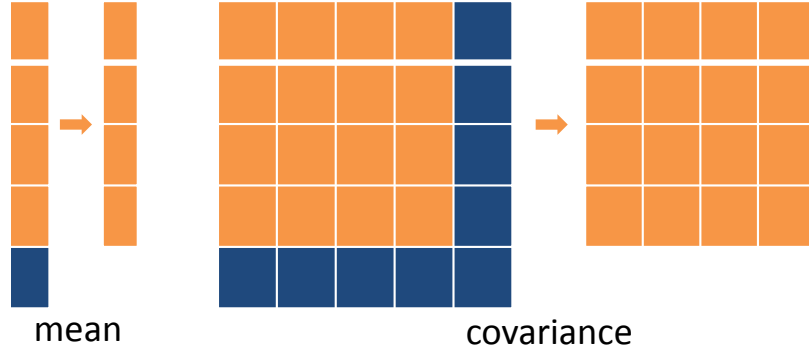


Figure 6.1: Conceptual diagram for the marginalization: the fifth feature has been marginalized by deleting the fifth mean and the fifth column and row of covariance.

6.2.2 Decoding with missingness classification based on the decoder

In this section, we formulate a generalized expectation maximization (GEM) procedure to estimate the missingness label for each feature component in every frame. In the process of GEM iteration, robust ASR can be performed as well. Therefore, the proposed method is considered jointly optimal not only for IBM estimation but also for ASR accuracy. Figure 1.3 shows a conceptual block diagram for the proposed method, where $\vec{O}$ is used for $\vec{x}$ in this chapter. Features are labeled as “missing” if doing so increases ASR performance; at the same time, ASR is performed aiming to increase the accuracy of the missingness labels. For the initialization procedure, we may use conventional IBM estimation methods, but we try to rely only on the speech model we already have except for the part of estimating the stationary noise level, if any.

Maximum likelihood estimation (MLE) of missing feature labels

Instead of following (1.21), in this chapter we replace the marginalization for the state sequence by a forward-backward algorithm with the best sequence by Viterbi decoding as shown in (6.5). Then, the expectation step will be less computationally expensive, and the maximization step can also be implemented with fewer for-loops, because we are given the states at each

frame, so all frames can be independently updated.

$$\begin{aligned}
& \hat{x}_L(t, f) \\
&= \arg \max_{x_L(t, f) \in \{0, 1\}} \sum_{i=1}^{N_C} p(C_i | \vec{x}(1 : T), \vec{x}_L(1 : T)^0) \\
&\cdot \log p(\vec{x}(t) | x_L(t, f), \vec{x}_L(t, \neg f) = \vec{x}_L(t, \neg f)^0, C_i, q_t^{C_i}),
\end{aligned} \tag{6.5}$$

where $q_t^{C_i}$ represents a best path state at time t in a given model C_i . For a fair comparison, some normalization procedure is needed to prevent the situation where we always choose $x_L(t, f) = 0$ in (6.5). To prevent this, we may want to claim that we should know a mismatch source (noise) model responsible for the missing components. However, because it is often not available, we introduce a method to resolve this issue without the necessity of those noise models in the following.

$$\hat{x}_L(t, f) = \arg \max_{x_L(t, f) \in \{0, 1\}} f(x_L(t, f)), \tag{6.6}$$

where $f(1)$ is the same as (6.5), but $f(0)$ is formulated as follows:

$$\begin{aligned}
& f(0) \\
&= \sum_{i=1}^{N_C} p(C_i | \vec{x}(1 : T), \vec{x}_L(1 : T)^0) \\
&\cdot [\log p(\vec{x}(t, \neg f) | \vec{x}_L(t, \neg f) = \vec{x}_L(t, \neg f)^0, C_i, q_t^{C_i}) \\
&+ \log p_n(x(t, f) | \vec{x}_L(t, \neg f) = \vec{x}_L(t, \neg f)^0, C_i, q_t^{C_i})],
\end{aligned} \tag{6.7}$$

where

$$\begin{aligned}
& p_n(x(t, f) | \vec{x}_L(t, \neg f) = \vec{x}_L(t, \neg f)^0, C_i, q_t^{C_i}) \\
&= \max_{x(t, f) \notin \text{CI}_\alpha} p(x(t, f) | \vec{x}_L(t, \neg f) = \vec{x}_L(t, \neg f)^0, C_i, q_t^{C_i}).
\end{aligned} \tag{6.8}$$

In statistics, a confidence interval (CI) is an interval estimate of a random parameter [113]. The probability that the interval includes the parameter is determined by the confidence level α . For example, if we simplify (6.8) to the case of two reliable feature components $x(t, f_1) = a$, $x(t, f_2) = b$ in a given model C and state q , then $p(x(t, f) | x(t, f_1) = a, x(t, f_2) = b, C, q)$ can

be described as follows:

$$p(x(t, f)|x(t, f_1) = a, x(t, f_2) = b, C, q) \sim N(\mu, \Sigma), \quad (6.9)$$

where

$$\begin{aligned} \mu &= E[x(t, f)] + \text{Cov}[x(t, f), [x(t, f_1) \ x(t, f_2)]] \\ &\cdot \text{Cov}[[x(t, f_1) \ x(t, f_2)]]^{-1}([a \ b] - E[x(t, f_1) \ x(t, f_2)])^T, \end{aligned} \quad (6.10)$$

and

$$\begin{aligned} \Sigma &= \text{Cov}[x(t, f)] - \text{Cov}[x(t, f), [x(t, f_1) \ x(t, f_2)]] \\ &\cdot \text{Cov}[[x(t, f_1) \ x(t, f_2)]]^{-1} \\ &\cdot \text{Cov}[[x(t, f_1) \ x(t, f_2)], x(t, f)]. \end{aligned} \quad (6.11)$$

Note that in our approach the confidence interval can be interpreted as a method of rejecting outliers from the perspective of the trained model [114].

Initialization

Before starting the iterative IBM estimation, a reasonably good starting point with GEM is important. The following schemes are designed to achieve a good starting point.

<Stationary background noise filtering>

We can easily obtain the stationary background noise statistics [48]. Therefore, spectral components below the stationary noise floor can be initially labeled as “missing” ($x_L(t, f)^0 = 0$).

<Confidence interval>

The second part of the initialization scheme is formalized as follows:

$$H_0 : x \in \cup_{\theta \in \Theta} \text{CI}_{\alpha_{high}, \alpha_{low}|\theta}. \quad (6.12)$$

$$\text{CI}_{\alpha_{high}, \alpha_{low}|\theta} = (x_L, x_H), \quad (6.13)$$

such that $P(x < x_L) = \alpha_{low}$ and $P(x > x_H) = \alpha_{high}$, where Θ includes all

states in all models. The basic object of this scheme is to detect the missing parts based on confidence scores of speech models, which have been already trained in a target environment, e.g. high-SNR environment.

<Speech model based voice activity detection (VAD)>

At the end of the initialization process, we apply a model based VAD, which is dependent on the relative number of reliable components in each frame compared with the maximum number of reliable components among all frames in a given utterance. We declare a frame with fewer reliable components than the threshold as missing, thereby removing frames with very few reliable feature components.

6.3 Experiment

To validate the proposed method, isolated digit HMM recognizers with a single Gaussian per state were built using 13 MFCCs transformed from 26 MFSCs. The hidden Markov model toolkit (HTK) [20] was used to train the 12 different models “one”, “two”, ..., “nine”, “zero”, “oh”, and “silence” using the TIDIGITS corpus [44]. After training the models, we transformed the models back to the spectral domain by following (6.4). When we train HMMs, we can exclude the silence model using forced alignment. Given only the digit models, the silence frames (presumably contaminated with background noise) can be classified as missing. The proposed method has been implemented based on the HMM toolbox for MATLAB [63].

Three different tests were performed to check the validity of (6.4), first using cepstral features, second using spectral features for the testing data, and third using part of the spectral features to simulate the missing feature cases. Note that, for the first set, we just need to use the original HMM, trained on 13 MFCCs. For the second and third sets, we need to use the linearly transformed HMM. As discussed previously, the global mean and variance have been used to fill the last 13 high-order MFCCs. For the third case, the 26th MFSC, which summarizes the energy between 3.6 and 4.0 kHz, is simulated as missing. Table 6.1 shows the recognition accuracy in all three cases. As expected, the transformed HMM produces exactly the same accuracy as the original HMM. Interestingly, the accuracy of clean-speech

Table 6.1: Recognition accuracy(%)

	MFCC	MFSC	MFSC-missing one
Accuracy(%)	94.24	94.24	94.64

TIDIGITS recognition is improved slightly when we marginalize out the last MFSC.

Figure 6.2 shows the ASR accuracy of the noisy speech with white noise as background noise, after applying enhancement by MMSE-logSA [48], and after MMSE-logSA together with the proposed scheme. Note that MMSE-logSA with the proposed method outperforms MMSE-logSA alone. Figure 6.3 (highly non-stationary dance music as background noise) shows more interesting results with the proposed scheme. A conventional speech enhancement algorithm deteriorates the ASR performance. However, the proposed method highly outperforms the baseline. Normally it is known to be very hard to achieve robust speech recognition with music-like non-stationary background noise [71], but the proposed method performs much better than the baseline in this kind of unexpected, untrainable mismatch condition.

6.4 Conclusions and future work

This chapter uses the knowledge of a trained ASR to classify the reliable feature components for missing-feature speech recognition, which in turn contributes to increase the accuracy of ASR. Signal enhancement methods such as MMSE-logSA [48] can be combined with the proposed scheme without hurting the performance of the front-end processing. In fact, having more reliable feature components through this separate front-end processing gives us a better chance to have more accurate speech recognition results. The GEM iteration iteratively approaches the utterance specific speech process models allowing better estimation of the binary mask. To summarize, the proposed method can provide better ASR performance not only for the case of stationary mismatch (by combining with pre-processing) but also for the case of the non-stationary mismatch, where the mismatch is not easily modeled or estimated.

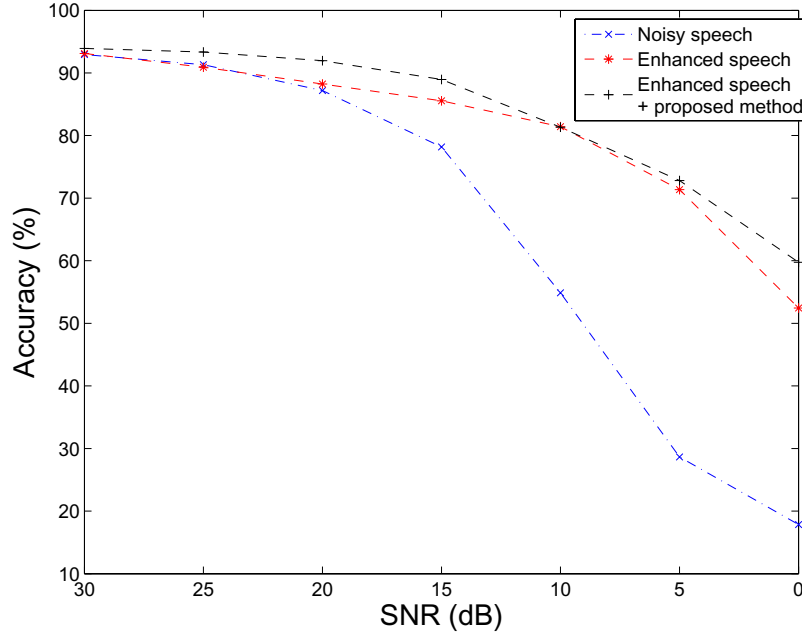


Figure 6.2: Recognition accuracy(%) vs SNR (dB) with white noise as background noise.

In the experimental study on understanding the cocktail party effect, the binaural aspect of human hearing has been emphasized [111]. This location-based multi-channel information will be combined to boost the IBM estimation accuracy with the proposed one-channel scheme given the correct estimate of the direction of arrival (DOA) of the target speech for each feature index [30].

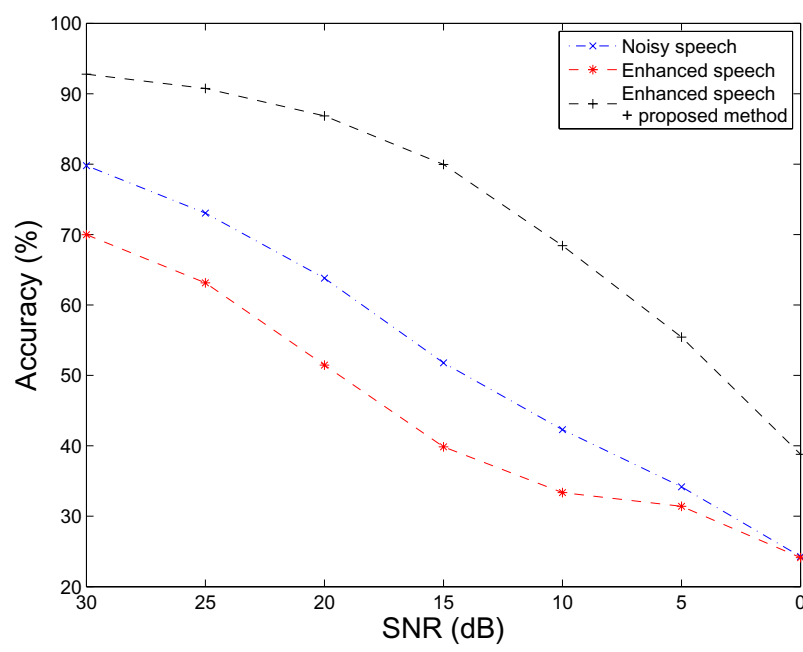


Figure 6.3: Recognition accuracy(%) vs SNR (dB) with dance music in the movie “Dream Girls” as background noise.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTION

In this thesis, a general unifying framework has been set up to tie together multi-microphone clean speech estimation and statistical model based robust speech recognition. With the generalized EM framework in the short-time Fourier transform domain, the speech estimation and ASR benefit each other through iterations: better speech estimation more closely fitted to the ASR model produces more accurate ASR performance, while a better specified statistical model for each specific utterance gives us a better chance of estimating more suitable speech components for those utterance specific speech models.

To achieve a near global optimum in this framework, not only initialization for GEM with accurate IBM estimation but also assuring as many reliable feature components as possible by optimal front-end processing is important. Regularized feed-forward network based ICA followed by postfiltering based on DOA per frequency-bin achieved good source separation, dereverberation, and de-noising performance for the simultaneously speaking speakers who are assumed to remain in a designated area. The EM beamforming demonstrated good potential for fast tracking with reasonable separation, dereverberation, and de-noising performance of target speech for each frame with a statistical model based generative model. With the assumption of estimated DOA, DOA estimation was extended to the estimation of distinct early responses. The remaining portion of the early response above a certain frequency, called Schroeder's frequency [107], was modeled as random variables, and a multi-microphone EM algorithm was derived to compensate for them. Robust speech recognition has been formulated with optimal estimation of IBM and showed good performance in an unexpected, non-stationary mismatch condition.

Speech estimation, which is to fill the unreliable portion after the convergence of the decoder oriented GEM algorithm, will be the direction of

ongoing future work. There exist several issues with utterance specific statistical model based speech recognition. First, we need to investigate how we apply phase information after estimating the unreliable spectral feature components. One straightforward way will be a filter bank approach in which we treat each mel-frequency bin as a mel-filter bank and apply the estimated spectral amplitude with noisy phase. This method can be regarded as applying a gain function as in the case of classical speech estimation, e.g. Wiener filtering. Another method is to apply the mechanism of bandwidth extension in the speech coding area, the main goal of which is to generate high-frequency band signal based on a low-frequency band signal and a statistical model such as HMM on speech [61]. As in this case, we may be able to apply the residual extension method to fill the missing phase part.

Another issue involves the model complexity of speech process. We have verified that we can use many more mel-filterbank outputs than the current number for speech recognition through the linear transformation from the cepstral domain HMM model with a tying process among the high-order spectral parts as in Chapter 6. To be more effective for speech estimation, we need to increase the number of mel-filterbank channels without much need of training data.

In Chapter 6, we focused on estimating IBM in the direction of increasing ASR performance. Once we estimate IBM, we apply a hard decision on the selection of reliable feature components. There has been another approach called the soft-decision based missing feature method [71], but this is only easily applicable when the covariance matrix of the spectral features can be modeled as a diagonal matrix and when the original and the corrupted feature components can be modeled with a linear relationship. To perform the analytically intractable multiple integration, a Monte Carlo based speech recognition scheme can be considered as in Appendix F, where the Monte Carlo method has been utilized to solve the integration over the non-Gaussian PDF. Since the part with the sample generation is not dependent on the different speech models, enough sample vectors per frame can be generated, and the likelihood function of clean speech is replaced with the likelihood function of the noisy speech by marginalizing out the clean feature in the joint PDF of both of them. The marginalization, in fact, can be approximated as the sample average with the PDF corresponding to the samples generated with the Monte Carlo method.

APPENDIX A

DIFFERENT FORMULATIONS FOR JOINT OPTIMIZATION OF SPEECH ESTIMATION AND ASR

In this appendix, three different problem formulations with (1.2) for joint optimization are discussed. In the end of each section, some comments why (1.2) fits better are addressed.

A.1 MLE of both model and missing features

In this case, we treat the model C and the missing feature $\vec{O}_m(1 : T)$ as deterministic parameters to be estimated as in the following.

$$[\hat{C}, \hat{\vec{O}}_m(1 : T)] = \arg \max_{[C, \vec{O}_m(1:T)]} \log p \left(\vec{O}_r(1 : T) | [C, \vec{O}_m(0 : T)] \right). \quad (\text{A.1})$$

Equation (A.1) can be reformulated as an EM setup by considering the hidden state sequence $\vec{q}$ as a latent variable, which is common in ASR using HMM [22].

$$[\hat{C}, \hat{\vec{O}}_m(1 : T)] = \arg \max_{[C, \vec{O}_m(1:T)]} Q([C, \vec{O}_m(1 : T)], [C, \vec{O}_m(1 : T)]^0), \quad (\text{A.2})$$

where

$$\begin{aligned} & Q([C, \vec{O}_m(1 : T)], [C, \vec{O}_m(1 : T)]^0) \\ = & E_{p(\vec{q}(1:T) | \vec{O}_r(0:T), [M, y_m(0:T)]^0)} \left[\log p(\vec{O}_r(1 : T), \vec{q}(0 : T) | C, \vec{O}_m(1 : T)) \right]. \end{aligned} \quad (\text{A.3})$$

Note that in this problem setting, we do not consider the posterior probability of the model; therefore, we cannot incorporate the soft information on all possible models to estimate the missing features.

A.2 MLE of model and missing features as latent variables

In this section, we only try to estimate C in a maximum likelihood sense with marginalization over missing features.

$$\hat{C} = \arg \max_C \log p \left(\vec{O}_r(1:T) | C \right). \quad (\text{A.4})$$

Equation (A.4) is also reformulated into an EM problem:

$$\hat{C} = \arg \max_C Q(C, C^0), \quad (\text{A.5})$$

where

$$Q(C, C^0) = E_{p(\vec{O}_m(1:T), \vec{q}(1:T) | \vec{O}_r(1:T), C^0)} \left[\log p(\vec{O}_r(1:T), \vec{O}_m(1:T), \vec{q}(1:T) | C) \right]. \quad (\text{A.6})$$

To avoid complex multi-dimensional integrations, we may use the maximum of $p(y_m(0:T))$ and $Q(0:T)$ at each expectation step instead of integrating over a *posteriori* PDF $p(\vec{O}_m(0:T), \vec{q}(0:T) | \vec{O}_r(0:T), C^0)$. Then, the problem setting becomes similar to that in Section A.1.

A.3 Both model and missing features as latent variables

For the previous cases, we estimated model or missing features in an MLE sense, but we can actually implement a Bayesian estimate for both of them such that we can obtain two *a posteriori* PDFs iteratively based on the data augmentation (DA) algorithm [115]. However, because it involves a Monte Carlo sampling, we do not address the derivation due to the high complexity. This formulation will be one of the possible extensions of this thesis.

APPENDIX B

SUFFICIENT STATISTIC CONSIDERING REVERBERATION PROCESS

Equation (1.10) is reformulated as following by considering a series of observation frames,

$$\begin{aligned}
& p(\vec{Y}, \vec{Y}^{\cdot-1}, \dots, \vec{Y}^{\cdot-(N_Y-1)} | \vec{S}, \vec{S}^{\cdot-1}, \dots, \vec{S}^{\cdot-(N_M-1)}, \mathbf{H}^0, \mathbf{H}^1, \dots, \mathbf{H}^{N_H-1}) \\
& \sim \exp \left(-\frac{1}{2} (\vec{Y}_{N_Y} - \mathbf{H}_{N_Y \times N_M} \vec{S}_{N_M})^H \mathbf{R}'_N (\vec{Y}_{N_Y} - \mathbf{H}_{N_Y \times N_M} \vec{S}_{N_M}) \right) \quad (\text{B.1}) \\
& \sim \exp \left(-\frac{1}{2} \left((\mathbf{H}_{N_Y \times N_M}^H \mathbf{H}_{N_Y \times N_M})^{-1} \mathbf{H}_{N_Y \times N_M}^H \vec{Y}_{N_Y} - \vec{S}_{N_M} \right)^H \right. \\
& \quad \cdot \left[(\mathbf{H}_{N_Y \times N_M}^H \mathbf{H}_{N_Y \times N_M})^{-1} \mathbf{H}_{N_Y \times N_M}^H \mathbf{R}'_N \mathbf{H}_{N_Y \times N_M} (\mathbf{H}_{N_Y \times N_M}^H \mathbf{H}_{N_Y \times N_M})^{-1} \right]^{-1} \\
& \quad \cdot \left. \left((\mathbf{H}_{N_Y \times N_M}^H \mathbf{H}_{N_Y \times N_M})^{-1} \mathbf{H}_{N_Y \times N_M}^H \vec{Y}_{N_Y} - \vec{S}_{N_M} \right) \right),
\end{aligned}$$

where superscript H means conjugate transpose, $\mathbf{R}'_N = E[\vec{N}_{N_Y} \vec{N}_{N_Y}^H]$, where $\vec{N}_{N_Y} = [\vec{N}^T \ \vec{N}^{\cdot-1T} \ \dots \ \vec{N}^{\cdot-(N_Y-1)T}]^T$, $\vec{Y}_{N_Y} = [\vec{Y}^T \ \vec{Y}^{\cdot-1T} \ \dots \ \vec{Y}^{\cdot-(N_Y-1)T}]^T$, $\vec{S}_{N_M} = [\vec{S}^T \ \vec{S}^{\cdot-1T} \ \dots \ \vec{S}^{\cdot-(N_M-1)T}]^T$, and

$$\mathbf{H}_{N_Y \times N_M} = \begin{bmatrix} \mathbf{H}^0 & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{H}^1 & \mathbf{H}^0 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{H}^2 & \mathbf{H}^1 & \mathbf{H}^0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{H}^{N_H-1} & \mathbf{H}^{N_H-2} & \mathbf{H}^{N_H-3} & \dots & \mathbf{H}^0 \\ \mathbf{0} & \mathbf{H}^{N_H-1} & \mathbf{H}^{N_H-2} & \dots & \mathbf{H}^1 \\ \vdots & \vdots & \dots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{H}^{N_H-1} \end{bmatrix}, \quad (\text{B.2})$$

where the number of observation frames $N_Y \geq N_M$, which is the number of the source frames due to the reverberation process in the STFT domain. Then, because the Gaussian PDF is a member of an exponential family, (B.1)

can be factored as

$$\begin{aligned}
& p(\vec{Y}, \vec{Y}^{\cdot-1}, \dots, \vec{Y}^{\cdot-(N_Y-1)} | \vec{S}, \vec{S}^{\cdot-1}, \dots, \vec{S}^{\cdot-(N_M-1)}, \mathbf{H}^0, \mathbf{H}^1, \dots, \mathbf{H}^{N_H-1}) \\
& = C(\vec{S}, \dots, \vec{S}^{\cdot-(N_M-1)}) h(\vec{Y}, \dots, \vec{Y}^{\cdot-(N_Y-1)}) g(\vec{S}_{N_M}^H T_{\mathbf{H}_{N_Y \times N_M}}(\vec{S}_{N_M}^H)),
\end{aligned}$$

where C , g , and h are arbitrary functions, and

$$\begin{aligned}
& T_{\mathbf{H}_{N_Y \times N_M}}(\vec{S}_{N_M}^H) \\
& \sim \left([(\mathbf{H}_{N_Y \times N_M}^H \mathbf{H}_{N_Y \times N_M})^{-1} \mathbf{H}_{N_Y \times N_M}^H \mathbf{R}'_N \mathbf{H}_{N_Y \times N_M} (\mathbf{H}_{N_Y \times N_M}^H \mathbf{H}_{N_Y \times N_M})^{-1}]^{-1} \right. \\
& \cdot \left. (\mathbf{H}_{N_Y \times N_M}^H \mathbf{H}_{N_Y \times N_M})^{-1} \mathbf{H}_{N_Y \times N_M}^H \vec{Y}_{N_Y} \right). \tag{B.3}
\end{aligned}$$

With $N_H = 1$, $T_{\mathbf{H}_{N_Y \times N_M}}(\vec{S}_{N_M}^H) \sim \mathbf{H}_{N_Y \times N_M}^H \mathbf{R}_N'^{-1} \vec{Y}_{N_Y}$, which corresponds to the original derivation by [25]. Also, note that if $\mathbf{R}_N = \mathbf{I}$, then $T_{\mathbf{H}_{N_Y \times N_M}}(\vec{S}_{N_M}^H) \sim (\mathbf{H}_{N_Y \times N_M}^H \mathbf{H}_{N_Y \times N_M})^{-1} \mathbf{H}_{N_Y \times N_M}^H \mathbf{I} \vec{Y}_{N_Y}$, where $(\mathbf{H}_{N_Y \times N_M}^H \mathbf{H}_{N_Y \times N_M})^{-1}$ is an linear transformation multiplied without loss of generality so that the sufficient statistic has a least squares solution form.

APPENDIX C

PROOF OF EQUATION (1.8)

In this appendix, we show that the proper representation of multi-channel convolutive mixing in a short-time Fourier transform (STFT) domain is a convolutive mixing even in a STFT domain by regarding each subband signal as a discrete time domain signal with complex variables from the overlap addition context [23].

Proposition 3. *Convolution in the time domain by a filter with longer length than a frame length used for STFT is equivalent to a subband domain convolution with delay of the frame shift.*

$$Y^r(e^{j\omega_k}) = \sum_{l=0}^{N_H-1} H^l(e^{j\omega_k}) S^{r-l}(e^{j\omega_k}), \quad (\text{C.1})$$

where ω_k is the angular frequency at frequency index k , superscript $\cdot -l$ stands for the frame, which is l frames before the current frame, and N_H stands for the total number of frames to include the convolution filter length sufficiently and the l^{th} tap subband domain convolution filter

$$H^l(e^{j\omega_k}) = \sum_{t=-\infty}^{\infty} h^l[t] e^{-j\omega_k t}, \quad (\text{C.2})$$

where

$$h^l[t] = h[t] \text{win}[lR - t], \quad -\infty < t < \infty, \quad (\text{C.3})$$

where $\text{win}[t]$ is a proper window function and from the overlap-add context

$$h[t] = \sum_{l=0}^{N_H-1} h^l[t]. \quad (\text{C.4})$$

Proof. The equivalence between the subband domain convolution operation and the original time domain convolution can be demonstrated by taking

an inverse Fourier transform on (C.1) and summing them over all possible frames, which is eventually shown to be equivalent to the results when we simply implement the convolution in the time domain with the original incoming signal and filter before having subband decomposition. In the time domain, a convolution can be performed using a filter $h[t]$ with an incoming signal $s[t]$.

$$y[t] = \sum_{\tau=-\infty}^{\infty} s[\tau]h[t - \tau], \quad (\text{C.5})$$

From the perspective of subband convolution, the convolved signal can be obtained as following.

$$\begin{aligned} & \hat{y}[t] \\ &= \sum_{r=-\infty}^{\infty} \frac{1}{K} \sum_{k=0}^{K-1} \left[\sum_{l=0}^{N_H-1} H^l(e^{j\omega_k}) S^{r-l}(e^{j\omega_k}) \right] e^{j\omega_k t} \\ &= \sum_{r=-\infty}^{\infty} \frac{1}{K} \sum_{k=0}^{K-1} \left[\sum_{l=0}^{N_H-1} H^l(e^{j\omega_k}) \left[\sum_{\tau=-\infty}^{\infty} s[\tau] \text{win}[(r-l)R - \tau] e^{-j\omega_k \tau} \right] \right] e^{j\omega_k t} \\ &= \sum_{\tau=-\infty}^{\infty} s[\tau] \sum_{l=0}^{N_H-1} \left[\frac{1}{K} \sum_{k=0}^{K-1} H^l(e^{j\omega_k}) e^{j\omega_k(t-\tau)} \right] \sum_{r=-\infty}^{\infty} \text{win}[(r-l)R - \tau] \\ &= \sum_{\tau=-\infty}^{\infty} s[\tau] \sum_{l=0}^{N_H-1} h^l[t - \tau] \sum_{r=-\infty}^{\infty} \text{win}[(r-l)R - \tau] \\ &= \sum_{\tau=-\infty}^{\infty} s[\tau] h[t - \tau]. \end{aligned} \quad (\text{C.6})$$

Note that with a careful choice of the window function, we can fulfill

$$\sum_{r=-\infty}^{\infty} \text{win}[(r-l)R - \tau] = 1. \quad (\text{C.7})$$

Therefore,

$$y[t] = \hat{y}[t], \quad (\text{C.8})$$

and the subband domain convolution represented in (C.1) is a correct way of implementing a convolution in the subband domain. $\square$

Note that the RIR has been chopped into N_H frames with the same amount of the frame shift as in the case of framing an incoming signal.

Proposition 4. *Multi-channel convolutive (reverberant) mixing in the time domain with longer room impulse responses length than a typical frame length used for subband decomposition is equivalent to a subband domain convolutive mixing with the delay of the frame shift.*

$$Y_i^r(e^{j\omega_k}) = \sum_{j=1}^{N_S} \sum_{l=0}^{N_H-1} H_{i,j}^l(e^{j\omega_k}) S_j^{r-l}(e^{j\omega_k}), \quad (\text{C.9})$$

where the subscript i means the i^{th} convolutive mixing output and j means the channel index. The number of the mixed sources N_S is not necessarily less than N_{mic} , but it is generally assumed so to perform source separation [2].

Proof. Proof is straightforward from Proposition 3. In the time domain, multi-channel convolutive mixing can be performed using a multi-channel filter $h_{i,j}[t]$ with an incoming signal $s_j[t]$.

$$y_i[t] = \sum_{j=1}^{N_S} \sum_{\tau=-\infty}^{\infty} s_j[\tau] h_{i,j}[t - \tau], \quad (\text{C.10})$$

$$\begin{aligned} & \hat{y}_i[t] \\ &= \sum_{r=-\infty}^{\infty} \frac{1}{K} \sum_{k=0}^{K-1} \left[\sum_{j=1}^{N_S} \sum_{l=0}^{N_H-1} H_{i,j}^l(e^{j\omega_k}) S_j^{r-l}(e^{j\omega_k}) \right] e^{j\omega_k t} \\ &= \sum_{r=-\infty}^{\infty} \frac{1}{K} \sum_{k=0}^{K-1} \left[\sum_{j=1}^{N_S} \sum_{l=0}^{N_H-1} H_{i,j}^l(e^{j\omega_k}) \left[\sum_{\tau=-\infty}^{\infty} s_j[\tau] \text{win}[(r-l)R - \tau] e^{-j\omega_k \tau} \right] \right] e^{j\omega_k t} \\ &= \sum_{j=1}^{N_S} \sum_{\tau=-\infty}^{\infty} s_j[\tau] \sum_{l=0}^{N_H-1} \left[\frac{1}{K} \sum_{k=0}^{K-1} H_{i,j}^l(e^{j\omega_k}) e^{j\omega_k(t-\tau)} \right] \sum_{r=-\infty}^{\infty} \text{win}[(r-l)R - \tau] \\ &= \sum_{j=1}^{N_S} \sum_{\tau=-\infty}^{\infty} s_j[\tau] \sum_{l=0}^{N_H-1} h_{i,j}^l[t - \tau] \sum_{r=-\infty}^{\infty} \text{win}[(r-l)R - \tau] \\ &= \sum_{j=1}^{N_S} \sum_{\tau=-\infty}^{\infty} s_j[\tau] h_{i,j}[t - \tau]. \end{aligned} \quad (\text{C.11})$$

Therefore, with careful choice of the window function win,

$$y_i[t] = \hat{y}_i[t], \quad i = 1, 2, \dots, N_{\text{mic}} \quad (\text{C.12})$$

and (C.9) is a correct way of implementing the multi-channel convolutive mixing in the subband domain. $\square$

APPENDIX D

PRE-AMP CIRCUIT DIAGRAM

Figure D.1 shows a block diagram of a pre-amp module for a microphone array. The same eight units are integrated by sharing only the +12 V power part denoted by J2. We ensured that each unit operates independently with similar gain. Maximum gain of up to 200 times amplification can be achieved and can be adjusted using a variable register $R3$. The linear voltage regulator $U1$ not only produces a stabilized constant voltage to the op amp, but also minimizes the inter-channel leakage due to the high input impedance. The actual self-made pre-amp is shown in Figure D.2.

APPENDIX E

CORPUS GENERATION FOR SOURCE SEPARATION EXPERIMENT

In this appendix, we explain how the entire corpus for source separation has been built. To make the situation affordable in terms of cost and time, we can generate the corpus under the linear time invariant system assumption about the acoustic space we measure. In other words, we measure room impulse responses from the mouth position to the microphone array and by convolving these with the sets of clean speech utterances, we assume that we can obtain the reverberated signal which can be similar enough to the real measurement of the speech in those designated positions. Natural ambient noise is also measured in the same space so that we can add some segment of it into the convolved signal to simulate the natural environmental noise condition with a clean reverberated signal.

E.1 Room impulse response measurement

In this section, we discuss measuring room impulse response, because it is an unavoidable process in accessing the proposed algorithm with reference to clean speech sources. We may be able to use some microphones close enough to the speakers when we capture distant speech situations, but still it would be hard to collect various different environmental conditions or speakers, etc. Therefore, in this section, under the assumption that the speaker is positioned at a designated place and the room is modeled as a linear time-invariant system, we performed a system identification process for the room. After obtaining the RIR, we can generate reverberant speech mixture with background noise based on (1.6). Figure E.1 shows a configuration to measure RIR, where the double line represents vector input or output.

To measure RIR accurately, we need to consider which type of source should be used since it is directly related to the quality of measured RIR. In

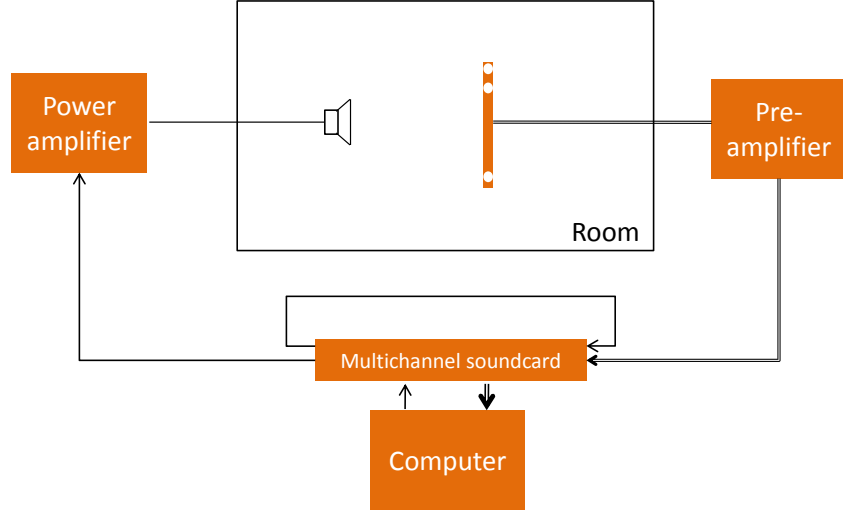


Figure E.1: RIR measurement configuration.

this thesis, we used a swept sine signal, which is known to reject nonlinear distortions, which may come from the loudspeaker, by using logarithmic sweep, robust to time variance [116]. Below is the equation to generate the logarithmic sweep [117]:

$$s(t) = \sin \left(\frac{2\pi f_{start}}{\log \left(\frac{f_{end}}{f_{start}} \right)} \cdot \left(\exp \left(\frac{t}{T} \log \frac{f_{end}}{f_{start}} \right) - 1 \right) \right), \quad (\text{E.1})$$

where f_{start} is a starting frequency, f_{end} is an ending frequency, and T is a time duration in seconds. Figure E.2 shows a generated signal according to (E.1) with $f_{start} = 22$ Hz, $f_{end} = 22$ kHz, $T = 1$, and the sampling rate $f_s = 48$ kHz. Before measuring RIR, we need to compensate for the loudspeaker response. In this thesis, the generated corpus is designed for the reverberant speech mixture; therefore, the loudspeaker used for RIR represents the mouth of the speaker. The mouth simulator B&K Type 4227 was used to simulate the directivity pattern of the human mouth. Note that since most of the recorded speech data in a clean environment embed the direct response of the speaker, not compensating for the loudspeaker response results in duplicating the direct response of the speaker. This compensation can be performed by deconvolving the response measured in an anechoic chamber. Instead, in a quietened room, by taking early response, we can substitute the speaker response, but there may exist poor compensation in the low-frequency range

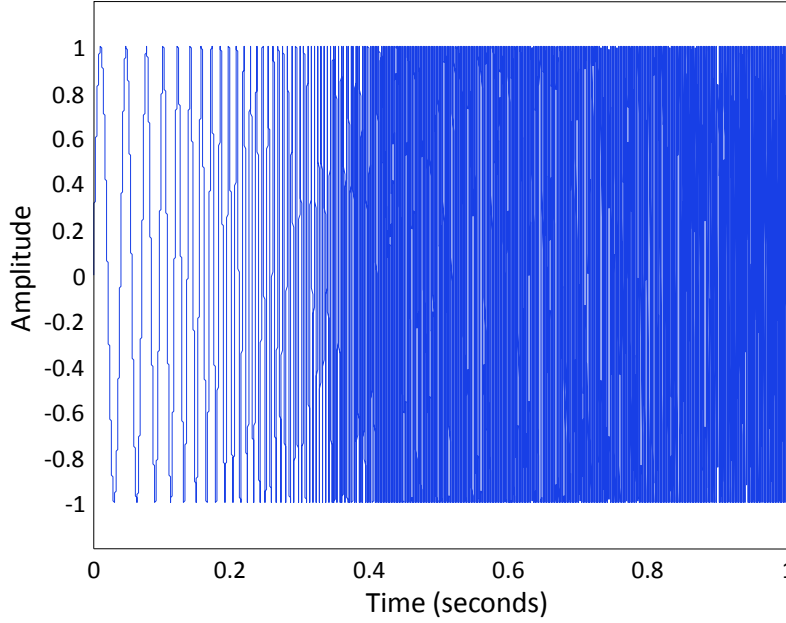


Figure E.2: Logarithmic sweep signal: time domain signal.

because of the short duration of the response.

The room impulse response can be calculated by deconvolving the measured sweep signal with the source sweep signal. To compensate a possible distortion by a sound card, we measure line output of source and use it as the source sweep signal. Note that because of background noise in the region of the no-energy frequency region, e.g. below 22 Hz and above 22 kHz in the case of Figure E.2, if we do not properly obtain the deconvolution filter, we may end up with unwanted boosting of background noise. Figure E.3 shows a conceptual diagram to obtain a deconvolution filter by dividing reference frequency response by the sweep sine signal. To determine the zero response frequency region, we need to set the threshold above the background noise level, not to boost it. After measuring the sweep sine response, we need to convolve the deconvolution filter to obtain RIR as shown in Figure E.4(b). We also can see the harmonic distortion response after the deconvolution by logarithmic sweep in Figure E.4(a).

We can make sure the quality of one time measurement by comparing it with series of measurements. For some reason, e.g. extensive attempts to access the network resource or a hard drive, an unexpected poor impulse response can be generated. Normally we get around 1/10 (20 dB) attenuation

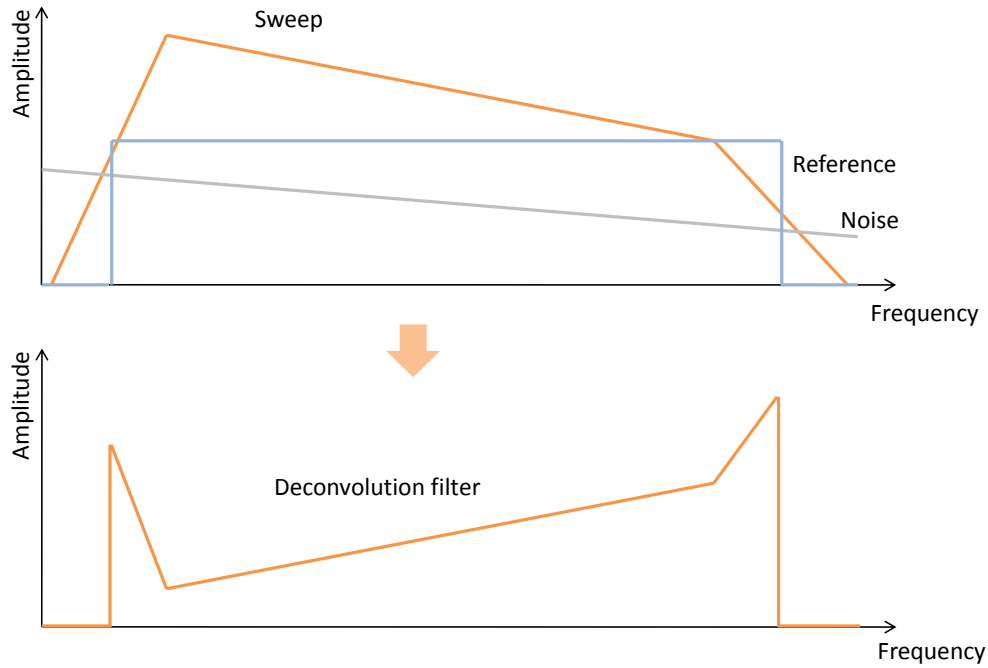
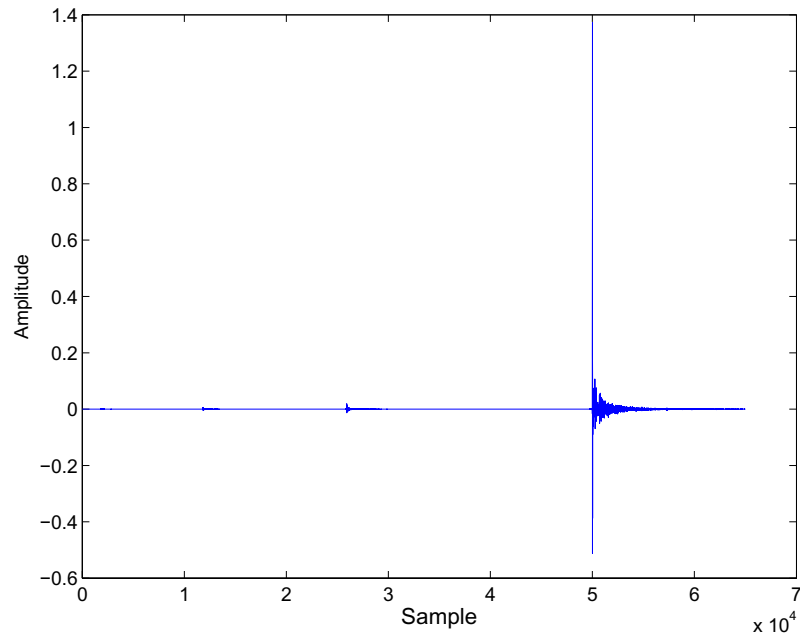
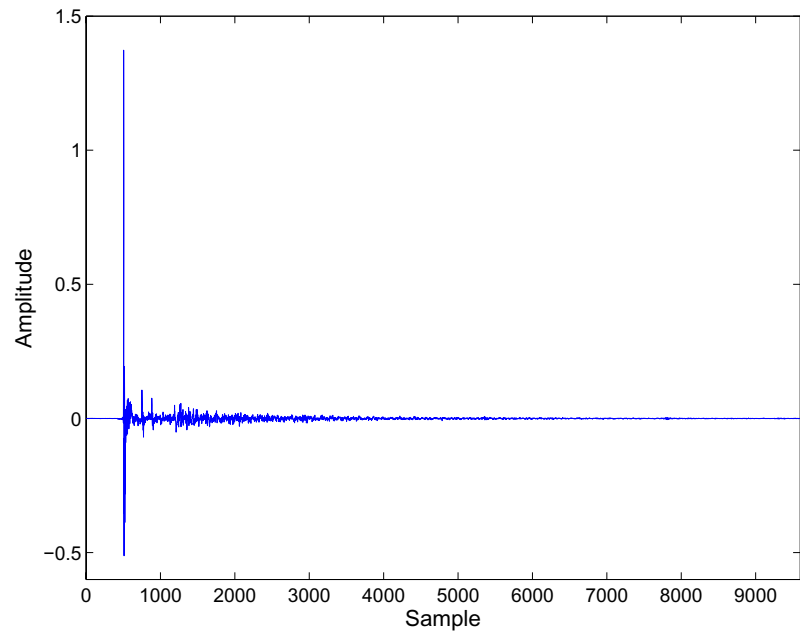


Figure E.3: Conceptual diagram for deconvolution filter

on average by just subtracting the generated signal by convolution from a series of measured ones. If not, we repeat the measurement. Figure E.5 shows the generated signals by convolving impulse responses, one of the series of measured signals, and difference between them. This is the case with 20 dB attenuation just by subtracting.



(a)



(b)

Figure E.4: Sample room impulse response measured with logarithmic sweep signal: (a) left smaller repetitions of RIR correspond to harmonic distortion. (b) RIR by excluding harmonic distortion.

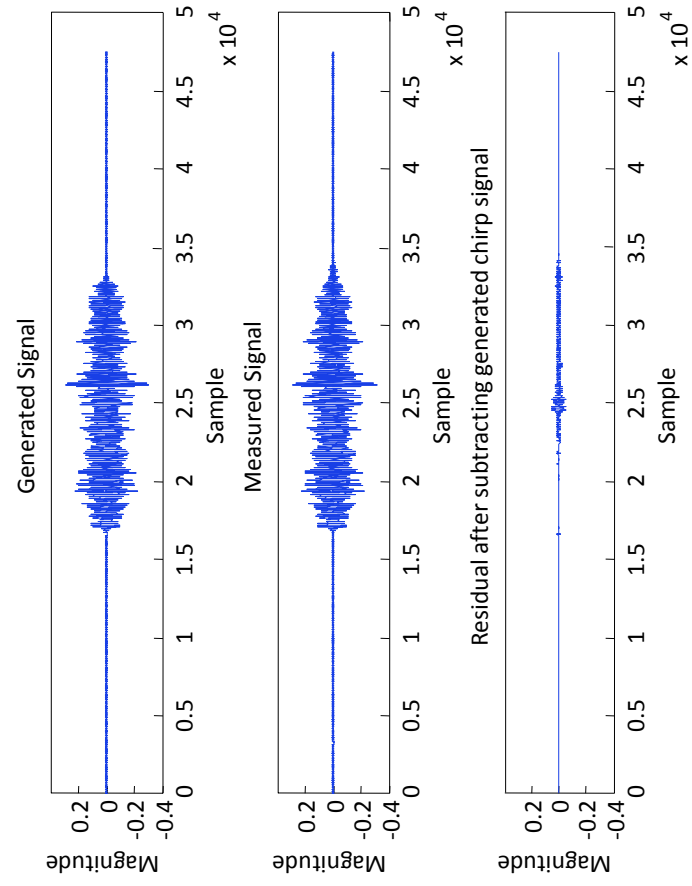


Figure E.5: Difference between generated and measured signal.

APPENDIX F

ROBUST SPEECH RECOGNITION BASED ON MONTE CARLO UNCERTAINTY DECODING

F.1 Problem formulation

In this section, we introduce a Monte Carlo based noisy likelihood given the following additive assumption on the clean speech and the acoustic noise.

$$y(t) = s(t) + n(t), \quad (\text{F.1})$$

where $s(t)$ is the clean speech, $n(t)$ is the noise, and $y(t)$ is the noisy observation at time t . Generally we train the speech recognizer using the overlapped series of the framed spectral domain representation such as MFCC [13] or PLP [14] with the window which can be short enough to meet the stationarity condition with that frame. This framing is indispensable, since the speech process is highly non-stationary. Spectral domain representation of (F.1) is given as following:

$$\vec{Y} = \vec{S} + \vec{N}, \quad (\text{F.2})$$

where $\vec{\cdot}$ represents the spectral vector of the frame. We also assume that the noise part can be estimated [9]. The HMM based speech recognition in a given situation can be described in a simple graphical model as in Figure F.1, where the empty rectangle represents the hidden state, the empty circle represents clean speech observation, and the shaded circle represents the noisy observation. If there is no noise, then we simply take the observation for obtaining the feature vector and perform the recognition by comparing the likelihood $p(\vec{Y}_1, \vec{Y}_2, \dots | M_i)$ and $i = 1, 2, \dots$, the number of models. However, if the noise is not zero, then $\vec{Y}_t$'s are not the same as $\vec{S}_t$'s any more, and there exists mismatch we should take into account. Instead of filling the missing part $\vec{S}_t$'s by taking an optimal value based on (F.2), we perform the marginalization over the $\vec{S}_t$'s in the joint PDF of $\vec{S}_t$'s and $\vec{Y}_t$'s given each

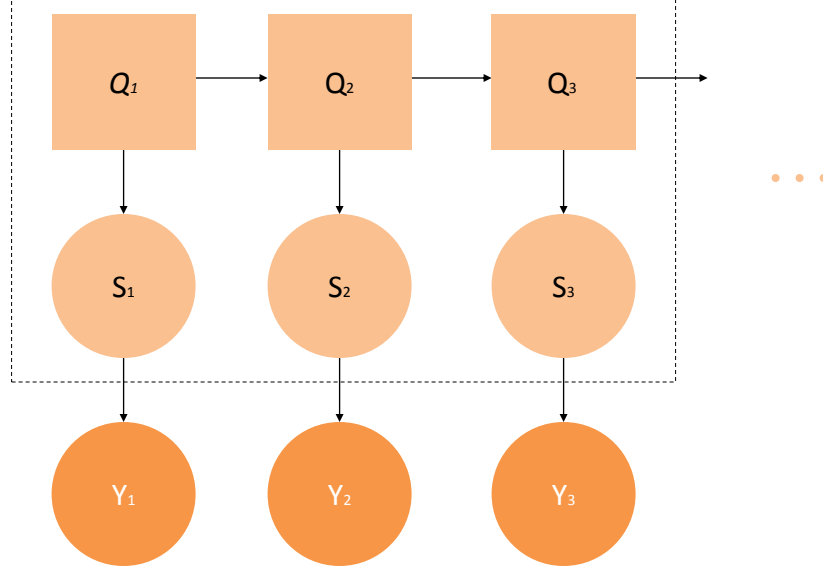


Figure F.1: Graphical representation of HMM based speech recognition in a noisy environment.

model as in the following:

$$p(\vec{Y}_1, \dots | M_i) = \int \dots \left(\int \dots p(\vec{Y}_1, \dots, \vec{S}_1, \dots, Q_1, \dots) d\vec{S}_1, \dots \right) dQ_1 \dots . \quad (\text{F.3})$$

Marginalization over the Q_i 's can be performed using the dynamic programming method called the forward algorithm [22] in a clean acoustic environment. The forward algorithm still can be used in this situation because the only different part in the induction process of obtaining $\alpha_{t+1}(Q_{t+1})$ is the noisy likelihood given the state:

$$\alpha_{t+1}(Q_{t+1}) = \left[\sum_{Q_t} \alpha_t(Q_t) p(Q_{t+1} | Q_t) \right] p(\vec{Y}_{t+1} | Q_{t+1}), \quad (\text{F.4})$$

and

$$\begin{aligned} p(\vec{Y}_t | Q_t) &= \int p(\vec{Y}_t | \vec{S}_t) p(\vec{S}_t | Q_t) d\vec{S}_t \\ &= \int \frac{p(\vec{S}_t | \vec{Y}_t) p(\vec{Y}_t)}{p(\vec{S}_t)} p(\vec{S}_t | Q_t) d\vec{S}_t. \end{aligned} \quad (\text{F.5})$$

Note that as we move forward in the forward algorithm, we marginalize the $\vec{S}_t$ only locally at time t .

F.2 Monte Carlo method

Because typically we assume that speech spectrum follows the complex Gaussian PDF and the noise spectrum is also assumed to follow the Gaussian PDF, the analytic solution of (F.5) can be obtained in a straightforward manner. However, the conventional speech recognizers are not trained by using the spectrum, but by using a nonlinearly transformed spectrum. Therefore, (F.5) can be rephrased as in the following (F.6), and the integration with the non-Gaussian PDF is not a simple matter any more. To deal with that, the Monte Carlo method is introduced as in (F.6). Note that interestingly $p(\vec{Y}_t)$ can be omitted since it is the same regardless of different models and states such that it will not affect the difference among the total likelihoods given each model. Basically, we are going to sample the $\rho(\vec{S}_t)$'s following $p(\rho(\vec{S}_t)|\vec{Y}_t)$ and calculate the sample mean of $\frac{1}{p(\rho(\vec{S}_t))}p(\rho(\vec{S}_t)|Q_t)$.

$$\begin{aligned}
p(\vec{Y}_t|Q_t) &= \int \frac{p(\rho(\vec{S}_t)|\vec{Y}_t)p(\vec{Y}_t)}{p(\rho(\vec{S}_t))}p(\rho(\vec{S}_t)|Q_t)d\rho(\vec{S}_t) \\
&= E_{p(\rho(\vec{S}_t)|\vec{Y}_t)} \left[\frac{p(\vec{Y}_t)}{p(\rho(\vec{S}_t))}p(\rho(\vec{S}_t)|Q_t) \right] \\
&\propto E_{p(\rho(\vec{S}_t)|\vec{Y}_t)} \left[\frac{1}{p(\rho(\vec{S}_t))}p(\rho(\vec{S}_t)|Q_t) \right] \\
&\approx \frac{1}{\text{Sample size}} \sum_{\rho(\vec{S}_t)} \frac{1}{\sum_{Q_t} p(\rho(\vec{S}_t)|Q_t)p(Q_t)} p(\rho(\vec{S}_t)|Q_t). \quad (\text{F.6})
\end{aligned}$$

As $\rho(\cdot)$, we normally use $\log(|\cdot|)$. The sampling problem comes down to sampling the values following $p(\log(|\vec{S}_t|)|\vec{Y}_t)$ because $p(\rho(\vec{S}_t)|Q_t)$ and $p(Q_t)$ are already given by the trained model. This sampling process can be simply replaced as the sampling process of $p(\vec{S}_t|\vec{Y}_t)$ and applying $\log(|\cdot|)$ on the generated samples. To make the sampling process as efficient as possible, we assume that the correlation among the frequency bins is zero as in the conventional assumption [9]; therefore, we sample each frequency bin by bin separately according to the following:

$$\begin{aligned}
&p(R(S_t)|R(Y_t)) \quad (\text{F.7}) \\
&\sim N\left(\frac{\text{Cov}(R(S_t), R(Y_t))}{\text{Cov}(R(Y_t))}R(Y_t), \text{Cov}(R(S_t)) - \frac{\text{Cov}(R(S_t), R(Y_t))^2}{\text{Cov}(R(Y_t))}\right),
\end{aligned}$$

where $R(\cdot)$ represents the real part of the spectrum and the imaginary part is also generated similarly. We can interpret (F.7) as following. If the variance of the noise is very small (near noise-free situation), then the generated sample will be just around the measured value of Y_t , but if the variance of the noise is very high (noise-dominating situation), then the state dependent effect is marginalized out.

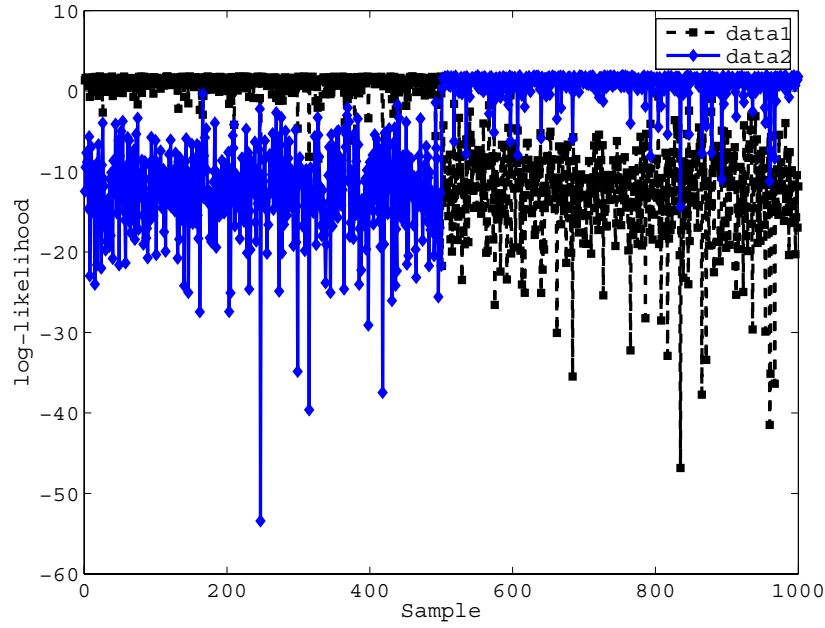
F.3 Experiment

To verify the core idea of using a Monte Carlo method, the speech recognition problem is downsized into a small classification problem of recognizing two different models with only one state. The model has been trained by using two-dimensional nonlinearly transformed $\log(| \cdot |)$ feature vectors generated from two independent Gaussians. For model 1, the first dimension has a high variance and the second dimension has a low variance. For model 2, the first dimension has a low variance and the second dimension has a high variance. As we commonly assume for the HMM based recognizer, we also assume that each model follows two-dimensional Gaussian PDF, even though that assumption is an approximation for the real PDF. (Actually this approach is working well especially when we use a Gaussian mixture model, but still it reasonably works well just by using one Gaussian with full-covariance.) By varying the level of the noise variance for the first feature dimension, we test the classification with or without the proposed method. Note that in this small problem we do not consider the time dynamics as in the HMM model for the sake of simplicity. For simplicity, we generated the test data; the first half are generated as in the training data for model 1 and the other half are generated as in the training data for model 2. Without significant noise, the likelihood of the correct model gives higher values than that of the other model. However, once the noise level becomes significantly high, the likelihood of the correct model does not always give higher values. Figure F.2 shows the case with a very low noise level, and we see that the result with or without the proposed method shows similar results. However, Figure F.3 shows the case with high noise level, and we definitely see the difference between the two cases; more specifically, the proposed model outperforms the case without it in the second half of the test data, because during the second

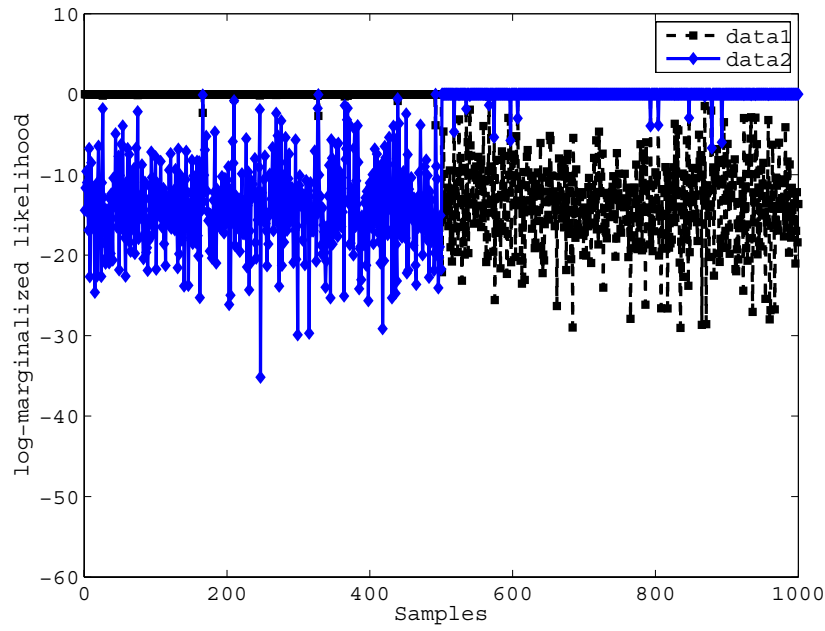
half the first dimension of the original clean data has low-level variance, which means by adding noise this part will be distorted a lot. By washing out the unreliable part by marginalizing, we can rely more on the reliable part in the feature vector, in this case the second feature component in the 2-D feature vector.

F.4 Conclusion

In this section, a Monte Carlo based speech recognition scheme has been proposed and tested in a down-sized simple problem. As we expect from the theory part, the experimental result shows that the proposed method produces noise-robust recognition (classification) results. Because the forward algorithm is shown to be utilized with the proposed method with locally generated samples, building a real speech recognizer using the proposed scheme will be part of our future work.

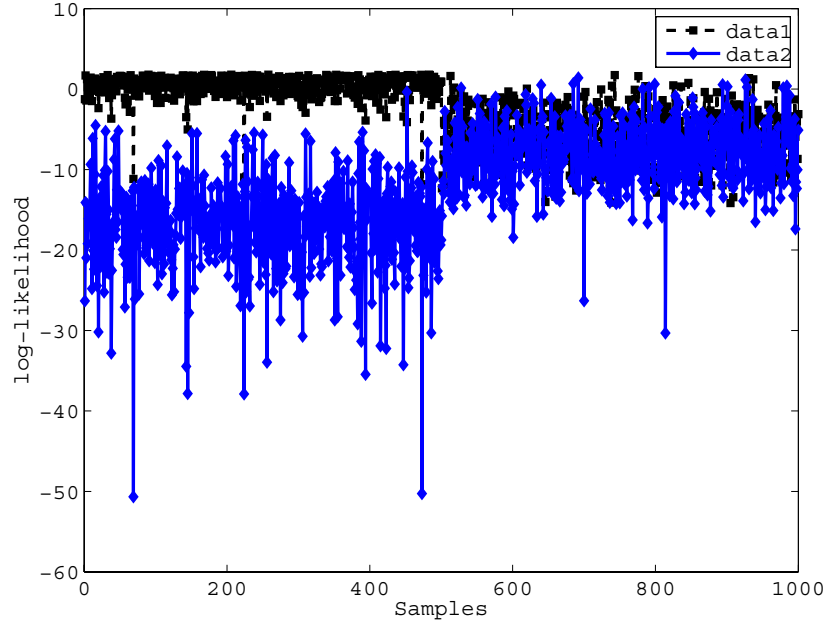


(a)

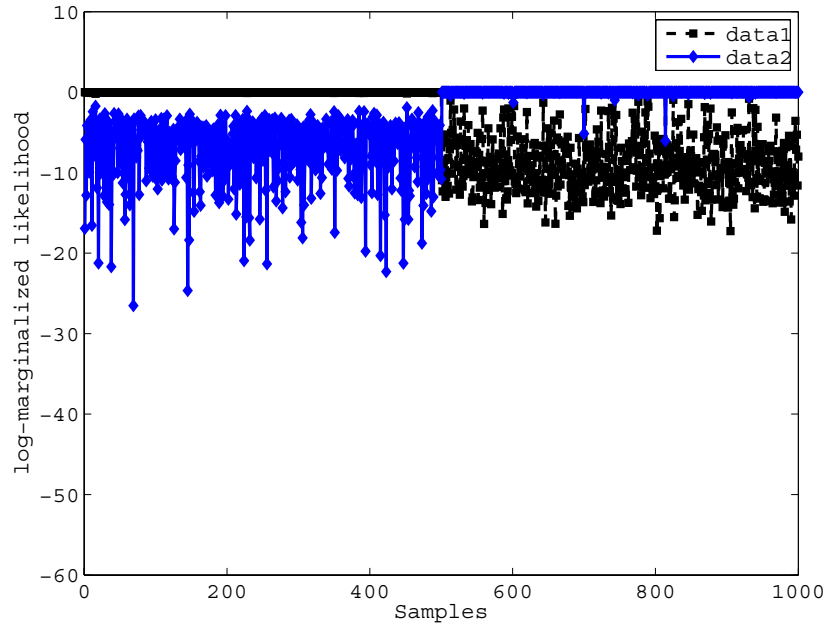


(b)

Figure F.2: (a) Log-likelihood given each model in the almost noise-free situation: blue is for model 1, red is for model 2. (b) Log-marginalized likelihood given by (F.6): blue is for model 1, red is for model 2.



(a)



(b)

Figure F.3: (a) Log-likelihood given each model in the high-level noise situation: blue is for model 1, red is for model 2. (b) Log-marginalized likelihood given by (F.6): blue is for model 1, red is for model 2.

REFERENCES

- [1] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, “Novel on-line adaptive learning algorithms for blind deconvolution using the natural gradient approach,” in *Proc. IEEE 11th IFAC SYSID-97*, 1997, pp. 1057–1062.
- [2] T.-W. Lee, *Independent Component Analysis Theory and Applications*. Boston, MA: Kluwer Academic Publishers, 1998.
- [3] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, NY: Wiley, 2001.
- [4] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, “Blind source separation based on a fast-convergence algorithm combining ICA and beamforming,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 666–678, 2006.
- [5] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, “Equalization in acoustic reverberant environment: Robustness results,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 311–319, 2000.
- [6] Y. Huang, J. Benesty, and J. Chen, “A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 882–895, 2005.
- [7] M. Delcroix, T. Hikichi, and M. Miyoshi, “Precise dereverberation using multichannel linear prediction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 430–440, December 2007.
- [8] J. S. Lim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, pp. 1586–1604, 1979.
- [9] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, pp. 1109–1121, December 1984.

- [10] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 40, pp. 725–735, April 1992.
- [11] A. Hyvarinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Computation*, vol. 11, no. 7, pp. 1739–1768, 1999.
- [12] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Communication*, vol. 34, pp. 3–12, 2001.
- [13] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, pp. 357–366, 1980.
- [14] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.
- [15] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 1996.
- [16] "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Computer Speech and Language*, vol. 23, no. 3, pp. 389–405, 2009.
- [17] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.
- [18] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.
- [19] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. Cambridge, MA: MIT Press, 1999, pp. 355–370.
- [20] *HTK Speech Recognition Toolkit*, Cambridge University Engineering Department, 2007. [Online]. Available: <http://htk.eng.cam.ac.uk/>
- [21] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009.

- [22] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, pp. 257–285, 1989.
- [23] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
- [24] B. Picinbono, “Second-order complex random vectors and normal distributions,” *IEEE Transactions on Signal Processing*, vol. 44, no. 10, pp. 2637–2640, October 1996.
- [25] R. Balan and J. Rosca, “Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase,” in *Proc. Sensor Array and Multichannel Signal Process. Workshop*, 2002, pp. 209–213.
- [26] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. Berlin, Germany: Springer, 1994.
- [27] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 1996.
- [28] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.
- [29] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 4th ed. New York, NY: Wiley, 2000.
- [30] L.-H. Kim, M. Hasegawa-Johnson, G. Potamianos, and V. Libal, “Joint estimation of DOA and speech based on EM beamforming,” in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2010, pp. 121–124.
- [31] H. L. V. Trees, *Optimum Array Processing Part IV of Detection, Estimation, and Modulation Theory*. New York, NY: Wiley, 2002.
- [32] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. T. Jr., “Measurement of correlation coefficients in reverberant sound fields,” *Journal of the Acoustical Society of America*, vol. 27, pp. 1072–1077, 1955.
- [33] X. Anguera, “Robust speaker diarization for meetings,” Ph.D. dissertation, Universitat Politècnica de Catalunya, Barcelona, Spain, 2006.
- [34] I. Tashev, *Sound Capture and Processing Practical Approaches*. West Sussex, UK: Wiley, 2009.
- [35] G. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, pp. 320–327, 1976.

- [36] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, March 1986.
- [37] Y. Bresler, S. Basu, and C. Couvreur, *Hilbert Spaces and Least Squares Methods for Signal Processing*. Class notes for ECE 513, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 2000.
- [38] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
- [39] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2002, pp. I1001–I1004.
- [40] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.
- [41] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech, Signal Processing Magazine*, vol. 5, pp. 4–24, 1988.
- [42] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 1495–1503, 1989.
- [43] M. Brandstein and D. Ward, Eds., *Microphone Arrays Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2001.
- [44] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 1984, pp. 328–331.
- [45] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice Hall, 1995.
- [46] O. Farooq and S. Datta, "Mel filter-like admissible wavelet packet structure for speech recognition," *IEEE Signal Processing Letters*, vol. 8, pp. 196–198, July 2001.
- [47] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Oxford, UK: Academic Press, 1999.

- [48] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, pp. 443–445, April 1985.
- [49] L.-H. Kim and M. Hasegawa-Johnson, "Optimal speech estimator considering room response as well as additive noise: Different approaches in low and high frequency range," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2008.
- [50] Y. Ephraim, "Speech enhancement using state dependent dynamical system model," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 1992, pp. 289–292.
- [51] X. Huang, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ: Prentice-Hall, 2001.
- [52] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press, 2001, pp. 793–799.
- [53] T. T. Kristjansson, "Speech recognition in adverse environments: A probabilistic approach," Ph.D. dissertation, University of Waterloo, Ontario, Canada, 2002.
- [54] M. J. Reyes-Gomez, B. Raj, and D. P. W. Ellis, "Multi-channel source separation by factorial HMMs," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2003, pp. I664–I667.
- [55] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 133–143, March 2004.
- [56] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley, 2006.
- [57] M. Fujimoto and S. Nakamura, "A non-stationary noise suppression method based on particle filtering and Polyak averaging," *IEICE Transactions on Information and Systems*, vol. E89-D, pp. 922–930, 2006.
- [58] B. Hajek, *An Exploration of Random Processes for Engineers*. <http://www.ifp.uiuc.edu/hajek/Papers/randomprocesses.html>: Class

notes for ECE 534, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 2008.

- [59] R. Martin, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, September 2005.
- [60] B. Chen and P. C. Loizou, “A Laplacian-based MMSE estimator for speech enhancement,” *Speech Communication*, vol. 49, no. 2, pp. 134–143, 2007.
- [61] G. Chen and V. Parsa, “HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies,” in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2006, pp. I709–I712.
- [62] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magami-Doss, and K. Saenko, “Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop,” in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, vol. 4, 2007, pp. IV621–IV624.
- [63] K. Murphy, “Bayes net toolbox for MATLAB,” 2007. [Online]. Available: <http://code.google.com/p/bnt/>
- [64] J. A. Bilmes and C. Bartels, “Graphical model architectures for speech recognition,” *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 89–100, 2005.
- [65] F. Kschischang, B. Frey, and H.-A. Loeliger, “Factor graphs and the sum-product algorithm,” *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, February 2001.
- [66] K. Murphy, “Dynamic Bayesian networks: Representation, Inference and Learning,” Ph.D. dissertation, University of California, Berkeley, CA, 2002.
- [67] A. N. Deoras and M. Hasegawa-Johnson, “A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel,” in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2004, pp. I861–I864.
- [68] M. Cooke, P. Green, C. Anderson, and D. Abberley, “Handling missing data in speech recognition,” in *Proc. Intern. Conf. on Speech and Language Processing*, 1994, pp. 1555–1558.

- [69] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study," in *Proc. Eurospeech 99*, 1999, pp. 2407–2410.
- [70] I. A. McCowan, A. C. Morris, and H. Bourlard, "Improving speech recognition performance of small microphone arrays using missing data techniques," in *Int. Conf. on Spoken Language Process.*, 2002, pp. 2181–2184.
- [71] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, 2004.
- [72] A. C. Morris, M. P. Cooke, and P. D. Green, "Some solutions to the missing feature problem in data classification, with application to noise robust ASR," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 1998, pp. 737–740.
- [73] S. J. Rennie, "Graphical models for robust speech recognition in adverse environments," Ph.D. dissertation, University of Toronto, Ontario, Canada, 2008.
- [74] B. J. Frey, L. Deng, A. Acero, and T. Kristjansson, "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *EUROSPEECH-2001*, 2001, pp. 901–904.
- [75] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.
- [76] L. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 352–362, 2002.
- [77] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 109–116, 2005.
- [78] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation*, T.-W. L. S. Makino and H. Sawada, Eds. Dordrecht, Netherland: Springer, 2007, pp. 47–78.
- [79] S. Y. Low, S. Nordholm, and R. Togneri, "Convolutive blind signal separation with post-processing," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 539–548, September 2004.

- [80] B.-J. Yoon, I. Tashev, and A. Acero, “Robust adaptive beamforming algorithm using instantaneous direction of arrival with enhanced noise suppression capability,” in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2007, pp. I–133–I–136.
- [81] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY: Wiley, 2001.
- [82] G. W. Taylor, M. L. Seltzer, and A. Acero, “Maximum a posteriori ICA: Applying prior knowledge to the separation of acoustic sources,” in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2008, pp. 1821–1824.
- [83] ITU, “ITU-T recommended p. 862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone network and speech codecs,” 2000. [Online]. Available: <http://www.itu.int/rec/T-REC-P.862/en>
- [84] H. S. Malvar, “A modulated complex lapped transform and its applications to audio processing,” in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 1999, pp. 1421–1424.
- [85] L.-H. Kim, M. Hasegawa-Johnson, and K.-M. Sung, “Generalized optimal multi-microphone speech enhancement using sequential minimum variance distortionless response (MVDR) beamforming and post-filtering,” in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2006, pp. III –III.
- [86] K. L. Bell, Y. Ephraim, and H. L. V. Trees, “A Bayesian approach to robust adaptive beamforming,” *IEEE Transactions on Signal Processing*, vol. 48, pp. 386–398, 2000.
- [87] CHIL, “Computers in the human interaction loop,” 2008. [Online]. Available: <http://chil.server.de>
- [88] T. Gustafsson, B. D. Rao, and M. Trivedi, “Source localization in reverberant environments: Modeling and statistical analysis,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 791–803, November 2003.
- [89] L.-H. Kim, M. Hasegawa-Johnson, J.-S. Lim, and K.-M. Sung, “Acoustic model for robustness analysis of optimal multipoint room equalization,” *Journal of the Acoustical Society of America*, vol. 123, pp. 2043–2053, 2008.
- [90] S. Gazor and W. Zhang, “Speech enhancement employing Laplacian-Gaussian mixture,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 896–904, September 2005.

- [91] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas Propagation*, vol. AP-30, pp. 27–34, January 1982.
- [92] M. Miyoshi and U. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 145–152, February 1988.
- [93] Y. Haneda, S. Makino, and Y. Kaneda, "Multi-point equalization of room transfer functions by using common acoustical poles," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 325–333, 1997.
- [94] S. L. Gay and J. Benesty, *Acoustic Signal Processing for Telecommunication*. Norwell, MA: Kluwer Academic Publishers, 2000.
- [95] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, pp. 926–935, August 1972.
- [96] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 1997, pp. 1167–1170.
- [97] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, November 2003.
- [98] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, pp. 1614–1626, August 2001.
- [99] I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1149–1160, 2004.
- [100] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 561–571, November 2004.
- [101] S. J. Elliott, L. P. Bhatia, F. S. Deghan, A. H. Fu, N. S. Stewart, and D. W. Wilson, "Practical implementation of low-frequency equalization using adaptive digital filters," *Journal of the Audio Engineering Society*, vol. 42, pp. 988–998, 1994.
- [102] F. Talantzis and D. B. Ward, "Robustness of multichannel equalization in an acoustic reverberant environment," *Journal of the Acoustical Society of America*, vol. 114, pp. 833–841, 2003.

- [103] S. Bharitkar and C. Kyriakakis, "Cascaded FIR filters for multiple listener low frequency room acoustic equalization," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2006, pp. V–V.
- [104] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, pp. 2677–2684, October 1999.
- [105] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: An audiovisual speech corpus in a car environment," in *Proc. Int. Conf. Spoken Language Processing*, 2004, pp. 2489–2492.
- [106] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. 11th IEEE Signal Process. Workshop*, 2001, pp. 496–499.
- [107] H. Kuttruff, *Room Acoustics*, 4th ed. New York, NY: Spon Press, 2000.
- [108] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, December 1977.
- [109] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.
- [110] D. R. Campbell, K. J. Palomaki, and G. J. Brown, "Roomsim, a MATLAB simulation of shoebox room acoustics for use in teaching and research," 2008. [Online]. Available: <http://media.paisley.ac.uk/campbell/Roomsim/>
- [111] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O Society*, vol. 12, pp. 35–50, 1992.
- [112] S. Srinivasan and D. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2130–2140, 2007.
- [113] S. Weisberg, *Applied Linear Regression*, 3rd ed. Hoboken, NJ: Wiley, 2005.
- [114] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Hoboken, NJ: Wiley, 2009.

- [115] M. A. Tanner and W. H. Wong, “The calculation of posterior distributions by data augmentation,” *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 528–540, 1987. [Online]. Available: <http://dx.doi.org/10.2307/2289457>
- [116] S. Muller and P. Masarani, “Transfer-function measurement with sweeps,” *Journal of the Audio Engineering Society*, vol. 49, pp. 443–471, 2001.
- [117] A. Farina, “Simultaneous measurement of impulse response and distortion with a swept-sine technique,” in *Proc. Audio Engineering Society 108th Convention, Preprint 5093*, 2000.