

INTEGRATING MONAURAL AND BINAURAL CUES FOR SOUND LOCALIZATION AND SEGREGATION IN REVERBERANT ENVIRONMENTS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of the Ohio State University

By

John Woodruff, M.Mus.

Graduate Program in Computer Science and Engineering

The Ohio State University

2012

Dissertation Committee:

Professor DeLiang Wang, Advisor

Professor Mikhail Belkin

Professor Eric Fosler-Lussier

Professor Nicoleta Roman

© Copyright by

John Woodruff

2012

ABSTRACT

The problem of segregating a sound source of interest from an acoustic background has been extensively studied due to applications in hearing prostheses, robust speech/speaker recognition and audio information retrieval. Computational auditory scene analysis (CASA) approaches the segregation problem by utilizing grouping cues involved in the perceptual organization of sound by human listeners. Binaural processing, where input signals resemble those that enter the two ears, is of particular interest in the CASA field. The dominant approach to binaural segregation has been to derive spatially selective filters in order to enhance the signal in a direction of interest. As such, the problems of sound localization and sound segregation are closely tied. While spatial filtering has been widely utilized, substantial performance degradation is incurred in reverberant environments and more fundamentally, segregation cannot be performed without sufficient spatial separation between sources.

This dissertation addresses the problems of binaural localization and segregation in reverberant environments by integrating monaural and binaural cues. Motivated by research in psychoacoustics and by developments in monaural CASA processing,

we first develop a probabilistic framework for joint localization and segregation of voiced speech. Pitch cues are used to group sound components across frequency over continuous time intervals. Time-frequency regions resulting from this partial organization are then localized by integrating binaural cues, which enhances robustness to reverberation, and grouped across time based on the estimated locations. We demonstrate that this approach outperforms voiced segregation based on either monaural or binaural analysis alone. We also demonstrate substantial performance gains in terms of multisource localization, particularly for distant sources in reverberant environments and low signal-to-noise ratios. We then develop a binaural system for joint localization and segregation of an unknown and time-varying number of sources that is more flexible and requires less prior information than our initial system. This framework incorporates models trained jointly on pitch and azimuth cues, which improves performance and naturally deals with both voiced and unvoiced speech. Experimental results show that the proposed approach outperforms existing two-microphone systems in spite of less prior information.

We also consider how the computational goal of CASA-based segregation should be defined in reverberant environments. The ideal binary mask (IBM) has been established as a main goal of CASA. While the IBM is defined unambiguously in anechoic conditions, in reverberant environments there is some flexibility in how one might define the target signal itself and therefore, ambiguity is introduced to the notion of the IBM. Due to the perceptual distinction between early and late reflections, we introduce the *reflection boundary* as a parameter to the IBM definition to allow target

reflections to be divided into desirable and undesirable components. We conduct a series of intelligibility tests with normal hearing listeners to compare alternative IBM definitions. Results show that it is vital for the IBM definition to account for the energetic effect of early target reflections, and that late target reflections should be characterized as noise.

Dedicated to my wife, Liz Celeste, and my children, Milo and Maeve Woodruff

ACKNOWLEDGMENTS

First and foremost, I owe my sincerest thanks to my advisor Professor DeLiang Wang. His unwavering support throughout my time at Ohio State helped me to develop as both a researcher and an individual. Dr. Wang leads by example, with a firm commitment to honest scientific exploration. He taught me sound research practices and kept me focused on worthwhile problems, and without his guidance, this work would not have been possible.

I would like to thank Professor Mikhail Belkin, Professor Eric Fosler-Lussier and Professor Nicoleta Roman for serving on my dissertation committee and for providing valuable feedback on this dissertation. I am also grateful to Professor Belkin and Professor Fosler-Lussier for participating in my candidacy exam and for offering excellent courses where I learned much about speech processing and machine learning. The studies included on ideal binary masking in reverberation could not have been completed without Professor Nicoleta Roman. I am grateful to Dr. Roman for taking the lead on finding and testing subjects and for numerous helpful discussions on both IBM processing and binaural tracking.

I would like to acknowledge my friends and lab mates in PNL. Soundarajan Srinivasan and Yang Shao were always willing to answer questions as I got started on my research. I worked closely with Yipeng Li and learned a great deal about music processing in doing so. Zhaozhang Jin was a tremendous resource for me and his work on pitch-based processing is an important component of this dissertation. Ke Hu and I began our careers at Ohio State in the same year, and he has been a wonderful ally throughout this process. I thank him for countless discussions and for providing a great example of how to conduct high-quality research. Kun Han, Arun Narayanan, Yuxuan Wang and Xiojia Zhao are inspiring to watch as they move forward with their research. It has been a joy to work alongside them, attempt to answer some of their challenging questions, and to take advantage of their expertise in many areas.

I also owe my gratitude to many friends and colleagues I have worked with over the last six years. Dr. Andrew Sabin is a great friend and in spite of being at different universities, he has consistently been a valuable resource when it comes to perception and psychoacoustics. William Hartmann, Preethi Jyothi, Dr. Jeremy Morris and Rohit Prabhavalkar were great travel companions at conferences and their expertise in speech and language processing was an asset. In particular, I would like to thank Rohit for his vital contribution to our work on binaural segregation using conditional random fields.

I would like to thank Dr. Wang, Dr. Ole Fogh Olesen and Dr. Søren Riis for making my research visit to Oticon in Copenhagen, Denmark, possible. I owe a special thanks to Dr. Ulrik Kjems and Dr. Michael Pedersen, with whom I worked closely

during my stay. I learned a tremendous amount about beamforming and multichannel signal processing from both Ulrik and Michael, and I very much appreciate their guidance on the project we conducted.

I of course owe much to my family. My sisters, Laura Jenz and Anne Anderson, and my parents, Fred and Barb Woodruff, have always given me the utmost support in any endeavor. My children, Milo and Maeve, are a wonderful and constant reminder that there is more to life than research.

Finally, I would like to thank my wife, Liz Celeste, for her patience and support over the last five years. We met in my first month at Ohio State and were married by my second year. She is the most amazing partner and mother that I can imagine, and it is difficult to find words that express my gratitude to her for everything that she has given me.

VITA

October, 7, 1978	Born in Battle Creek, MI, USA
2002	B.F.A. in Performing Arts and Technology, The University of Michigan
2004	B.S. in Mathematics, The University of Michigan
2006	M.M. in Music Technology, Northwestern University

PUBLICATIONS

J. Woodruff and B. Pardo “Active source estimation for improved source separation,” *Technical Report* NWU-EECS-06-01, Department of Electrical Engineering and Computer Science, Northwestern University, 2006.

J. Woodruff, B. Pardo, and R. Dannenberg, “Remixing stereo music with score-informed source separation,” In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2006.

J. Woodruff and B. Pardo, “Using pitch, amplitude modulation and spatial cues for separation of harmonic instruments from stereo music recordings,” *EURASIP J. Adv. Signal Proc.*, vol. 2007, pp. 1–10, 2007.

A.D. Shamma, B. Pardo, and J. Woodruff “MusicStory: an autonomous, personalized music video creator,” In *Intelligent Music Information Systems: Tools and Methodologies* J. Shen, J. Shepherd, B. Cui, L. Liu, Eds., 2007

- J. Woodruff, Y. Li and D. L. Wang, “Resolving overlapping harmonics for monaural musical sound separation using pitch and common amplitude modulation,” In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2008.
- Y. Li, J. Woodruff and D. L. Wang, “Monaural musical sound separation using pitch and common amplitude modulation,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, pp. 1361–1371, 2009.
- J. Woodruff and D. L. Wang, “On the role of localization cues in binaural segregation of reverberant speech,” In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.
- J. Woodruff and D. L. Wang, “Integrating monaural and binaural analysis for localizing multiple reverberant sound sources,” In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.
- J. Woodruff and D. L. Wang, “Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, pp. 1856–1866, 2010.
- J. Woodruff, R. Prabhavalkar, E. Fosler-Lussier and D. L. Wang, “Combining monaural and binaural evidence for reverberant speech segregation,” In *Proceedings of INTERSPEECH*, 2010.
- J. Woodruff and D. L. Wang, “Directionality-based speech enhancement for hearing aids,” In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.
- N. Roman and J. Woodruff, “Intelligibility of reverberant noisy speech with ideal binary masking,” *J. Acoust. Soc. Amer.*, vol. 130, pp. 2153–2161, 2011.
- J. Woodruff and D. L. Wang, “Binural speech segregation based on pitch and azimuth tracking,” In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.
- J. Woodruff and D. L. Wang, “Binaural localization of multiple sources in reverberant and noisy environments,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, pp. 1503–1512, 2012.

FIELDS OF STUDY

Major Field: Computer Science and Engineering

Specialization: Artificial Intelligence

TABLE OF CONTENTS

	Abstract	ii
	Dedication	v
	Acknowledgments	vi
	Vita	ix
	List of Tables	xv
	List of Figures	xvi
CHAPTER		
	PAGE	
1	Introduction	1
	1.1 Motivation	1
	1.2 Objectives	12
	1.3 Organization of Dissertation	14
2	Background	17
	2.1 Binaural Localization and Segregation	17
	2.2 Alternatives to Time-Frequency Masking	24
	2.3 DOA Estimation and Tracking	27
	2.4 Integrating Multiple Acoustic Cues	30
	2.5 Summary	33
3	Simultaneous and Sequential Organization	34
	3.1 Introduction	34
	3.2 System Overview	37
	3.3 Simultaneous Organization	39
	3.4 Binaural Processing	40
	3.4.1 Binaural Cue Extraction	42
	3.4.2 Azimuth-Dependent Likelihood Functions	43

3.4.3	Cue Weighting	46
3.5	Localization and Sequential Organization	48
3.6	Evaluation and Comparison	51
3.6.1	Training and Mixture Generation	52
3.6.2	Localization Performance	54
3.6.3	Simultaneous and Sequential Organization Performance	58
3.7	Discussion	64
4	Multisource Localization in Adverse Conditions	67
4.1	Introduction	67
4.2	Binaural Pathway	70
4.2.1	Auditory periphery and binaural feature extraction	70
4.2.2	Azimuth-dependent binaural model	70
4.2.3	Model Training	73
4.3	Monaural Pathway	76
4.3.1	Multipitch Tracking	76
4.3.2	Pitch-based Grouping	77
4.3.3	Onset/offset Based Segmentation	78
4.3.4	Onset-based Weights	79
4.4	Localization Framework	80
4.5	Evaluation Methodology	81
4.5.1	Binaural Impulse Responses	82
4.5.2	Evaluation Data	83
4.5.3	Training Data	84
4.5.4	Binaural Models	85
4.5.5	Comparison Systems	87
4.5.6	Evaluation Metrics	88
4.6	Evaluation Results	88
4.6.1	Experiment 1: Influence of Monaural Cues	89
4.6.2	Experiment 2: Comparison on KEMAR Evaluation Set	91
4.6.3	Experiment 3: Comparison on HATS Evaluation Set	96
4.6.4	Experiment 4: Source Detection	98
4.7	Discussion	101
5	Defining the Ideal Binary Mask in Reverberation	103
5.1	Introduction	103
5.2	IBM Definition	107
5.3	Experiment 1: The Effect of IBM Processing on Reverberant Speech Mixed with Speech-shaped Noise	108
5.3.1	Method	109
5.3.2	Results	113

5.4	Experiment 2: The Effect of IBM Processing on Reverberant Speech Mixed with a Competing Talker	116
5.4.1	Method	116
5.4.2	Results	117
5.5	Discussion of Experiments 1 and 2	118
5.6	Experiment 3: Interaction Between Reflection Boundary and SNR Threshold	119
5.6.1	Method	121
5.6.2	Results	123
5.7	Experiment 4: The Effect of IBM Processing on Reverberant Speech	125
5.7.1	Method	128
5.7.2	Results	131
5.8	Discussion	134
6	Binaural Detection, Localization and Segregation	137
6.1	Introduction	137
6.2	Overview	139
6.3	Feature Extraction	143
6.4	Hidden Markov Model Framework	144
6.4.1	T-F Unit Assignment	146
6.4.2	Observation Likelihood	147
6.4.3	State Predictor	150
6.4.4	Pitch and Azimuth Modules	151
6.5	Segregation	153
6.6	Evaluation Methodology	157
6.6.1	Binaural simulation	157
6.6.2	Evaluation Database	157
6.6.3	Model training	160
6.7	Evaluation	162
6.7.1	Experiment 1: Simultaneous and sequential organization	162
6.7.2	Experiment 2: Comparison with ground truth information	167
6.7.3	Experiment 3: Comparison to existing systems	173
6.7.4	Experiment 4: Detection and Localization	177
6.7.5	Analysis: Tracking 3 Pitches	180
6.8	Discussion	182
7	Contributions and Future Work	185
7.1	Contributions	185
7.2	Future Work	188
	Bibliography	190

LIST OF TABLES

TABLE	PAGE
3.1	Labeling accuracy as a function of spatial separation (in $^{\circ}$) 63
4.1	Recall (%) for the KEMAR set for alternative T-F integration methods. 90
4.2	Recall (%) and fine error ($^{\circ}$) for the KEMAR set. 95
4.3	Recall (%) and fine error ($^{\circ}$) for the HATS set. 97
6.1	Single source state transition probabilities. Rows 1, 2 and 3 list transitions out of voiced, unvoiced and inactive states, respectively. Columns 1, 2 and 3 list transitions into voiced, unvoiced and inactive states, respectively. 150
6.2	Simultaneous and sequential organization performance 165
6.3	Average Hit-FA (%) on evaluation set 1 for variants of the proposed system with ground truth (GT) and estimated (E) pitch/azimuth and ideal or azimuth-based sequential organization. Target is placed at 0° for all mixtures and performance is shown as a function of interference azimuth. 170
6.4	Avg. Δ SNR (in dB) for the proposed system and three comparison systems using measured impulse responses from four room conditions. The T_{60} for each room (in s) is listed in parenthesis. 176
6.5	Detection and localization performance of the proposed and two comparison systems on a subset of mixtures from evaluation set 1. 178

LIST OF FIGURES

FIGURE	PAGE	
1.1	Target (a), interference (b), and mixture (c) cochleagrams with corresponding IBM (d) are shown for a mixture of two simultaneous talkers. Unmasked T-F units are shown in white, masked T-F units shown in black.	4
1.2	Illustration of a localization-based grouping system. Mixture cochleagram (a), template of ITD cues for target and source azimuths (b), observed ITD values (c) and estimated binary mask (d) are shown for a mixture of two simultaneous talkers.	6
1.3	Observed ITD cues for a mixture of two simultaneous talkers in an anechoic (a) and a reverberant (b) environment.	8
1.4	Illustration of source segregation based on separate simultaneous and sequential organization stages. Target (a), interference (b) and mixture (c) for a mixture of two simultaneous talkers in reverberation. T-F regions dominated by the same underlying speaker over continuous voiced and unvoiced time intervals are shown with the same color in (d), and grouped into corresponding target and interference streams in (e).	9
2.1	Direct-path ITD and ILD cues as a function of azimuth and frequency as measured from the HRTFs of a KEMAR mannequin [67].	19
3.1	Schematic diagram of the proposed system. Cochlear filtering is applied to both the left and right ear signal of a binaural input. Monaural processing generates simultaneous streams from the <i>better ear</i> signal. Azimuth-dependent cues are extracted using a set of models trained on between-ear level and timing differences. Simultaneous streams and azimuth-dependent cues are combined in a final stage to achieve localization and sequential organization.	37

3.2	Example of multipitch detection and simultaneous organization using the tandem algorithm. (a) Cochleagram of a two-talker mixture. (b) Ground truth pitch points (solid lines) and detected pitches (circles and squares). Different pitch contours are shown by alternating between circles and squares. (c) Simultaneous streams corresponding to different pitch contours are shown with different gray levels.	41
3.3	Examples of ITD-ILD likelihood functions for azimuth 25° at frequencies of 400, 1000 and 2500 Hz. Each example shows the log-likelihood as a surface with projected contour plots that show cross sections of the function at equally spaced intervals.	45
3.4	Azimuth estimation error averaged over 200 two-talker mixtures, or 400 utterances, for various reverberation times. Results are shown using the proposed approach with and without cue weighting, and three alternative approaches.	56
3.5	Labeling accuracy of the proposed and comparison systems shown as a function of reverberation time for (a) two-talker and (b) three-talker mixtures.	61
4.1	Marginal ITD (a) and ILD (b) likelihoods, DRR prior (c), and equal contour plots of the ITD-ILD log-likelihood distributions (d) and (e) for $\theta = 70^\circ$ at 1000 Hz. The distribution in (d) uses the descending prior (squares) from (c), and the distribution in (e) uses the ascending prior (circles) from (c).	74
4.2	Recall (%) shown over the two-talker KEMAR set as a function of (a) integration time, (b) distance and (c) noise level. In (b) and (c), we show results for a 2 s integration time. The legend in (a) is applicable to all figures shown.	92
4.3	Recall (%) shown over the three-talker KEMAR set as a function of (a) integration time, (b) distance and (c) noise level. In (b) and (c), we show results for a 2 s integration time. The legend in (a) is applicable to all figures shown.	93
4.4	Recall (%) as a function of noise level for the HATS evaluation set with an integration time of 2 s.	96
4.5	Recall vs. false estimate rate for three comparison methods with unknown number of sources. Recall and false estimate rate for the case with known number of sources are shown with filled symbols.	100

5.1	Average SRTs measured for ten test conditions with SSN interference (a) and a female talker interference (b). In both (a) and (b), data is grouped according to T_{60} time. Gray-scale values indicate the processing method used. Black corresponds to the unprocessed condition ('Unp'), dark gray to IBM-DS, light gray to IBM-ER and white to IBM-R. A lower SRT corresponds to better performance. Error bars indicate 95% confidence intervals around the mean values.	115
5.2	Average percentage of correctly recognized sentences for the two unprocessed conditions and twenty-one IBM-processed conditions tested in Experiment 3. Recognition shown as a function of RC (a) and LC (b). Error bars in (a) indicate standard deviation.	126
5.3	Average percentage of correctly recognized sentences for the unprocessed condition and twenty-three IBM-processed conditions tested in Experiment 4. Recognition shown as a function of RC for T_{60} equal to 2 (a), 3 (b), and 30 s (c). Error bars indicate standard deviation.	130
5.4	Average percentage of correctly recognized sentences for the unprocessed condition and twenty-three IBM-processed conditions tested in Experiment 4. Recognition shown as a function of LC for T_{60} equal to 2 (a), 3 (b), and 30 s (c).	133
6.1	Schematic diagram of the proposed system. Cochlear filtering is applied to both the left and right ear signal of a binaural input. Correlogram features and binaural features are generated and fed to independent pitch and azimuth modules. Features along with both pitch and azimuth candidates are passed to the HMM framework. Viterbi decoding generates simultaneous streams and corresponding pitch and azimuth contours. Azimuth-based sequential organization groups simultaneous streams to form T-F masks, azimuth estimates and pitch estimates for each source.	142
6.2	Illustration of the HMM framework. Multisource states are shown with large dashed oval. Computation of observation likelihoods are illustrated inside of the dashed rectangle.	145
6.3	Example output of simultaneous organization and T-F mask estimation using the proposed system. Mixture from Set 2 with two male talkers placed at -90° and -25° in a simulated environment with T_{60} equal to 0.6 s.	156
6.4	Example mixture from Set 2 with two male talkers placed at -90° and -25° in a simulated environment with T_{60} equal to 0.6 s.	159

6.5	Example IBMs (a), estimated masks (b), ground truth (c) and estimated (d) azimuth, and ground truth (e) and estimated pitch (f) for a mixture of two male talkers placed at -90° and -25° in a simulated environment with T_{60} equal to 0.6 s. Target mask, azimuth and pitch are shown in blue, interference in green. Spurious estimates are shown in gray.	169
6.6	Example of posterior probability (MLP output) based on ITD and ILD alone (a) and based jointly on ITD, ILD and correlogram features (b) using ground truth pitch and azimuth for the target source. Mixture from evaluation set 2 with two male talkers placed at -90° and -25° in a simulated environment with T_{60} equal to 0.6 s.	171
6.7	Δ SNR the proposed algorithm and three comparison methods on evaluation sets 1 (a) and 2 (b,c).	175
6.8	F-score (%) for the proposed and comparison systems on the two- and three-talker mixtures from set 2. Results for two-talker mixtures are shown with solid lines while those for three-talker mixtures are shown with dashed lines.	179
6.9	Log likelihood ratios comparing one- vs. two-pitch states in two-pitch frames, and two- vs. three-pitch states in three-pitch frames.	183

CHAPTER 1

INTRODUCTION

1.1 Motivation

As the father of a two year old, something I find myself saying to my son with some frequency is, “listen to your mommy”. Whether she is coaxing him out of a standing position on top of a playground slide or requesting that he cease throwing objects across the room, my plea actually has little to do with listening, but rather with his behavior in response to his mother’s instruction. My expectation is that her directive was understood, and the fact that our son acquiesces now and then supports this expectation. While any parent may marvel at those instances when their toddler behaves in accordance with their wishes, we often take for granted the true listening skills that make these interactions possible. Our verbal instructions rarely reach his ears in isolation. They most often occur in combination with the sounds of other children at the playground, music playing on the stereo or perhaps his little sister’s cries during a diaper change. Fundamental to any such communication with our son

is then the capacity of his auditory system to isolate, or *segregate*, the sound of our voices from the mixture of sounds that reaches his ears.

Individuals with normal hearing excel at segregating sounds of interest from an acoustic background in order to discern relevant information. This capability facilitates awareness of our environment, such as what produced a sound and where it came from, and as the above example illustrates, allows for communication in spite of interfering sounds. Numerous existing technologies would benefit from a similar segregation capability. It is well recognized that while hearing aids provide an improvement in terms of speech audibility, the benefit in terms of intelligibility is limited and thus users are often dissatisfied in complex, multi-source settings [53]. Similarly, current cochlear implant technology limits the fidelity with which acoustic signals can be transmitted and patients can have difficulty isolating sounds of interest in difficult conditions. Automatic source segregation would also facilitate more robust speech recognition, multimedia search and information retrieval, and allow for the development of novel audio and video production tools.

Given the breadth of potential application areas, source segregation has received considerable attention from the research community. Fundamental to any approach is the need to identify acoustic properties that distinguish the source of interest from interfering sources. Speech enhancement methods often assume different statistical distributions or temporal characteristics for speech and background noise [115]. Beamforming methods capitalize on the assumption that the target source arises from

a different spatial location relative to interference and create spatially-dependent attenuation patterns in order to enhance the signal from a particular direction [15]. Many blind source separation (BSS) methods assume that sources are both separated in space and statistically independent [22]. Computational auditory scene analysis (CASA) is a promising approach to the segregation problem that utilizes the acoustic cues involved in the perceptual organization of sound by human listeners [178], such as periodicity [19, 36, 85, 98, 140, 177, 181], onset synchrony [84, 88], common amplitude modulation [82, 109], common frequency modulation [37], spectral modulation features [74, 87, 103, 161] or interaural differences [76, 119, 124, 138, 149].

Consistent with principles of auditory scene analysis (ASA) [17], the goal of CASA-based segregation is to allocate sound components of the mixed signal to individual sources. Typically, a mixture is passed through a bank of frequency selective filters, where each filter output is then divided into short time frames to create a time-frequency (T-F) representation known as a *cochleagram* [178]. A T-F unit then refers to an elemental sound component from one frame and one filter channel. The *ideal binary mask* (IBM) has been established as the main computational goal of CASA-based segregation [176]. With access to the individual source signals before they are mixed, the IBM labels those T-F units in which the signal-to-noise ratio (SNR) of a specified *target* source exceeds a predetermined threshold as 1, and labels all other T-F units as 0. Use of the IBM as a segregation goal is motivated by principles of machine perception, ASA, and by the fact that an acoustic masker can render a target stimulus inaudible within a critical band [128]. Accordingly, we refer to those

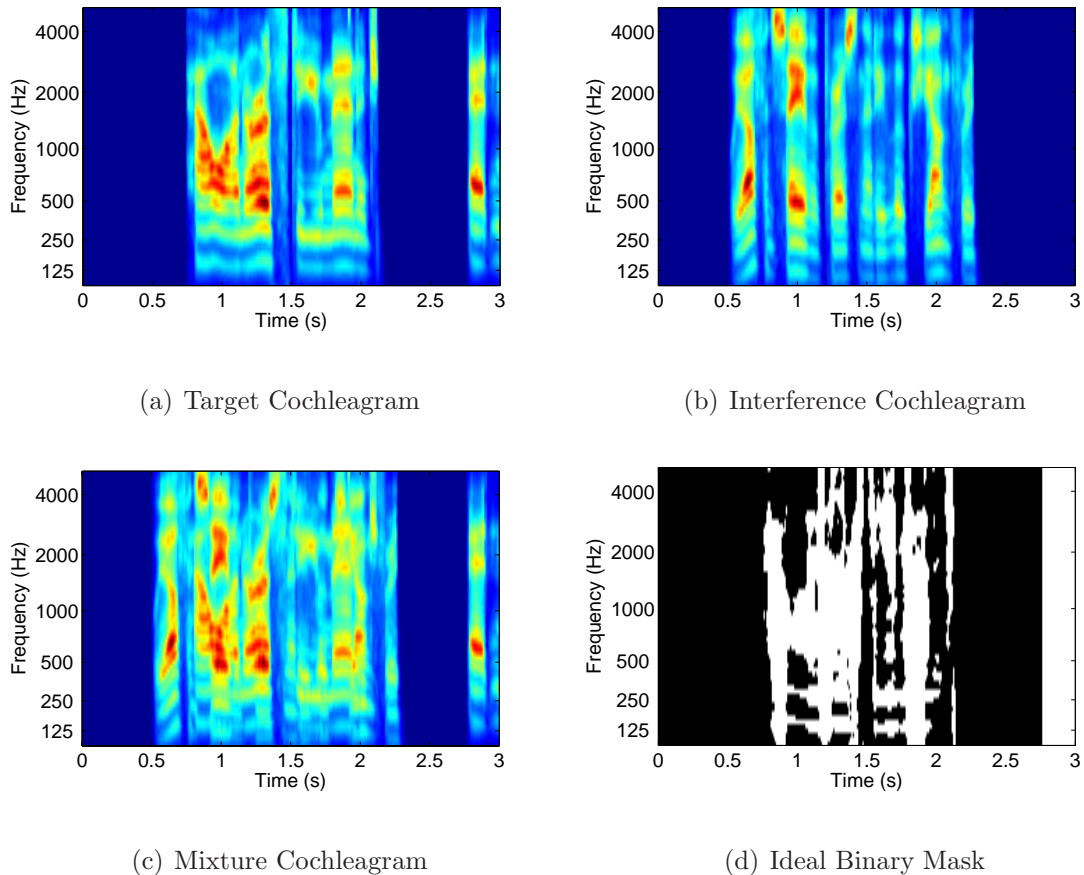


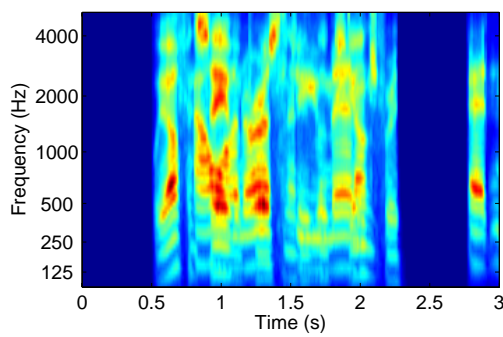
Figure 1.1: Target (a), interference (b), and mixture (c) cochleograms with corresponding IBM (d) are shown for a mixture of two simultaneous talkers. Unmasked T-F units are shown in white, masked T-F units shown in black.

T-F units above threshold as *unmasked* and units below threshold as *masked*. With the IBM representing the performance upper bound, the goal of CASA algorithms is to generate a binary T-F mask using only the observed mixture signal(s). We illustrate the generation of the IBM for a target talker mixed with an interfering talker in Figure 1.1.

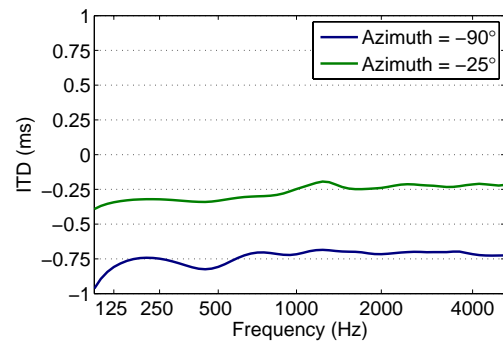
Binaural processing, where input signals resemble those that enter the two ears,

is of particular interest in the CASA field. Most binaural CASA systems measure interaural (between ear) cues to estimate the IBM using a process called localization-based grouping (see e.g. [64, 119, 138, 149]). While there are numerous differences between existing localization-based grouping systems, as will be discussed in more detail in Chapter 2, the high-level approach is as follows. First, both left and right mixture signals are transformed into the T-F domain. Interaural cues, such as *interaural time difference* (ITD) and *interaural level difference* (ILD), are then extracted from each pair (left and right ear) of T-F units. Source locations are estimated by integrating these cues across time and frequency, where often the number of sources is assumed to be known. Once source locations are identified, predetermined models or templates of interaural cues for the estimated source locations are used to identify the mixture T-F units that are consistent with the target location. We illustrate the main components of the localization-based grouping approach for a mixture of two simultaneous talkers in Figure 1.2. Figure 1.2(a) shows a mixture cochleagram. Figure 1.2(b) shows the expected ITD cues for the two source locations, while Figure 1.2(c) shows the ITD cues extracted from each pair of mixture T-F units. Note the clearly delineated boundaries in the measured ITDs, which are due to shifts in the locally dominant source. Finally, Figure 1.2(d) shows a binary mask generated by identifying those observed ITD values that are more consistent with the target location than the interference location.

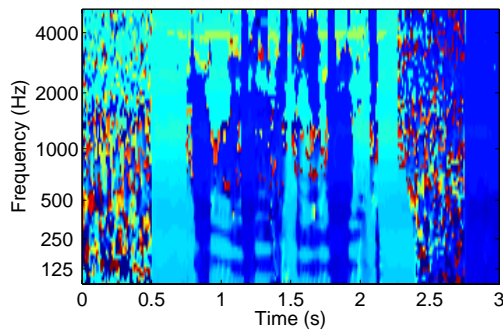
As illustrated by comparing the estimated mask in Figure 1.2(d) and the ideal mask in Figure 1.1(d), the localization-based grouping approach can be extremely



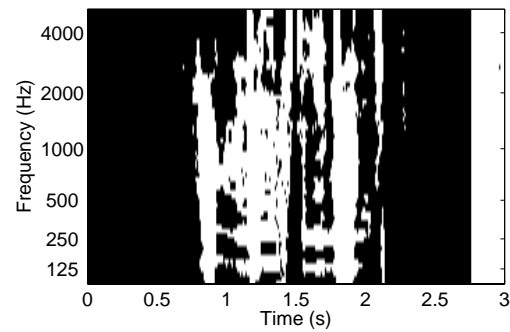
(a) Mixture Cochleagram



(b) Expected ITD Cues for Source Locations



(c) Observed ITD Cues



(d) Estimated Binary Mask

Figure 1.2: Illustration of a localization-based grouping system. Mixture cochleagram (a), template of ITD cues for target and source azimuths (b), observed ITD values (c) and estimated binary mask (d) are shown for a mixture of two simultaneous talkers.

effective in certain conditions, however, there are several shortcomings. First, since segregation is based on spatial information, much like beamforming and spatial BSS methods (discussed further in Chapter 2), this approach requires sufficient spatial separation between sources. When sources are co-located or even closely spaced, the method can fail outright. Further, substantial performance degradation is incurred in reverberant environments. Rigid surfaces reflect a sound source incident upon them, and hence, even isolated sounds reach the microphones via multiple paths in an enclosed space. This causes measured cues to deviate from predicted interaural cues, which can greatly influence the effectiveness of localization-based grouping. We illustrate this in Figure 1.3. In Figure 1.3(a) we show ITD cues extracted from the same mixture as shown in Figures 1.1 and 1.2, where source signals are simulated in an anechoic environment. In Figure 1.3(b) we show ITD cues for the same mixture in a reverberant environment. While some T-F units of the reverberant mixture still exhibit ITD near that of the anechoic mixture, many are corrupted by reflected sound energy.

Another drawback of the localization-based grouping paradigm is that it conflicts with known aspects of human auditory perception. First, this exclusively binaural approach ignores many monaural cues that are important in ASA, such as pitch, onset synchrony, amplitude modulation and spectral modulation [17]. While spatial cues do benefit segregation in some circumstances [35, 40], listeners are capable of achieving segregation in the absence of spatial cues, and performance of human listeners does not deteriorate in reverberant or co-located conditions in the way that

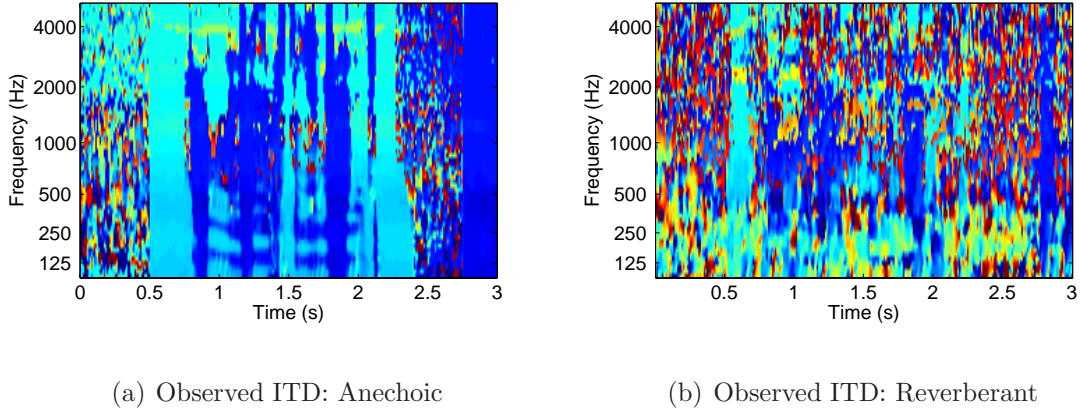
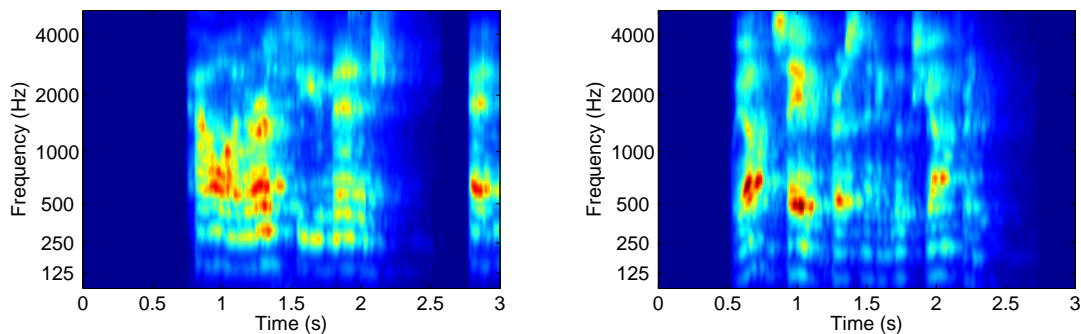


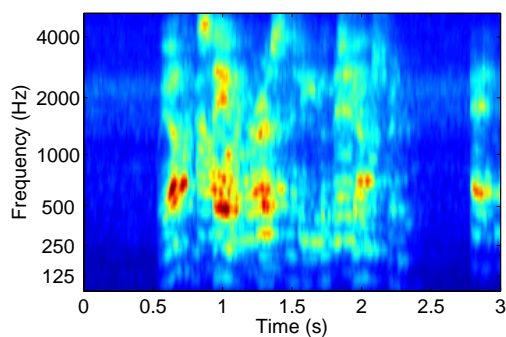
Figure 1.3: Observed ITD cues for a mixture of two simultaneous talkers in an anechoic (a) and a reverberant (b) environment.

localization-based grouping does [40]. Further, research in psychoacoustics has shown that spatial cues are relatively weak for across-frequency grouping [41, 156], particularly when compared to grouping on the basis of fundamental frequency or onset synchrony [7, 41, 49, 89, 157]. So it is not just the case that there are situations in which spatial cues provide little grouping information, but that spatial cues are likely secondary to monaural cues for *simultaneous organization*, or grouping of sound components across frequency over continuous time intervals. However, listening studies have shown that spatial cues are powerful for *sequential organization*, or grouping across time [6, 46, 47, 65, 77, 102]. Taken together, these studies suggest that an effective computational strategy should favor monaural (non-spatial) cues in terms of simultaneous organization, but rely more heavily on spatial cues for sequential organization. To illustrate the role of each grouping stage, we show an idealized simultaneous and sequential organization for a mixture of two concurrent talkers in

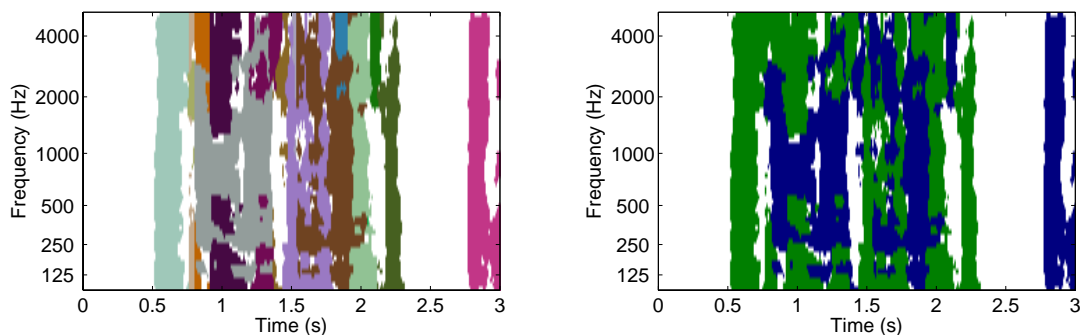


(a) Target Cochleagram

(b) Interference Cochleagram



(c) Mixture Cochleagram



(d) Simultaneous Organization

(e) Target and Interference Streams

Figure 1.4: Illustration of source segregation based on separate simultaneous and sequential organization stages. Target (a), interference (b) and mixture (c) for a mixture of two simultaneous talkers in reverberation. T-F regions dominated by the same underlying speaker over continuous voiced and unvoiced time intervals are shown with the same color in (d), and grouped into corresponding target and interference streams in (e).

reverberation in Figure 1.4. T-F regions dominated by the same underlying speaker over continuous voiced and unvoiced time intervals are shown with the same color in Figure 1.4(d). The task of sequential organization is then to group the color coded regions into separate streams corresponding to each talker, as shown in 1.4(e).

Finally, we also note that the approach to localization taken in most binaural CASA systems cannot account for observed phenomena in human localization judgements. Localization systems typically integrate binaural cues across the entire frequency range in order to estimate one or more source locations in a given time interval [11, 112, 127, 149]. While there is substantial support from the psychoacoustical literature for across-frequency integration [60, 170, 193], integration is influenced by monaural grouping cues [9, 78, 81, 170]. One well supported interpretation of this research is that the auditory system performs grouping using multiple features, and that localization judgements are formed by integrating spatial features within these larger auditory “objects” [9, 43]. Essentially, engineering systems treat localization as a means to perform segregation, and expect to localize multiple sources without explicitly segregating them, while the psychoacoustics literature suggests that localization is likely the consequence of (at least partial) segregation on the basis of multiple acoustic cues.

To address the performance limitations of existing binaural methods and to develop a computational framework that is more compatible with human auditory perception, the focus of this dissertation is on the development of algorithms to perform automatic sound segregation and localization based jointly on monaural and binaural

cues. We first propose a strategy motivated by the psychacoustical studies discussed above. T-F regions are formed on the basis of monaural cues, then localized by integrating within-region binaural cues and finally, grouped across time based on the estimated locations. We show that this approach can lead to improved segregation and localization performance relative to exclusively binaural methods. We then build on this approach to develop a framework that integrates monaural and binaural cues at a more fundamental level for simultaneous organization. Perceptual studies have shown that spatial cues can supplement monaural cues to improve simultaneous segregation [40,157], and that spatial cues can influence across-frequency grouping when monaural evidence is ambiguous (i.e. monaural evidence supports grouping a component into two competing streams) [44,45]. Further, in ideal circumstances, spatial cues alone have been shown to induce across-frequency grouping in the absence of monaural grouping cues [56]. To reconcile the observation that monaural cues are *stronger* than spatial cues for simultaneous organization, but that spatial cues may contribute when circumstances allow (e.g. low reverberation, well separated sources, ambiguous monaural cues), we learn the relative contribution of each cue through training. The algorithms presented represent an important step toward a system that, much like human listeners, can perform segregation even in the absence of useful spatial information, but that can benefit from spatial information when available. We outline the main objectives of the dissertation in the following section and conclude this chapter with a description of how the dissertation is organized.

1.2 Objectives

The primary goal of this dissertation is the development of a framework for binaural segregation and localization based jointly on multiple acoustic cues. In order to achieve a robust solution we focus on realistic acoustic environments with multiple reverberant sources and background noise. Due to the many applications in which speech is the sound of interest, we focus on mixtures of simultaneous talkers, although none of the methods discussed are necessarily restricted to speech processing. Our final system detects the unknown and time-varying number of sources across time, localizes each source, tracks the voicing characteristics of each source (including pitch) and segregates a specified target signal. To achieve this goal we focus on the following important objectives:

- *Simultaneous and Sequential Organization.* Most existing binaural CASA methods do not make a distinction between simultaneous and sequential organization and perform grouping based on spatial cues alone. As stated above, the psychoacoustics literature suggests that the role of spatial cues may differ between these grouping processes. Guided by such observations and by recent advances in pitch-based simultaneous organization, we first develop a framework to integrate pitch and azimuth cues for segregation of voiced speech. In this framework, pitch cues are used for simultaneous organization, while azimuth cues are used for sequential organization.
- *Multisource Localization in Adverse Conditions.* Multisource localization is an

important problem in many application areas, and is an important subproblem for segregation that incorporates spatial cues. The psychoacoustics literature supports the perspective that monaural cues influence localization judgements by human listeners. To analyze whether monaural cues can improve automatic source localization and to facilitate segregation based jointly on monaural and binaural cues, we extend the framework discussed above to localize multiple sources in noisy and reverberant conditions. To achieve this end we develop a novel azimuth-dependent model of binaural cues that is considerably more flexible than existing models.

- *Defining the Ideal Binary Mask in Reverberant Environments.* The IBM has been established as a main computational goal of CASA systems. In anechoic environments, the IBM can be defined unambiguously. However, in reverberant environments one can choose to treat reflections due to the target signal as either *desirable* or *undesirable*. We formalize this point by introducing a parameter to the IBM definition called the *reflection boundary*, which is a time boundary to divide early and late target reflections. We conduct a set of subjective tests to identify how the reflection boundary parameter should be set in order to improve speech intelligibility in noisy and reverberant conditions.
- *Detection, Localization and Segregation.* Our final objective is the development of a binaural segregation system based jointly on monaural and binaural cues.

We extend our initial system to handle mixtures with an unknown and time-varying number of sources and for segregation of both voiced and unvoiced speech. To do so we develop a novel hidden Markov model (HMM) framework to track the number of sources, the azimuth of each active source, and the voicing characteristics of each active source (including pitch). The framework implicitly performs simultaneous organization such that segregation of a desired source can be readily achieved by identifying either the pitch or azimuth characteristics of the target source. In this case, simultaneous organization is based jointly on pitch and azimuth cues, whereas our first systems utilize only monaural cues for simultaneous organization. As discussed above, while the psychoacoustics literature shows that monaural cues may be stronger than spatial cues for across-frequency grouping, there is evidence that spatial cues supplement grouping when they provide useful information. This final system is capable of taking full advantage of both types of cues.

1.3 Organization of Dissertation

The rest of this dissertation is organized as follows. In Chapter 2 we provide a thorough review of the literature relevant to the problems of both binaural segregation and localization. We also review existing work that has considered strategies for segregation and localization that integrate multiple acoustic cues.

In Chapter 3 we analyze the capacity of both monaural and binaural cues to perform simultaneous and sequential organization. Using an existing system for pitch-based simultaneous organization [85], we develop a framework for joint localization and segregation of voiced speech. We compare the performance of pitch-based simultaneous organization to azimuth-based simultaneous organization as a function of the level of reverberation and number of sources. We then compare the performance of a monaural sequential organization approach based on speaker-dependent features to a binaural, azimuth-based approach.

In Chapter 4 we extend the system described in Chapter 3 and provide a thorough analysis of localization performance in reverberant and noisy conditions. To this end, we develop flexible azimuth-dependent model of binaural cues and incorporate additional monaural grouping cues. We directly analyze the impact of monaural grouping on localization estimates, and compare localization performance of the proposed method to existing two-microphone methods. We also measure the robustness of the proposed method in the case when measured impulse responses are used. We finally perform one experiment to test the capacity of the proposed and comparison methods to both detect and localize sources in adverse conditions.

In Chapter 5 we consider how best to define the ideal binary mask in reverberant settings from the perspective of human speech intelligibility. We parameterize the IBM using a boundary point between early and late reflections and run a set of subject tests to compare the intelligibility of IBM processed reverberant and noisy speech. We first test three candidate IBM definitions on reverberant and noisy speech,

where we consider two different types of additive interference. We then provide a more thorough analysis of the interaction between the reflection boundary parameter and the local SNR threshold.

In Chapter 6 we develop a framework for detection, localization and segregation of speech based on pitch and azimuth cues. These problems are handled jointly using a novel hidden Markov model framework. This final system is considerably more flexible and requires less prior information than the systems presented in Chapters 3 and 4. We first perform an analysis to demonstrate improvements in simultaneous organization relative to a pitch-based approach. We then analyze segregation performance using various amounts of ideal information to understand the key factors that impact performance. We compare the proposed approach to two state-of-the-art two-microphone systems in a variety of acoustic conditions, using both simulated and measured impulse responses. We finally compare azimuth detection and localization to two binaural baseline systems.

We conclude with a discussion of the main contributions of this dissertation and outline directions for future work in Chapter 7.

CHAPTER 2

BACKGROUND

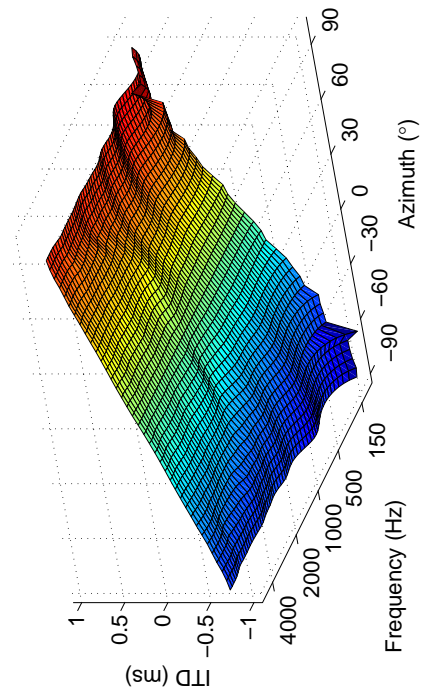
In this chapter we review existing work relevant to the problems of binaural segregation and localization. We first discuss the main approaches taken in the CASA literature. We then cover alternative array signal processing approaches to the related problems of speech enhancement, blind source separation, time difference of arrival estimation and acoustic source tracking. We conclude with a discussion of existing work that incorporates both monaural and binaural cues.

2.1 Binaural Localization and Segregation

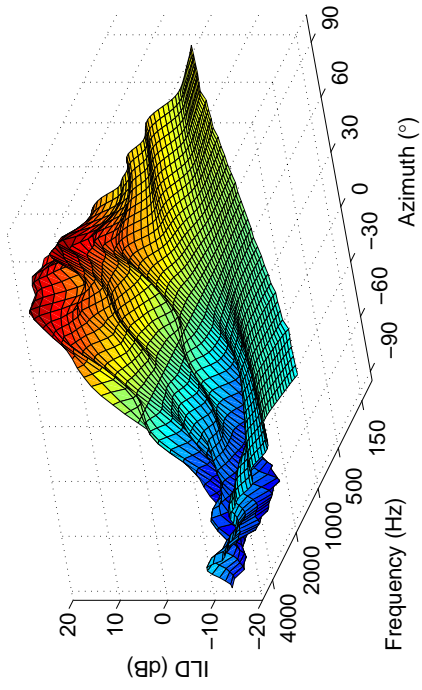
Research on binaural localization and segregation has largely been conducted along two fronts. Much of the literature focuses on the development of computational models to account for experimental data on binaural perception [167]. Alternatively, due to potential applications in binaural hearing aids, spatial sound reproduction and mobile robotics, many application-oriented binaural segregation and localization systems have also been proposed. Our primary interest is in the latter and thus we focus our attention in this area. However, as there is much overlap between the

methods used in each case, we also review some of the influential binaural models. For more thorough coverage of the literature from the behavioral perspective, see the reviews provided in [34, 167].

We begin by noting that for human listeners, sound emitted in space is altered by reflection and diffraction of the head, torso and pinnae before entering the ear canals. These effects are captured by what is known as the *head-related transfer function* (HRTF) [10]. The characteristics of the HRTF are listener dependent and change as a function of azimuth, elevation and, to some extent, distance of the source. As a result, sound emitted from a given source position in anechoic environment produces a frequency-dependent pattern of ITDs and ILDs due to the listener’s HRTFs. We refer to this azimuth- and frequency-dependent pattern of cues as *direct-path* cues, because they are measured assuming only direct propagation from source to microphone. We illustrate the direct-path ITD and ILD cues for a given listener (in this case, measured from KEMAR mannequin [67]) in Figure 2.1. Cues are shown for as a function of azimuth, between -90° and 90° , with 0° elevation. The plots show that ITD is largely frequency-independent and monotonic as a function of azimuth (with the exception of a few anomalous low frequency measurements). In fact, as one might expect, ITD can be predicted well using only the distance between ears [2, 167]. In contrast, ILD is frequency dependent and the relationship between azimuth and ILD is highly listener dependent [2]. Another characteristic that is clear from Figure 2.1(b) is that ILDs provide little azimuth-dependent information at low frequencies due to the relatively large wavelengths as compared to the size of the head. A characteristic that is not



(a) ITD



(b) ILD

Figure 2.1: Direct-path ITD and ILD cues as a function of azimuth and frequency as measured from the HRTFs of a KEMAR mannequin [67].

shown in these plots is that ITD cannot be measured unambiguously for pure tones with wavelength smaller than the distance between ears. This results in what is called *spatial aliasing* for tones with frequency above roughly 1500 Hz.

While humans are capable of localizing sounds in three dimensions, listeners exhibit the most acuity in terms of azimuth [10], and thus, both binaural models and binaural systems often focus on sources in the frontal horizontal plane (i.e. between -90° and 90° azimuth with 0° elevation). Many models of lateralization and azimuth estimation are rooted in the Jeffress hypothesis [94]. Jeffress postulated a neural mechanism that measures coincidences between time delayed versions of the signals entering each ear. A source's lateral position could then be encoded by a set of coincidence detectors, each sensitive to a different ITD. The Jeffress hypothesis is typically realized via computation of a short-time cross-correlation between the ear signals [154]. Similarly influential is the equalization-cancellation (EC) model, originally proposed to account for binaural masking level differences [59]. The EC model equalizes the signals arriving at each ear by accounting for the ITD and ILD for a given stimulus position, and then subtracts the two signals. Signals arriving from the specified position will be cancelled, while those with a different ITD and ILD will remain (or be reinforced).

While recent work has proposed several extensions to these models to better account for an increased understanding of the physiology involved in binaural perception (see e.g. [16, 33, 52, 110, 158, 168]), the majority of models focus on predicting subjective data for fairly simple stimuli in controlled acoustic conditions [167]. Part of

the reason for the divergence between literature on binaural models and literature on application-oriented systems is that approaches to machine localization and segregation must deal with complex mixtures of sound and additional distortions due to reverberation or background noise. In real-world applications, the problems of *multisource* localization and segregation are paramount and closely tied.

As discussed in Chapter 1, the localization-based grouping paradigm has been the primary approach to binaural segregation in the CASA field [64]. Again, the main strategy has been to first localize sources by integrating binaural cues, then utilize templates or models of interaural cues (such as those shown in 2.1) to identify T-F units that match the estimated target location. An early localization-based grouping system designed for concurrent speech signals was proposed by Lyons in [119]. In keeping with the Jeffress hypothesis, this system computes a running cross-correlation in individual frequency bands, dubbed the “cross-correlogram”, in which multiple ITD peaks can be identified via across-frequency summation. Real-valued functions that measure how well ITDs measured from individual bands match one of the ITD peaks are then used to perform segregation. Bodden proposed a similar approach in [11], however sub-band time lags are first mapped to azimuth based on supervised learning (see Figure 2.1(a)), and across-frequency summation is weighted based on a learned band-importance function. Roman *et al.* introduced a method to sharpen the resolution of the resulting azimuth-dependent response function [149]. Peaks in the cross-correlogram are detected and convolved with a Gaussian kernel prior to across-frequency integration to form the so-called “skeleton” cross-correlogram,

which overcomes some of the inherent limitations in terms of spatial resolution in low frequency channels. Another contribution of [149] is the use of supervised learning to perform segregation. Probabilistic models of ITD and ILD are trained for each configuration of source azimuths for both two- and three-talker conditions. After the azimuths of both target and interfering sources are identified from the skeleton cross-correlogram, the appropriate models are used to group T-F units consistent with the azimuth of the target source. A related approach is taken in [138] where again, target and interference azimuths are first identified from the skeleton cross-correlogram. T-F units consistent with the target azimuth are selected using a set of heuristics that compare correlogram values at the estimated target and interference azimuths and ensure consistency between the target azimuth and observed ILD using a set of azimuth-dependent templates based on the HRTFs of the binaural setup (again, see Figure 2.1). The segregation result is a binary T-F mask used in a missing data framework for robust speech recognition [38]. Harding *et al.* adopted the supervised training approach of [149] to generate binary T-F masks for missing data speech recognition [76]. In this case, training is performed in simulated reverberation to account for small room acoustics.

Variants of the localization-based grouping approach that avoid the use of prior training or the use of templates specific to a given microphone setup have also been proposed [93, 123, 124, 136, 152]. If given the number of sources, clusters of T-F units can be identified in the ITD, interaural phase difference (IPD), and/or ILD feature space. T-F masks can easily be generated for a given source by simply zeroing out

those units contained in different clusters. This approach is often referred to as *spatial clustering*. Many spatial clustering systems are designed for two closely spaced microphones, and thus are only applicable over a limited frequency range in the binaural case due to spatial aliasing [93, 136, 152]. Among these methods, the MESSL system of Mandel *et al.* is a state-of-the-art approach that iteratively fits Gaussian mixture models (GMMs) of IPD and ILD to the observed mixture data using an EM procedure [124]. Across frequency integration is handled by tying GMMs in individual frequency bands to a principal ITD. The system is initialized by estimating the ITD of a known number of sources. Other variants of the localization-based grouping approach have also been proposed (see e.g. [113, 144, 147])

The systems discussed so far have primarily considered localization simply as a means to perform segregation. However, binaural localization also has applications in hearing prostheses, spatial sound reproduction and mobile robotics. Several studies have explored both azimuth and elevation estimation [48, 86, 100, 101, 107, 133] or even distance estimation [118] from a binaural input. As cues for elevation and distance are influenced by the sound source more so than cues for azimuth, these studies focus on localization of an individual source. Willert *et al.* propose a method to learn so-called “activity maps” corresponding to sources presented at different azimuths [184]. An activity map captures average correlation responses and level differences, as a function of both frequency and time lag, for each trained position and a probabilistic method for localization of individual sound sources is developed based on the trained activity maps. Parametric models of ITD and ILD are proposed in [143], where the focus is on

developing a generic model for azimuth estimation based on analysis of a set of HRTFs measured from human subjects [2]. May *et al.* study the use of a GMM of ITD and ILD for azimuth estimation of multiple sources in a reverberant environment [127]. The study provides a thorough evaluation of several interaural timing cues (ITD, IPD, interaural envelope difference) and on robustness of the proposed method to mismatch between the training and testing position of the binaural microphone in a simulated room.

Very little work has dealt with the related problems of binaural tracking of moving sources or detecting the number of sources. In [148], an HMM framework based on ITD and ILD cues is proposed to estimate the number of sources and azimuth of each active source in each frame, however the system was primarily tested in anechoic conditions. The study of [52] is concerned with physiologically plausible cue extraction and integration across frequency, however the authors briefly discuss incorporating the model in a particle filter-based tracking framework, although tracking is not systematically evaluated. May *et al.* consider source detection by estimating the azimuth of the most dominant source per frame, and subsequently setting a threshold to ensure an utterance-level azimuth is only estimated for sources that were dominant in a sufficient number of frames [127].

2.2 Alternatives to Time-Frequency Masking

In keeping with monaural CASA processing, the computational goal of the binaural segregation systems discussed in the previous section is to estimate a T-F mask (most

often binary). While there is substantial evidence that a binary T-F mask is sufficient to improve speech intelligibility in adverse conditions (as will be discussed at length in Chapter 5), considerable effort has gone toward microphone-array based techniques with different enhancement objectives. We now review the main multi-microphone alternatives to T-F masking seen in the literature.

The most ubiquitous approach to array-based enhancement is beamforming, which filters and sums the received signals in order to create a spatially-dependent attenuation pattern [15]. Fixed beamformers assume a certain direction for the target signal and spatial distribution for interference energy to generate a fixed attenuation pattern. Often interference energy is assumed to be equally likely to arrive from any direction and thus attenuation increases gradually as the direction of arrival (DOA) deviates from the target direction. In order to achieve more substantial interference attenuation in a variety of conditions, beamformers have been developed to adapt across time based on the spatial characteristics of the observed signal [26, 66, 72, 183]. Provided the target direction is known or can be detected, the advantage of an adaptive beamformer is that sharper nulls can be steered in the direction of interfering sources. In principal, it is possible for a beamformer to achieve interference attenuation without any signal distortion in the direction of interest, and thus beamformers designed with this constraint are said to have a minimum-variance distortionless response (MVDR). This is in contrast to the T-F masking approach, where distortion of the target signal is unavoidable whenever attenuation is applied to a T-F unit that contains some target energy.

It is possible to further increase SNR by applying a post-filter (essentially a real-valued T-F mask) to the output of a beamformer. Ideally, the beamformer achieves some interference attenuation without distorting the target signal, then interference can be further reduced using single-channel enhancement methods. The multichannel Wiener filter (MWF), which cascades a MVDR beamformer and a single-channel Wiener post-filter, is the optimal multichannel linear filter in terms of mean-square error (MSE) under the assumption that the statistical distribution of both speech and noise are Gaussian [163, 165]. Although the technique is not new, there has been considerable interest in the MWF as an enhancement method for digital hearing aids in recent years [39, 55]. Following substantial work in single-channel speech enhancement [62, 117, 125], it has been shown that the cascade of a MVDR beamformer and a post-filter based on non-Gaussian priors is MSE optimal under alternative statistical assumptions [79].

Independent component analysis (ICA) is another well-studied alternative to T-F masking that exploits the assumption that the mixture is comprised of a known number of statistically independent sources in distinct spatial positions [22, 91]. While fundamentally relying on many of the same principles as beamforming [28], the main advantages of ICA are that no prior knowledge of the source or microphone positions are required, updates to the demixing system can be performed even if multiple sources are active simultaneously, and that higher order statistics can be used to exploit the non-Gaussianity of each source [22]. Two major drawbacks, however, are that the number of sources must be known *a priori* and that in many formulations,

the number of microphones must be equal to or greater than the number of assumed sources [22,27,91,164]. To overcome this constraint on the number of sources, methods often perform separation in individual frequency bands and then attempt to resolve the resulting across-frequency permutation ambiguity [5,58,139]. The most common approach to resolving the permutation ambiguity is to estimate the DOA of each separated signal in each frequency, then group those signals across frequency based on DOA (see e.g. [153]). Thus, although the sub-band separation mechanism may differ from the T-F masking systems presented in the previous section, there is still a close relationship to the localization-based grouping paradigm.

2.3 DOA Estimation and Tracking

In Section 2.1, we focused our attention on binaural approaches to source localization. Much like in the previous section where we discussed alternatives to binaural T-F masking, we now provide some background on array-based source localization and tracking methods that do not assume a binaural input. Such methods are closely related to many of those discussed for binaural localization. One of the primary differences being that, since no effect of the head is assumed, array-based methods often assume the principal cue for DOA estimation is the relative difference in arrival time between microphone pairs due to different propagation distances, referred to as the time difference of arrival (TDOA). Note that the term DOA is used rather than azimuth, because many array methods assume more than two microphones and thus

DOA may capture both azimuth and elevation, and the term TDOA is used rather than ITD, because no listener is assumed.

The generalized cross correlation (GCC) method is a well-known approach for TDOA estimation that assumes ideal single-path propagation of an individual source [105]. By this we mean that the model accounts only for direct propagation from the source location to the microphone and ignores any reflected energy. In GCC, the two received signals are multiplied and summed over an integration window with various time lags applied to one signal. The time lag that produces the most correlated signals is assumed to reflect the principal TDOA and, based on knowledge of the microphone spacing, can be used to estimate the DOA. Note that the cross-correlogram based methods discussed in Section 2.1 are closely related to GCC. Alternatively, one can find the time lag that minimizes the average magnitude difference function (AMDF) [42]. As the underlying model for GCC and AMDF does not account for the effect of reverberation or background noise, several methods have been proposed to increase robustness in real environments [14, 29, 51, 166]. Methods that more effectively model source propagation in reverberant environments [8, 30] or reverberant environments with background noise [54] have also been proposed.

The above methods are formulated to estimate the DOA of a single sound source, where key differences are the result of differing assumptions about environmental factors such as source propagation and background noise. For localization of multiple sound sources, methods also differ in how they handle source activity, interaction and source movement across time. If it can be assumed that sources are in a fixed

spatial position over a given time interval, a simple approach is to integrate the frame-level response of a DOA method across time and select multiple peaks in the resulting function [1, 112] (much like localization in [127, 149] discussed above). This approach implicitly assumes non-stationary sources in that it requires that different sources dominate different time periods. It can be effective with sufficient separation between sources and time integration, but can perform poorly when one source is dominant over the majority of the integration period. As was true for binaural segregation methods, there is an inherent relationship between multisource localization and separation, and as such, the separation methods discussed in the previous section implicitly extract information about the location and propagation of each separated source. The demixing filters estimated in an ICA-based approach contain the TDOA of each source [23], and focusing on localization rather than separation allows one to handle under-determined mixtures [116]. Similarly, the covariance matrices obtained to separate each source in [58] and the models used to estimate T-F masks in [124] contain estimates of source TDOAs.

While the above methods can handle localization of one or more sources, none explicitly deal with tracking the position of a source across time. Tracking is vital to many applications where sources may move, the number of sources may, or even the microphone array may move (e.g. microphones mounted on a hearing aid or mobile robot). The field of multitarget tracking is well developed [122], however most effort has gone towards tracking in SONAR or RADAR applications. Methods for tracking the position of one or more acoustic sources from a set of microphones have been

proposed in [121, 126, 171, 180, 197]. The method proposed in [121] extends the single-source methods proposed in [171, 180]. GCC-based TDOA estimates generated from multiple microphone pairs are used to construct a multitarget Bayes filter using the formalism of random finite sets [122]. Source birth, death and movement are naturally captured with a transition model and the multitarget posterior is approximated using a particle filter. The method proposed in [126] is related to the ICA-based approach of [116], but the system incorporates a statistical framework to propagate information across time and uses a “glimpsing model” to handle a time-varying number of sources. Because separation is handled independently in frequency sub-bands, it is possible to both separate and localize more sources than sensors, although similar to the separation systems mentioned above, this causes the system to be sensitive to aliasing because of across-frequency permutation ambiguity.

2.4 Integrating Multiple Acoustic Cues

In this section we discuss relevant literature that incorporates both non-spatial and spatial cues to perform either localization, tracking or segregation. We first note that out of convenience, we often refer to non-spatial cues as monaural cues. We point this out to make clear that we are not referring to monaural spatial cues due to the outer ear, which are important for three-dimensional sound localization [10].

First, while most existing approaches to array-based segregation and enhancement cannot function in a condition without spatial separation between sources, it is important to point out that such systems do not ignore monaural, source-dependent cues.

Much like single-channel speech enhancement techniques, multichannel enhancement techniques that incorporate a post-filter also take advantage of assumed statistical distributions for both speech and noise. Similarly, by maximizing independence between output signals in ICA-based separation, the optimization criteria used exploits non-Gaussian characteristics of each source [22].

Inspired by single-channel systems that incorporate prior training of spectral models for speech [151], multichannel systems have also been developed to perform separation of a known number of speech sources based jointly on spatial cues and pre-trained speech models [134, 135, 145, 182, 185]. With speaker-independent models [134, 135, 182, 185], such systems can provide a benefit relative to using spatial cues alone by enforcing consistency between the estimated signals and the trained models, but still fundamentally rely on spatial cues. Systems that incorporate speaker-dependent models [145, 182] could potentially function even with co-located sources (in which case performance would correspond to monaural processing), but require knowledge or detection of the speakers contained in the mixture. Methods that combine multichannel enhancement and speech recognition have also been proposed (see e.g. [146, 155]). In this case, knowledge of the target word sequence can be used to design an objective function for a filter and sum beamformer that maximizes the likelihood of that word sequence, leading to improved enhancement of the target signal.

The above systems incorporate either speaker-dependent or speaker-independent spectral models to complement spatial cues. Numerous studies have also considered

integrating periodicity to improve array-based localization or segregation. Several studies have noted that both pitch and TDOA are well represented in the cross-spectrum between two microphone signals and have thus proposed methods for joint estimation of both features [31,95,99,131]. However, these methods do not provide a systematic framework for dealing with multiple sources, where multiple pitches and TDOAs must be tracked across time and paired consistently with the same underlying source. The system proposed in [73] extends the “position-pitch” algorithm of [99] to the case with multiple speakers in a reverberation environment, but a large microphone array is used. In [14,32], pitch information is used to improve frame-level TDOA estimation of a dominant source in reverberation. Under the assumption that sources have strong harmonic components, a method to localize a fixed number of sources based on phase cues extracted from sinusoidal tracks is proposed in [196]. Segregation of two talkers based on joint estimation of pitch and location using a recurrent timing neural network was proposed in [195], however the authors focus on anechoic conditions. The system proposed in [194] derives separate target speech estimators based on both pitch and localization cues, where estimates are then combined based on confidence scores derived from consistency of the pitch and azimuth estimates across time. Tracking of the time delay and pitch of the dominant source is handled implicitly by the system. In [50,130,159], localization cues are used to improve pitch estimation and across-time assignment of pitch points to one of two sources. The system proposed in [120] combines both pitch and azimuth cues in a framework for fragment-based speech recognition.

2.5 Summary

The review above illustrates that both localization and segregation are well studied problems and that there is much overlap between the techniques available for each task. While systems have been developed from different perspectives, the physical cues underlying different approaches to spatial processing are largely the same - between microphone timing and level differences. Although many incorporate monaural information in some capacity, most existing approaches to binaural segregation, multichannel speech enhancement and BSS fundamentally rely on the spatial cues for each source to be sufficiently different. Little work has systematically compared the capacity of monaural and binaural cues to perform simultaneous and sequential organization or studied the potential of monaural grouping to improve multisource localization. In the next four chapters we present our proposed approaches to address these important computational problems.

CHAPTER 3

SIMULTANEOUS AND SEQUENTIAL ORGANIZATION

In this chapter we analyze the capacity of both monaural and binaural cues to perform simultaneous and sequential organization. We develop a maximum likelihood framework for joint localization and sequential organization of voiced speech that incorporates an existing system for pitch-based simultaneous organization [85]. Preliminary studies with this framework were published in [188–190].

3.1 Introduction

As outlined in the previous chapters, existing approaches to array-based speech segregation and enhancement utilize spatial cues [15] and consequently, rely on sufficient spatial separation between sources and limited reverberation and background noise. Binaural CASA systems utilize spatial cues within frameworks for localization-based grouping [64, 149], whereby one or more sound sources are first localized, then T-F units are grouped according to their level of consistency with the identified locations. As discussed in Chapter 2, these systems are closely related to spatial clustering approaches to BSS [93, 123, 136, 152].

While significant effort has been invested in increasing the robustness of localization-based grouping or spatial clustering to reverberation [76,124,138,147], these methods are limited by the discriminative power of spatial cues. In this chapter we propose an alternative framework that integrates monaural and binaural analysis to achieve robust localization and segregation of voiced speech in reverberant environments. Adopting the language of ASA [17], our proposed system uses monaural cues to achieve simultaneous organization, or grouping sound components of the mixture across frequency and short, continuous time intervals. This allows locally extracted, unreliable binaural cues to be integrated over large T-F regions. Integration over such regions enhances localization robustness in reverberant conditions and in turn, we use robust localization to achieve sequential organization, or grouping sound components of the mixture across disparate intervals of time. The proposed framework is motivated in part by the psychoacoustics literature discussed in Chapter 1, which suggests that binaural cues may play a limited role in simultaneous organization [41,156], but are important for sequential organization [6,46,47,65,77,102].

Utilizing binaural cues to handle sequential organization is attractive because monaural features alone may not be able to solve the problem. For example, in a mixture of two male speakers with a similar pitch range, pitch-based features cannot be used for grouping components that are distant in time. As a result, feature-based monaural systems have largely avoided sequential organization by focusing on short utterances of voiced speech [174] or assuming prior knowledge of the target signal's

pitch [96], or achieved sequential organization by assuming speech mixed with non-speech interference [84].

Shao and Wang explicitly addressed sequential organization in a monaural system using a model-based approach [161]. They use pitch-based monaural processing to perform simultaneous organization of voiced speech, and speaker identification to perform sequential organization of the already formed time-frequency segments. They provide extensive results on sequential organization performance in co-channel speech mixtures as well as speech mixed with non-speech intrusions. The study of [187] also utilizes speaker-dependent models to perform sequential organization of pitch estimates using a factorial hidden Markov model. Speaker-independent clustering of pitch-based T-F segments based on cepstral features is proposed in [87]. However, these studies do not address sequential organization in reverberant environments.

In the following section we provide an overview of the proposed architecture. In Section 3.3 we discuss monaural simultaneous organization of voiced speech. Section 3.4 outlines our methods for extraction of binaural cues, for calculating azimuth-dependent cues, and a mechanism for weighting cues based on their expected reliability. In Section 3.5, we formulate joint sequential organization and localization in a probabilistic framework. We assess both simultaneous and sequential organization performance, and compare the proposed system to existing methods in Section 3.6. We conclude with a discussion in Section 3.7.

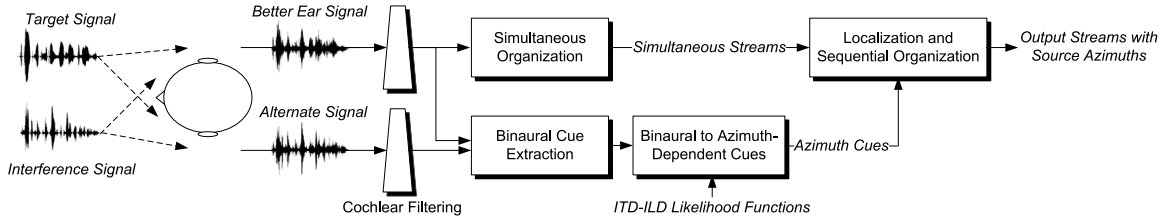


Figure 3.1: Schematic diagram of the proposed system. Cochlear filtering is applied to both the left and right ear signal of a binaural input. Monaural processing generates simultaneous streams from the *better ear* signal. Azimuth-dependent cues are extracted using a set of models trained on between-ear level and timing differences. Simultaneous streams and azimuth-dependent cues are combined in a final stage to achieve localization and sequential organization.

3.2 System Overview

The proposed system integrates monaural and binaural analysis to achieve segregation of voiced speech. A diagram is provided in Figure 3.1. The input to the system is a binaural recording of a speech source mixed with one or more interfering signals. The recordings are assumed to be made with two microphones inserted in the ear canals of a human listener or dummy head, and we will refer to the two mixture signals as the left ear and right ear signals, denoted by $u^L[n]$ and $u^R[n]$ respectively.

When processing a given mixture, the system first passes both the left and right signals through a bank of 128 gammatone filters [141] with center frequencies from 50 to 8000 Hz spaced on the equivalent rectangular bandwidth (ERB) scale [70]. As source signals are originally sampled at 16 kHz, the filterbank captures the entire speech bandwidth. Each bandpass filtered signal is divided into 20 ms time frames with a frame shift of 10 ms to create a cochleagram [178] of T-F units. A T-F unit is

an elemental sound component from one frame, indexed by m , and one filter channel, indexed by c . We denote a T-F unit as $u_{c,m}^E$ where $E \in \{L, R\}$ indicates the left or right ear signal.

In the first stage of the system, the tandem algorithm of Hu and Wang [85] is used to form *simultaneous streams* from the T-F units of the *better ear* signal. By better ear signal, we mean the signal in which the input SNR is higher, as determined from the signals before mixing. A simultaneous stream refers to a collection of T-F units over a continuous time interval that are thought to be dominated by the same source. In the CASA literature, a *stream* typically corresponds to the set of T-F units dominated by a specific source. A simultaneous stream refers to a continuous part of a stream that is grouped through simultaneous organization (i.e. through across frequency grouping and temporal continuity). The tandem algorithm generates simultaneous streams for voiced speech using harmonicity and amplitude modulation cues. Unvoiced speech presents a greater challenge for monaural systems and is not dealt with in this Chapter (see e.g. [84, 88]).

Binaural cues are extracted that measure differences in timing and level between corresponding T-F units of the left and right ear signals. A set of trained, azimuth-dependent likelihood functions are then used to map from timing and level differences to cues related to source location. Azimuth cues are integrated within simultaneous streams in a probabilistic framework to achieve sequential organization and to estimate the underlying source locations. The output of the system is a set of streams, one for each source in the mixture, and the azimuth angles of the underlying sources.

3.3 Simultaneous Organization

Simultaneous organization in CASA systems forms simultaneous streams, each of which may contain disconnected T-F segments across frequency but span a continuous time interval. We use the tandem algorithm proposed in [85] to generate simultaneous streams for voiced regions of the better ear mixture. The tandem algorithm iteratively estimates a set of pitch contours and associated simultaneous streams. In a first pass, T-F segments that contain voiced speech are identified using cross-channel correlation of correlogram responses. The correlogram is a normalized running auto-correlation performed in each frequency channel for each time frame [178]. Up to two pitch points per time frame are estimated by finding peaks in the summary correlogram, created from only the selected, voiced T-F segments. For each pitch point found, T-F units that are consistent with that pitch are identified using a set of trained multi-layer perceptrons (MLPs), one for each frequency channel. Pitch points and associated sets of T-F units are linked across continuous time intervals to form pitch contours and associated simultaneous streams using a criterion that measures pitch deviation and spectral continuity. Pitch contours and simultaneous streams that span only a single time frame are discarded. Finally, the pitch contours and associated simultaneous streams are iteratively refined until convergence.

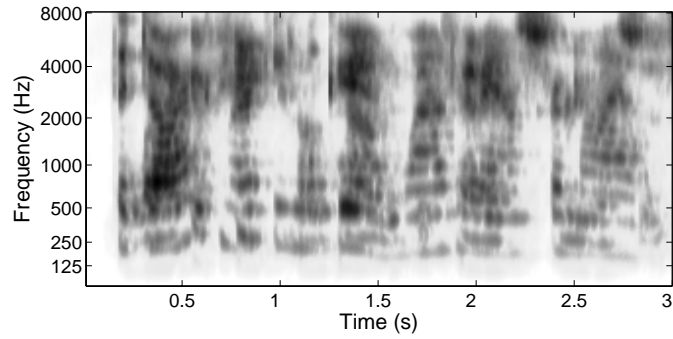
We focus on multi-talker mixtures in reverberant environments, and find that in this case the criterion used in the tandem algorithm for connecting pitch points and simultaneous streams across continuous time intervals is too liberal. For this

reason, we break pitch contours and simultaneous streams when the pitch deviation between time frames is large. Specifically, let γ_1 and γ_2 be pitch periods from the same contour in neighboring time frames. If $|\log_2(\gamma_1/\gamma_2)| > 0.08$, the contour and associated simultaneous streams are broken into two contours and two simultaneous streams. The value of 0.08 was selected on the basis of informal analysis, and was not specifically tuned for optimal performance on the data set discussed in Section 3.6.

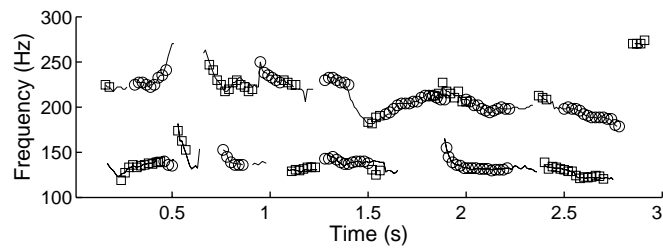
An example set of pitch contours and simultaneous streams are shown in Figure 3.2. The plots are generated using the better ear mixture of a female talker placed at -15° azimuth and a male talker placed at 30° azimuth in a reverberant environment with 0.4 s reverberation time (T_{60}). There are a total of 27 contour and simultaneous stream pairs shown. The energy of each T-F unit in the cochleagram of the mixture is shown in Figure 3.2(a). In Figure 3.2(b), detected pitch contours are shown by alternating between circles and squares, while ground truth pitch points generated from the reverberant signals prior to mixing are shown as solid lines. In Figure 3.2(c), each gray level corresponds to a separate simultaneous stream. One can see that simultaneous streams may contain multiple segments across frequency but are continuous in time.

3.4 Binaural Processing

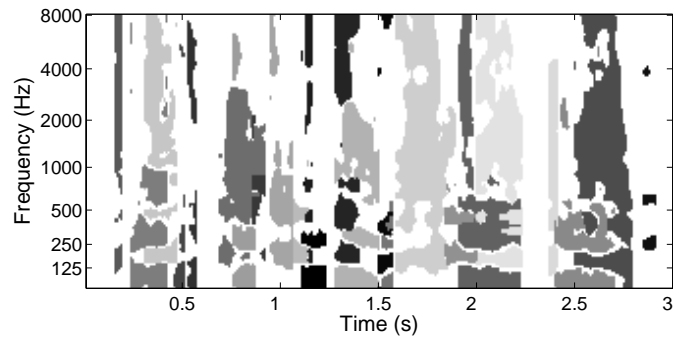
In this section we describe how binaural cues are extracted from the mixture signals and propose a mechanism to translate these cues into information about the azimuth



(a) Cochleagram



(b) Detected pitch contours



(c) Simultaneous streams

Figure 3.2: Example of multipitch detection and simultaneous organization using the tandem algorithm. (a) Cochleagram of a two-talker mixture. (b) Ground truth pitch points (solid lines) and detected pitches (circles and squares). Different pitch contours are shown by alternating between circles and squares. (c) Simultaneous streams corresponding to different pitch contours are shown with different gray levels.

of the underlying source signals. We also discuss a method to weight binaural cues according to their expected reliability.

3.4.1 Binaural Cue Extraction

As described in Chapter 2, two primary binaural cues used by humans for localization of sound sources are interaural time and level differences, or ITD and ILD, respectively. We calculate ITD in individual frequency bands by first computing the normalized cross-correlation,

$$C(c, m, \tau) = \frac{\sum_n u_{c,m}^L[n] u_{c,m}^R[n - \tau]}{\sqrt{\sum_n u_{c,m}^L[n]^2} \sqrt{\sum_n u_{c,m}^R[n - \tau]^2}}, \quad (3.1)$$

where $\tau \in [-44, 44]$ is the time lag for the correlation and summations are performed over the corresponding interval of a T-F unit. The ITD is then defined as the time lag that produces the maximum peak in the normalized cross-correlation function, or,

$$\tau_{c,m} = \arg \max_{\tau \in U} C(c, m, \tau), \quad (3.2)$$

where U denotes the set of peak lags in $C(c, m, \tau)$.

ILD corresponds to the energy ratio in dB between corresponding T-F units, calculated as,

$$\lambda_{c,m} = 10 \log_{10} \left(\frac{\sum_n u_{c,m}^L[n]^2}{\sum_n u_{c,m}^R[n]^2} \right). \quad (3.3)$$

3.4.2 Azimuth-Dependent Likelihood Functions

As discussed in Section 2.1, sound emitted from a given source position in anechoic environment produces a frequency-dependent set of ITDs and ILDs due to the listener’s HRTFs. Again, we refer to this azimuth- and frequency-dependent pattern of cues as *direct-path* cues (see Figure 2.1). In order to effectively integrate interaural information across frequency for a given position, the direct-path cues must be taken into account. Further, integration of ITD and ILD cues extracted from reverberant and multisource mixtures should account for deviations from the direct-path cues.

To alleviate some of the complexity associated with multisource localization and segregation, we restrict sound sources to be in front of the listener with 0° elevation. As a result, source localization reduces to azimuth estimation in the interval $[-90^\circ, 90^\circ]$. To translate from raw ITD-ILD information to azimuth, we train a joint ITD-ILD likelihood function, $P_c(\tau, \lambda|\theta)$, for each azimuth, θ , and frequency channel, c . Likelihood functions are trained on single-source speech in various room configurations and reverberation conditions using kernel density estimation [162]. The room size, listener position, source distance and reflection coefficients of the wall surfaces are randomly selected from a pre-defined set of 540 possibilities (see Section 3.6.1 for more details). Following Roman *et al.* [149], we use Gaussian kernels for density estimation and choose smoothing parameters using the least-squares cross-validation method [162]. For a more detailed description, see [149].

An ITD-ILD likelihood function is generated for each of 37 azimuths, $[-90^\circ, 90^\circ]$

spaced by 5° , and for each of the 128 frequency channels. With these functions, we can translate the ITD-ILD values measured from a given T-F unit pair into an azimuth-dependent response. Due to reverberation, we do not expect the maximum of the response for each T-F unit pair to be a good indication of the dominant source’s azimuth, but hope that a good indication of the dominant source’s azimuth emerges through integration over a simultaneous stream.

The set of likelihood distributions for a specific azimuth captures both the frequency-dependent pattern of ITDs and ILDs for that azimuth and the multi-peak ambiguities present at higher frequencies where signal wavelengths are shorter than the distance between microphones. Each distribution has a peak corresponding to the direct-path cues for that angle, but also captures common deviations from the direct-path cues due to reverberation. We show three distributions in Figure 3.3 for azimuth 25° . Note that, in addition to the above points, the azimuth-dependent distributions capture the complementary nature of localization cues [10] in that ITD provides greater discrimination between angles at lower frequencies (note the large ILD variation in the 400 Hz example) and ILD improves discrimination between angles at higher frequencies where spatial aliasing hinders discrimination by ITD alone.

Our approach is adapted from the one proposed in [149]. In that system two ITD-ILD likelihood functions are trained for each frequency channel, $P_c(\tau_{c,m}, \lambda_{c,m}|H_0)$ and $P_c(\tau_{c,m}, \lambda_{c,m}|H_1)$, where H_0 denotes the hypothesis that the target signal is stronger than the interference signal, and H_1 that the target is weaker. The distributions $P_c(\tau_{c,m}, \lambda_{c,m}|H_0)$ and $P_c(\tau_{c,m}, \lambda_{c,m}|H_1)$ are trained for each target/interference angle

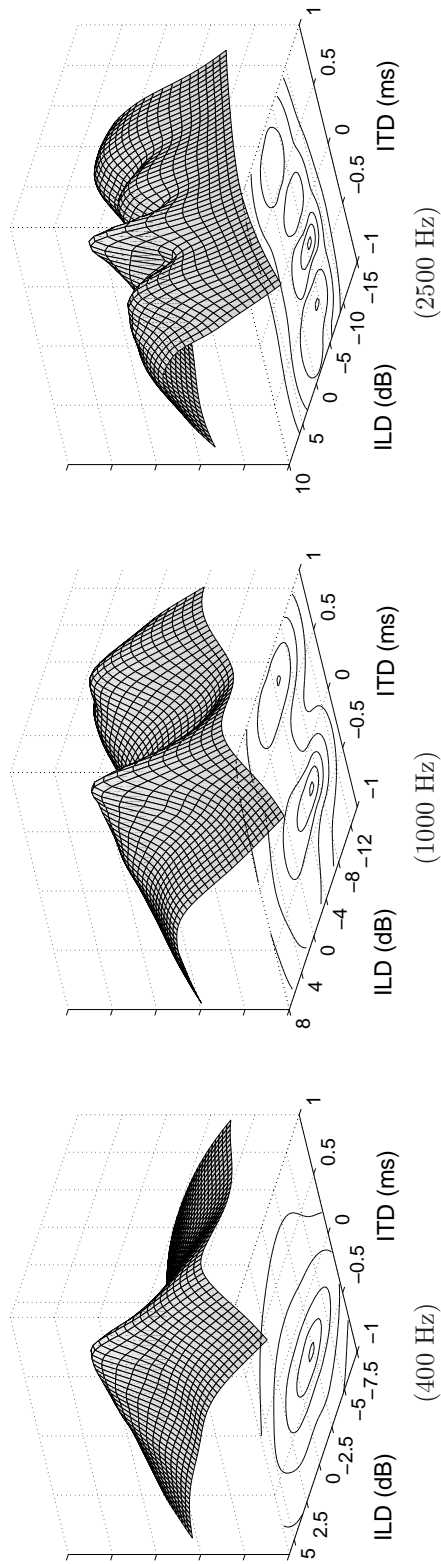


Figure 3.3: Examples of ITD-ILD likelihood functions for azimuth 25° at frequencies of 400, 1000 and 2500 Hz. Each example shows the log-likelihood as a surface with projected contour plots that show cross sections of the function at equally spaced intervals.

configuration. The ITD search space is limited around the expected direct-path target ITD in both training and testing to avoid the multi-peak ambiguity in higher frequency channels. For a test utterance, the azimuths of both target and interference sources are estimated, the appropriate set of likelihood distributions is selected and the maximum *a posteriori* decision rule is used to estimate a binary mask for the target source.

There are two primary reasons for altering the method in [149] to the one proposed here. First, our proposed approach lowers the training burden because likelihood functions are trained for each angle individually, rather than as combinations of angles. Second, the fact that we do not limit the ITD search space in training allows us to use the likelihood functions in estimation of the underlying source azimuths, rather than requiring a preliminary stage to estimate the angles. Because we do not limit the ITD search space, our approach does not attempt to resolve the multi-peak ambiguity inherent in high frequency ITD calculation at the T-F unit level. For frequency channels in which the wavelength of the signal is shorter than the spacing between microphones, multiple peaks are captured by the likelihood functions (see Figure 3.3). Spatial aliasing in these channels is naturally resolved by integrating across frequency within a simultaneous stream.

3.4.3 Cue Weighting

In reverberant environments, many T-F units will contain cues that differ significantly from direct-path cues. Although these deviations are incorporated in the training of

the ITD-ILD likelihood functions described above, including a weighting function or cue selection mechanism that indicates when an azimuth cue should be reliable can improve localization performance. Motivated by the *precedence effect* [111], we incorporate a simple cue weighting mechanism that identifies strong onsets in the mixture signal. When a large increase in energy occurs, and shortly thereafter, the azimuth cues are expected to be more reliable. We therefore generate a weight, $w_{c,m}^E$, associated with $u_{c,m}^E$ that measures the change in signal energy over time. We first extract the signal envelope for each frequency channel of the left and the right signal by squaring and passing each sub-band through a first-order IIR filter with a time constant of 10 ms. The resulting envelope signals are then decimated to a sample rate of 100 Hz (to match the frame rate of the other processing stages). Finally we compute,

$$w_{c,m}^E = \frac{e_c^E[m] - e_c^E[m-1]}{e_c^E[m-1]}, \quad (3.4)$$

as the weight for unit $u_{c,m}^E$. Here $e_c^E[m]$ denotes the sample of the decimated envelope signal corresponding to $u_{c,m}^E$.

In preliminary testing, we have found better performance by keeping only those weights above a specified threshold. The difficulty with a fixed threshold however, is that one may end up with a simultaneous stream with no unit above the threshold. To avoid this we set a threshold for each simultaneous stream so that the set of T-F units exceeding the threshold retain 25% of the signal energy in the simultaneous stream. $w_{c,m}$ is set to 0 for all T-F units below the selected threshold. We have found that

the system is not particularly sensitive to the value of 25% and that values between about 15% and 40% give similar performance in terms of localization accuracy.

Alternative selection mechanisms have been proposed in the literature [32, 63, 186]. Faller and Merimaa proposed *interaural coherence* as a cue selection mechanism [63], although in preliminary experiments we found the proposed method to outperform selection methods based on interaural coherence. The method proposed in [186] uses ridge regression to learn a finite-impulse response filter that predicts localization precision for single-source reverberant speech in stationary noise. This method essentially identifies strong signal onsets, as does our approach, but requires training. The study in [32] finds that a precedence motivated cue weighting scheme performs similarly to two alternatives on a database of two-talker mixtures in a small office environment.

3.5 Localization and Sequential Organization

As described above, the first stage of the system generates simultaneous streams for voiced regions of the better ear mixture and extracts azimuth-dependent cues from all T-F unit pairs. In this section we describe the source localization and sequential organization process. The goal of sequential organization is to generate a target or interference label for each of the simultaneous streams, thereby grouping the simultaneous streams across time. Our approach jointly determines the source azimuths

and sequential organization (simultaneous stream labeling) that maximizes the likelihood of the binaural data. This approach is inspired by the model-based sequential organization scheme proposed in [160].

Let K be the number of sources in the mixture, and I be the number of simultaneous streams formed using monaural analysis. Denote the set of all possible azimuths as Θ and the set of simultaneous streams as $G = \{g_1, g_2, \dots, g_I\}$, where g_i is an individual simultaneous stream, or, a collection of T-F units. Let Y be the set of all K^I sequential organizations, or labelings, of the set G and y be a specific organization. We seek to maximize the joint probability of a set of angles and a sequential organization given the observed data, Z . This can be expressed as,

$$\hat{\theta}_0, \dots, \hat{\theta}_{K-1}, \hat{y} = \arg \max_{\theta_0, \dots, \theta_{K-1} \in \Theta, y \in Y} P(\theta_0, \dots, \theta_{K-1}, y | Z). \quad (3.5)$$

For simplicity, assume that $I = 2$ and apply Bayes rule to get,

$$\begin{aligned} \hat{\theta}_0, \hat{\theta}_1, \hat{y} &= \arg \max_{\theta_0, \theta_1 \in \Theta, \theta_0 \neq \theta_1, y \in Y} \frac{P(Z | \theta_0, \theta_1, y) P(\theta_0, \theta_1, y)}{P(Z)}, \\ &= \arg \max_{\theta_0, \theta_1 \in \Theta, \theta_0 \neq \theta_1, y \in Y} P(Z | \theta_0, \theta_1, y), \end{aligned} \quad (3.6)$$

assuming that all angle combinations and sequential organizations are equally likely (with the exception that $P(\theta_0 = \theta_1) = 0$). We note that the assumption that all sequential organization are equally likely (i.e. $P(y)$ is uniform) is made to derive a computationally efficient solution and does not necessarily hold. We provide more discussion regarding this point in Section 3.7.

Now, let G_0 be the set of simultaneous streams associated with θ_0 and G_1 be the set of simultaneous streams associated with θ_1 by y . Using ITD and ILD as the

observed mixture data, and assuming independence between simultaneous streams and between T-F units of the same simultaneous stream, we can express Equation (3.6) as,

$$\hat{\theta}_0, \hat{\theta}_1, \hat{y} = \arg \max_{\theta_0, \theta_1 \in \Theta, \theta_0 \neq \theta_1, y \in Y} \left[\prod_{g_i \in G_0} \prod_{u_{c,m} \in g_i} P_c(\tau_{c,m}, \lambda_{c,m} | \theta_0) \cdot \prod_{g_j \in G_1} \prod_{u_{c,m} \in g_j} P_c(\tau_{c,m}, \lambda_{c,m} | \theta_1) \right] \quad (3.7)$$

where P_c denotes a probability function defined for frequency channel c (see Section 3.4.2). Note that we have dropped the superscript $E \in \{L, R\}$ for T-F unit notation since monaural grouping is performed over the better ear signal, which is mixture dependent.

One can express the above equation as two separate equations that can be solved simultaneously in one polynomial-time operation as,

$$\hat{y}_i = \arg \max_{y_i \in \{0,1\}} \left[\sum_{u_{c,m} \in g_i} \log(P_c(\tau_{c,m}, \lambda_{c,m} | \theta_{y_i})) \right], \quad (3.8)$$

$$\hat{\theta}_0, \hat{\theta}_1 = \arg \max_{\theta_0, \theta_1 \in \Theta, \theta_0 \neq \theta_1} \left[\sum_{i=1}^I \sum_{u_{c,m} \in g_i} \log(P_c(\tau_{c,m}, \lambda_{c,m} | \theta_{\hat{y}_i})) \right], \quad (3.9)$$

where \hat{y}_i denotes the label assigned to g_i . The key assumption in moving to Equations (3.8) and (3.9) is the independence between simultaneous streams expressed in Equation (3.7).

Incorporating the weighting parameter defined in Section 3.4.3, Equations (3.8) and (3.9) become,

$$\hat{y}_i = \arg \max_{y_i \in \{0,1\}} \left[\sum_{u_{c,m} \in g_i} w_{c,m} \log(P_c(\tau_{c,m}, \lambda_{c,m} | \theta_{y_i})) \right], \quad (3.10)$$

$$\hat{\theta}_0, \hat{\theta}_1 = \arg \max_{\theta_0, \theta_1 \in \Theta, \theta_0 \neq \theta_1} \left[\sum_{i=1}^I \sum_{u_{c,m} \in g_i} w_{c,m} \log(P_c(\tau_{c,m}, \lambda_{c,m} | \theta_{\hat{y}_i})) \right]. \quad (3.11)$$

For the case with $K > 2$, use $y_i \in \{0, 1, \dots, K-1\}$ rather than $y_i \in \{0, 1\}$ in Equation (3.10) and $\{\theta_0, \theta_1, \dots, \theta_{K-1}\} \in \Theta, \theta_i \neq \theta_j$ in Equation (3.11). The complexity of the search space is $I \binom{|\Theta|}{K}$, which is reasonable when the number of sources of interest is relatively small and the size of the azimuth space is moderate. In our experiments in Section 3.6, $|\Theta| = 37$ and $K \leq 3$. We provide a more thorough discussion regarding search complexity and independence assumptions in Section 3.7.

3.6 Evaluation and Comparison

In this section we evaluate source localization, localization-based sequential organization, and segregation of voiced speech using the proposed integration of monaural and binaural processing. We analyze localization performance with and without the cue weighting mechanism discussed in Section 3.4.3 and compare the proposed method to two existing methods in various reverberation conditions. We evaluate sequential organization performance in various reverberation conditions through comparison to a model-based approach and to a method that incorporates prior knowledge. Finally,

we evaluate voiced speech segregation of the full system through comparison to an exclusively binaural approach and to identify the conditions in which integration of monaural and binaural analysis can outperform binaural analysis alone.

3.6.1 Training and Mixture Generation

We use the ROOMSIM package [25] to generate impulse responses that simulate binaural input at human ears. This package uses measured HRTF data from a KE-MAR mannequin [67] in combination with the image method for simulating room acoustics [3]. We generate a training and an evaluation library of binaural impulse responses (BIRs) for 37 direct sound azimuths between -90° and 90° spaced by 5° , and 7 T_{60} times between 0 and 0.8 s. For the training library, 3 room size configurations, 3 source distances from the listener (0.5, 1 and 1.5 m) and 5 listener positions in the room are used. For the evaluation library, 2 room size configurations (different from those in training), 3 source distances from the listener (same as those in training) and 2 listener positions (different from those in training) are used.

In order to train the ITD-ILD likelihood distributions, speech signals randomly selected from the 8 dialect regions in the training portion of the TIMIT database [68] are upsampled to 44.1 kHz and convolved with a randomly selected BIR from the training library (for a specified angle). Training is performed over 100 reverberant signals for each of the 37 azimuths (see Section 3.4.2).

For evaluation mixtures we select target and interference speech signals from the TIMIT database, upsample the signals to 44.1 kHz, pass the signals through a BIR

from the evaluation library for a desired azimuth and T_{60} time, and sum the resulting binaural target and interference signals to create a binaural mixture. We generate 200 two-talker mixtures and 200 three-talker mixtures for each of the reverberation conditions. Room dimensions, source distance and listener position are randomly selected and applied to all sources for each mixture. For the two-talker mixtures, source azimuths are selected randomly to be between 10° and 125° apart. For the three-talker mixtures, source azimuths are selected randomly to be at least 10° apart. The average azimuth spacing over each set of two-talker mixtures is 53° , whereas the average spacing from the target source to the closest interference source is 41° for each set of three-talker mixtures. Speech utterances, azimuths and room conditions remain constant across different T_{60} times. Only the reflection coefficient of the wall surfaces was changed to achieve the selected T_{60} . The SNR of each mixture is set to 0 dB using the dry, monaural TIMIT utterances. This results in better ear mixtures that average 2.8 dB in anechoic conditions down to 1 dB in 0.8 s T_{60} for the two-talker case, and -0.4 dB in the anechoic mixtures down to -1.6 dB in 0.8 s T_{60} for the three-talker case. Mixture lengths are determined using the target utterance with the interference signals either truncated or concatenated with themselves to match the target length. In order to make a comparison to the model-based approach (discussed further in Section 3.6.3), the speakers used for the test mixtures are drawn from the set of 38 speakers in the DR1 dialect region of the TIMIT training database.

3.6.2 Localization Performance

In this section we analyze the localization accuracy of the method described in Section 3.5. Specifically, we measure average azimuth estimation error with and without cue weighting. We also compare localization performance to two existing methods for localization of multiple sound sources, as proposed in [51, 112], and to an exclusively binaural system that incorporates the azimuth-dependent likelihood functions described in Section 3.4.2, but labels each T-F unit independently.

The approach proposed by Liu et al. in [112], termed the *stencil filter*, performs coincidence detection for each frequency bin and time frame and counts the detected ITD as evidence for a particular azimuth if it falls along the azimuth’s “primary” or “secondary” traces. The primary trace is simply the predicted ITD for that angle, while the secondary traces are due to ambiguity at higher frequencies. For comparison on the database described, some changes were necessary to account for the (somewhat) frequency-dependent nature of ITDs as detected by a binaural system and the discrete azimuth space. Further, because angles are assumed constant over the length of the mixture, azimuth responses from the stencil filter were integrated over all time frames for added accuracy and the two most prominent peaks were selected as the underlying source angles.

The system proposed in [51], denoted SRP-PHAT, is a steered beamformer that incorporates the phase transform (PHAT) weighting to increase robustness in reverberant conditions. Our implementation measures the response power over 20 ms time

frames that overlap by 50%. We integrate over frequencies up to 8 kHz, since the TIMIT sources do not have energy beyond this frequency, sum the responses across time and select the K most prominent peaks as the source azimuths. We consider the same set of azimuths used in the proposed method and use the direct-path interaural phase differences of the KEMAR HRTFs for beam steering.

The exclusively binaural system treats each T-F unit independently and jointly estimates source azimuths and time-frequency masks. Specifically, for a given set of angle hypotheses $\{\hat{\theta}_0, \dots, \hat{\theta}_{N-1}\}$, each T-F unit is given a source assignment, $y_{c,m}$, using the azimuth-dependent likelihood functions. The azimuth set that maximizes the likelihood after integration over all T-F units is selected. This can be expressed with a slight alteration of Equations (3.8) and (3.9),

$$\hat{y}_{c,m} = \arg \max_{y_{c,m} \in \{0, \dots, N-1\}} P_c(\tau_{c,m}, \lambda_{c,m} | \theta_{y_{c,m}}), \quad (3.12)$$

$$\hat{\theta}_0, \dots, \hat{\theta}_{N-1} = \arg \max_{\theta_0, \dots, \theta_{N-1} \in \Theta} \sum_{u_{c,m}} P_c(\tau_{c,m}, \lambda_{c,m} | \theta_{\hat{y}_{c,m}}). \quad (3.13)$$

This approach is similar in spirit to [93, 123, 124] in that source azimuths and time-frequency masks are jointly estimated, allowing localization cues to be integrated over a subset of T-F units in the mixture. One key difference is that the binaural system presented here takes advantage of the pre-trained, non-parametric likelihood functions whereas [93, 123, 124] fit parametric models directly to the observed mixture. It is important to note that we do not incorporate the voiced simultaneous streams in any way, thus unlike the proposed system, the binaural localization system makes use of both voiced and unvoiced speech.

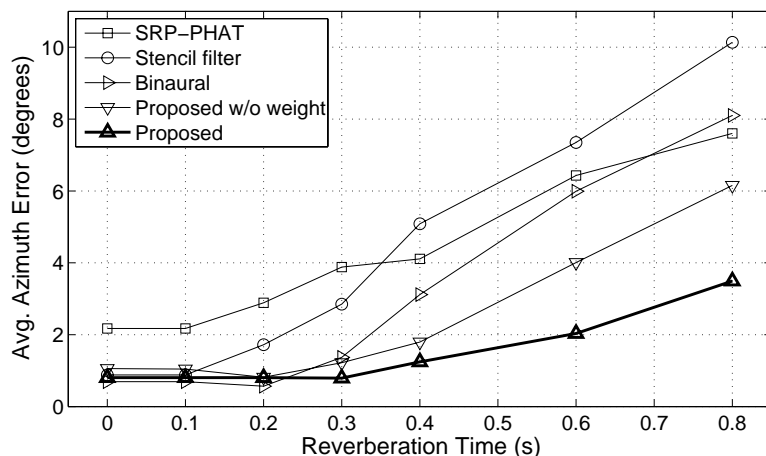


Figure 3.4: Azimuth estimation error averaged over 200 two-talker mixtures, or 400 utterances, for various reverberation times. Results are shown using the proposed approach with and without cue weighting, and three alternative approaches.

Average azimuth error on the two-talker mixtures is shown in Figure 3.4. Estimation is performed for 400 source signals (2 in each of the 200 two-talker mixtures) and for 7 T_{60} times. The results indicate that including weights associated with signal onsets improves azimuth estimation of the proposed method when significant reverberation is present. We can also see that both proposed methods outperform the existing methods for T_{60} of 300 ms or larger. The improvement relative to the stencil filter method averages 5.18° over the T_{60} range of 400 ms to 800 ms, 3.74° relative to the SRP-PHAT approach, and 3.51° relative to the exclusively binaural approach.

The difference in performance between the methods is largely captured by how well they localize *both* sources in the mixtures. If we consider only the source that was localized with the most precision, the average azimuth error of all methods was

near or below 2° in all T_{60} times. However, the proposed method was able to localize the second source with far more accuracy than the alternative methods. When T_{60} ranges from 400 ms to 800 ms, the proposed method decreased the average azimuth error of the less accurately localized source by between 60% and 70% relative to the alternative systems.

Performance on the three-talker mixtures followed the same trends, with the proposed system providing an accuracy improvement of 33%, 41% and 48% over the binaural, SRP-PHAT and stencil filter methods, respectively, over the T_{60} range of 300 ms to 800 ms. The proposed system achieved about 5° azimuth error on this set of reverberant mixtures, averaged over the 600 sources (3 in each of the 200 mixtures) localized in each of the 4 T_{60} times.

The key advantage of both the proposed system and the binaural system is that azimuth-dependent cues for a particular source are not integrated over the entire mixture, as they are in the stencil filter and SRP-PHAT approaches. The comparison between the proposed method without cue weighting and the binaural method shows that monaural grouping alone facilitates more accurate localization as T-F units are not treated completely independent of one another. Selecting a subset of the T-F units using a mechanism for cue weighting is also advantageous in terms of localization accuracy. We extend the proposed system and more thoroughly evaluate localization in adverse conditions in Chapter 4.

3.6.3 Simultaneous and Sequential Organization Performance

We analyze the quality of both simultaneous and sequential organization using the IBM. As the proposed system only deals with voiced speech, we evaluate simultaneous organization in voiced speech regions by finding the percentage of mixture energy (in dB) contained in the simultaneous streams that is correctly labeled by an estimated mask, where ground truth labeling of a T-F unit in a simultaneous stream is generated using the IBM of the better ear mixture. We refer to this metric as the labeling accuracy. To evaluate sequential organization, we compare performance against a “ceiling” measure that incorporates ideal knowledge and to a recent model-based system [161]. We refer to the ceiling performance measure as *ideal sequential organization* (Ideal S.O.). In this case, a target/interference decision is made for each simultaneous stream based on whether the majority of the mixture energy is labeled target or interference by the IBM.

The model-based system uses pre-trained speaker models to perform sequential organization of simultaneous streams for voiced speech [161]. Speaker models are trained using an auditory feature, gammatone frequency cepstral coefficients [161], and the system incorporates missing data reconstruction and uncertainty decoding to handle simultaneous streams that do not cover the full frequency range. The system is designed for anechoic speech trained in matched acoustic conditions. To account for both the azimuth-dependent HRTF filtering and reverberation contained in the mixture signals used in our database, some adjustments were made. First,

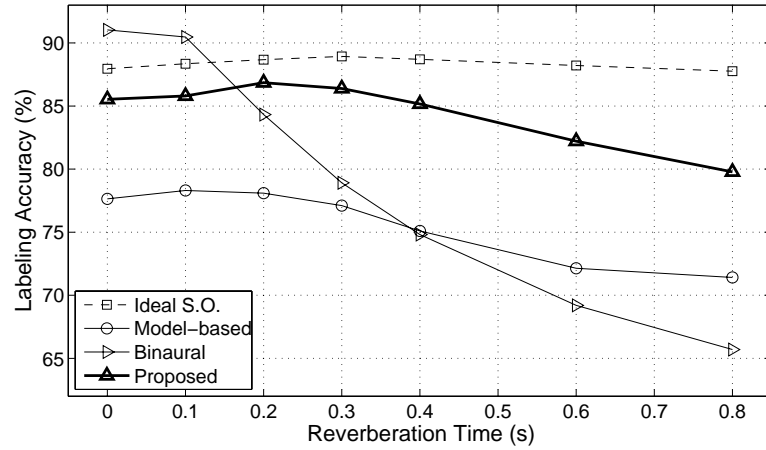
we train speaker models for each of the reverberation conditions that will be seen in testing. For each of the 38 speakers, we select 7 out of 10 utterances for training, generate 10 variations of each of these utterances with randomly selected azimuths for each of the 7 reverberation times. This helps to minimize the mismatch between training and testing conditions, although as mentioned above, the impulse responses used in training are different from those in testing. We found this approach to give better performance than feature compensation methods (e.g. cepstral mean and variance normalization) for mismatched training and testing conditions. In [161], a background model is used to allow the system to process speech mixed with multiple speech intrusions or non-speech intrusions. Since we focus on the two and three-talker cases, we found that assuming all speakers are known *a priori* produces better results than using a generic background model. Incorporating this prior knowledge ensures that we are comparing to a high level of performance potentially achievable by the model-based system.

To identify the conditions in which the proposed integration of monaural and binaural analysis can improve segregation relative to binaural analysis alone, we compare performance to the exclusively binaural system described in Equations (3.12) and (3.13). For the purpose of comparison, we continue to measure the labeling accuracy *within* the simultaneous streams, even though the exclusively binaural approach is able to generate a binary mask for the entire mixture.

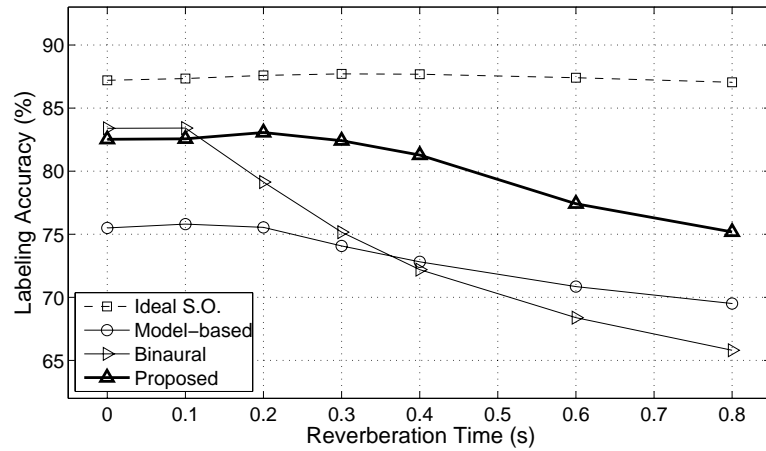
As previously stated, the exclusively binaural system has much in common with

the systems proposed in [93, 123, 124]. The key difference is that the binaural system presented here uses pre-trained, non-parametric likelihood functions rather than fitting parametric models to the observed mixture. To test whether models that are tuned to capture the reverberation condition of a specific mixture improves performance, we trained alternative non-parametric likelihood functions tuned for each T_{60} time of the test database. On our two-talker database we found little benefit in using the T_{60} -specific models for either the exclusively binaural or the proposed system (0.3% better on average for both systems). In training the likelihood functions as described in Section 3.6.1, we have generated a binaural model that, while specific to the binaural microphone (or listener) used for training, provides good performance across a variety of room conditions.

In Figure 3.5 we show the performance of the proposed system, the model-based system, the binaural system and the ideal sequential organization scheme on the two- and three-talker mixtures. The performance achieved by Ideal S.O. indicates the quality of the monaural simultaneous organization. Any decrease below 100% reflects that the simultaneous streams are not exclusively dominated by target or interference. On the two-talker mixtures shown in Figure 3.5(a), labeling error due to monaural analysis averages 11.6% across all T_{60} times, and is largely consistent across reverberation conditions. The performance difference between Ideal S.O. and the model-based or proposed systems reflects errors due to sequential organization. Model-based sequential organization introduces an additional 12.7% labeling error, averaged over all T_{60} times. The error introduced by localization-based sequential organization ranges



(a) Two-Talker Mixtures



(b) Three-Talker Mixtures

Figure 3.5: Labeling accuracy of the proposed and comparison systems shown as a function of reverberation time for (a) two-talker and (b) three-talker mixtures.

from 1.8% in low reverberation conditions, up to almost 8% in the most reverberant condition. The relative performance improvement over the model-based system ranges between 9.5% and 14%, depending on the T_{60} time. This is notable, especially considering that the model-based results incorporate prior knowledge of the speaker identities contained in the mixture and the T_{60} time of the mixture. The proposed system outperforms the model-based approach on the three-talker mixtures as well (see Figure 3.5(b)), although the gap is not as large.

In comparing the proposed system to the Ideal S.O. system, one can see that the proportion of labeling error attributable to localization-based sequential organization increases with both T_{60} time and the number of talkers, suggesting that an increase in the number of talkers or the reverberation time has a larger impact on the binaural sequential organization than on the accuracy of the monaural grouping. However, since all results are obtained from voiced speech only, as generated from the tandem algorithm’s simultaneous streams, these measures do not penalize the simultaneous organization stage for what one might call misses, or T-F units that contain primarily voiced energy from one of the source signals, but are not captured by any of the simultaneous streams. We note that the proportion of total mixture energy (both voiced and unvoiced) that is captured by a simultaneous stream is 57% in the two-talker anechoic case, decreases to 35% averaged over the two-talker mixtures between 300 ms and 800 ms T_{60} and 33% averaged over the three-talker mixtures between 300 ms and 800 ms T_{60} . This suggests that using monaural simultaneous organization

Table 3.1: Labeling accuracy as a function of spatial separation (in $^{\circ}$)

	Two-Talker Mixtures			Three-Talker Mixtures		
	≤ 30	35 – 60	> 60	≤ 30	35 – 60	> 60
Binaural	63.3%	74.8%	79.8%	66.8%	73.1%	79.0%
Proposed	79.9%	85.1%	85.9%	77.5%	81.1%	82.8%

developed specifically for reverberant environments [96] may improve performance using the proposed framework.

One can see a strong influence of the reverberation time on the binaural system. For the two-talker mixtures in which there is little reverberation present, i.e. with T_{60} of 0 and 100 ms, the binaural system outperforms even the Ideal S.O. system. This suggests that in these cases the binaural cues are more powerful than pitch-related cues for achieving simultaneous organization. However in the three-talker case and in even moderate amounts of reverberation, simultaneous organization achieved by monaural processing improves performance over exclusively binaural grouping. The gap between the Ideal S.O. system and the binaural systems increases with both the amount of reverberation and the number of talkers, indicating that the potential gain of integrating monaural and binaural processing is greater as the mixture complexity increases.

It is clear from Figure 3.5 that the proposed system represents a significant improvement over the binaural system, and that the margin between the two increases as a function of T_{60} . The performance margin is also dependent on spatial separation between sources. Table 3.1 shows the average labeling accuracy of the proposed and

binaural system as a function of spatial separation between the target source and the closest interference source for mixtures with T_{60} between 300 ms and 800 ms. One can see that our system’s performance does not degrade as severely as the binaural system for closely spaced sources.

Due to the nature of the monaural processing used in this study, there is some influence of source gender on performance of the proposed system. For the two-talker mixtures with T_{60} between 300 ms and 800 ms, the average labeling accuracy is 81.7% for mixtures where talkers have the same gender and 85.3% when talkers have different genders. This effect is even more pronounced for the model-based system where average accuracy is 80.2% when talkers have different genders and only 68.2% for same-gender mixtures. In our two-talker database, 46% of the mixtures have sources with different genders. The difference in performance between the proposed system and comparison systems is similar for male-male and female-female mixtures.

3.7 Discussion

The results in the previous section illustrate that integration of monaural and binaural analysis allows for robust localization performance, which enables sequential organization of speech in environments with considerable reverberation. The localization-based sequential organization outperforms model-based sequential organization that utilizes only monaural cues, and the proposed integration of monaural and binaural analysis outperforms an exclusively binaural approach in terms of voiced speech segregation on two- and three-talker reverberant mixtures. We have also shown that,

in addition to improving segregation performance, incorporation of monaural grouping improves localization performance over three exclusively binaural methods. We address multisource localization in adverse conditions more thoroughly in Chapter 4.

The discrete azimuth space used in this study avoids two potential issues. First, the azimuth-dependent ITD-ILD likelihood functions are manageable in number (37 for each frequency channel in this study). Second, the joint search over all possible azimuths is computationally feasible. In the case of a more finely sampled or continuous azimuth space, or a localization space that includes elevation, one would need to carefully consider how to overcome both issues. To overcome the need for training an unwieldy amount of likelihood functions in a variety of acoustical conditions, parametric likelihood functions could be used without considerable performance sacrifice. In analyzing the trained ITD-ILD likelihood functions, clear patterns emerge that could be utilized to formulate a parametric model. Certain key parameters, such as the primary peak locations and spread of the distributions, could be learned from training data from a discrete set of source positions and extrapolated to a continuous space. We develop a model along these lines in Section 4.2.2. The second issue of joint search over all possible angles in a finely sampled or continuous space could be avoided by doing an initial search in a discretized space (such as the one used here), then refining the source positions in a limited range.

The development in Section 3.5 assumes that all sequential organizations are equally likely. For mixtures in which the input SNR is significantly different from 0 dB, improved performance may be achieved by allowing simultaneous stream labels

to favor one source. Further, simultaneous stream labels are not truly independent. While this may be true for simultaneous streams that are separated in time, this assumption is questionable when two simultaneous streams overlap in time. In the majority of cases, simultaneous streams that overlap in time are due to different sources. Further, it may be possible to capture common relationships between simultaneous streams nearby in time due to regularities in speech spectra. The framework developed in Chapter 6 takes an alternative approach to simultaneous organization to address this issue.

Finally, since the proposed system only processes voiced speech, it is essential to develop methods to handle unvoiced speech. Binaural cues are likely a powerful tool for handling unvoiced speech, which is challenging with only monaural cues (see [84]). In Chapter 4, we incorporate unvoiced speech by adding additional monaural cues to the localization procedure. In Chapter 6 we develop a full segregation system that handles both voiced and unvoiced speech.

CHAPTER 4

MULTISOURCE LOCALIZATION IN ADVERSE CONDITIONS

The focus of this chapter is on localization of multiple sources from a binaural input. We extend the system described in Chapter 3 and provide a thorough analysis of localization performance in reverberant and noisy conditions. We propose a novel azimuth-dependent model of binaural cues and incorporate additional monaural grouping cues. A preliminary version of this chapter was published in [191].

4.1 Introduction

As outlined in Section 2.1, binaural localization has received significant attention in CASA due to a desire to understand and model the underlying computational mechanisms involved in human sound localization, and because automatic localization has applications in hearing prostheses, spatial sound reproduction and mobile robotics. As we have now mentioned in Chapters 1, 2 and 3, the two main physical cues for sound localization in terms of azimuth are ITD and ILD. As discussed in Sections

2.1 and 2.3, the main differences between the observation models used in localization methods are a result of different assumptions about environmental factors such as source propagation, background noise or the microphone setup. For multisource localization, methods also differ in how spatial information is integrated across time and frequency, where differences are largely a function of assumptions about source activity and interaction.

In this chapter we focus on localization of a known number of spatially fixed sources. As such, integration of spatial cues across time can be handled more simply than when sources (or microphones) are moving (see e.g. [121, 148]). When sources are assumed to be fixed over a given interval of time, a simple approach is to first integrate azimuth information across frequency, then average across time and select multiple peaks in the resulting azimuth-dependent response function [1, 112, 127, 149]. These methods can be effective with sufficient separation between sources and time integration, but can perform poorly when one source is dominant over the majority of the integration period. By assuming source sparsity in a time-frequency (T-F) representation, spatial clustering methods have been proposed to jointly segregate and localize a known number of spatially stationary sources [93, 123, 124, 136]. In this case, localization could potentially be improved by integrating features over a subset of T-F units, however the demonstrated benefit of recent systems is in terms of segregation rather than localization [124].

We propose a localization method where, similar to spatial clustering methods, azimuth estimates are derived from only those T-F units in which a given source is

thought to be dominant. In contrast to existing spatial clustering methods, segregation is performed on the basis of both monaural and binaural cues and we demonstrate that this improves azimuth estimation in reverberant and noisy conditions. The proposed approach is motivated by psychoacoustics studies on *binaural interference*, which show that spectrally remote interfering signals can impact lateralization and ITD discrimination of a target signal [60, 170, 193]. The degree to which the interfering signals influence subjective judgements, however, is influenced by the degree to which monaural cues support grouping between the target and interference signals [9, 78, 81, 170]. One well supported interpretation of this research is that the auditory system performs grouping using multiple features, and that localization judgements are formed by integrating spatial features within these larger auditory “objects” [9, 43]. Existing approaches that assume full integration across frequency (e.g. [1, 8, 105, 127, 149]) are inconsistent with binaural interference studies because maskers would have the same impact on localization independent of the support for monaural grouping. Existing spatial clustering approaches are also inconsistent with binaural interference studies because they implicitly assume object formation on the basis of spatial cues - thus no binaural interference should be expected.

In Section 4.2 we describe extraction of binaural features and propose a novel azimuth-dependent binaural model and associated training procedure. We summarize the monaural CASA methods used in Section 4.3. In Section 4.4 we describe how binaural and monaural cues are integrated within the proposed framework for the purpose of multisource localization. We describe the evaluation methodology in

Section 4.5 and discuss the results of several experiments using both simulated and measured binaural impulse responses in Section 4.6. Section 4.7 concludes the paper with a discussion of the insights gained from the evaluation and future work.

4.2 Binaural Pathway

4.2.1 Auditory periphery and binaural feature extraction

As in Chapter 3, we assume a binaural input signal sampled at a rate of 44.1 kHz. The binaural signal is analyzed using a bank of 64 gammatone filters [141] with center frequencies from 80 to 5000 Hz spaced on the ERB scale. Each bandpass filtered signal is divided into 20 ms time frames with a frame shift of 10 ms to create a cochleagram [178] of T-F units. Again, we denote a T-F unit as $u_{c,m}^E$ where $E \in \{L, R\}$ indicates the left or right ear signal.

The binaural pathway consists of a low-level feature extraction stage where we calculate the ITD, denoted $\tau_{c,m}$, and ILD, denoted $\lambda_{c,m}$, for each T-F unit pair. We calculate ITD and ILD as in Equations (3.2) and (3.3), respectively. We then map ITD-ILD value pairs to azimuth-dependent features using the trained probabilistic models described below.

4.2.2 Azimuth-dependent binaural model

In Chapter 3 we described a set of non-parametric models to map ITD and ILD measurements to an azimuth-dependent response for each T-F unit pair (see Section

3.4.2). Models were trained on simulated reverberant speech using kernel density estimation. Shortcomings of this approach are that models are not easily adaptable to a new binaural setup (listener) or a new environment, and that binaural impulse responses in a reverberant environment must either be measured or simulated (e.g. using [25]). In this chapter, we develop a simple and flexible azimuth-dependent GMM of ITD and ILD. The model independently captures the frequency-dependent pattern of ITD and ILD values due to direct-path propagation, which we again refer to as direct-path cues, and the statistical effect of environmental factors such as noise and reverberation. As a result, the model is easily adaptable to different binaural setups and acoustic conditions. We propose a training approach that avoids the necessity of reverberant BIRs, which allows for use of the model when only anechoic HRTFs are available.

We again denote the likelihood of observing a pair of ITD and ILD values in frequency channel c given energy from a point source with azimuth θ using $P_c(\tau, \lambda|\theta)$. In order to model the direct-path ITD and ILD independently of variance due to the acoustic conditions, we introduce the direct-to-residual ratio (DRR) for a point source as a latent variable. We calculate DRR, denoted $r_{c,m}$, within a pair of T-F units $u_{c,m}^L$ and $u_{c,m}^R$ as,

$$r_{c,m} = \frac{\sum_n (x_{c,m}^L[n]^2 + x_{c,m}^R[n]^2)}{\sum_n (x_{c,m}^L[n]^2 + x_{c,m}^R[n]^2 + v_{c,m}^L[n]^2 + v_{c,m}^R[n]^2)} \quad (4.1)$$

where n indexes a signal sample, $x_{c,m}^E$ denotes the component of $u_{c,m}^E$ in response to the direct-path of the target source, and $v_{c,m}^E = u_{c,m}^E - x_{c,m}^E$. Each summation is over

the interval of the corresponding T-F unit. Note that our use of DRR differs from the common use as an acronym for direct-to-reverberant ratio.

Given the DRR, r , and the direct-path ITD and ILD associated with azimuth θ , denoted τ_θ and λ_θ , we approximate the joint ITD-ILD observation likelihood for an individual frequency channel using,

$$P_c(\tau, \lambda|\theta) \approx \sum_r P_c(\tau|r, \tau_\theta)P_c(\lambda|r, \lambda_\theta)P_c(r), \quad (4.2)$$

where $P_c(r)$ denotes the prior probability of DRR. Here, we assume that r is independent of τ_θ and λ_θ and that the observed ITD and ILD values are conditionally independent given the DRR and direct-path cues. We also approximate integration over r by summation over a discrete set of values.

Due to spatial aliasing, the probability space for observed ITDs in higher frequency channels is multi-modal. We therefore use a mixture of Gaussians to capture $P_c(\tau|r, \tau_\theta)$, or,

$$P_c(\tau|r, \tau_\theta) = \sum_{k=1}^{K_c} \rho_{c,k}(r, \tau_\theta) \mathcal{N}(\tau|\mu_{c,k}(r, \tau_\theta), \sigma_{c,k}(r, \tau_\theta)), \quad (4.3)$$

where K_c is determined based on the channel center frequency, the direct-path ITD, and the range of observable ITD values (between -1 and 1 ms in this study). The ILD likelihood is well described by a single Gaussian, $P_c(\lambda|r, \lambda_\theta) = \mathcal{N}(\lambda|\mu_c(r, \lambda_\theta), \sigma_c(r, \lambda_\theta))$. Finally, letting R be the number of discretized values for r , $P_c(r)$ is a set of R scalar values. Given that each component of the model is either a set of Gaussians or a scalar, the full model can be written as a two-dimensional GMM with $R \cdot K_c$ components.

We show example models for $\theta = 70^\circ$ at 1000 Hz in Figure 4.1. Figures 4.1(a) and 4.1(b) show the marginal likelihoods of ITD and ILD, respectively, Figure 4.1(c) shows two different DRR priors, and Figures 4.1(d) and 4.1(e) show the two resulting log-likelihood distributions with r marginalized. The joint log-likelihood functions in Figures 4.1(d) and 4.1(e) are shown as equal contour plots, where 4.1(d) is generated using the descending prior (squares), and 4.1(e) is generated using the ascending prior (circles). $R = 5$ in this example. While each function exhibits two peaks, the primary peak in 4.1(e) is much higher and sharper than the primary peak in 4.1(d) and is more selective in terms of ILD. Also note that the secondary peak in 4.1(d) has a slightly different ITD location and ILD much closer to 0 than the secondary peak in 4.1(e).

4.2.3 Model Training

Recent approaches to training binaural models of ITD and ILD incorporate simulations of multisource pickup in a reverberant environment [127], as described in Section 3.4.2, and thus may be sensitive to deviation from the room configuration or acoustic conditions used in training. In this work we generate training mixtures by combining a point source with a simulated diffuse noise, and in doing so, avoid capturing environment-specific effects. We assume only the HRTFs of the binaural setup are known. We simulate a point source by filtering monaural signals using the HRTF for a given azimuth. The diffuse noise is created by passing uncorrelated noise signals

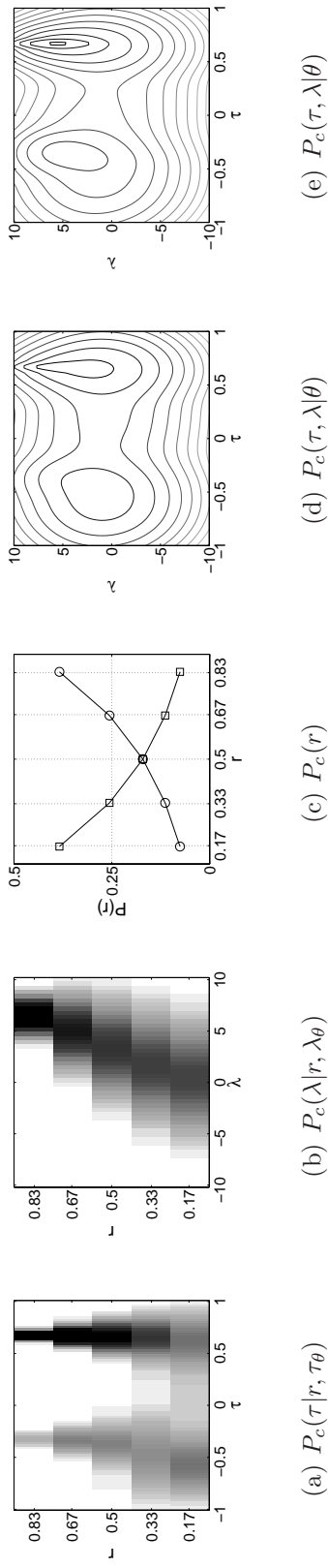


Figure 4.1: Marginal ITD (a) and ILD (b) likelihoods, DRR prior (c), and equal contour plots of the ITD-ILD log-likelihood distributions (d) and (e) for $\theta = 70^\circ$ at 1000 Hz. The distribution in (d) uses the descending prior (squares) from (c), and the distribution in (e) uses the ascending prior (circles) from (c).

through each of the HRTFs for the binaural setup and then adding them together. We provide more detail on the generation of training data in Section 4.5.3.

Given a set of training data for a specific azimuth, we measure τ and λ from each pair of mixture T-F units and calculate r using Equation (4.1). Since the simulated target includes only direct-path propagation, $x_{c,m}^E$ and $v_{c,m}^E$ are simply the premixed target and diffuse noise signals. We discretize the r values into R equally spaced bins. In this study we let $R = 5$ and have found the procedure to be relatively insensitive to the number of bins, provided a sufficient number (roughly 3 or more) is used. The total number of Gaussian components in the resulting model is proportional to R , thus choosing a small number limits the complexity of the model.

For each frequency channel, azimuth and DRR bin we learn the GMM parameters for the ITD dimension, $\{\rho_{c,k}(r, \tau_\theta), \mu_{c,k}(r, \tau_\theta), \sigma_{c,k}(r, \tau_\theta)\}$, using the EM algorithm, where $k \in \{1, \dots, K_c\}$. We set the number of components, K_c , by determining the number of peaks in the range between -1.1 to 1.1 ms (to capture some edge effects) assuming that the cross-correlation function used to calculate ITD is periodic with the channel center frequency and that a peak exists at τ_θ . We then add one extra component to give the model more flexibility. The expected number of peaks in the cross-correlation function, and therefore K_c , increases systematically with center frequency. For each frequency channel, azimuth and DRR bin we also measure the sample mean and variance for the ILD dimension, $\{\mu_c(r, \lambda_\theta), \sigma_c(r, \lambda_\theta)\}$. Finally, we calculate the number of data points that fall into each DRR bin for $P_c(r)$, although

in order to remove the influence of training conditions, these values may be unused. We discuss how $P_c(r)$ is set for the experiments in this study in Section 4.5.4.

4.3 Monaural Pathway

Both harmonicity and onset synchrony are known to be strong cues for across frequency grouping in ASA [17] and have been shown to influence localization judgements by human listeners [9, 81]. Motivated by this work and recent advances in monaural source segregation [178], the proposed framework incorporates a monaural pathway that uses a pitch-based and an onset/offset analysis to group T-F units dominated by the same underlying source. The grouping is used to constrain the integration of binaural cues for azimuth estimation.

We use existing algorithms for multipitch tracking [97] and onset/offset based segmentation [83]. We also incorporate a pitch-based grouping method that is similar to the approach described in [96]. In this section we provide only a brief description of these methods and discuss their role in the proposed system. The interested reader is referred to the cited papers for more details.

4.3.1 Multipitch Tracking

In order to group T-F units based on pitch information, we incorporate a recent multipitch tracking system designed for reverberant and noisy speech [97]. This system estimates up to two pitch periods per time frame using a hidden Markov model

(HMM) tracking framework. The state space of the HMM is a collection of subspaces corresponding to the cases with zero, one or two voiced sources. The one- and two-source subspaces consist of all allowable single pitch and pitch combinations, respectively (covering the frequency range from 80 to 500 Hz). The model is allowed to jump between subspaces (i.e. the number of voiced sources can change), and pitch dynamics within a subspace are modeled by pitch transition probabilities. The observed data used in computation of state likelihoods is based on the correlogram [178]. The Viterbi algorithm is used to find the optimal path through the pitch subspaces, thereby estimating both the number of voiced sources and the corresponding pitch periods in each time frame.

We use this system to generate pitch estimates from both the left and right signals independently. Once pitch estimates are generated, we link pitch points across time when the change in pitch is below a predetermined threshold. We refer to a set of linked pitch points as a *pitch contour*. We use a threshold of 7% relative change in pitch frequency.

4.3.2 Pitch-based Grouping

Pitch contours are used as the basis for grouping T-F units dominated by the same voiced source. For each individual pitch estimate, we use a supervised learning approach to identify T-F units across frequency that exhibit periodicity consistent with that of the estimate. Since the pitch estimates have already been linked across time intervals into pitch contours, T-F units associated with each pitch estimate are also

grouped across time to form sets of T-F units, which we refer to as *simultaneous streams*.

Specifically, we use a MLP to model the posterior probability that the dominant source in a T-F unit is consistent with a hypothesized pitch period. The features used as input to the MLP are extracted from the correlogram and envelope correlogram, calculated from both the left and right signals. We use a low-pass filter with 500 Hz cutoff frequency and a Kaiser window to extract signal envelopes. We train a separate MLP for each frequency channel, which consists of a hidden layer with 30 nodes. Training is accomplished using a generalized Levenberg-Marquardt backpropagation algorithm. We train the MLPs using a set of mixtures described in Section 4.5.3. For each training mixture we extract the correlogram and envelope correlogram features, calculate the IBM and generate the *ground truth* pitch of the target signal by running the pitch estimation method proposed in [12] on the premixed signal. The IBM is used to provide the true classification label for each T-F unit and the ground truth pitch points are used to select the correlogram features corresponding to the pitch period of the target source. A more detailed description of models and training for pitch-based grouping can be found in [96].

4.3.3 Onset/offset Based Segmentation

To capture unvoiced speech regions, the monaural pathway also incorporates the onset/offset segmentation approach proposed in [83]. The method first identifies onsets (increases in signal energy) and offsets (decreases in signal energy) across

time within gammatone sub-bands. Detected onsets and offsets are linked across frequency into onset and offset fronts based on synchrony. Onset fronts are grouped with corresponding offset fronts based on frequency overlap. The set of T-F units between a pair of onset and offset fronts forms a T-F segment. Segmentation is performed with three different scales of across-time and across-frequency smoothing. Segments generated using the different smoothing scales are then integrated into a single set of T-F segments.

We use this segmentation system to generate T-F segments for the left and the right mixture independently. We make three changes to the implementation relative to that presented in [83]. First, to match the peripheral processing of the binaural pathway we implement the segmentation algorithm using 64 frequency channels, rather than 128. Second, we adjust the standard deviation of the Gaussian kernels used for across-frequency smoothing to account for the change from 128 to 64 channels. Third, in preliminary experiments we have found that pitch-based grouping is more reliable than the onset/offset segmentation in voiced speech regions. With this in mind we eliminate T-F units from the segments if they are members of a pitch-based simultaneous stream.

4.3.4 Onset-based Weights

As described in Section 3.4.3, we find it beneficial to weight the contribution of T-F units to the localization decision so as to minimize the effect of units that are likely dominated by noise. We therefore include the same procedure to generate onset-based

weights, denoted $w_{c,m}^E$, for each T-F unit $u_{c,m}^E$. However, in Section 3.4.3 we used a hard threshold to keep the 25% of T-F units with highest weight per simultaneous stream. In the system presented in this chapter, we simply use half-wave rectification, denoted $[\cdot]^+$, to create a set of non-negative weights,

$$w_{c,m}^E = \left[\frac{e_c^E[m] - e_c^E[m-1]}{e_c^E[m-1]} \right]^+, \quad (4.4)$$

where again, $e_c^E[m]$ denotes the sample of the decimated envelope signal corresponding to $u_{c,m}^E$.

4.4 Localization Framework

The binaural pathway extracts azimuth-dependent information from each T-F unit pair while the monaural pathway groups T-F units that are likely to be dominated by the same source. The final stage of the proposed system then integrates this information and produces a set of K azimuth estimates. In Section 3.5, we developed a maximum likelihood framework for joint localization and labeling of pitch-based simultaneous streams. We take a similar approach here, but now deal with both voiced and unvoiced speech, and also use simultaneous streams and T-F segments generated from both the left and right mixture.

Conceptually speaking, to perform localization we first postulate a set of K possible azimuths, where we assume K is known. For each simultaneous stream or T-F

segment we find the most likely azimuth from the postulated set and integrate likelihood scores over all streams and segments. The process generates a total likelihood for each postulated set of azimuths, and we choose the set that maximizes this value.

Formally, let I^E be the total number of simultaneous streams and T-F segments from ear signal E . Each individual simultaneous stream or T-F segment, denoted g_i^E , is a collection of T-F units. Assuming conditional independence between T-F units, the weighted log-likelihood for g_i^E is then,

$$\beta_i^E(\theta) = \sum_{c,m \in g_i^E} w_{c,m}^E \ln(P_c(\tau_{c,m}, \lambda_{c,m}|\theta)). \quad (4.5)$$

We search for the most likely set of K azimuths using,

$$\hat{\Theta} = \arg \max_{\Theta} \left(\sum_{i=1}^{I^L} \beta_i^L(\theta_{\hat{y}_i^L}) + \sum_{j=1}^{I^R} \beta_j^R(\theta_{\hat{y}_j^R}) \right), \quad (4.6)$$

where $\Theta = \{\theta_0, \theta_1, \dots, \theta_{N-1}\}$ denotes a set of K azimuths and,

$$\hat{y}_i^E = \arg \max_{y \in \{0,1,\dots,N-1\}} \beta_i^E(\theta_y). \quad (4.7)$$

4.5 Evaluation Methodology

We conduct three experiments to evaluate the effectiveness of the proposed method relative to existing systems. This section provides necessary details regarding the generation of training and evaluation data, and the binaural models, comparison systems and metrics used in the evaluation.

4.5.1 Binaural Impulse Responses

We use two different sets of binaural impulse responses (BIRs) in this study. One set is simulated and one set is measured in real environments. Each set assumes a different binaural setup, and we will refer to them according to the assumed setup.

The simulated BIRs, which we refer to as the KEMAR set, are generated using the ROOMSIM package [25]. This software combines the image method for reverberation [3] with HRTF measurements [67] made using a KEMAR dummy head. BIRs generated in this way represent a reasonable simulation of pickup by a KEMAR in real environments while allowing control of array and source placement, as well as characteristics of the room. We create a library of BIRs by generating 10 room configurations, where room size, array position and array orientation are set at random. We then generate BIRs for azimuths between -90° and 90° , spaced by 5° , at distances of 1, 2, and 4 m (where available in the room configuration). Reflection coefficients of the wall surfaces are set to be equal and to be the same across frequency, such that the reverberation time (T_{60}) is approximately 600 ms. These BIRs are used to generate the evaluation mixtures used in Experiment 1-3. In order to train binaural models, as described in Section 4.2.3, we generate anechoic BIRs for the same azimuths using the HRTF measurements directly (i.e. no room simulation).

The other set includes publicly available measured BIRs, which are described in [90]. Impulse responses are measured using a head and torso simulator (HATS) in five different environments. Four environments are reverberant (rooms A, B, C and D),

with different sizes, reflective characteristics and reverberation times. Measurements are also made in an anechoic environment. In all cases, BIRs are measured for azimuths between -90° and 90° , spaced by 5° , at a distance of 1.5 m. We use the BIRs from the three most reverberant rooms (B, C and D) to generate an evaluation database, where the T_{60} times are listed as 0.47, 0.68 and 0.89 s, respectively. We use the anechoic measurements to train binaural models. We refer to this set of BIRs as the HATS set.

4.5.2 Evaluation Data

We create two evaluation sets, one from the KEMAR BIR set and one from the HATS BIR set. In the KEMAR evaluation set we consider 2 or 3 target talkers, source distances of 1, 2 and 4 m, and infinite, 6 and 0 dB speech-to-noise ratios (SNR) for a total of 18 conditions. We generate 100 binaural mixtures for each condition. Azimuths are selected randomly such that sources are spaced by 10° or more. The SNR is set by summing the energy of all speech sources relative to a simulated diffuse noise. The energy of both left and right channels is summed prior to SNR calculation. Speech sources are simulated by filtering monaural utterances, drawn randomly from the TIMIT database [68], by a selected KEMAR BIR. Monaural utterances, originally sampled at 16 kHz, are upsampled to 44.1 kHz to match the rate of the KEMAR BIRs. The diffuse noise is created by filtering uncorrelated speech-shaped noise signals through each of the anechoic KEMAR BIRs and then adding them together. We create the speech-shaped filter by averaging the amplitude spectra

of 200 speech utterances drawn from TIMIT at random. Each mixture has a length of 2 s, where monaural speech utterances are concatenated so that they are sufficiently long (if needed). We employ an energy threshold to eliminate silence at the beginning and end of the monaural utterances in order to ensure that speech sources are active in the majority of time frames.

We create the HATS evaluation set in the same way. In this case we consider 2 target talkers in 3 rooms (B, C and D), and infinite, 6 and 0 dB SNRs, giving us a total of 9 conditions. All other details are as described for the KEMAR set.

4.5.3 Training Data

To train binaural models we generate data using the anechoic KEMAR and HATS BIRs. In each case we generate 250 speech plus noise mixtures where, as described in Section 4.2.3, we simulate anechoic speech using a BIR for a selected azimuth and simulate diffuse speech-shaped noise as described in Section 4.5.2. Speech utterances are drawn randomly from TIMIT. The only factors varying between mixtures are the speech utterances used and the input SNR, which is selected randomly to be -24 , -12 , -6 , -3 , 0 , 3 , 6 , 12 , or 24 dB.

In order to evaluate how well the proposed scheme for training binaural models compares to a more ideal training scenario, we also generate a training set using the reverberant HATS BIRs. We generate 250 mixtures for each azimuth and for each of the 3 rooms seen in the HATS evaluation set. The procedure used to generate these training mixtures is identical to that used for the evaluation mixtures, however, each

training mixture generated for a specific azimuth contains one speech source placed at that azimuth.

Finally, we generate a set of 100 mixtures to train the MLPs used for pitch-based grouping. Each mixture contains a dominant speech source corrupted by a multi-talker babble consisting of 10, 15 or 20 interfering speech sources. Monaural speech utterances are drawn randomly from the TIMIT database and filtered by a selected KEMAR BIR. The azimuth of all sources is selected randomly between -90° and 90° and the SNR between the dominant talker and the multi-talker interference is set at random between -6 and 12 dB (in 3 dB steps).

4.5.4 Binaural Models

Using the training procedure outlined in Section 4.2.3 along with the anechoic speech plus diffuse noise data described in the previous subsection, we create KEMAR and HATS models. In addition to using the HATS models trained from anechoic measurements, we generate a set of models for the HATS evaluation set that we refer to as *matched*. The matched models are created using the second set of training mixtures described in the previous subsection. A separate model is trained for each room. In this case, target signals are simulated by convolution with a measured, reverberant impulse response. It is therefore necessary to approximate direct-path propagation of the target in order to calculate the DRR. To accomplish this we identify the approximate location of the direct-path component by finding the largest peak in the

BIR, then truncate the impulse response 10 ms after the start of the direct-path component. For the HATS BIRs used in this study, we have found that choosing 10 ms ensures capture of the full direct-path component, while minimizing the number of reflections included. This parameter may vary for different measurements, but is not necessary to train models based on measurements made in a controlled environment.

The choice of values for the DRR prior, $P_c(r)$, will influence the shape of the resulting likelihood distribution (see Figure 1). If $P_c(r)$ is set empirically (i.e. by counting the number of training data points that fall into each DRR bin), the distributions will reflect the acoustic conditions seen in training. If one desires to minimize the influence of training data, $P_c(r)$, can be set according to some assumptions about the acoustics that will be seen in practice. As described in Section 4.2.3, we discretize DRR into 5 bins, corresponding to values of 0.83, 0.67, 0.5, 0.33 and 0.17, or approximately 7, 3, 0, -3 and -7 in dB. For the KEMAR and HATS models, we set $P_c(0.17) = 0.6$, $P_c(0.33) = 0.1$, $P_c(0.5) = 0.1$, $P_c(0.67) = 0.1$, $P_c(0.83) = 0.1$ for all frequencies and azimuths. We chose these values to inject limited knowledge of the evaluation set acoustics. Specifically, this prior reflects an assumption that a given T-F unit is more likely to be dominated by the residual signal (noise or reverberation) than the direct-path of a speech source. These specific values were chosen by an informal analysis of a small number of mixtures that resemble those seen in the evaluation set. Since the matched models for the HATS evaluation set are trained using data that perfectly matches the conditions that will be seen in testing, we set $P_c(r)$ empirically for the matched models.

4.5.5 Comparison Systems

In the experiments below, we compare performance of the proposed method with two existing methods from the literature [51, 124]. The first comparison system is SRP-PHAT, described in Section 3.6.2. The second comparison system used is the joint localization and segregation approach presented in [124], dubbed MESSL, and is representative of the spatial clustering approach to localization. We use an implementation of MESSL provided by the algorithm’s author. The system requires specification of the number of sources and iteratively fits GMMs of IPD and ILD to the observed data using an EM procedure. Across frequency integration is handled by tying GMMs in individual frequency bands to a principal ITD. Based on the model fits, we find the most likely ITD for each source and map this to an azimuth estimate using the the group delay of the anechoic KEMAR or HATS BIRs, depending on the evaluation set. MESSL is initialized using the PHAT-Histogram method [1], where we use the group delay of the anechoic KEMAR or HATS BIRs to specify the ITD bins for the histogram. Mixture signals are first downsampled from 44.1 kHz to 16 kHz because the original TIMIT sources were sampled at 16 kHz.

We selected these methods from a set of candidates that also included the systems proposed in [1, 112, 196]. We found that in most conditions, the performance of MESSL and PHAT-Histogram [1] was comparable, but that MESSL outperformed PHAT-Histogram for short integration times. We also found the stencil filter method in [112] to perform similarly, but systematically worse than the SRP method. Finally,

we found the clustering method proposed in [196] to perform poorly on our data set. The system was unable to localize sources at angles more lateral than 45° even in single-source anechoic conditions, due to the large number of frequencies in which spatial aliasing was present.

4.5.6 Evaluation Metrics

In Experiments 1-3 we assume oracle knowledge of the number of speech sources. With this knowledge we seek to estimate the azimuth angle of each source based on a fixed amount of observed data. We evaluate the different localization systems using two metrics. For each evaluation mixture, we consider a source to be detected if there is an azimuth estimate within (and including) 10° . We then measure the *recall* as the percentage of detected sources. We also measure the average azimuth error of those estimates that were within the 10° threshold and refer to this as the *fine error*. Note that a single azimuth estimate cannot be used to detect more than one source.

4.6 Evaluation Results

In this section we present the results from four experiments. The first experiment analyzes the impact of monaural cues on localization. The second experiment provides a comparison of the proposed method to existing systems using simulated impulse responses. The third experiment tests generalization of the system to measured impulse responses and robustness using mismatched binaural models. The fourth experiment considers both detection and localization of speech sources.

4.6.1 Experiment 1: Influence of Monaural Cues

In this experiment we analyze the influence of monaural cues on localization performance. We compare performance to two binaural baselines that use the proposed azimuth-dependent models but do not incorporate monaural cues. The first baseline, denoted Binaural-Hist, uses the procedure proposed in [127]. This approach estimates the dominant azimuth in each frame according to,

$$\hat{\theta}_m = \arg \max_{\theta} \sum_c P_c(\tau_{c,m}, \lambda_{c,m} | \theta), \quad (4.8)$$

then generates an across-time histogram of the frame-level azimuth estimates and selects the K largest histogram peaks as the source azimuths. The second baseline method, denoted Binaural-ML, is a maximum likelihood procedure similar to the proposed method, but does not incorporate monaural grouping. In this case, azimuth estimates are derived using Equations (3.12) and (3.13) from Section 3.6.2. The Binaural-ML system performs segregation on the basis of binaural cues, similar to [93,123,124,136], and derives each azimuth estimate from a subset of T-F units. Along with the binaural baselines, we evaluate three variations of the proposed system, where we consider only pitch-based grouping, only onset/offset segmentation, and the full proposed system. Performance differences between the two baselines and different variations of the proposed system are entirely due to how binaural information is integrated across time and frequency.

Table 4.1: Recall (%) for the KEMAR set for alternative T-F integration methods.

	Two-talker	Three-talker
Binaural-Hist	90.3	84.1
Binaural-ML	91.7	86.9
Binaural + Pitch-based Grouping	96.2	93.1
Binaural + Onset/offset Segmentation	95.6	92.3
Proposed	97.0	94.6

Table 4.1 shows the recall over the entire set of two- and three-talker KEMAR mixtures. We first note that that the Binaural-ML method provides a small improvement over the Binaural-Hist approach. This gain can be attributed to the fact that evidence for multiple sources can be extracted from even a single time frame, which is not possible with the Binaural-Hist approach. However, the rather marginal gain suggests that while it is conceptually appealing to perform joint segregation and localization, there appears to be little improvement in localization when the segregation process is based entirely on binaural cues. In contrast, all systems that incorporate monaural grouping achieve substantial gains relative to the binaural baselines. The best performance is achieved by the full system that incorporates both types of monaural grouping and onset-based weights, where we see a nearly 8% absolute gain in recall relative to the Binaural-ML approach on the three-talker mixtures.

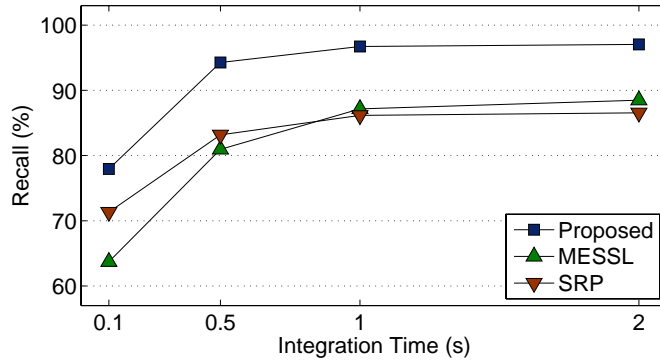
We also note that in addition to the constraints enforced on T-F grouping, the monaural mechanisms select a subset of the T-F units for binaural integration. On the KEMAR data set, about 84% of T-F units are selected. The number of talkers and the source distance appear to have a very small influence on this percentage, while

decreasing the SNR can substantially reduce the percentage of T-F units selected. On average, the percentage of T-F units selected decreases from roughly 91% at infinite SNR to 79% at 0 dB SNR.

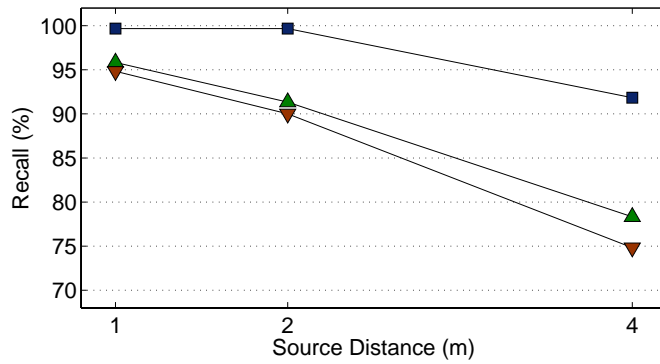
4.6.2 Experiment 2: Comparison on KEMAR Evaluation Set

In this experiment we compare localization performance of the proposed system to the two comparison methods from the literature [51, 124] on the KEMAR evaluation set. We present the recall for various experimental conditions in Figures 4.2 and 4.3. We show results considering integration times of 0.1, 0.5, 1 and 2 s in Figures 4.2(a) and 4.3(a). We do so by providing each system the mixture signals from beginning to the specified time. Results for different integration times are averaged over all distances and SNRs. We show results as a function of source distance in Figures 4.2(b) and 4.3(b). In this case we generate results using the entire mixture (2 s) and average results over SNRs. Similarly, we show results as a function of SNR in Figures 4.2(c) and 4.3(c) using the entire mixture and average over source distances. As one would expect, all systems perform better as more data is used for the estimate, while there is a systematic decrease in performance as sources become more distant or the background noise level increases.

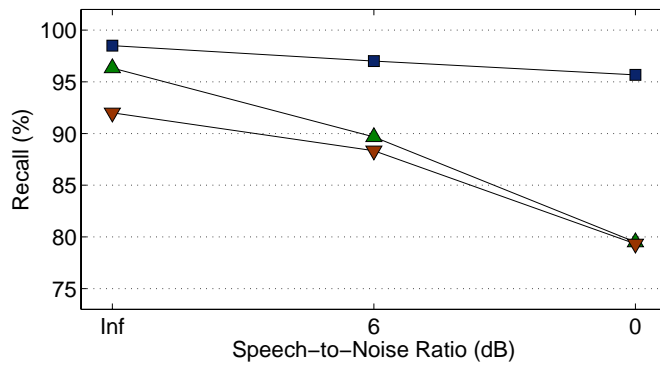
We can see that the proposed system outperforms the comparison methods in terms of recall for all evaluation conditions. MESSL outperforms SRP when the integration time is 1 s or longer. On the shortest integration time, 0.1 s, the initialization of MESSL by PHAT-Histogram [1] is poor, and the algorithm is more likely to



(a) Recall vs. Integration Time

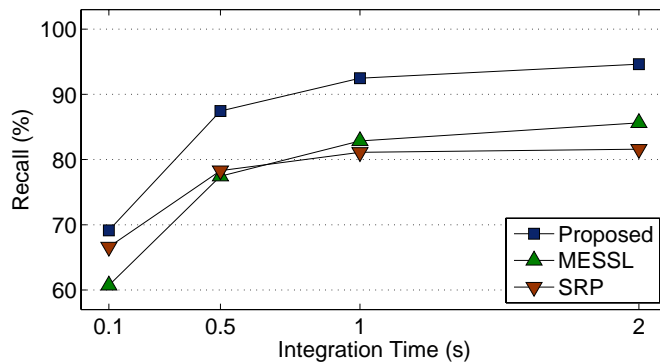


(b) Recall vs. Distance

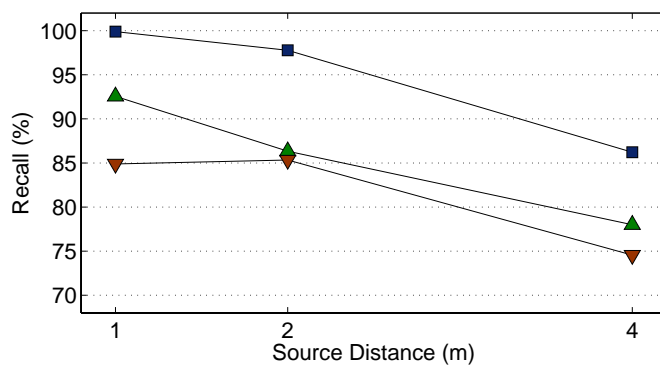


(c) Recall vs. Noise Level

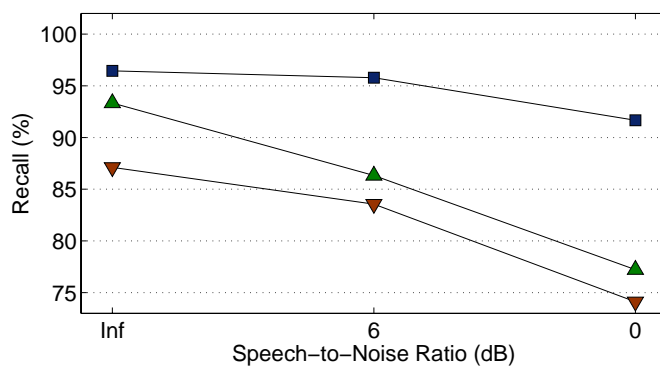
Figure 4.2: Recall (%) shown over the two-talker KEMAR set as a function of (a) integration time, (b) distance and (c) noise level. In (b) and (c), we show results for a 2 s integration time. The legend in (a) is applicable to all figures shown.



(a) Recall vs. Integration Time



(b) Recall vs. Distance



(c) Recall vs. Noise Level

Figure 4.3: Recall (%) shown over the three-talker KEMAR set as a function of (a) integration time, (b) distance and (c) noise level. In (b) and (c), we show results for a 2 s integration time. The legend in (a) is applicable to all figures shown.

have large errors than SRP. The improvement in recall by the proposed system over MESSL is 8.8% (absolute), calculated over the entire two- and three-talker evaluation set. The improvement in recall relative to SRP is 11.7% over the entire evaluation set. In Figures 4.2(b), 4.2(c), 4.3(b) and 4.3(c), we see that the improvement achieved by the proposed system tends to be larger in the more difficult conditions with distant sources and strong background noise. For example, on the two-talker evaluation set with sources at 4 m and 0 dB SNR, the improvement in recall is about 23% relative to both MESSL and SRP.

In Table 4.2 we show the recall and the fine error on the full two- and three-talker data sets when using a 2 s integration time. As previously stated, the recall using the proposed method is higher than for the comparison methods on both the two- and three-talker data and we can also see that the fine error is lower. Since the proposed system utilizes prior training, the performance increase relative to comparison methods is due to both the inclusion of monaural cues and the prior knowledge captured by the binaural model. Although there are numerous differences between the Binaural-ML system (see Section 4.6.1) and the comparison methods, some indication of the relative contribution of monaural cues and the binaural model can be gained by noting that the Binaural-ML system achieves a 2.3% and 5.2% gain in recall relative to MESSL and SRP, respectively, while the proposed method achieves the 8.8% and 11.7% gains noted above.

To test the necessity of prior training with HRTFs of the binaural setup that will be seen in testing, we also performed tests with binaural models trained on HRTFs

Table 4.2: Recall (%) and fine error ($^{\circ}$) for the KEMAR set.

	Recall		Fine Error	
	Two-talker	Three-talker	Two-talker	Three-talker
Proposed	97.0	94.6	1.0	1.3
MESSL	88.5	85.6	1.5	1.9
SRP	86.6	81.6	1.4	1.8

that simulate microphone pickup on the surface of a rigid sphere [57]. We found degradation in terms of recall to be only 3.4% and 4.5% on the two- and three-talker data sets, respectively. Degradation in terms of fine error was larger, from 1.0° with the KEMAR models to 3.3° with the sphere-based models on the two-talker set, and from 1.3° to 3.1° on the three-talker set. These results indicate that the proposed method can still perform well even with no prior knowledge of the binaural setup to be used in practice.

As one might expect from studies of localization acuity in human subjects [10], the azimuth error is lower near the median plane than to the side of the head when using the proposed method. Across the entire two-talker data set, the average error (error over all estimates, not the fine error) for sources with azimuth between -30° and 30° is 0.6° , whereas it increases to 4° for sources with azimuth more lateral than 60° . We also note that recall is lower in test cases where sources are spaced more closely.

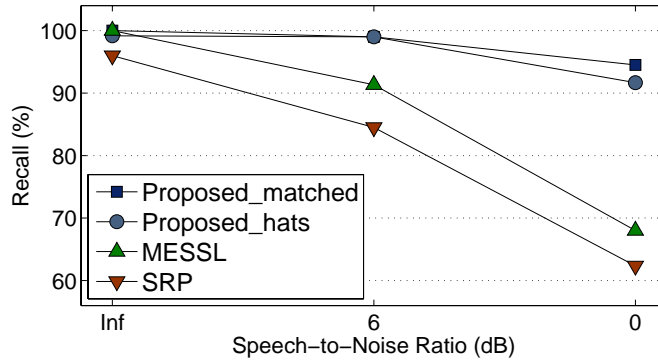


Figure 4.4: Recall (%) as a function of noise level for the HATS evaluation set with an integration time of 2 s.

4.6.3 Experiment 3: Comparison on HATS Evaluation Set

In this experiment we compare localization performance of the proposed system to the two comparison methods on the HATS evaluation set, which uses measured BIRs from real room environments. We also compare the performance achieved using the HATS models trained on anechoic measurements to the matched models trained on the BIRs seen in testing. We assume that using the matched models will provide a performance upper bound and are interested in the amount of degradation due to using mismatched models. Performance using the HATS models on this evaluation set should give the best indication of how the system would perform in a practical setting where calibration measurements may be assumed, but extensive training in real environments would not be available.

We present the recall as a function of SNR in Figure 4.4, where results are averaged over all rooms and an integration time of 2 s is used. Notable is the fact that the

Table 4.3: Recall (%) and fine error (°) for the HATS set.

	Recall			Fine Error		
	B	C	D	B	C	D
Proposed_matched	98.5	97.5	97.5	0.5	0.7	0.5
Proposed_HATS	97.2	97.0	95.7	1.0	0.8	1.7
MESSL	87.3	87.5	84.5	0.9	1.2	1.1
SRP	82.3	80.8	79.7	1.1	1.0	1.9

difference in recall between the matched models and the HATS models is 1.1% or less for the infinite and 6 dB mixtures and 3.2% for the 0 dB mixtures. Consistent with Experiment 2, the performance improvement achieved by the proposed system relative to the comparison methods increases as the level of background noise increases.

In Table 4.3 we show the recall and the fine error for all four systems on each room in the HATS set separately, with a 2 s integration time. We see that the HATS models perform comparably to the matched models in terms of recall, and the proposed system with HATS models achieves a recall about 10% higher than MESSL and about 15% higher than SRP. However, we can see that the fine error is consistently lower when using the matched models. The fine error is similar for all 3 realizable systems, with MESSL achieving the lowest fine error on average. The larger fine error for the proposed system with HATS model and the SRP system on the Room D data is due to a systematic discrepancy between the direct-path cues of the anechoic measurements and the direct-path cues of the Room D measurements.

4.6.4 Experiment 4: Source Detection

In Experiments 1-3 we assumed the number of source signals was known. In this experiment, we analyze the capability of the proposed method to both detect the number of sources and estimate each source’s azimuth. We compare performance to the Binaural-ML and Binaural-Hist systems described in Section 4.6.1. In this case, we evaluate the different systems using two metrics. We again measure the recall as the percentage of detected sources (with 10° tolerance). If an estimate is not within 10° of any source, it is labeled as a false estimate. We measure the *false estimate rate* by dividing the number of false estimates by the total number of estimates.

To allow the proposed and Binaural-ML systems to both detect and localize sources, we introduce a penalty term to Equations (4.6) and (3.13) as follows. In the proposed system, we change Equation (4.6) to,

$$\hat{\Theta} = \arg \max_{\Theta} \left(\sum_{i=1}^{I^L} \beta_i^L(\theta_{\hat{y}_i^L}) + \sum_{j=1}^{I^R} \beta_j^R(\theta_{\hat{y}_j^R}) - \sum_E \sum_{u_{c,m}} w_{c,m}^E \alpha(K) \right), \quad (4.9)$$

where $\alpha(K)$ is a scalar penalty whose values depends on K . We include the term $\sum_E \sum_{u_{c,m}} w_{c,m}^E$ so that the same penalty can be used in spite of integration over a different number of T-F units or total weight. Without the penalty term the system is biased toward over-estimating the number of sources because as K is increased, there is an increased flexibility in the assignment of simultaneous streams and T-F segments using Equation (4.7). The penalty acts to balance over hypotheses with different numbers of sources, similar to well-known model selection criteria such as Akaike information criterion or minimum description length [24].

Similarly, we change Equation (3.13) to,

$$\hat{\Theta} = \arg \max_{\Theta} \sum_{c,m} \ln (P_c(\tau_{c,m}, \lambda_{c,m} | \theta_{\hat{y}_{c,m}})) - \alpha(K), \quad (4.10)$$

where we note that the values for the penalty used in Equations (4.9) and (4.10) are not the same.

In [127] it was proposed to use the Binaural-Hist approach for detecting the number of sources. In this case, a threshold is included such that rather than choosing the K most prominent peaks, any peak above threshold is assumed to be due to a true source.

In Figure 4.5 we show recall vs. false estimate rate on the entire one-, two-, and three-talker KEMAR data set. Curves are generated for each method by systematically varying parameters that control detection sensitivity. For the proposed and Binaural-ML system, we consider a range of positive values for each of $\alpha(1)$, $\alpha(2)$, and $\alpha(3)$. For the Binaural-Hist system, azimuth histograms are normalized and the largest peaks above a detection threshold are used as source estimates. We vary the detection threshold between 0 and 1, and do not allow the system to generate more than 3 azimuth estimates (in keeping with the proposed an Binaural-ML implementations). As a point of reference, we also include the recall and false estimate rates for each system assuming a known number of sources.

We first note that the proposed method achieves a higher recall for a given false estimate rate relative to both the Binaural-ML and Binaural-Hist methods. When the number of sources is known, the recall and false estimate rate are roughly 95% and

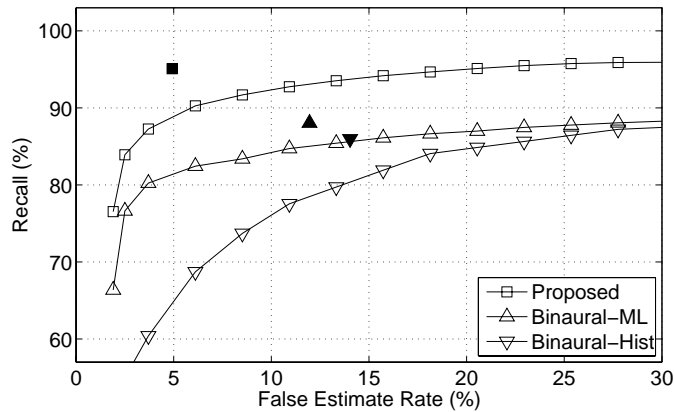


Figure 4.5: Recall vs. false estimate rate for three comparison methods with unknown number of sources. Recall and false estimate rate for the case with known number of sources are shown with filled symbols.

5%, respectively for the proposed system. While maintaining a 5% false estimate rate, recall drops to about 89% for the case when sources must be detected. This recall percentage exceeds that of Binaural-ML and Binaural-Hist even when the number of sources are provided. When sources must be detected by the comparison systems, gross accuracies for a 5% false estimate rate are roughly 82% and 65% for the Binaural-ML and Binaural-Hist methods, respectively. It is also interesting to note that while the maximum recall for the Binaural-ML and Binaural-Hist methods are similar, the Binaural-ML approach achieves a much higher recall for low false estimate rates, suggesting that sources are more easily detected using this approach.

4.7 Discussion

The results in Section 4.6 demonstrate the effectiveness of the proposed localization system. By integrating monaural CASA methods with an azimuth-dependent model of ITD and ILD, we are able to accurately localize multiple sources in adverse conditions. The method yields a significant improvement over baseline methods that do not incorporate monaural grouping. The results from Experiment 1 support the perspective that monaural segregation can facilitate localization. The results from Experiment 2 show that localization improvement is largest in adverse conditions and for distant sources and the results from Experiment 3 establish the robustness of the proposed method when using impulse responses measured in real room environments. Experiment 4 shows that the proposed method is also capable of detecting the number of sources by adding a penalty term to the maximum likelihood azimuth estimation.

We have also proposed a flexible binaural model that can be easily adapted to different binaural setups and acoustic conditions. Results from Experiments 2 and 3 indicate that robust performance can be achieved with only anechoic measurements of the binaural setup, and thus the simulations used to train the models proposed in [127, 190] may be unnecessary. Although only briefly discussed here, preliminary results for generalization to unseen binaural setups are promising.

Since we generate pitch-based simultaneous streams and onset/offset based segments from both the left and right signals, some of the resulting sets of T-F units will

overlap in time and frequency, thus the independence assumption made in order to derive Equations (4.6) and (4.7) is clearly violated. Considering dependencies between simultaneous streams and T-F segments will increase computational complexity of the system; however, it is possible that doing so could improve performance.

An important extension of the proposed framework is to the case with a time-varying number of sources. The results presented in this chapter suggest that incorporating monaural cues improves assignment of T-F units to source signals, and as such, monaural cues could potentially benefit detection and tracking. We address this topic in Chapter 6.

CHAPTER 5

DEFINING THE IDEAL BINARY MASK IN REVERBERATION

In this chapter we consider how best to define the ideal binary mask in reverberant settings to maximize speech intelligibility. We parameterize the IBM using a boundary point between early and late reflections and conduct four experiments to compare the intelligibility of reverberant and noisy speech processed with alternative IBM definitions. The results presented in Experiments 1 and 2 were previously published in [\[150\]](#).

5.1 Introduction

Being that one of the main goals of this work is the development of a binaural segregation system, it is important to define a concrete computational objective so that performance can be appropriately measured and comparisons between systems can be made. Several different objectives and corresponding metrics have been used in the development of speech enhancement, BSS and CASA-based segregation systems.

For both single-channel and multichannel speech enhancement, researchers have often sought to design optimal estimators based on assumed statistical distributions for speech and noise [18, 61, 62, 79, 117, 125, 165]. Most often, the optimization criteria is MSE. While the theoretical guarantees of MSE optimality are appealing, strong assumptions regarding the distribution of both speech and noise are necessary and methods must be developed to estimate important model parameters. To measure the performance achieved in practice, researchers often utilize metrics such as the gain in SNR, speech distortion, noise attenuation or variants designed to better reflect human speech intelligibility [71]. This requires a target signal to be defined such that direct comparison between the estimated and desired target signal can be made. The goal of BSS methods is to separate *each* of a known number of signals from a mixture. In this case, a set of target signals must be defined and again, measures such as SNR or measures proposed specifically for source separation [172, 173] can be used.

Research in CASA has progressed with a number of objectives in mind [176, 178]. Systems have been developed to perform source segregation, speech recognition and model behavioral data. As such, the computational goal of CASA systems is not obvious. In [176], the IBM is proposed as a main computational goal of research in CASA. Wang argues that while the BSS goal of separating each signal may be the “gold standard”, it is likely unrealistic from an engineering perspective and is not consistent with auditory perception. Rather than separate each signal, the IBM performs a figure-ground segregation based on a predefined target signal. Specifically,

given both a mixture and target signal, the IBM retains T-F units of the mixture in which the local SNR exceeds a predetermined threshold and attenuates those T-F units in which the SNR falls below threshold. The formulation of the IBM as the computational goal of CASA is motivated by principles of machine perception, ASA and auditory masking. The psychoacoustics literature shows that an acoustic masker can render a target stimulus inaudible within a critical band [128]. The local SNR threshold in the IBM definition, dubbed the *local criterion* (LC), then serves to label T-F units as either *masked* or *unmasked*, where the mixture components contained in masked T-F units are attenuated under the assumption that they are detrimental to the perception of the target signal.

Several studies have now firmly established the potential of binary T-F masking to improve intelligibility of target speech corrupted by additive noise [4, 20, 21, 104, 108, 179]. The study of Wang et al. [179] reports 7.4 dB and 10.5 dB decreases in speech reception threshold (SRT) for normal hearing listeners with speech corrupted by speech shaped noise (SSN) and cafeteria noise, respectively, where SRT corresponds to the SNR at which 50% word recognition is achieved. For hearing-impaired listeners, reductions in SRT of 9.2 dB and 15.6 dB are reported for the SSN and cafeteria noise conditions, respectively. Large gains in intelligibility have also been observed using different corpora and recognition tasks [4, 20], different noise conditions [4, 20, 21, 108], and alternative IBM definitions [4, 104].

While the IBM is defined unambiguously using the LC parameter in anechoic conditions, in reverberant environments there is some flexibility in how one might

define the target signal itself and therefore, ambiguity is introduced to the notion of the IBM. CASA systems have generally treated reverberation due to the target source as part of the target signal [98, 142, 147]. In contrast, some researchers treat only the direct sound (anechoic) component of the target source as the target signal [124, 137]. However, it is known that early reflections are integrated by the auditory system and thus contribute to speech perception [13, 114, 175], while late reflections are detrimental and act as masking noise. In cases where the direct-path energy is low relative to reflected energy, early reflections can provide a substantial benefit due to an increase in the effective SNR of the target source [13]. Given the division between early and late reflections in terms of perceptual significance, one can create a third IBM definition by treating early reflections of the target source as a part of the target signal, while treating late reflections as part of the noise component. While the division between early and late reflections is assumed to be somewhere between about 50 and 80 ms [80], the exact boundary depends on the signal and environment. We therefore propose to introduce a second parameter to the IBM definition, the *reflection boundary*. The existing IBM definitions that treat either the direct-path target component or the fully reverberant target as the desired signal are then captured by setting the reflection boundary to the two extremes of 0 ms and the length of the reverberant impulse response, respectively.

In this chapter we describe a set of subjective experiments to analyze the effects of IBM processing on speech corrupted by both noise and reverberation. The following section provides a precise working definition of the IBM parameterized by both

the reflection boundary and LC threshold. In Sections 5.3 and 5.4 we describe two experiments to measure the SRT of IBM-processed reverberant and noisy speech, where we use a fixed LC and consider three alternative IBM definitions. In Section 5.6 we describe a third experiment to analyze the effects of changing both the reflection boundary and LC parameters. In Section 5.7 we describe a final experiment to measure intelligibility of IBM-processed reverberant speech without background noise. We consider various reflection boundary values near the range suggested by the psychoacoustics literature. We conclude the chapter with a discussion of the experimental results and how they influence the computational goal of our final system described in Chapter 6

5.2 IBM Definition

We now formalize the concepts described above and specify the processing details used to compute IBMs. We first define the mixture signal as,

$$u[n] = h[n] * s[n] + \epsilon[n] \quad (5.1)$$

where $s[n]$ denotes the (anechoic) target signal, $h[n]$ denotes the room impulse response between the source location and microphone and $\epsilon[n]$ denotes any additional additive noise. We define the *desired* signal as,

$$x_b[n] = h_b[n] * s[n] \quad (5.2)$$

where $h_b[n]$ denotes the part of $h[n]$ up to reflection boundary b . We use the term desired signal rather than target signal to make clear that both the anechoic target

signal and some room reflections may be considered beneficial to the listener, and thus desirable. The *residual* signal is then,

$$v_b[n] = u[n] - x_b[n]. \quad (5.3)$$

In this set of experiments the mixture is analyzed using a bank of 64 gammatone filters [141] with center frequencies from 50 to 8000 Hz spaced on the equivalent rectangular bandwidth scale. Each bandpass filtered signal is divided into 20 ms time frames with a frame shift of 10 ms to create a cochleagram [178] of T-F units. We let $X_b(c, m)$ and $V_b(c, m)$ denote the energy due to the desired and residual signals, respectively, in T-F unit $u_{c,m}$. The IBM can then be defined as,

$$\text{IBM}_b(c, m) = \begin{cases} 1, & \text{if } 10 \log_{10} \left(\frac{X_b(c, m)}{V_b(c, m)} \right) > \text{LC} \\ 0, & \text{otherwise} \end{cases} \quad (5.4)$$

where LC is the local SNR threshold expressed in dB.

5.3 Experiment 1: The Effect of IBM Processing on Reverberant Speech Mixed with Speech-shaped Noise

In Section 5.1 we discussed that existing work in CASA and BSS has treated either the reverberant target as the desired signal [98, 142, 147] or the direct sound of the target as the desired signal [124, 137] to define the IBM. We refer to these masks as IBM-R and IBM-DS, respectively. The psychoacoustics literature motivates a third IBM definition, IBM-ER, that includes early reflections of the target source as part of the desired signal, but treats late arriving reflections as interference. In this experiment

we measure SRTs for IBM-processed reverberant and noisy speech with these three alternative definitions. For the IBM-ER mask, we use a reflection boundary of 50 ms [13]. The IBM-DS and IBM-R definitions correspond to reflection boundaries of 0 and ∞ , respectively.

In this experiment we measure sentence-level SRTs with a male speaker corrupted by both reverberation and speech-shaped noise (SSN). We consider three different T_{60} times: 0, 0.4 and 0.8 s. With 0 s T_{60} , there is no difference between mask definitions, thus we test only an IBM-processed and unprocessed condition. With 0.4 and 0.8 s T_{60} , we measure SRTs for each IBM definition and an unprocessed condition. In total, we calculate SRTs for ten different conditions.

5.3.1 Method

Materials

The target speech signals used in this experiment all contain the same male speaker reading individual sentences from the HINT corpus [132]. The HINT corpus contains 25 lists of 10 sentences that are phonetically balanced and equated for naturalness, difficulty, length, and reliability. Sentences within each list follow a predictable subject-verb-object syntactic structure, and range in length between four and seven words. Monaural signals were recorded in a controlled environment and digitized at 22.05 kHz with 16-bit quantization. The SSN signal used as interference in this experiment was generated by filtering white noise with the long-term average spectrum of the male talker used as target speech.

Impulse responses to simulate room reverberation were generated using the image method [3] with the ROOMSIM package [25]. The parameters of the simulation were as follows. The room size was set to 15 m \times 13 m \times 3.3 m. The sound source and microphone positions were set to [9.5 m, 11 m, 1.2 m] and [9.5 m, 7 m, 1.2 m], respectively. As such, the sound source was positioned directly in front of the microphone (0° azimuth and 0° elevation) at a distance of 4 m. The reflective characteristics of the room surfaces were set to be frequency-independent and to be the same at each surface so that a single parameter controlled the T_{60} time. Impulse responses were generated with this configuration for T_{60} equal to 0, 0.4 and 0.8 s. Note that monaural impulse responses were generated assuming a single, omni-directional microphone.

Stimuli

In order to generate test stimuli, a specified target speech utterance was convolved with a room impulse response for a given T_{60} time. The root-mean-square (RMS) level of the reverberant target speech was normalized to match the RMS level of a 64 db SPL white noise signal. An interference signal was then created by convolving the SSN signal with the same impulse response used for the target speech utterance. The level of the reverberant interference signal was adjusted to achieve a specified SNR relative to the reverberant target.

Desired and residual signals were then generated as in Equations (5.2) and (5.3). For the IBM-DS definition, the desired signal was generated by convolving the target

speech signal with the first impulse of the selected impulse response, which corresponds to the direct sound. For the IBM-ER definition, the desired signal was generated by convolving the target speech signal with the first 50 ms (relative to the direct sound component) of the selected impulse response. For the IBM-R definition, the desired signal corresponds to the reverberant target speech utterance. Given a mixture and specified desired signal, the residual signal was generated by subtracting the desired signal from the mixture.

The cochleagram representation of both desired and residual signals were then generated (see Section 5.2) and the energy of both desired and residual signals in each T-F unit was calculated and used in Equation (5.4) to generate an IBM. The LC parameter was fixed at -6 dB, as suggested in [20, 179]. Masks were then applied to the mixture cochleagram in a synthesis stage to generate time-domain test stimuli [178]. Note that test stimuli for the unprocessed conditions were generated by applying an “all-one” mask to mixture cochleagrams in the synthesis stage.

Procedure

An operator controlled the experiment using a PC running Matlab software. Subject and operator were seated inside of a sound attenuating booth. Stimuli were presented diotically with Sennheiser HD 280 Pro headphones. Subjects were given sufficient time to repeat or guess the sentence content and the operator recorded whether or not the sentence was correct. A sentence was considered correct if all the keywords were correct. The only substitutions allowed were: a/the, an/the, is/was, are/were,

has/had and have/had. Each trial lasted approximately an hour and consisted of a training phase followed by SRT testing in each of the ten conditions outlined above. The training phase was performed with two lists (twenty sentences) of unprocessed HINT sentences (i.e. no reverberation, noise or IBM processing) to familiarize listeners with the test procedure and to ensure audibility of the target speech. All listeners obtained 100% recognition in this phase.

A one-up one-down adaptive procedure was used to measure SRTs at 50% sentence-level accuracy. Twenty-five sentences were used for each test condition. The first five were randomly selected from three held-out HINT lists and used to converge on an initial SRT, while the final twenty sentences (two HINT lists) were unique to each condition and used to calculate the final SRT. The first sentence was presented at a fixed initial SNR (determined in a pilot study) and repeated while increasing the SNR by 2 dB with each presentation until at least half of the words in the sentence were recognized. After the first sentence, the SNR was decreased by 2 dB when the subject repeated the previous sentence correctly, and increased by 2 dB otherwise. A Latin square design was used to generate the sequence of test conditions for each subject and specify the unique HINT lists to be used for each condition.

Subjects

Twelve normal hearing, native speakers of American English participated in the experiment with ages varying between 18 and 27 with an average of 22. The subjects

were paid for their participation. Although their audiograms were not evaluated, the subjects reported that they were unaware of any hearing problems.

5.3.2 Results

Figure 5.1(a) shows the average SRT values (in dB) for each of the ten test conditions. Results are grouped by T_{60} time and gray-scale values correspond to different processing methods. In the anechoic condition ($T_{60} = 0$), we label the IBM-processed result with IBM-DS. Again, in this case no reflected energy is included in the mixture so the desired signal for all three IBM definitions corresponds to only the direct sound component of the target signal.

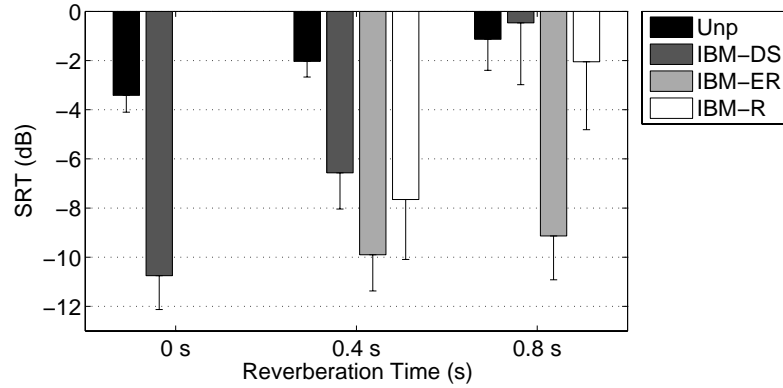
We first note that the mean SRTs obtained for the 0 s T_{60} condition are in good agreement with the existing literature. The SRT for the unprocessed mixtures is -3.41 dB, which is slightly lower than the value of -2.92 dB reported in [132]. The mean SRT for the IBM-processed mixtures is -10.75 dB, meaning that the benefit of the IBM is 7.3 dB. Wang *et al.* reported a benefit of 7.4 dB in a similar condition, although a different speech database and recognition task were used.

For the 0.4 s T_{60} condition, we see that each IBM definition is able to lower the SRT relative to the value of -2 dB obtained with unprocessed mixtures. The IBM-ER mask yields the largest benefit of 7.9 dB, whereas the benefit of the IBM-DS and IBM-R masks is 4.5 dB and 5.6 dB, respectively. However, in the 0.8 s T_{60} condition, only IBM-ER achieves an improvement relative to the unprocessed case. The average

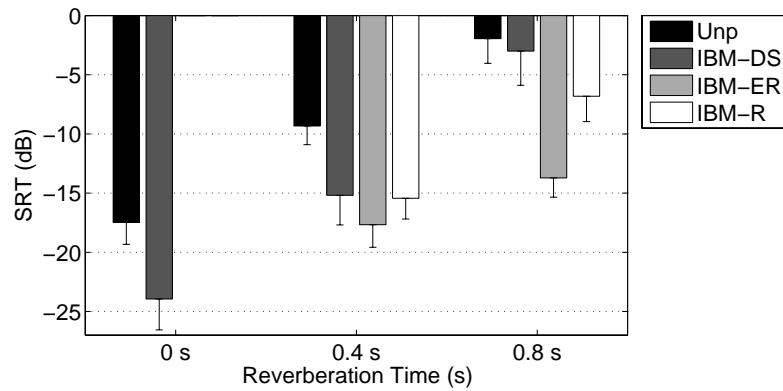
SRT obtained with IBM-ER is -9.1 dB as compared to -1.1 dB for the unprocessed mixtures.

A two-way analysis of variance (ANOVA) with repeated measures was performed on the measured SRTs. The ANOVA revealed a significant effect due to processing type [$F(3, 33) = 145.15, p < 0.001$], T_{60} time [$F(2, 22) = 108.12, p < 0.001$] and interaction between the two [$F(6, 66) = 37.06, p < 0.001$]. Paired t -tests with a Bonferroni *post-hoc* correction showed a significant difference between unprocessed and IBM-processed conditions for each T_{60} time. For T_{60} equal to 0.4 s and 0.8 s, IBM-ER significantly lowered SRTs as compared to unprocessed mixtures and the IBM-DS and IBM-R definitions.

We also note the slight increase in SRT as the reverberation time is increased for both the unprocessed mixtures and the best performing IBM definition. There is an intuitive explanation for this trend. As noted above in Section 5.3.1, the input SNR for each mixture reflects the energy of the fully reverberant target relative to the SSN interference. Consistent with the motivation behind the IBM-ER definition, some of the energy contained in the reverberant target signal is detrimental to perception of the target signal. Thus, for a fixed input SNR, as the level of reverberation is increased, the target signal becomes more difficult to recognize and correspondingly, SRTs increase slightly.



(a) SSN Masker



(b) Female Masker

Figure 5.1: Average SRTs measured for ten test conditions with SSN interference (a) and a female talker interference (b). In both (a) and (b), data is grouped according to T_{60} time. Gray-scale values indicate the processing method used. Black corresponds to the unprocessed condition ('Unp'), dark gray to IBM-DS, light gray to IBM-ER and white to IBM-R. A lower SRT corresponds to better performance. Error bars indicate 95% confidence intervals around the mean values.

5.4 Experiment 2: The Effect of IBM Processing on Reverberant Speech Mixed with a Competing Talker

In this experiment we measure sentence-level SRTs with a male speaker corrupted by both reverberation and reverberant female talker. As in Experiment 1, we consider three alternative IBM definitions and three reverberation times for the same ten test conditions.

5.4.1 Method

The procedures used in this experiment are nearly identical to those used in Experiment 1. The same set of HINT utterances recorded with a male speaker were used as the target signals. In this case, however, female speech signals were used as interference. The female speech signals were recorded monaurally in a controlled environment and digitized at 22.05 kHz with 16-bit quantization. A set of 40 sentences from the Harvard Sentence List [92] was used. The interference utterance was selected randomly for each test stimulus.

Mixture signals were generated by passing both target and interference speech signals through a selected impulse response and summing the resulting reverberant signals, as described in Section 5.3.1. Again, the target speech level was controlled to match the RMS level of a 64 dB SPL white noise signal, and the reverberant interference was scaled to achieve the specified input SNR. Desired signals, residual signals, IBMs and test stimuli were generated as described in Section 5.3.1.

The procedure used for measuring SRT is identical to that described in Section 5.3.1 with one exception. The fixed initial SNR for the first presentation of the first sentence was lowered relative to the one used in Experiment 1 based on a pilot experiment.

Twelve normal hearing, native speakers of American English that did not participate in Experiment 1 participated in this experiment. Their ages varied between 19 and 23 with an average of 21. As before, the subjects were paid for their participation and they reported no problems with their hearing. Again, all the participants obtained 100% recognition during the training phase.

5.4.2 Results

Figure 5.1(b) shows the average SRT values (in dB) for each of the ten test conditions. Results are grouped by T_{60} time and gray-scale values correspond to different processing methods. As described above for Figure 5.1(a), we label the IBM-processed result with IBM-DS for the anechoic condition.

We first note the large increase in SRT for the unprocessed mixtures as T_{60} is increased. Whereas in Experiment 1 with the SSN masker we saw a slight increase due to the effect of reverberation on the target signal, the increase in SRT for the unprocessed mixtures is 15.5 dB from the anechoic to the most reverberant condition. This large change is due the fact that late reverberation of both talkers creates a SSN-like background that reduces the “glimpsing” opportunities for perception of the target talker.

A two-way ANOVA with repeated measures was performed on the measured SRTs. As in Experiment 1, the ANOVA revealed a significant effect due to processing type [$F(3, 33) = 85.2, p < 0.001$], T_{60} time [$F(2, 22) = 798.6, p < 0.001$] and interaction between the two [$F(6, 66) = 21.52, p < 0.001$]. Paired t -tests with a Bonferroni *post-hoc* correction again showed a significant difference between unprocessed and IBM-processed conditions for each T_{60} time. For 0.4 s T_{60} , the difference between IBM-DS and IBM-ER was not significant ($p = 0.014$), while the IBM-ER definition achieved significantly lower SRTs than unprocessed, IBM-DS and IBM-R for T_{60} equal to 0.8 s.

5.5 Discussion of Experiments 1 and 2

The results from Experiments 1 and 2 are consistent in that IBM-ER achieved the lowest SRTs for both T_{60} times and interference types. The large decrease in SRT suggest that the IBM-ER definition effectively characterizes signal energy as either beneficial or harmful to speech intelligibility. While the IBM-DS and IBM-R definitions improved intelligibility for the 0.4 s condition, this benefit was either eliminated or substantially reduced for both definitions in the longer T_{60} time of 0.8 s. The degradation of the IBM-DS mask can be understood by recognizing that early reflections are treated as undesirable, or harmful to intelligibility. As a result, early reflections decrease the effective SNR of the desired signal in the IBM-DS definition and T-F units that contain beneficial speech information may be attenuated. Simply put, the IBM-DS masks become too sparse in the 0.8 s condition. On the other hand,

the IBM-R definition treats late reflections as desirable. In this case T-F units that do not contain beneficial speech information may be selected, resulting in perceptual artifacts (i.e. musical noise) during the reverberation tail of the target signal and less effective suppression of noise energy.

While we expect the observations from Experiments 1 and 2 to generalize well to other acoustic conditions, the magnitude of those effects may vary. Essentially, while we expect to observe that IBM-ER outperforms IBM-DS and IBM-R in all conditions, the direct-to-reverberant energy ratio and the T_{60} time will influence the degree of dissimilarity between these masks.

In these first two experiments, we considered only a single reflection boundary value of 50 ms between the two extremes of 0 ms and infinite. While this value is motivated from existing literature [13], it may not be optimal for ideal binary masking. We explore the effect of changes to the reflection boundary more thoroughly in our fourth experiment discussed in Section 5.7.

5.6 Experiment 3: Interaction Between Reflection Boundary and SNR Threshold

In Experiment 1 we measured SRTs for IBM-processed reverberant speech mixed with a SSN masker. Results showed that the IBM-ER definition achieved the lowest SRT in both reverberant conditions and that both the IBM-DS and IBM-R definitions failed to achieve an appreciable benefit in the 0.8 s T_{60} condition. As discussed

above in Section 5.5, degradation of the IBM-DS performance is due to a decrease in *effective* SNR caused by treating reflections as part of the residual signal. With a fixed LC of -6 dB, as used in Experiment 1, the resulting IBM-DS masks are too sparse and therefore attenuate too much target speech energy. For the IBM-R definition, the relatively larger effective SNR results in masks that do not effectively attenuate detrimental energy when using the -6 dB LC.

While the choice of -6 dB LC is supported by the studies of [20,179], it is possible that this LC favored the IBM-ER definition. To explore the interaction between mask definition and local SNR threshold, in this experiment we measure intelligibility of IBM-processed reverberant and noisy speech for the same IBM definitions over a range of SNR threshold values. To do so, we fix the input SNR and reverberation time across all test stimuli and measure the percent of correctly recognized sentences for each processing condition. We set T_{60} to be 0.8 s, as this time produced the largest differences between mask definitions in Experiment 1. We set the input SNR to be -1 dB to match the SRT for the unprocessed condition at this T_{60} time. Thus, we expect sentence recognition to be near 50% for the unprocessed condition in this experiment, while improvements due to IBM processing should increase recognition scores.

To ensure that an appropriate range of local SNR thresholds are considered for each mask definition, we employ the concept of the *relative criterion* (RC) in this experiment [104]. The RC is equal to the LC minus the input SNR of the mixture, and was motivated by the observation that co-varying the input SNR and LC does

not change the resultant IBM (assuming a linear filterbank) [20, 104]. As such, for a fixed RC, changes to the input SNR will have no effect on the IBM generated. In the current experiment we do not explicitly change the input SNR (i.e. the level of SSN is fixed), however increasing the reflection boundary will systematically increase the effective SNR due to shifting some reflections of $h[n]$ from being included in $v_b[n]$ to being included in $x_b[n]$. In this case, fixing the RC across different mask definitions does not ensure precisely the same IBM, but ensures that the mask density (balance between 1s and 0s) will be similar across IBM definitions.

We evaluate intelligibility with the IBM-DS, IBM-ER and IBM-R definitions for seven different RC values: -30 dB, -15 dB, -9 dB, -6 dB, -3 dB, 0 dB and 6 dB. Based on the study of [104], we expect the best performance for each mask definition to occur in the RC range between -12 and 0 dB, so we have focused on this range. This gives twenty-one IBM-processed conditions to which we add two unprocessed conditions. First, intelligibility of unprocessed mixtures is measured, where as noted above, we expect to see roughly 50% recognition. Second, intelligibility of unprocessed reverberant target speech is measured. In this case no SSN is added and we can directly assess the impact of reverberation on target speech intelligibility. In total we evaluate twenty-three test conditions.

5.6.1 Method

The materials used and generation of mixture signals is identical to that of Experiment 1. In this case however, we consider only T_{60} equal to 0.8 s and fix the input SNR

at -1 dB. The process used to generate IBMs differed from Experiment 1 due to the use of the RC rather than a fixed LC. Specifically, for a given mask definition, desired and residual signals were created as described in Section 5.3.1. Given a desired and residual signal, the effective SNR was measured as,

$$\text{SNR}_b = 10 \log_{10} \left(\frac{\sum_n x_b[n]^2}{\sum_n v_b[n]^2} \right). \quad (5.5)$$

For a specified RC, the LC used in Equation (5.4) to generate the IBM was then set as $\text{LC} = \text{RC} + \text{SNR}_b$. Note that while there was some consistency in the effective SNR across mixtures for a given IBM definition, the effective SNR and, therefore, the LC was mixture dependent. As in Experiment 1, the cochleagram representation was used to generate IBMs and masks were applied to mixture cochleagrams in a synthesis stage to generate a time-domain stimuli. Again, stimuli for unprocessed conditions were generated by applying an “all-one” mask to mixture or reverberant target cochleagrams.

The physical setup of the experiment was identical to Experiment 1. In this case however, each trial consisted of a short training phase followed by testing in the twenty-three test conditions described above. Each trial lasted about an hour and subjects were told that breaks were available as needed. A single list of HINT sentences was used for the training phase for all subjects to ensure audibility of target speech and familiarize subjects with the procedure. Unprocessed clean HINT utterances were used and all listeners obtained 100% recognition in the training phase. A single list of HINT sentences was then used to obtain a recognition score for each

of the twenty-three test conditions. The sequence of test conditions for each subject and the unique HINT list used for each condition were randomized.

Seven normal hearing, native speakers of American English that did not participate in Experiment 1 or 2 participated in this experiment. Their ages varied between 21 and 33 with an average of 24. As before, the subjects were paid for their participation and they reported no problems with their hearing.

5.6.2 Results

Figure 5.2(a) shows the average percentage of correctly identified sentences for the two unprocessed and twenty-one IBM-processed conditions. Unprocessed conditions are shown on the left with a square (unprocessed mixtures, “Unp”) and circle (unprocessed reverberant target, “UnpR”) marker. Processed conditions are shown as bars grouped by RC with the three alternative mask definitions indicated by different gray-scale values. Standard deviations are shown with error bars.

We first note that average recognition for the unprocessed mixtures is 42.9%, which is slightly lower than the expected 50% predicted by Experiment 1, although 50% is well within a single standard deviation around the measured average. Subjects achieved an average accuracy of 91.4% on the unprocessed reverberant target speech. As all subjects achieved 100% during the training phase, this reveals some degradation of intelligibility due to reverberation alone and is in agreement with the [129], which reports 92.5% accuracy for a similar condition.

A two-way ANOVA with repeated measures was performed on the rationalized

arcsine transform of recognition percentages from all IBM-processed conditions. This revealed a significant effect due to RC [$F(6, 36) = 72.0, p < 0.001$], reflection boundary [$F(2, 12) = 106.47, p < 0.001$] and interaction between the two [$F(12, 72) = 6.29, p < 0.001$].

Paired *t*-tests with a Bonferroni *post-hoc* correction showed that both IBM-DS and IBM-ER significantly improve recognition as compared to the unprocessed mixtures for RC values between -15 dB and 0 dB. The peak scores for both mask definitions exceed 95% recognition. The average effective SNR for the IBM-DS definition over the entire corpus of 250 target sentences is -12.2 dB. As the range of effective RC values is between -15 dB and 0 dB, the range of effective LC values for this mixture condition is then between about -27 dB and -12 dB, substantially lower than the -6 dB LC used in Experiment 1. The average effective SNR for the IBM-ER definition is -5.5 dB. This suggests that the optimal LC range for this mixture condition is between about -20 dB and -5 dB, which contains the -6 dB LC used in Experiment 1.

These results indicate that the IBM-DS mask *can* be effective provided that the impact of reverberation on the effective SNR is accounted for. The similar performance between the IBM-DS and IBM-ER definitions in this experiment suggests that one can account for the impact of early reverberation by either lowering the LC used with the IBM-DS definition, or increasing the reflection boundary value with the IBM-ER definition. Consistent with the results from Experiment 1 and 2, this

experiment shows that it is vital that the impact of early reverberation is accounted for to increase intelligibility.

Paired t -tests with the Bonferroni correction showed that the IBM-R definition achieved a significant improvement relative to unprocessed mixtures for $RC = -9$ dB. As the effective SNR for this definition is equal to the -1 dB input SNR, the LC in this case is -10 dB, fairly close to the -6 dB threshold used in Experiment 1. While the average recognition in this case is 64.3%, 21.4% higher than for unprocessed mixtures, this was significantly lower than recognition using IBM-DS, IBM-ER and recognition of the unprocessed reverberant target speech. This shows that the poor performance in Experiment 1 with the IBM-R definition was not a result of the -6 dB LC used. Since intelligibility of the unprocessed reverberant target signals is significantly higher than the intelligibility with the IBM-R definition, performance cannot be explained by retention of reverberant target energy alone.

We illustrate results as a function of LC in Figure 5.2(b), where for each mask definition, we shift the performance as a function of RC by the average effective SNR for the definition.

5.7 Experiment 4: The Effect of IBM Processing on Reverberant Speech

Experiment 3 revealed that increases in intelligibility are possible with both the IBM-DS and IBM-ER definitions provided that an appropriate SNR threshold is used, but

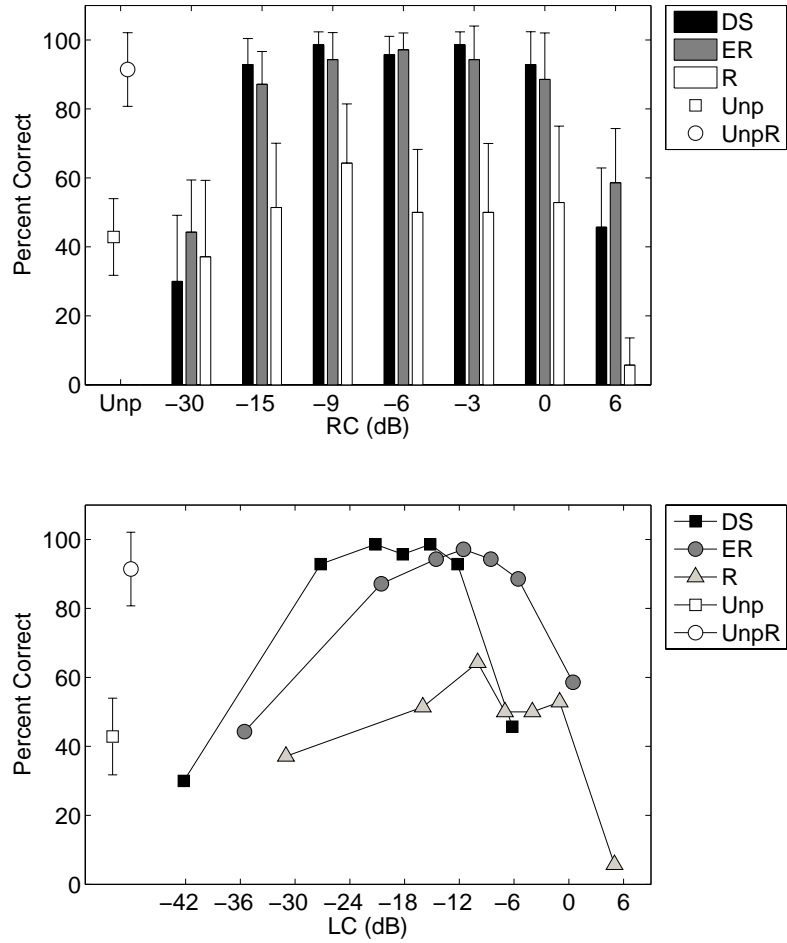


Figure 5.2: Average percentage of correctly recognized sentences for the two unprocessed conditions and twenty-one IBM-processed conditions tested in Experiment 3. Recognition shown as a function of RC (a) and LC (b). Error bars in (a) indicate standard deviation.

that the IBM-R definition was not able to improve speech intelligibility regardless of the SNR threshold. This suggests that a reflection boundary of 50 ms effectively characterizes mixture energy as either beneficial or detrimental to speech perception, but that including all reflections in the desired signal does not. In this experiment we analyze the effect of four reflection boundary settings to better understand the point in time for which the reflection boundary value no longer creates IBMs that benefit intelligibility. We focus on the case with reverberation only (i.e. no additive noise) to highlight differences due to the reflection boundary. We consider long T_{60} times so that intelligibility of the unprocessed speech is expected to be degraded relative to anechoic speech. In keeping with Experiment 3, we test intelligibility as a function of local SNR threshold using RC.

Intelligibility results were measured for three T_{60} times: 2 s, 3 s and 30 s. The condition with $T_{60} = 2$ s corresponds to the speech reception reverberation threshold (SRRT) at 50% accuracy [69]. The condition with $T_{60} = 3$ s was chosen to exaggerate the effect of both IBM processing and the effect of reflection boundary values. In the $T_{60} = 30$ s condition, the unprocessed signal becomes essentially speech shaped noise. This condition was chosen to validate the IBM-processed noise condition presented in [179].

For each T_{60} time tested, we considered four reflection boundaries: 5 ms, 50 ms, 100 ms and 200 ms. For the 5 ms and 50 ms reflection boundaries we tested RC values of -30 dB, -15 dB, -9 dB, -6 dB, -3 dB, 0 dB and 6 dB. A subset of these values were tested for the 100 ms and 200 ms conditions due to a limitation in the

number of unique HINT lists. For each reverberation time, an unprocessed condition was also tested giving us a total of twenty-four conditions for each T_{60} time.

5.7.1 Method

The experimental method used was similar to the one described in Section 5.6.1. As in Experiments 1-3, the same set of HINT utterances recorded with a male speaker were used as the target signals. In this experiment, no additive interference was incorporated. The main difference from the previous experiments was the use of exponentially decaying impulse responses generated as described in [69]. While impulse responses generated in this manner are admittedly a crude approximation of room reverberation, we felt it was important to follow existing literature on speech perception in reverberation [69] and more complex simulations such as the image method require setting many additional parameters such as source and microphone positions, room geometry and the reflective characteristics of wall surfaces.

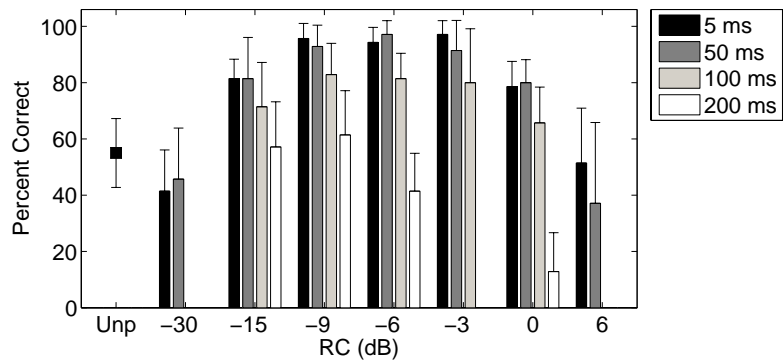
Specifically, we let $h[n] = \epsilon[n]e^{-6.91n/T_{60}}$, where $\epsilon[n]$ is a white noise signal and the time constant for the envelope decay is equal to $T_{60}/6.91$ [106]. Impulse responses were truncated to have length equal to T_{60} . Impulse responses were generated with a sampling frequency of 22.05 kHz to match the sampling frequency of the target speech corpus. To generate mixture signals, multiple copies of the target speech signal were concatenated before convolving with the impulse response and the last copy was used as the reverberant speech signal. This ensured that the impact of reverberation was present throughout the duration of the target speech signal. As in Experiments 1-3,

the RMS level for the reverberant speech was fixed across all stimuli and set to match the RMS level of a 64 dB white noise signal.

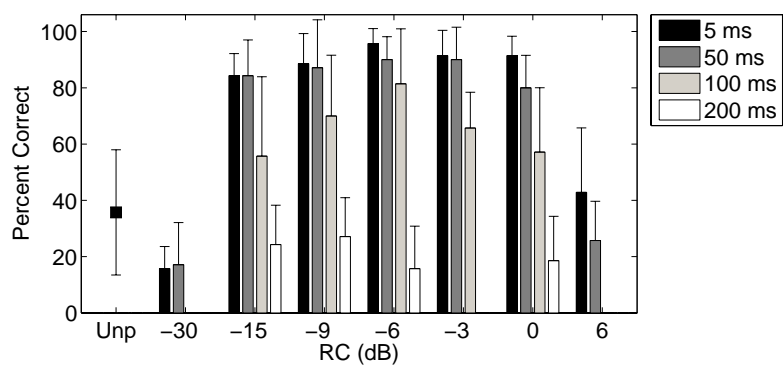
The desired and residual signals used to generate IBMs were created as described for the IBM-ER definition in Section 5.3.1. IBMs were created given a specified RC value as described in Section 5.6.1. Again, the cochleagram representation was used to generate IBMs, masks were applied to mixture cochleagrams in a synthesis stage to generate time-domain stimuli and stimuli for the unprocessed condition were generated by applying “all-one” masks to the mixture cochleagrams.

The procedure used to measure intelligibility was identical to Experiment 3 (see Section 5.6.1), where each trial followed training on one list of clean HINT sentences with testing on each of the twenty-four experimental conditions, with one unique HINT list for each condition. Again, all listeners obtained 100% recognition in the training phase, and the sequence of test conditions for each subject and the unique HINT list used for each condition were randomized.

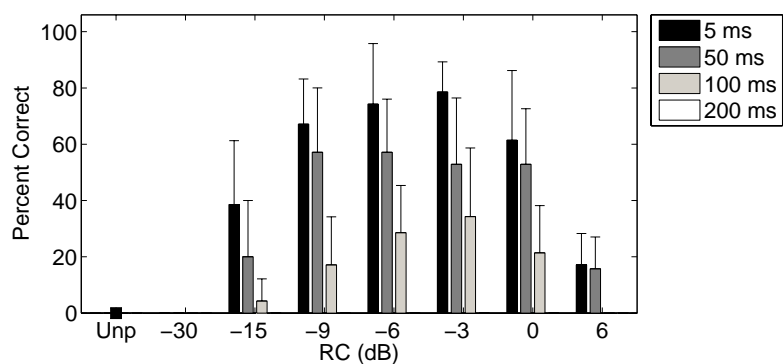
Twenty-one normal hearing, native speakers of American English that did not participate in Experiments 1-3 participated in this experiment. Seven subjects participated for each T_{60} condition. Their ages varied between 19 and 32 with an average of 22. As before, the subjects were paid for their participation and they reported no problems with their hearing.



(a) $T_{60} = 2$ s



(b) $T_{60} = 3$ s



(c) $T_{60} = 30$ s

Figure 5.3: Average percentage of correctly recognized sentences for the unprocessed condition and twenty-three IBM-processed conditions tested in Experiment 4. Recognition shown as a function of RC for T_{60} equal to 2 (a), 3 (b), and 30 s (c). Error bars indicate standard deviation.

5.7.2 Results

Figure 5.3 shows the average percentage of correctly identified sentences for all test conditions and T_{60} times. Unprocessed conditions are shown on the left with a square marker. Processed conditions are shown as bars grouped by RC with the results for different reflection boundary values indicated by different gray-scales. Standard deviations are shown with error bars.

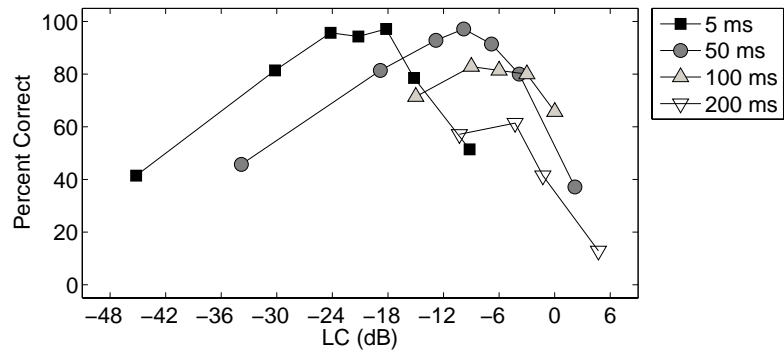
For the 2 s T_{60} conditions shown in Figure 5.3(a), we note that average recognition for the unprocessed mixtures is 55%, which is in good agreement with the 50% predicted by the SRRT [69]. Paired *t*-tests with a Bonferroni *post-hoc* correction on the rationalized arcsine transform of recognition percentages were performed to identify IBM-processed conditions that achieved significant intelligibility improvements relative to the unprocessed condition. These tests showed that significant improvements were achieved for reflection boundary of 5, 50 and 100 ms, while no significant improvement was obtained for reflection boundary of 200 ms. The highest recognition scores for the 5, 50 and 100 ms reflection boundaries were obtained with RCs between -9 dB and -3 dB. There was no significant difference observed in this range between the 5 and 50 ms reflection boundaries, while recognition with both 5 and 50 ms reflection boundaries were significantly higher than recognition with the 100 ms boundary at RC equal to -6 dB.

Similar results were obtained for the 3 s T_{60} conditions. In this case, average accuracy on the unprocessed mixtures was 35.7%, whereas peak accuracy for the 5,

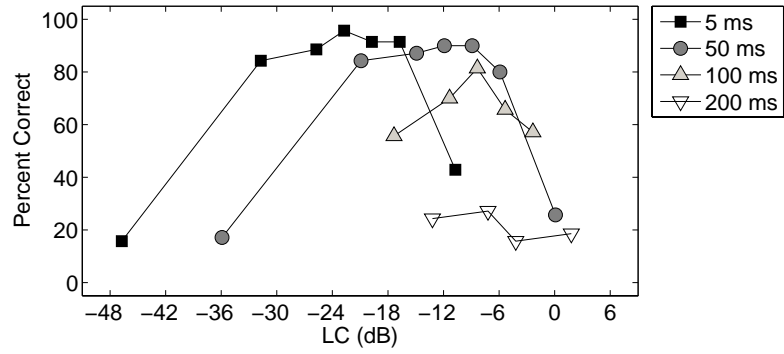
50, 100 and 200 ms reflection boundaries were 95.7%, 90%, 81.4% and 27.1%, respectively. As stated above, the 30 s T_{60} conditions serves as a follow up to [179], which showed high intelligibility for IBM-processed noise. Wang *et al.* illustrated that the spectro-temporal pattern of the binary mask carries sufficient information for speech recognition. In this study we do not explicitly mask SSN, but as noted above, speech reverberated with such an unrealistically long T_{60} becomes quite similar to SSN. In fact, subjects informally reported hearing only noise for unprocessed mixtures in this condition. As expected, recognition on the unprocessed mixtures was 0%. Consistent with [179], intelligible speech could be induced by IBM-processing. In this case, recognition was highest with the 5 ms reflection boundary, where peak accuracy exceeded 75%. Performance between each reflection boundary was significantly different, as indicated by paired *t*-tests with a Bonferroni *post-hoc* correction on the rationalized arcsine transform of mean recognition percentages.

A three-way ANOVA across data from all conditions and T_{60} times revealed that all three main effects of T_{60} time, reflection boundary and RC value were significant [$F(2, 414) = 312.59, p < 0.001$; $F(3, 414) = 169.7, p < 0.001$; $F(6, 414) = 132.24, p < 0.001$]. There was also a significant interaction between the T_{60} time and the reflection boundary [$F(6, 414) = 7.73, p < 0.001$], and between the T_{60} time and the RC value [$F(12, 414) = 5.45, p < 0.001$].

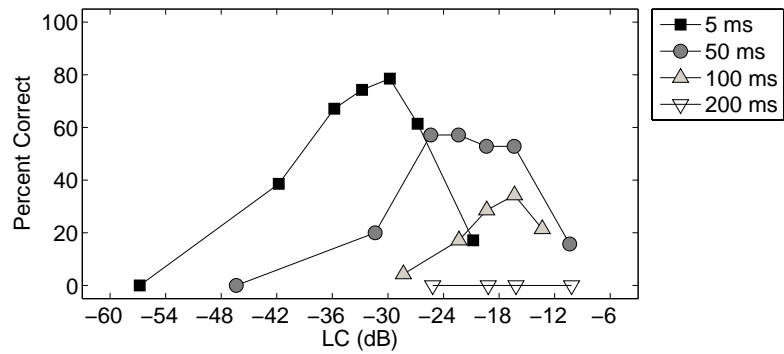
As in Experiment 3, we show results as a function of LC, rather than RC, in Figure 5.4. Again, for each reflection boundary, we shift the performance as a function of RC by the average effective SNR over 250 mixtures created with each of the



(a) $T_{60} = 2$ s



(b) $T_{60} = 3$ s



(c) $T_{60} = 30$ s

Figure 5.4: Average percentage of correctly recognized sentences for the unprocessed condition and twenty-three IBM-processed conditions tested in Experiment 4. Recognition shown as a function of LC for T_{60} equal to 2 (a), 3 (b), and 30 s (c).

HINT sentences. The average effective SNR for the 5, 50, 100 and 200 ms reflection boundaries is -15.2 dB, -3.8 dB, 0 dB and 4.7 dB, respectively, for T_{60} equal to 2 s. For T_{60} equal to 3 s, the effective SNRs drop to -16.7 dB, -5.9 dB, -2.3 dB and 1.8 dB, and for T_{60} equal to 30 s, effective SNRs are -26.8 dB, -16.4 dB, -13.3 dB and -10.8 dB, respectively.

5.8 Discussion

The ideal binary mask has been proposed as a main computational goal of CASA systems and has been commonly used as a performance upper bound for segregation based on T-F masking. Although the IBM has been extensively studied for anechoic signals with additive interference, no existing work has investigated extending the IBM to reverberant conditions. As discussed throughout this chapter, several definitions of the IBM are possible when dealing with reverberant speech. We introduce the concept of the reflection boundary to the IBM definition in order to parameterize the treatment of target speech reflections. The experiments presented in this chapter analyze the intelligibility of IBM-processed speech with various reflection boundaries and local SNR thresholds. Several conclusions can be drawn from the results presented here.

First, it is clear that, provided the mask is defined appropriately, binary T-F masking can improve intelligibility of target speech in reverberant and noisy conditions.

The experiments presented are, to our knowledge, the first studies that firmly establish this point. This outcome is crucial for engineering systems that seek to improve speech intelligibility in reverberant environments based on binary T-F masking.

Second, as discussed in Section 5.1, one common choice in the CASA field has been to treat the reverberant target as the desired signal in the definition of the IBM. Experiments 1-3 show that this definition is at best suboptimal and in many cases, does not lead to improved speech intelligibility. Particularly important is the outcome of Experiment 3, which shows that intelligibility of speech processed by IBM-R is significantly worse than intelligibility of unprocessed reverberant speech. This suggests that the IBM is not capable of effectively restoring the perception of a reverberant target signal by removing additive noise.

Third, all four experiments establish that the effect of early reflections must be accounted for in the IBM definition. The reflection boundary parameter allows one to do so in a manner that is most consistent with the literature on speech perception in reverberation and to utilize an SNR threshold in the range that is functional for anechoic speech with additive noise. Results from Experiments 3 and 4 show that it is also possible to improve intelligibility using the direct-sound based IBM (or a very short reflection boundary), but in this case, one must account for the substantial reduction in effective SNR caused by reverberation by lowering the SNR threshold. This suggests that another common choice of IBM used in the literature, IBM-DS with 0 dB LC, is a poor choice in conditions with low direct-to-reverberant ratios.

When considering the IBM as a performance upper bound for CASA algorithms,

the appropriate definition may, to some extent, be a matter of perspective. The use of IBM-R is motivated by the viewpoint that human listeners do not perform “dereverberation” of the stream of interest, and thus CASA should not seek to remove target reverberation. However, the psychoacoustics literature has established that late reverberation is not integrated into perception of the target and acts as masking noise. We argue then that if the purpose of CASA is to segregate a target signal consistent with a perceptual stream, both late reverberation and additive interference should be removed. Similarly, perceptual studies show that early reflections are integrated into the perception of target speech. Thus, while it is clearly possible to improve intelligibility using the IBM-DS definition, the characterization of mixture energy as either beneficial or detrimental is inconsistent with auditory perception. We contend that utilizing a reflection boundary in a reasonable range (e.g. 50 - 100 ms) is most consistent with human speech perception and, therefore, the most conceptually appealing alternative.

CHAPTER 6

BINAURAL DETECTION, LOCALIZATION AND SEGREGATION

In this chapter we propose a binaural system for joint localization and segregation of an unknown and time-varying number of sources. The proposed system is considerably more flexible and requires less prior information than the systems presented in Chapters 3 and 4. A preliminary study with this system was published in [192].

6.1 Introduction

In this chapter we propose a binaural system for joint localization and segregation of an unknown and time-varying number of sources. In keeping with the theme of this dissertation and with the systems presented in Chapters 3 and 4, we incorporate both monaural and binaural cues. Whereas the systems described in previous chapters performed simultaneous organization using monaural cues and sequential organization using binaural cues in a two-stage process, pitch and azimuth cues are

considered jointly for simultaneous organization by the system proposed in this chapter. This approach retains the benefit of pitch-based grouping in the formation of simultaneous streams, but allows for improved performance when pitch continuity alone leads to incorrect grouping across continuous time intervals. Further, by training models jointly on pitch and azimuth cues, the relative contribution of each type of cue is learned and the system naturally deals with both voiced and unvoiced speech. This approach has the potential to reconcile the observation that monaural cues are stronger than spatial cues for simultaneous organization [41, 89, 156], but that spatial cues may contribute when circumstances allow (e.g. low reverberation, well separated sources, ambiguous monaural cues) [40, 44, 45, 56, 157].

The proposed system incorporates many of the concepts presented in previous chapters. We utilize the multipitch tracking algorithm described in Section 4.3.1, incorporate the azimuth-dependent models presented in Section 4.2.2 and extend the penalized maximum likelihood method outlined in Section 4.6.4 to handle detection of an unknown number of sources across time. To assess performance we use an IBM that includes early reflections in the definition of the desired signal, as proposed in Chapter 5. We integrate these methods using a novel hidden Markov model (HMM) framework to estimate the number of active sources across time, estimate the azimuth of each active source per frame, assign pitch estimates to the corresponding azimuth, and generate a binary T-F mask for each active source. We focus on segregation of sources in fixed spatial positions, however the framework is amenable to the situation with moving sources through inclusion of a motion model.

As outlined in Section 2.4, several existing systems have considered joint estimation of pitch and TDOA from a microphone pair [31,95,99,131], although these studies do not provide a framework for dealing with multiple sources. Two-microphone segregation based on pitch and spatial cues has been investigated in [50,130,159,194,195], however these methods assumes a known and fixed number of sources (usually two) [50,130,195], or track only the pitch and azimuth of the dominant source [159,194]. While many of these multi-cue approaches are relevant, we are not aware of an existing system that can perform localization, pitch tracking and segregation of an unknown and time-varying number of sources.

In the following section we describe the front-end processing, define the computational goal and provide an overview of the proposed framework. In Section 6.3 we outline the acoustic features used. We introduce each component of the HMM framework in Section 6.4 and describe how estimates of the target signal are generated in Section 6.5. Finally, we outline the evaluation methodology and results in Sections 6.6 and 6.7, and conclude with a discussion in Section 6.8.

6.2 Overview

We utilize the same auditory front-end described in Section 4.2.1. Again, we denote a T-F unit as $u_{c,m}^E$ where $E \in \{L, R\}$ indicates the left or right ear signal, m indexes time frames and c indexes frequency channels.

The goal of the proposed system is to estimate the IBM. To asses performance we utilize an IBM definition that includes early reflections in the desired signal, as

presented in Chapter 5. As the formulation of the IBM with reflection boundary parameter (see Section 5.2) dealt with monaural signals, we reiterate the concepts here and propose an IBM definition suitable for binaural signals.

We model each T-F unit as,

$$u_{c,m}^E = \sum_k x_{k,c,m}^E + v_{c,m}^E, \quad (6.1)$$

where $x_{k,c,m}^E$ contains both the direct-path and early reflections of source k received by microphone E , and $v_{c,m}^E$ denotes the combination of late reflections from all sources and any additional background noise. Note that we utilize a reflection boundary of 50 ms, but have omitted the subscript b for clarity.

Given this signal model, the so-called useful-to-detrimental ratio (UDR) [114] for source k in T-F unit $u_{c,m}^E$ can be defined as,

$$\text{UDR}_k^E(c, m) = 10 \log_{10} \left(\frac{\sum_n (x_{k,c,m}^E[n])^2}{\sum_n (u_{c,m}^E - x_{k,c,m}^E[n])^2} \right), \quad (6.2)$$

where summations are over the interval of the corresponding T-F unit. Note that the UDR corresponds to the T-F unit-level *effective SNR*, as discussed in Chapter 5. We then let $\text{UDR}_k(c, m) = (\text{UDR}_k^L(c, m) + \text{UDR}_k^R(c, m))/2$ and define the IBM for source k as,

$$\text{IBM}_k(c, m) = \begin{cases} 1, & \text{if } \text{UDR}_k(c, m) > \text{LC} \\ 0, & \text{otherwise} \end{cases} \quad (6.3)$$

As studied in Chapter 5, the appropriate choice of LC depends on the effective SNR, which is a function of both the input SNR and the reflection boundary. One of the appealing properties of using a 50 ms reflection boundary is that this allows for

the use of LC values in the range that are common to anechoic settings. To facilitate comparison to existing segregation and enhancement methods that seek to maximize the output SNR, we set LC to 0 dB. Note that we average the UDR from the left and right signals so that each pair of T-F units, $u_{c,m}^L$ and $u_{c,m}^R$, are given the same assignment by $\text{IBM}_k(c, m)$. It is important to point out that this is only one possible choice of binaural IBM. Alternatively, independent IBMs for the left and right signals or an alternative method for combining information across ears could be used.

We utilize both spatial and periodicity information to estimate $\text{IBM}_k(c, m)$. To do so, we track the pitch and azimuth of up to three concurrent sources across time. We formulate the tracking problem such that we attempt to identify the most probable *multisource state* in each time frame, where a multisource state encodes the number of active sources, the azimuth of each active source, and the voicing characteristics of each active source. For each possible multisource state and time frame, we assign T-F units to one of the active sources using a set of trained MLPs. By identifying a path through the multisource state space across time, we generate a solution to the detection, localization, pitch-azimuth correspondence and simultaneous organization problems. Finally, azimuth-based sequential organization is performed to generate a T-F mask for each active source.

As will be discussed in Section 6.4, the cardinality of the full multisource state space is prohibitively large. In order to make computation feasible, we incorporate independent pitch and azimuth modules to identify a set of pitch and azimuth candidates to be considered by the HMM in each frame. We first introduce the main

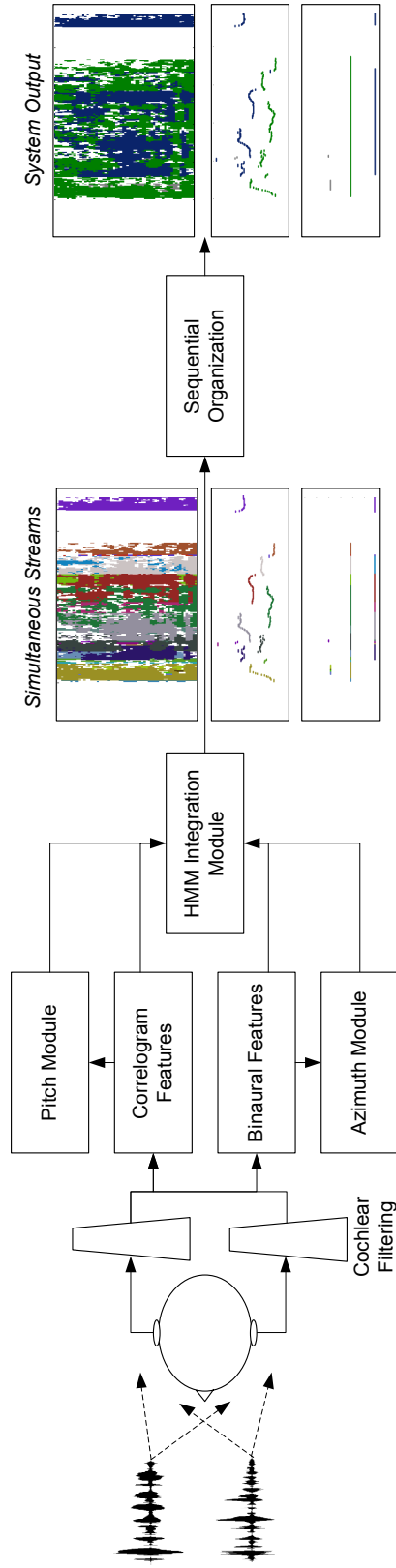


Figure 6.1: Schematic diagram of the proposed system. Cochlear filtering is applied to both the left and right ear signal of a binaural input. Correlogram features and binaural features are generated and fed to independent pitch and azimuth modules. Features along with both pitch and azimuth candidates are passed to the HMM framework. Viterbi decoding generates simultaneous streams and corresponding pitch and azimuth contours. Azimuth-based sequential organization groups simultaneous streams to form T-F masks, azimuth estimates and pitch estimates for each source.

components of the HMM in Sections 6.4.1-6.4.3, and then describe how the independent modules are used to generate candidate states in Section 6.4.4. A schematic diagram of the proposed framework is shown in Figure 6.1.

6.3 Feature Extraction

From each T-F unit pair, $u_{c,m}^L$ and $u_{c,m}^R$, we extract a set of pitch- and azimuth-related features as observations within the tracking framework described in Section 6.4. The pitch-related features are based on the correlogram and envelope correlogram [178]. We denote the correlogram and envelope correlogram as $A^E(c, m, \gamma)$ and $\bar{A}^E(c, m, \gamma)$, respectively. We use a low-pass filter with 500 Hz cutoff frequency and a Kaiser window to extract signal envelopes. Also note that we downsample the left and right signals to 16 kHz before computation of correlograms. We then let $\chi_{c,m}(\gamma) = \{A^L(c, m, \gamma), A^R(c, m, \gamma), \bar{A}^L(c, m, \gamma), \bar{A}^R(c, m, \gamma)\}$ denote the set of four pitch-related features for channel c , frame m and lag γ . We use X_m to denote the full set of pitch features for frame m .

As in Chapters 3 and 4, the binaural features calculated are the ITD, denoted $\tau_{c,m}$, and ILD, denoted $\lambda_{c,m}$. Again, we calculate ITD and ILD as in Equations (3.2) and (3.3), respectively. We use T_m and Λ_m to denote the full set of ITD and ILD features, respectively, for frame m . Finally, we use $Z_m = \{T_m, \Lambda_m, X_m\}$ to denote the entire set of observed data for frame m .

6.4 Hidden Markov Model Framework

We seek to model the posterior probability of a multisource state in each time frame based on the observed features described in Section 6.3. A multisource state, denoted $S = \{\theta_1, \theta_2, \theta_3, \gamma_1, \gamma_2, \gamma_3\}$, is a collection of individual pitch and azimuth states for three sources. We consider a discrete grid of azimuths in steps of 5° from -90° to 90° and allow sources to be inactive such that the azimuth state space for a single source is $\theta_k \in \{\emptyset, -90^\circ, -85^\circ, \dots, 90^\circ\}$. We consider a discrete grid of pitch lags from 32 to 200 samples (16 kHz sample rate), which correspond to frequencies between 80 and 500 Hz. Since sources may also be unvoiced, the pitch state space for a single source is $\gamma_k \in \{\emptyset, 32, \dots, 200\}$.

The posterior probability of a multisource state given the observed data can be expressed as,

$$p(S_m|Z_{1:m}) \propto p(Z_m|S_m)p(S_m|Z_{1:m-1}) \quad (6.4)$$

where subscript $_{1:m}$ denotes a collection of features from frame 1 through frame m . In the subsections below we discuss computation of the observation likelihood, $p(Z_m|S_m)$, and the state predictor, $p(S_m|Z_{1:m-1})$. In keeping with the assumption made by the IBM that each T-F unit can be assigned to at most one source, we include a data association stage where T-F units are assigned to an individual active source for each hypothesized multisource state. This assignment is the mechanism by which binary T-F masks are generated and facilitates computation of the multisource likelihood without modeling interaction between sources.

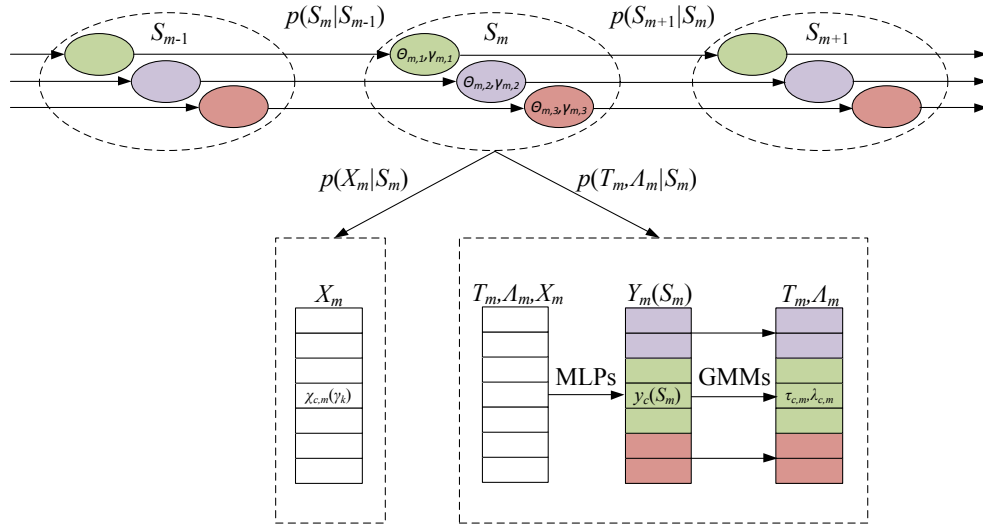


Figure 6.2: Illustration of the HMM framework. Multisource states are shown with large dashed oval. Computation of observation likelihoods are illustrated inside of the dashed rectangle.

We illustrate the proposed model in Figure 6.2. Multisource states that encode the characteristics (azimuth, voicing, pitch) of up to three active sources are shown with the large dashed ellipses. Multisource state transitions assume independence between sources. The process for computation of the observation likelihoods is illustrated inside of the dashed rectangle. Here, MLPs trained jointly on pitch and azimuth features are used to assign T-F units to one of the underlying active sources. Unit assignments, indicated by different colors, corresponding to one of the sources in the multisource state allow the frame-level likelihood to be decomposed into the product of unit-level likelihoods. We refer to this illustration as individual components of the model are described below.

6.4.1 T-F Unit Assignment

One of the principal advantages of the binary T-F masking approach to speech segregation is that it opens up a class of supervised learning algorithms to perform classification. Recent methods have yielded promising results using binaural features [149], pitch-based features [97] and modulation spectral features [103]. In keeping with this work, we incorporate a set of trained MLPs to assign T-F units to individual sources based on the azimuth and pitch information contained in a multisource state. As noted above, by solving this association problem for each possible multisource state, estimating a path through the multisource state space naturally generates a solution to the simultaneous organization problem.

Specifically, let $H_{k,c,m}$ denote the hypothesis that a source with azimuth θ_k and pitch γ_k satisfies the criteria necessary to be labeled 1 by the IBM. We then let $p_c(H_{k,c,m}|z_{k,c,m})$ denote the posterior probability of $H_{k,c,m}$ given the monaural and binaural observations, $z_{k,c,m} = \{\tau_{c,m}, \lambda_{c,m}, \chi_{c,m}(\gamma_k)\}$. For a given multisource state, we perform data association according to,

$$y_c(S_m) = \arg \max_{k \in \{1,2,3\}} [p_c(H_{k,c,m}|z_{k,c,m})], \quad (6.5)$$

where $y_c(S_m) \in \{1, 2, 3\}$ is an assignment to one of the sources contained in S_m .

Following the approach presented in Section 4.3.2, we train a set of MLPs to model $p_c(H_{k,c,m}|z_{k,c,m})$. We train two MLPs for each frequency channel and azimuth, one for unvoiced source states and one for voiced source states. For unvoiced states (i.e. $\theta \neq \emptyset$ and $\gamma = \emptyset$), the models ignore the pitch features and consider only ITD

and ILD. For voiced states (i.e. $\theta \neq \emptyset$ and $\gamma \neq \emptyset$), the models consider the full set of features. Since the pitch features are themselves a function of a specified pitch lag, the same models are used independent of γ_k . Details regarding the training procedures, training data and MLP topology are described in Section 6.6.3.

The T-F unit assignment procedure is illustrated on the left side of the large dashed rectangle in Figure 6.2. Each T-F unit, illustrated with small rectangles, is given a different color corresponding to one of the sources in the multisource state. Here, we use $Y_m(S_m)$ to denote the full set of T-F unit assignments for state S_m .

6.4.2 Observation Likelihood

The multisource observation likelihood captures the probability that the observed features were generated by a set of sources with azimuth and pitch characteristics specified by the multisource state, S_m . We first assume that binaural features and pitch-related features are conditionally independent such that, $p(Z_m|S_m) = p(T_m, \Lambda_m|S_m)p(X_m|S_m)$. Further, we assume conditional independence across frequency channels so that the T-F unit assignment described in the previous section allows for the decomposition of frame-level observation likelihoods conditioned on the characteristics of multiple sources into the combination of unit-level likelihoods conditioned on the characteristics of a single source.

Accordingly, we let,

$$p(T_m, \Lambda_m|S_m) = \alpha(S_m) \left(\prod_c p_c(\tau_{c,m}, \lambda_{c,m}|\theta_{y_c(S_m)}) \right)^\xi, \quad (6.6)$$

where $\alpha(S)$ is used to adjust the likelihoods based on the number of active sources contained in S and the term ξ is used to overcome the so-called probability overshoot phenomenon [75]. Multisource states with more active sources will produce systematically higher likelihoods due to increased flexibility in the T-F unit assignment expressed by Equation (6.5). Much like well-known model selection criteria, e.g. Akaike information criterion or minimum description length [24], the penalty term serves to minimize any systematic bias towards overestimation of the number of sources (also see Section 4.6.4). Probability overshoot results from the fact that, due to the overlapping passbands of gammatone channels, observations in individual channels are not entirely independent. We set $\alpha(S)$ to 1, 1, 0.4, and 0.25 for the cases with 0, 1, 2 and 3 active sources contained in S , respectively, and set $\xi = \frac{1}{16}$. These values were determined from a validation set.

While both the pitch and azimuth states are incorporated in the T-F unit assignments, the likelihood of a specific pair of ITD and ILD values, $\tau_{c,m}$ and $\lambda_{c,m}$, is assumed to be independent of the pitch states. We use the azimuth-dependent GMMs presented in Section 4.2.2 for $p_c(\tau_{c,m}, \lambda_{c,m}|\theta)$. Models are trained for each frequency channel and azimuth as described in Section 4.2.3 using the data described in Section 4.5.3. Finally, we set $p(T_m, \Lambda_m|\emptyset, \emptyset, \emptyset) = 0.02$, again based on a small validation set.

As will be discussed further in Section 6.4.4, we incorporate individual pitch and azimuth modules to supply a small set of candidate multisource states to the HMM. Multipitch and multiazimuth likelihoods, or $p(T_m, \Lambda_m|\Theta_m)$ and $p(X_m|\Gamma_m)$, are computed within each module to independently explore the full space of azimuth and pitch

combinations. In these modules, likelihood functions are symmetric about azimuth- or pitch-to-source assignments. A key part of likelihood computation within the HMM used for pitch-azimuth integration is identifying a correspondence between pitch and azimuth in source assignments. Since the joint dependence on pitch and azimuth is already captured by $p(T_m, \Lambda_m | S_m)$, we find it unnecessary to compute $p(X_m | S_m)$ in the same manner. We therefore set $p(X_m | S_m) = p(X_m | \Gamma_m)$, which is described in Section 6.4.4. Essentially, $p(X_m | S_m)$ captures the overall salience of a given set of pitches, independent of how they are paired with azimuths in S_m , and $p(T_m, \Lambda_m | S_m)$ then validates both the salience of an azimuth set and the pitch-azimuth correspondence specified by S_m .

In Figure 6.2 we illustrate computation of $p(T_m, \Lambda_m | S_m)$ in the rightmost large dashed rectangle. Given a set of T-F unit assignments, the set of binaural features calculated from the units assigned to each source are evaluated using that source’s azimuth-dependent GMM. This is indicated by the three separate arrows connecting the T-F unit labels, $Y_m(S_m)$, and binaural features, T_m, Λ_m . The leftmost dashed rectangle indicates that T-F unit assignments are not incorporated in computation of $p(X_m | S_m)$.

Table 6.1: Single source state transition probabilities. Rows 1, 2 and 3 list transitions out of voiced, unvoiced and inactive states, respectively. Columns 1, 2 and 3 list transitions into voiced, unvoiced and inactive states, respectively.

	$p(\theta, \gamma \cdot)$	$p(\theta, \emptyset \cdot)$	$p(\emptyset, \emptyset \cdot)$
$p(\cdot \theta', \gamma')$	$(1 - P_d)P_{vv}g(\gamma \gamma')f(\theta \theta')$	$(1 - P_d)(1 - P_{vv})f(\theta \theta')$	P_d
$p(\cdot \theta', \emptyset)$	$(1 - P_d)(1 - P_{uu})p(\gamma)f(\theta \theta')$	$(1 - P_d)P_{uu}f(\theta \theta')$	P_d
$p(\cdot \emptyset, \emptyset)$	$P_bP_vp(\gamma)p(\theta)$	$P_b(1 - P_v)p(\theta)$	$1 - P_b$

6.4.3 State Predictor

The state predictor captures the probability of a given multisource state given the posterior probabilities from the previous frame and state transition probabilities,

$$p(S_m|Z_{1:m-1}) = \sum_{S_{m-1}} p(S_m|S_{m-1})p(S_{m-1}|Z_{1:m-1}). \quad (6.7)$$

To allow estimation of the optimal *path* through the multisource state space using Viterbi decoding, we choose to approximate the predictor using,

$$p(S_m|Z_{1:m-1}) \approx \max_{S_{m-1}} [p(S_m|S_{m-1})p(S_{m-1}|Z_{1:m-1})]. \quad (6.8)$$

The key component of the predictor is the set of state transition probabilities. We assume independence between sources and define the multisource state transition probabilities according to,

$$p(S_m|S_{m-1}) = \prod_k p(\theta_{m,k}, \gamma_{m,k}|\theta_{m-1,k}, \gamma_{m-1,k}). \quad (6.9)$$

Source independence is illustrated in Figure 6.2 by the separate arrows connecting individual source states across time.

We list individual state transition probabilities in Table 6.1 where P_b and P_d are birth and death probabilities, respectively, $f(\theta|\theta')$ denotes the azimuth transition probability, $g(\gamma|\gamma')$ the pitch transition probability, P_v the prior probability of a source being voiced, and P_{vv} and P_{uu} are the voiced-voiced and unvoiced-unvoiced transition probabilities, respectively. P_b , P_d , $f(\theta|\theta')$ and $p(\theta)$ are highly situation dependant, as they are related to source activity, source motion and listener movements. In contrast, P_v , P_{vv} , P_{uu} , $g(\gamma|\gamma')$ and $p(\gamma)$ capture general properties of speech and should be relatively consistent across conditions. In the evaluation of this study we consider spatially fixed sources, and thus set $f(\theta|\theta') = \delta(\theta - \theta')$ and $p(\theta) = \frac{1}{|\Theta| - 1}$. Based on a validation set we set $P_b = 0.03$ and $P_d = 0.01$. Based on a small set of clean utterances from the TIMIT corpus [68], we set $P_v = 0.71$, $P_{vv} = 0.97$, $P_{uu} = 0.91$. Following [97] we use a Laplacian distribution with mean 0.4 and standard deviation 2.4 for $g(\gamma|\gamma')$. The choice of a Laplacian and values of 0.4 and 2.4 were validated on the same TIMIT corpus.

6.4.4 Pitch and Azimuth Modules

As noted in Section 6.2, a full search through the HMM state space is not tractable. The cardinality of S is equal to $((|\Theta| - 1)|\Gamma| + 1)^3 > 10^{11}$, or roughly 250 billion states. To make computation feasible we incorporate independent pitch and azimuth HMMs to identify a set of pitch and azimuth candidates for each frame.

We use the multipitch tracking system [97] described in Section 4.3.1 as the independent pitch module. We let $p(\Gamma_m|X_{1:m})$ denote the posterior probability of a

multipitch state in frame m , and let $p(X_m|\Gamma_m)$ denote the multipitch likelihood. Note that we choose to use notation consistent with this chapter rather than the notation used in [97].

The azimuth module is essentially a simplified version of the full HMM that ignores correlogram features. We compute,

$$p(\Theta_m|T_{1:m}, \Lambda_{1:m}) \propto p(T_m, \Lambda_m|\Theta_m) \max_{\Theta_{m-1}} [p(\Theta_m|\Theta_{m-1})p(\Theta_{m-1}|T_{1:m-1}, \Lambda_{1:m-1})], \quad (6.10)$$

where $p(T_m, \Lambda_m|\Theta_m)$ is calculated according to Equation (6.6),

$$p(\Theta_m|\Theta_{m-1}) = \prod_k p(\theta_{m,k}|\theta_{m-1,k}), \quad (6.11)$$

and

$$p(\theta_{m,k}|\theta_{m-1,k}) = \sum_{\gamma_m} \sum_{\gamma_{m-1}} p(\theta_{m,k}, \gamma_{m,k}|\theta_{m-1,k}, \gamma_{m-1,k})p(\gamma_{m-1,k}). \quad (6.12)$$

The T-F unit assignment in this case is still computed using Equation (6.5), however only the binaural MLPs are incorporated.

Once $p(\Gamma_m|X_{1:m})$ and $p(\Theta_m|T_{1:m}, \Lambda_{1:m})$ are computed for frame m , we use them to identify a set of candidate multisource states. Since one of the more challenging aspects of multipitch tracking is accurately detecting the number of pitched sources, we select the best 1- and 2-pitch candidates in each frame according to $p(\Gamma_m|X_{1:m})$, and also allow for the possibility of no voiced sources in each frame. We then select the best multiazimuth candidates according to $p(\Theta_m|T_{1:m}, \Lambda_{1:m})$. We select a total of 70 multipitch candidates and 150 multiazimuth candidates, yielding 10500 multisource

candidate states in each frame, a reduction of over 7 orders of magnitude relative to the full multisource state space. In preliminary experiments we found the system to be relatively insensitive to the number of multipitch and multiazimuth candidates considered, and that good tracking and segregation performance could be achieved even with such a severe reduction in the search space.

Note that the Jin and Wang multipitch tracking system deals with up to only two, rather than three simultaneous voiced sources. As a result the proposed framework, while it can localize and segregate up to three sources, is capable of assigning a pitch two at most two sources. This is a limitation of the current implementation, although we have studied extending the system to the three pitch case. We discuss this point in more detail in Section [6.7.5](#).

6.5 Segregation

The proposed HMM framework was developed with flexibility in mind. Various assumptions about source activity and source motion can be embedded in the prior and transition probabilities. While online tracking and segregation of moving sources are possible, we focus on offline segregation of an unknown number of sources in fixed spatial positions. This facilitates comparison to existing BSS methods that assume a known number of fixed sources and utilize the full mixture to localize each source [[58](#), [124](#)].

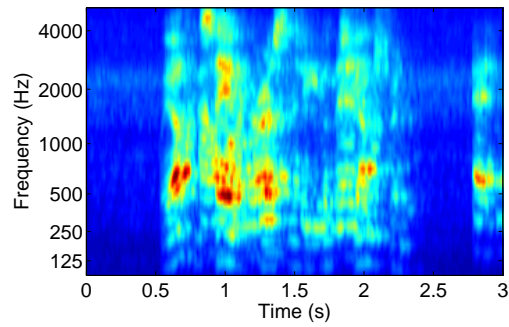
To perform segregation, we first determine the optimal path through the multi-source state space using Viterbi decoding. A selected state sequence encodes when

individual sources become active and inactive, and encodes the azimuth, voicing characteristics and set of T-F units associated with each source while it is active. Essentially, the HMM estimates a solution to the simultaneous organization problem in that grouping is performed across frequency within continuous time intervals. Specifically, when moving across time through the identified state path, we begin a new simultaneous stream, pitch contour and azimuth contour from the frame-level T-F mask, pitch estimate and azimuth estimate associated with the new source in the multisource state. The stream and associated contours are propagated across time using the frame-level masks, pitch estimates and azimuth estimates from subsequent frames until the source becomes inactive. Note that local connectivity within a simultaneous stream during voiced intervals is based on both pitch and azimuth. Connectivity between voiced and unvoiced regions is handled by azimuth alone. In cases when azimuth information is unreliable (e.g. co-located or closely spaced sources), it is in these regions that the system is most likely to wrongly group frame-level masks across time. For this reason, we break simultaneous streams at voiced to unvoiced or unvoiced to voiced transitions. This is only done to facilitate the analysis in Sections [6.7.1](#) and [6.7.2](#) using alternative sequential organization strategies. As will be discussed in the next paragraph, the proposed system perform sequential organization based on azimuth, and thus stitches the broken simultaneous streams back together.

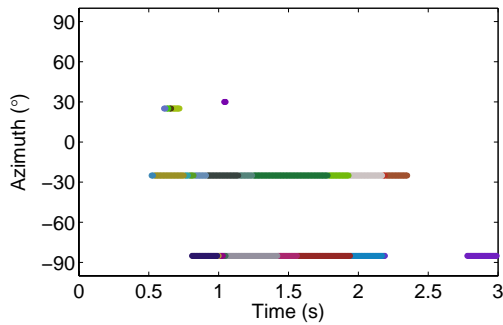
If an active source becomes silent and then reappears at a later time, the model in its current form is agnostic as to whether the two periods of activity are due to the same source. In other words, as indicated in [Figure 6.1](#), the HMM itself does not

perform sequential organization. A solution to the sequential organization problem is highly application dependent. Since we assume sources are in fixed spatial positions in the experiments in this chapter, azimuth is a powerful cue for sequential organization (as shown in Chapter 3). As such, subsequent to the formation of simultaneous streams, with their associated pitch and azimuth contours, we label streams as target dominant when their estimated azimuth is within a specified error tolerance around the known target azimuth. Thus we encode a “look direction” in order to perform sequential organization. We show the output of the proposed system on a mixture of two talkers from evaluation set 2 in Figure 6.3.

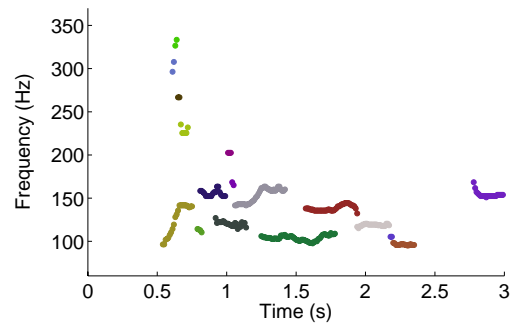
Numerous alternative approaches are possible. If online segregation is necessary, one could embed assumptions about the source of interest using a target-specific azimuth or pitch prior distribution. In this way the “look direction” or gender-specific information could be encoded directly in the HMM. Alternatively, one could estimate the total number of unique azimuths seen in a post processing stage, then group each stream with the closest detected azimuth. This approach would be similar in design to some of the binaural segregation and BSS approaches discussed in Chapter 2 [11, 119, 124, 136, 149], however localization and segregation would benefit from incorporating monaural cues in simultaneous organization. Finally, monaural speaker-dependent [161, 187] or speaker-independent [87] sequential organization could be used rather than relying on azimuth cues. We compare the proposed azimuth-based sequential organization to a speaker-independent clustering approach in Section 6.7.1.



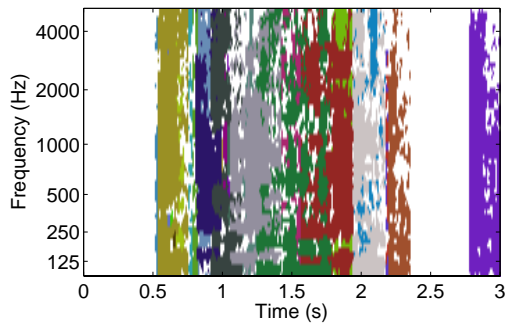
(a) Mixture Cochleagram



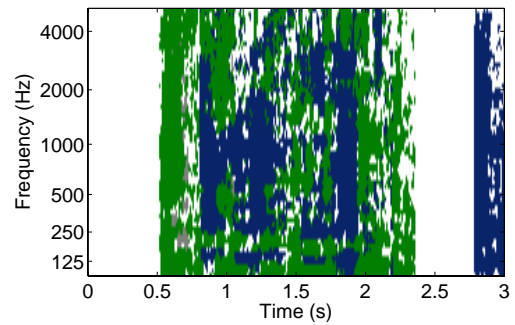
(b) Azimuth Contours



(c) Pitch Contours



(d) Simultaneous Streams



(e) Estimated Masks

Figure 6.3: Example output of simultaneous organization and T-F mask estimation using the proposed system. Mixture from Set 2 with two male talkers placed at -90° and -25° in a simulated environment with T_{60} equal to 0.6 s.

6.6 Evaluation Methodology

6.6.1 Binaural simulation

For both the training and evaluation databases, we generate binaural mixtures that simulate pickup of multiple speech sources in a reverberant space. Speech signals are drawn from the TIMIT database [68] and passed through a BIR for a specified angle and room condition. We use both simulated and measured BIRs. We simulate BIRs with the ROOMSIM package [25]. We generate BIRs with T_{60} equal to 0.2, 0.4 and 0.6 s, where for each T_{60} we create 15 room environments where room size, microphone position and microphone orientation are selected randomly and the reflection coefficients of wall surfaces are set to be equal and the same across frequency. For each T_{60} and environment we create BIRs for source positions between -90° and 90° , spaced by 5° , where the source is placed 2 m from the microphone array. Anechoic HRTF measurements from a KEMAR mannequin [67] are used in the simulation, so we refer to the simulated set of BIRs as the KEMAR set. The measured BIRs, referred to as the HATS set, are described in Section 4.5.1. Again, this set consists of measurements made in four reverberant environments (rooms A, B, C and D) with different sizes, reflective characteristics and reverberation times.

6.6.2 Evaluation Database

To evaluate the proposed system we generate three sets of mixtures that cover a variety of acoustic conditions. Since an important component of the proposed system

is estimating the number of active speech sources across time, we interlace monaural utterances from the same TIMIT speaker with periods of silence to form an individual speech source. Specifically, for each source we randomly choose an initial silence period between 0.1 and 1 s, a speech duration between 1 and 2 s and a gap duration between 0.1 and 1.5 s. Given these values a source is created by first placing zeros in the signal for the initial silence, then alternating between speech and silence periods until a 3 s signal has been created. Random utterances (without duplication) from the same speaker are used for all speech periods of the same source, but TIMIT speakers are chosen at random for each mixture. This process is carried out with monaural TIMIT signals prior to spatialization using the BIRs, and ensures that each mixture contains a time-varying number of sources. We show an example mixture from Set 2, described below, in Figure 6.4.

Set 1

For evaluation set 1 we simulate two speech sources at five different angular separations. For all mixtures we use the KEMAR BIRs with T_{60} set to 0.4 s and place a target source at 0° . We place the interference source at 0° (co-located), 5° , 10° , 15° , or 30° . The spatialized sources are set to have equal power when summated across left and right signals. To simulate a small amount of diffuse background noise, we filter uncorrelated speech-shaped noise through the anechoic BIRs for each azimuth (-90° to 90°) and sum them together. We create the speech-shaped filter by averaging the amplitude spectra of 200 speech utterances drawn from TIMIT at random. We then

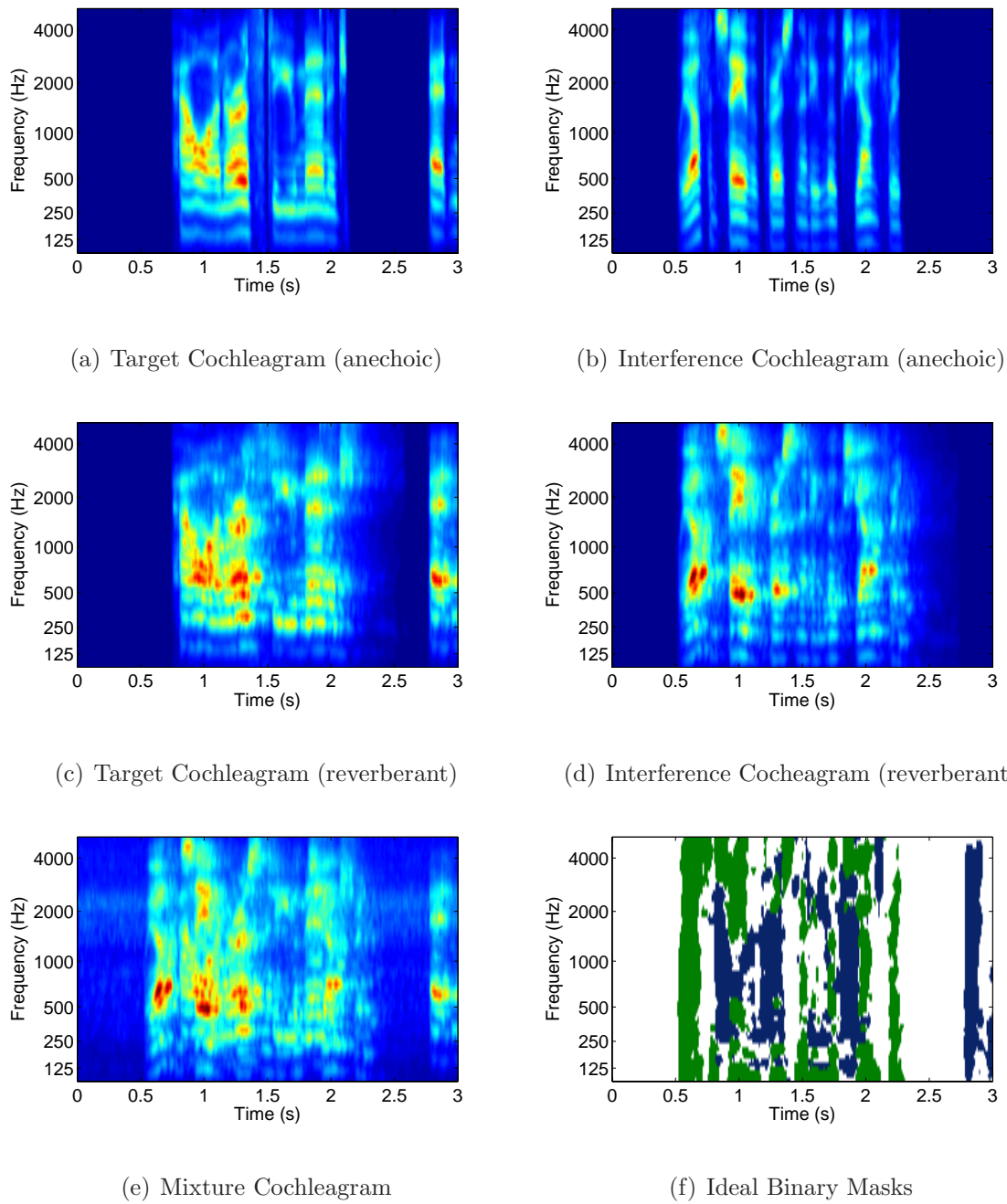


Figure 6.4: Example mixture from Set 2 with two male talkers placed at -90° and -25° in a simulated environment with T_{60} equal to 0.6 s.

add the diffuse noise to each mixture such that the total speech-to-noise ratio is 24 dB.

Set 2

For evaluation set 2 we generate both two- and three-talker mixtures where the azimuth of each source is selected randomly such that sources are spaced by 10° or more. We again use simulated BIRs in order to control T_{60} . We generate 100 mixtures for the two- and three-talker cases with T_{60} equal to 0.2, 0.4 and 0.6 s. Sources distances are set to 2 m for all sources and again, spatialized sources are set to have equal power when summated across left and right signals. We add diffuse noise so that speech-to-noise is 24 dB.

Set 3

To evaluate the system using real impulse responses we generate 50 two-talker mixtures for each room environment contained in the HATS BIR set. Azimuths are selected randomly such that sources are spaced by 10° or more. Again, spatialized sources have equal power and diffuse noise is added to achieve 24 dB speech-to-noise ratio.

6.6.3 Model training

The proposed system utilizes trained models in both the observation likelihood (see Section 6.4.2) and generation of T-F masks (see Section 6.4.1). We use the KEMAR and HATS models described in Section 4.5.4 to compute observation likelihoods.

Similarly for the MLPs used in T-F mask generation, we train models for both the KEMAR and HATS BIRs. For the KEMAR BIRs we generate 100 mixtures for each azimuth between -90° and 90° where the number of interfering talkers, interference azimuths, source distances and mixture T_{60} are selected randomly. The room environments used for the simulation are different from those used in the evaluation set. For the HATS BIRs we generate 100 mixtures for each azimuth and each room condition. The number of interfering talkers and interference azimuths are selected randomly. Separate models are trained for each room condition (A, B, C and D) on the data from the three alternative rooms so that the impulse responses used in an evaluation utterance have not been seen in training.

For each mixture we generate the observed binaural and monaural features (see Section 6.3), calculate the IBM according to Equation (6.3) and extract the ground truth pitch of the target source from the premixed signals. We use the pitch estimation method proposed in [12]. The IBM and pitch information are used to classify mixture T-F units as either unmasked and unvoiced, unmasked and voiced, and masked. The ground truth pitch information is used to select the appropriate correlogram features. Using all unmasked T-F units, we train a separate binaural MLP of ITD and ILD for each azimuth and frequency (used for unvoiced states). Using all unmasked and voiced T-F units, we train a separate joint MLP of ITD, ILD and correlogram features for each azimuth and frequency (used for voiced states).

For simplicity each MLP has the same network topology consisting of a hidden layer with 20 nodes, and hyperbolic tangent sigmoid transfer functions for both hidden

and output nodes. Training is accomplished using a generalized Levenberg-Marquardt backpropagation algorithm.

6.7 Evaluation

6.7.1 Experiment 1: Simultaneous and sequential organization

In this first experiment we compare simultaneous organization performance to the pitch-based method used in Chapters 3 and 4 in order to validate we achieve improved performance. Further, as a follow up to the sequential organization comparison made in Chapter 3, we compare azimuth-based sequential organization to a recent monaural method based on speaker-independent clustering of cepstral features [87]. To allow for comparison to pitch-based simultaneous organization, we do not keep multiple pitch candidates as described in Section 6.4.4, but rather utilize the multipitch tracker output and consider only voiced frames. For simplicity we also assume the number of sources and source azimuths are known. In this case, the HMM simply generates the across-time correspondence between pitch estimates and known azimuths.

The pitch-based simultaneous organization described in Chapter 4 first links pitch estimates from the multipitch tracker across time based on pitch deviation. When neighboring pitch frequencies are within 7% of each other, they are joined to form pitch contours (see Section 4.3.1). Once pitch contours are formed, corresponding simultaneous streams are generated using MLPs trained on correlogram features (see Section 4.3.2). There are two key differences between the pitch-based approach and

the HMM framework proposed in this chapter. First, grouping across frequency within a time frame is based jointly on correlogram and binaural features in the proposed system. This has the potential to achieve more effective grouping of mixed voiced and unvoiced speech (e.g. due to coarticulation or temporal smearing caused by reverberation) or when competing sources have similar pitch. Second, grouping between time frames is handled implicitly within the HMM and is based jointly on pitch and azimuth, which has the potential to improve local grouping in time when the pitch contours of competing speakers overlap.

To evaluate whether we achieve a performance gain due to either across-frequency grouping based jointly on binaural and correlogram features or due to across-time grouping based jointly on pitch and azimuth, we compare three alternative simultaneous grouping strategies. We refer to these based on the method used for grouping in time and frequency. The first is the pitch-based strategy. The second is an intermediate strategy that incorporates the joint MLPs for across-frequency grouping, but operates on pitch contours generated with the pitch-based strategy (i.e. pitch for grouping in time, joint for grouping in frequency). Specifically, pitch contours are created from the multipitch estimates using the same 7% relative change criteria, then each contour is assigned one of the known source azimuths based on the binaural features within the corresponding pitch-based simultaneous streams. We then use the combined pitch and azimuth information to update the set of simultaneous streams with the joint MLPs. Finally, the third alternative uses the simultaneous streams generated by the proposed system (joint in both time and frequency). We

remove T-F units from the simultaneous streams in time frames that are identified as unvoiced by the system.

We also compare three different sequential grouping methods. First, to analyze the ceiling performance achievable by the simultaneous methods, we perform ideal sequential organization by labeling simultaneous streams using the IBMs. Second, we use the monaural clustering approach proposed in [87]. Third, we perform azimuth-based sequential organization.

We show results for each of the nine combinations of simultaneous and sequential organization strategies on evaluation set 1 in Table 6.2. We measure performance using the percentage of correctly labeled target-dominant units (Hit), the percentage of incorrectly labeled interference-dominant units (FA), and the difference between the two (Hit-FA). Percentages are averaged over all 25 mixtures and both talkers and shown as a function of azimuth separation. First, to compare simultaneous organization performance, we focus on the case with ideal sequential organization. We note that including binaural features for across-frequency grouping using the joint MLPs (Pitch+Joint) increases the Hit rate by 10.5%, averaged across conditions, while increasing the FA rate by 2.5%, resulting in an 8% increase in Hit-FA on average. Hit-FA also increases using the full proposed system (Joint+Joint), by 10.3% on average relative to the entirely pitch-based. The improvement relative to the Pitch+Joint approach is 2.2% on average. Notable is the fact that the performance of the Pitch+Joint and Joint+Joint approaches are comparable in the case with co-located sources. This shows that the HMM successfully defaults to a pitch-based

Table 6.2: Simultaneous and sequential organization performance

Simultaneous Organization		Sequential Organization														
		Ideal			Azimuth-based			Monaural Clustering								
Method (Time, Frequency)	Metric	0°	5°	10°	15°	30°	0°	5°	10°	15°	30°	0°	5°	10°	15°	30°
Pitch, Pitch	Hit	62.5	62.8	61.5	61.6	61.6	-	60.4	60.7	60.6	61.1	54.5	56.4	54.4	54.3	55.4
	FA	8.6	8.4	9.1	9.2	9.0	-	9.3	9.6	9.6	9.2	12.0	10.9	12.2	12.6	11.9
	Hit-FA	53.9	54.5	52.9	52.3	52.7	-	51.1	51.1	51.0	51.9	42.5	45.5	42.2	41.7	43.5
Pitch, Joint	Hit	72.8	73.2	72.3	71.6	73.3	-	70.9	71.1	71.1	73.0	62.1	63.8	62.4	64.0	65.2
	FA	11.9	11.4	11.6	11.4	10.3	-	12.3	12.1	11.7	10.4	16.3	15.2	15.7	15.0	13.9
	Hit-FA	60.9	61.8	60.6	60.2	63.0	-	58.5	59.1	59.4	62.6	45.8	48.6	46.7	49.0	51.3
Joint, Joint	Hit	73.8	75.0	74.6	75.0	75.5	-	71.8	72.0	73.4	74.3	64.5	64.1	65.9	66.4	65.8
	FA	12.2	11.5	11.4	10.8	10.3	-	12.9	12.6	11.6	10.9	16.0	16.0	15.1	14.4	14.3
	Hit-FA	61.6	63.5	63.3	64.2	65.1	-	58.9	59.4	61.8	63.3	48.5	48.2	50.8	51.9	51.5

strategy when azimuth information is not beneficial. As one might expect, performance of the proposed system (Joint+Joint) increases as the separation between sources increases, from 61.6% to 65.1%.

In comparing sequential organization performance using the simultaneous streams generated by the proposed system, the azimuth-based approach clearly outperforms the monaural method, although the azimuth-based approach is not applicable in the co-located condition. Averaged over the other conditions, the drop in Hit-FA of the azimuth-based approach relative to ideal sequential organization is 3.2%, while the drop for the monaural method is 13.5%. It is important to note that this comparison method was not designed with reverberation in mind, although since no pre-trained models are incorporated, it is reasonable to expect some robustness to reverberation. Closer analysis of the results shows that the monaural system is actually quite competitive for mixtures with different genders. Averaged over mixtures with separation between sources of 5° or larger, Hit-FA on different gender mixtures is 64.4%, 62% and 61.2% for ideal, azimuth-based and monaural sequential organization, respectively. For same gender mixtures, Hit-FA scores drop to 63.6%, 59.6% and 39.1% for the three methods, respectively. Since azimuth-based sequential organization may not be possible in certain circumstances (the extreme case being co-located sources), it is promising that this monaural approach can achieve good results when source characteristics are sufficiently different.

6.7.2 Experiment 2: Comparison with ground truth information

In this experiment we validate the fundamental assumption that segregation based jointly on pitch and azimuth outperforms segregation based on azimuth alone. To do so we compare the quality of binary T-F masks generated using the MLPs that consider both correlogram and binaural features versus the MLPs that rely only on ITD and ILD. We show results assuming various amounts of ground truth information to both establish the ceiling performance achievable by the proposed mask estimation methods and to analyze the amount of degradation due to estimating the number of sources and the corresponding pitches and azimuths across time.

We perform this set of experiments on evaluation set 1, however, we exclude the co-located mixtures so that the proposed system with azimuth-based sequential organization and an exclusively binaural approach are applicable. For each mixture we generate the IBM for each source according to Equation (6.3). We also use the pitch tracking approach proposed in [12] to generate ground truth pitch for each source using the premixed signals. The IBM, ground truth pitch and known azimuth allow us to generate ground truth frame-level labels for each source. Specifically, we consider a source to be active in a frame when at least one T-F unit in the source’s IBM is labeled 1. Each active frame is labeled with the source’s known azimuth to generate ground truth azimuth for each source. For each active frame, we label the frame as either voiced or unvoiced depending on whether a pitch has been detected. The ground truth pitch for each source is then associated with that source’s voiced

frames. We show the IBMs, ground truth source activity and azimuth, and ground truth pitch points for an example mixture with two male talkers in Figures 6.5(a), 6.5(c) and 6.5(e), respectively.

In Table 6.3 we show the average Hit-FA for the proposed system along with various alternatives that incorporate ground truth pitch and/or azimuth and use either ideal or azimuth-based sequential organization. On the left side of Table 6.3 we list three columns: azimuth, pitch and sequential organization (S.O.). The proposed system seeks to estimate these three main properties for each source from the mixture signal alone. Each instantiation of the system shown either estimates (E) azimuths and/or pitches or utilizes the ground truth information (GT). Note that we treat detection of the number of sources and azimuth estimation together in the azimuth category.

The ceiling performance achievable by the proposed MLPs is shown in row 1, where both azimuth and pitch are generated based on ground truth information. Note that in this case there is no need for a separate sequential organization stage because the true azimuths and pitches for all sources are known in all frames. We can see that while there is some degradation as the interference source is placed more closely to the target source, the decrease in Hit-FA is less than 1.5%. We show the performance achieved using only ground truth azimuth (i.e. pitch is ignored) in row 8. Note that for systems based only on azimuth (rows 8 and 9), grouping across time and frequency are based entirely on azimuth, and thus there is no distinction between simultaneous and sequential organization. We can see that Hit-FA is systematically lower for the

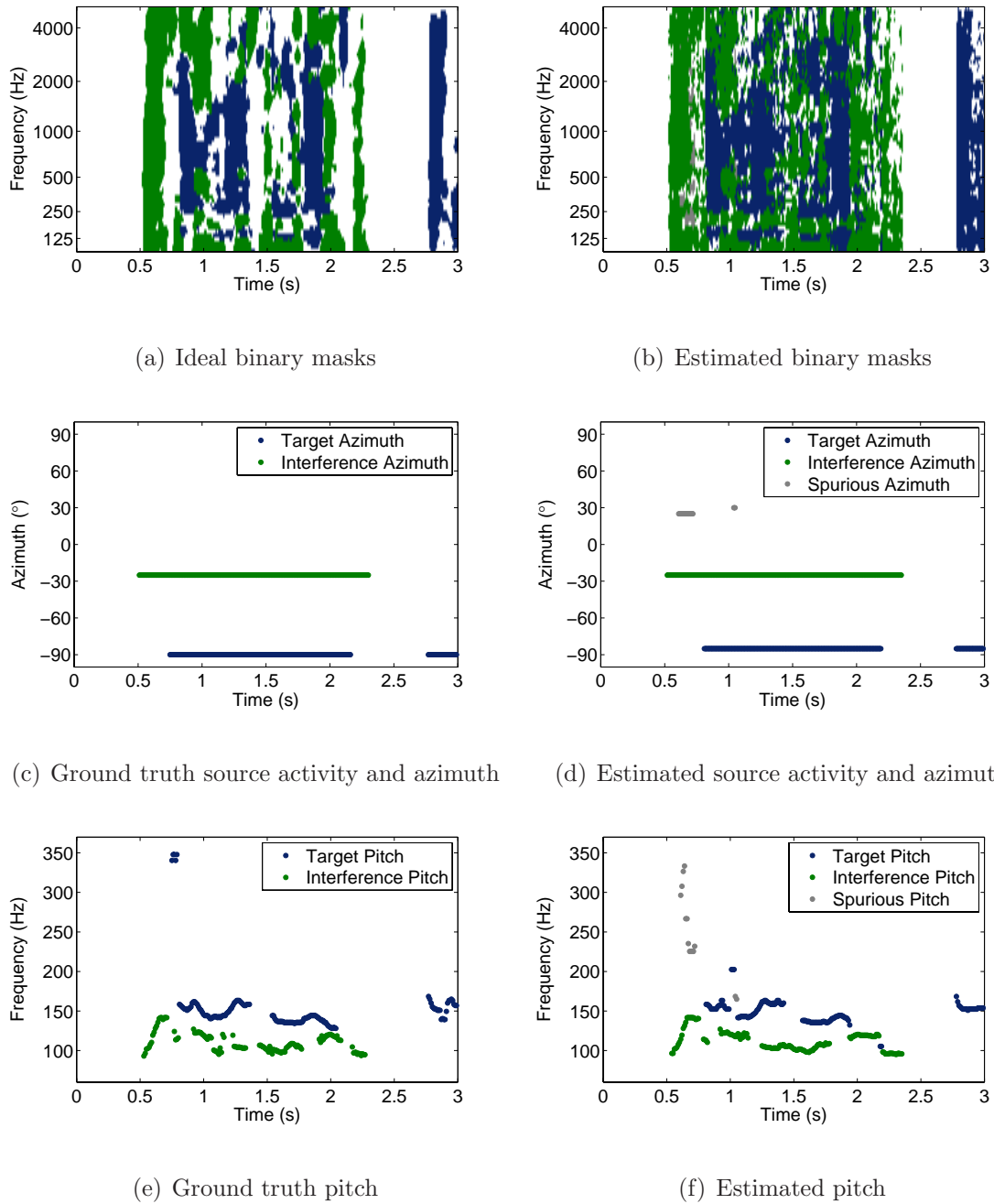


Figure 6.5: Example IBMs (a), estimated masks (b), ground truth (c) and estimated (d) azimuth, and ground truth (e) and estimated pitch (f) for a mixture of two male talkers placed at -90° and -25° in a simulated environment with T_{60} equal to 0.6 s. Target mask, azimuth and pitch are shown in blue, interference in green. Spurious estimates are shown in gray.

Table 6.3: Average Hit-FA (%) on evaluation set 1 for variants of the proposed system with ground truth (GT) and estimated (E) pitch/azimuth and ideal or azimuth-based sequential organization. Target is placed at 0° for all mixtures and performance is shown as a function of interference azimuth.

	Azimuth	Pitch	S.O.	Interference Azimuth				Avg.
				5°	10°	15°	30°	
1	GT	GT	-	74.4	75.0	75.9	77.5	75.7
2	E	GT	Ideal	73.3	74.4	75.2	76.7	74.9
3	GT	E	Ideal	72.1	74.1	75.0	76.5	74.4
4	E	E	Ideal	70.0	73.3	74.1	75.6	73.3
5	E	GT	Azimuth	69.9	73.1	74.2	75.9	73.3
6	GT	E	Azimuth	70.0	73.3	74.3	76.1	73.4
7	E	E	Azimuth	67.3	72.4	73.5	75.2	72.1
8	GT	-	-	59.4	63.6	65.7	69.6	64.6
9	E	-	-	54.9	61.6	63.8	67.5	61.9

azimuth-only system (11.1% drop from row 1 to row 8), and the degradation between the 5° case and 30° case exceeds 5%.

We show example posterior probability estimates generated by the binaural MLPs and the joint correlogram+binaural MLPs using ground truth information in Figure 6.6. While in this example there is very good azimuth separation between sources (65°), the improved discrimination due to adding correlogram features is clear, particularly at low frequencies. In general we find that the two sets of features are complimentary. Binaural features tend to perform poorly at low frequencies where wavelengths are large relative to the microphone spacing, but can be quite powerful in unvoiced and high frequency regions, whereas correlogram features are very powerful in low frequency channels with resolved harmonics, but are more easily corrupted in

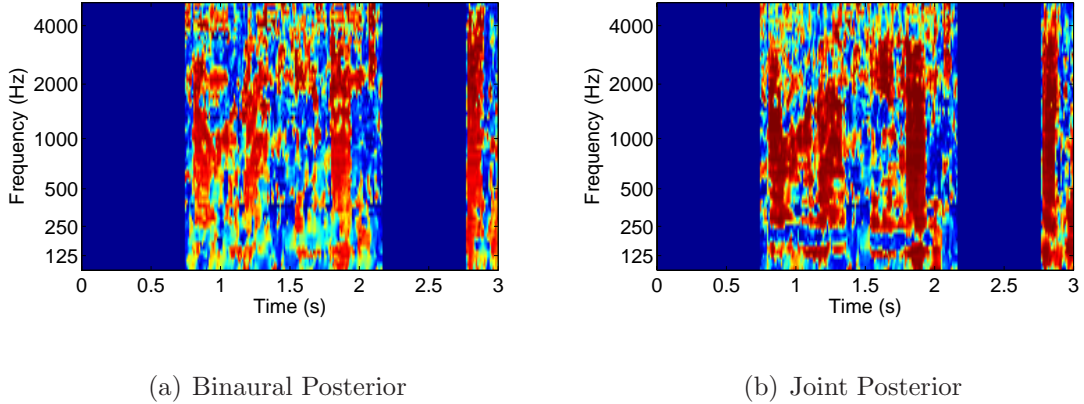


Figure 6.6: Example of posterior probability (MLP output) based on ITD and ILD alone (a) and based jointly on ITD, ILD and correlogram features (b) using ground truth pitch and azimuth for the target source. Mixture from evaluation set 2 with two male talkers placed at -90° and -25° in a simulated environment with T_{60} equal to 0.6 s.

high frequencies with unresolved harmonics and cannot be utilized in unvoiced speech regions.

In row 7 we show the Hit-FA achieved by the proposed system, where the number of sources and corresponding pitches and azimuths are estimated, and azimuth-based sequential organization is used. In row 9 we show the Hit-FA achieved by an estimated version of the azimuth-only approach. The azimuth-only version is the azimuth module described in Section 6.4.4. We first note that as in the ground truth case, the proposed system based on both pitch and azimuth achieves a systematic improvement relative to the azimuth-only system. We see an improvement of up to 12.4% in the 5° case and 10.2% on average.

The drop in performance of the proposed system relative to the full ground truth

system (compare row 7 to row 1) is between 2.5% and 7.1%, depending on the amount of separation between sources. In this case, degradation can be due to source detection, azimuth and pitch estimation (including generating the correct correspondence between azimuth and pitch in the HMM), and azimuth-based sequential organization. To analyze the impact of each of these factors, we show alternatives that incorporate partial ground truth information in rows 2 - 6. Consistent with the results in Experiment 1, azimuth-based sequential organization achieves very near ideal performance when sources are well separated (compare rows 2-4 to rows 5-7), but is a major contributor to the degradation for very closely spaced sources. We also see that both the pitch estimation and azimuth detection/estimation contribute similarly to the performance degradation (compare rows 2 and 3 to row 4, and rows 5 and 6 to row 7), however in most cases the drop in performance is not additive (i.e. drop due to pitch estimation plus drop due to azimuth estimation is typically more than drop due to estimation of both). In general we see that there is not one primary source of error in the proposed system.

We show the output of the proposed system on an example mixture from evaluation set 1 in Figures 6.5(b), 6.5(d) and 6.5(f), alongside the corresponding IBMs and ground truth information. This illustrates the high degree of accuracy in both mask estimation and pitch/azimuth estimation. The most notable errors in mask estimation are due to falsely detected T-F units, primarily in regions dominated by reverberation (and thus labeled 0 in the IBMs). Reverberation tends to smear the

periodic speech components across time and thus some T-F units in the reverberation tail are incorrectly assigned to the detected sources.

6.7.3 Experiment 3: Comparison to existing systems

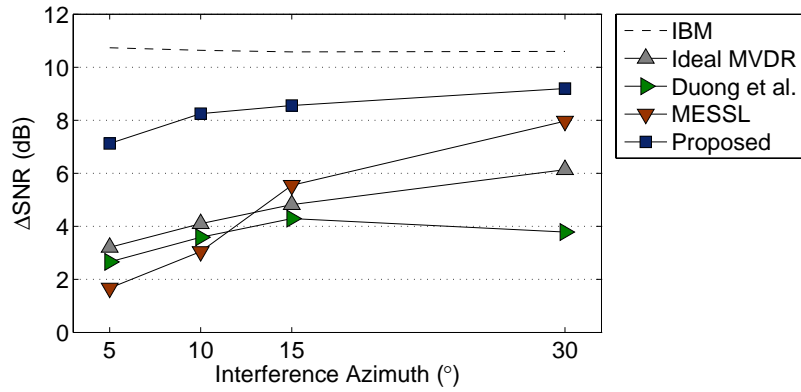
In this experiment we compare the proposed approach to three two-microphone systems from the literature. The first is an idealized MVDR beamformer [15]. In our implementation we calculate target and interference covariances from the clean target and residual signals, and thus this method represents the upper bound performance obtainable by a beamformer alone. We process 16 kHz noisy mixture signals through a 256 channel linear filterbank with a decimation factor of 64 samples. We also compare our method to the recent segregation methods presented in [58, 124]. Both of these methods assume the number of sources are known *a priori* and that sources are in a fixed spatial location. Although not required by the proposed approach, we provide these comparison methods with the number of speech signals contained in each mixture. We note that the method proposed in [58] was not explicitly designed to handle binaural mixtures, and thus is sensitive to spatial aliasing caused by a large microphone spacing. This approach is representative of a class of BSS methods that handle underdetermined mixtures by performing separation independently in each frequency band and then seek to resolve the across-frequency permutation ambiguity (see Section 2.2). We include these results to illustrate the difficulty the binaural case poses to solving the permutation problem.

In Figure 6.7, we show the change in SNR (Δ SNR) achieved by the proposed

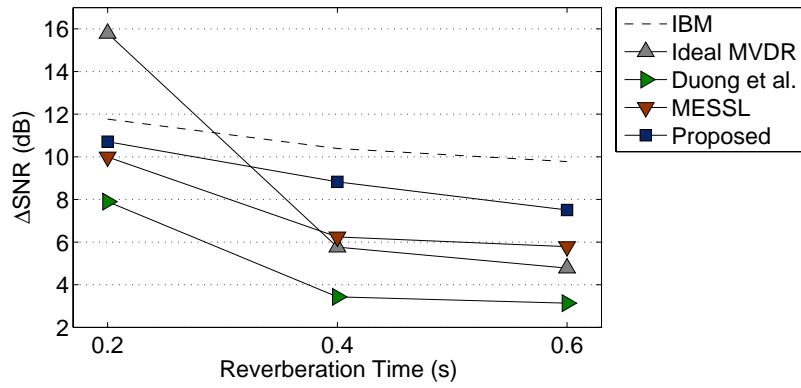
and comparison systems on evaluation sets 1 and 2. Since the comparison systems do not seek to estimate the IBM, we use the premixed target signal as the reference in calculation of SNR. Note that, in keeping with our definition of the IBM, we include early reflections as a part of the target signal. In Figure 6.7(a) we see that the proposed approach achieves an improvement in terms of ΔSNR relative to the comparison methods in all cases. The improvement is largest for the 5° and 10° mixtures, where it exceeds 3 dB. In Figures 6.7(b) and 6.7(c) we show ΔSNR achieved on evaluation set 2 as a function of T_{60} for two- and three-talker mixtures, respectively. The proposed system achieves the largest SNR gains in nearly all cases.

As one would expect, the ideal MVDR is able to achieve much larger SNR gains for mixtures with two talkers, particularly when there is little reverberation, because it is able to create a single null in the interference direction. As reverberation increases, sources are spaced more closely or the number of talkers is increased, the beamformer is less effective. In preliminary experiments we have also found an MVDR estimated from the mixture signal based on target detection is considerably less effective. However, since performance is influenced by numerous factors such as the activity detection method used, the degree of overlap between target and interference, and the amount of averaging used to derive the beam pattern, we include only the ideal MVDR results in this comparison.

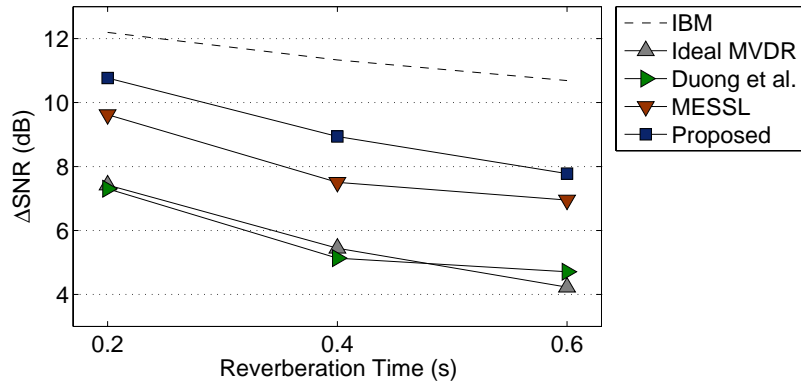
The Duong et al. system is an iterative implementation of the multichannel Wiener filter that combines a beamformer and post-filter. This system does not perform well on our evaluation set due to the large distance between microphones.



(a) SNR vs. Interference Azimuth



(b) SNR vs. T_{60} , Two-talkers



(c) SNR vs. T_{60} , Three-talkers

Figure 6.7: Δ SNR the proposed algorithm and three comparison methods on evaluation sets 1 (a) and 2 (b,c).

Table 6.4: Avg. Δ SNR (in dB) for the proposed system and three comparison systems using measured impulse responses from four room conditions. The T_{60} for each room (in s) is listed in parenthesis.

	A (0.32)	B (0.47)	C (0.68)	D (0.89)	Avg.
Proposed	8.6	7.9	8.7	6.9	8.0
Ideal MVDR	5.5	4.4	5.4	4.1	4.9
Duong et al.	3.5	3.1	3.9	3.5	3.5
MESSL	5.8	5.3	6.6	6.4	6.0
IBM	11.3	10.0	10.8	9.6	10.4

As mentioned above, it is important to note that the authors did not design the system for such a large microphone spacing and thus our result should not be too surprising, but it does illustrate the challenge in resolving the across-frequency permutation ambiguity for a binaural input.

The MESSL system clearly outperforms the other comparison methods and is capable of achieving large gains in SNR when sources are well separated in space. This is notable particularly because MESSL requires very little prior training and is still capable of handling spatial aliasing.

In Table 6.4 we show Δ SNR achieved by the proposed and comparison systems on evaluation set 3, which uses measured BIRs in four different room conditions. Consistent with the results when using simulated BIRs, the proposed system achieves the best performance in all room conditions.

6.7.4 Experiment 4: Detection and Localization

In this experiment we analyze the azimuth detection and localization performance of the proposed system and compare to two binaural baseline methods. The first is the SRP-PHAT approach, denoted “SRP”, described in Section 3.6.2. Rather than integrate across all frames as done in previous comparisons, in this experiment we recursively smooth azimuth-dependent responses across time frames to allow for detection of a time-varying number of sources. We then select peaks in the smoothed function in each frame that exceed a threshold. Smoothing constant and threshold are determined using a training set of 50 two-talker and 50 three-talker mixtures with $0.4 \text{ s } T_{60}$, generated as described for set 2. The second comparison system, denoted “SRP+Kalman”, pairs azimuth detections from SRP-PHAT with a set of Kalman filters using the data association techniques proposed in [169]. A second-order autoregressive model is used for source motion in each Kalman filter. Detections are associated with existing tracks when they fall within an acceptance region that accounts for measurement noise and possible target motion. New tracks are initialized when a detection cannot be attributed to an existing track. Tracks are terminated if there is an absence of detected azimuths within the acceptance region for multiple consecutive frames. System parameters are tuned based on the same set of 50 two-talker and 50-three talker mixtures.

We compute a number of metrics to evaluate detection and localization performance. For each frame of an evaluation mixture, we consider a source to be detected

Table 6.5: Detection and localization performance of the proposed and two comparison systems on a subset of mixtures from evaluation set 1.

	Proposed			SRP			SRP+Kalman		
	10°	15°	30°	10°	15°	30°	10°	15°	30°
Recall	88.4%	91.0%	91.7%	65.5%	85.3%	91.8%	70.2%	88.0%	92.6%
Precision	96.2%	96.8%	96.1%	81.7%	84.8%	85.3%	87.7%	88.8%	89.1%
F-score	92.1%	93.8%	93.8%	72.7%	85.0%	88.4%	78.0%	88.4%	90.8%
Fine Error	0.02°	0.05°	0.0°	1.27°	0.68°	0.23°	1.72°	1.31°	0.98°

if there is an azimuth estimate within (and including) 10° . Any estimate that cannot be attributed to an existing source is considered a false estimate. Note that a single azimuth estimate cannot be used to detect more than one source. We then measure both *precision* and *recall*. Precision is the percent of correct estimates out of the total number of estimates generated by the system. Recall is the percentage of detected sources out of the total number of true azimuths. We then calculate the *F-score* as the geometric mean of precision and recall. Comparison methods were optimized to maximize F-score. Finally, we measure the *fine error* as the average azimuth error of the correct estimates.

We show performance of the proposed and comparison systems on a subset of mixtures from set 1 in Table 6.5. Because the comparison methods utilize peak picking in the frame-level azimuth responses, they perform poorly for sources spaced more closely than 10° . While the proposed system does not suffer from this shortcoming, we show results only for mixtures with 10° of separation or more between sources. We see that the proposed system outperforms the comparison methods in terms of F-score

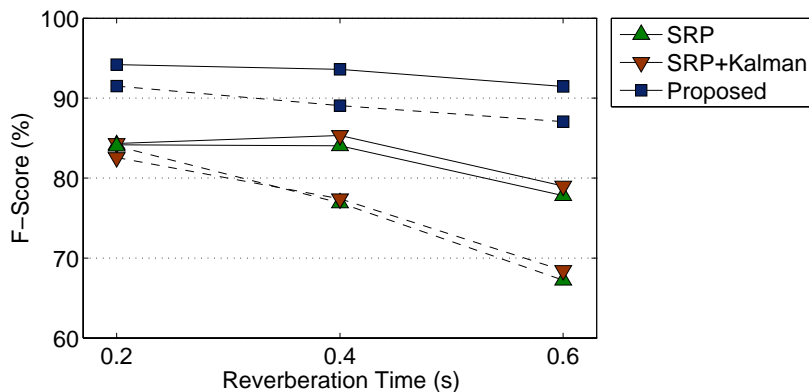


Figure 6.8: F-score (%) for the proposed and comparison systems on the two- and three-talker mixtures from set 2. Results for two-talker mixtures are shown with solid lines while those for three-talker mixtures are shown with dashed lines.

for each azimuth separation condition. The improvement is largest for mixtures with 10° separation, where we observe an absolute improvement of 18% relative to SRP and 12.7% relative to SRP+Kalman. We also see that the addition of the Kalman filter and data association techniques improve performance relative to SRP alone. Since sources are placed in fixed spatial positions, improvement is mainly due to a decrease in false estimates through inclusion of heuristics for track initialization and deletion. We also note that the fine error for the proposed method is less than or equal to 0.05° in all cases, suggesting that azimuth estimation is extremely accurate for detected sources.

In Figure 6.8 we show the F-scores achieved by each system on the two- and three-talker mixtures from set 2. Results for two-talker mixtures are shown with solid lines while those for three-talker mixtures are shown with dashed lines. Consistent with the

results on set 1, the proposed approach provides a substantial improvement relative to the comparison methods. Absolute improvement relative to SRP+Kalman, the more competitive comparison method, is between 8% and 12% on the two-talker mixtures and between 9% and 18% on the three-talker mixtures.

6.7.5 Analysis: Tracking 3 Pitches

As briefly mentioned in Section 6.4.4, the current implementation of the proposed system can track up to only two pitches per frame. Since the HMM framework is designed to deal with up to three simultaneous sources, we explored extending the Jin and Wang system to track three pitches. We first note that based on our small training set, we calculated the prior probability of a source being voiced in a given frame as 0.71. If all sources were active in every frame of a three-talker mixture, then all three sources would be voiced in roughly 35% of time frames. In our test database however, not all three sources are active in every frame. For the three-talker evaluation mixtures from set 2, we calculate that all three sources are active in roughly 39% of time frames, while all three sources are voiced in roughly 13% of time frames. This is not an insignificant number and thus it stands to reason that tracking three pitches would be beneficial to performance.

We analyzed the possibility of tracking three simultaneous pitches using the Jin and Wang system using a set of 100 three-talker, anechoic mixtures (constructed as in evaluation set 2, but where no reverberation or background noise was added). For good performance, we must be able to both accurately track all three pitches

and correctly discriminate between two- and three-pitch frames. To determine how well three simultaneous pitches could be tracked, we first extended the system to handle the possibility of three sources and performed tracking over sequences of frames in which all three sources were identified as voiced from the premixed signals. To ignore the complexity of detecting the number of pitches, we ran the system assuming (correctly) that it should always remain in the three-pitch subspace. We found that on average, 66.5% of pitches were estimated correctly (within a tolerance of 10% of the true pitch). In other words, even with knowledge that there were three pitches present in each frame, on average, the system could only track two correctly. The system correctly identified all three pitches in only 17.6% of the three-pitch frames.

While those numbers are not promising, we also analyzed whether it would be possible to discriminate between the two- and three-pitch subspaces using the method for computing multipitch likelihoods described in [97]. To avoid the complexity of the HMM, we simply analyzed the likelihood of the true three-pitch state in three-pitch frames and compared it to the best competing two-pitch likelihood. By this we mean that if the three active pitches were γ_1 , γ_2 and γ_3 , we compared $p(X|\gamma_1, \gamma_2, \gamma_3)$ to the maximum of $p(X|\gamma_1, \gamma_2)$, $p(X|\gamma_1, \gamma_3)$, and $p(X|\gamma_2, \gamma_3)$. Ignoring any penalty on the likelihood (see Equation (11) in [97]), the three-pitch likelihood will equal or exceed the best competing two-pitch likelihood, but in order to pull the HMM into the three-pitch subspace, the value of the three-pitch likelihood must be sufficiently larger. In Figure 6.9 we show a histogram of the log likelihood ratio comparing the true three-pitch likelihood to the best competing two-pitch likelihood in all three-pitch

frames of the analysis set. As a point of comparison, we also show a histogram of the log likelihood ratio comparing the true two-pitch likelihood to the best competing one-pitch likelihood in all two-pitch frames of the analysis set. As we know that the multipitch tracker does a reasonable job of discriminating between one and two pitches, although this discrimination is the largest source of error reported in [97], we can clearly see that discriminating between two and three pitches will be more challenging because the likelihoods of the two- and three-pitch states are much closer.

These small experiments show that both accurately tracking three simultaneous pitches and correctly identifying the frames in which three pitches are present are quite difficult using the Jin and Wang framework. Nevertheless, we did attempt to optimize a version of the system to handle up to three pitches and found that, on our anechoic data set, all three pitches were correctly detected in only about 6% of the three-pitch frames. Further, we tested a version of the proposed system that incorporated the three-pitch tracker on the three-talker mixtures of evaluation set 2, and found there was no performance gain. As the computational complexity of the multipitch tracker is increased by over four orders of magnitude by moving from two to three pitches, we concluded that using the original Jin and Wang system was preferable.

6.8 Discussion

The evaluation results show that the proposed integration of pitch and azimuth cues achieves more robust segregation than considering either monaural or binaural cues

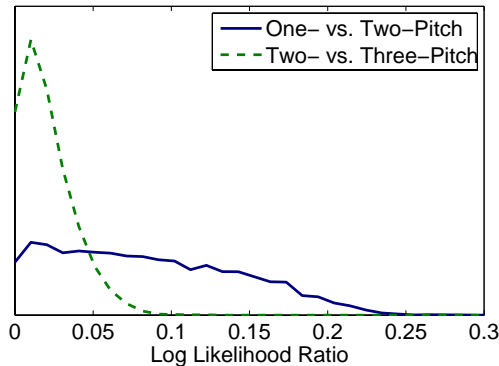


Figure 6.9: Log likelihood ratios comparing one- vs. two-pitch states in two-pitch frames, and two- vs. three-pitch states in three-pitch frames.

in isolation. Consistent with the observations made in Chapter 3, we observe that incorporating pitch cues for simultaneous organization is beneficial to binaural segregation but that, given spatially fixed and separated sources, azimuth-based sequential organization outperforms a monaural comparison system. Results also show that the proposed HMM framework improves simultaneous organization relative to the pitch-based methods proposed in Chapters 3 and 4. Improvement is primarily due to jointly utilizing spatial and periodicity for across-frequency grouping, although some benefit of across-time grouping based jointly on pitch and azimuth is also observed.

We also show that the proposed method outperforms three existing two-microphone systems. Improvement relative to the comparison methods of [58, 124] is particularly notable given that these methods assume that the number of sources is known, while sources are detected by the proposed method. Further, while these comparison methods fundamentally assume sources are in a fixed position, the proposed method

is capable of processing mixtures with moving sound sources through inclusion of a motion model. The results from Experiment 4 illustrate the high degree of accuracy achieved in terms of detection and localization and indicate that the proposed HMM framework may be a promising approach to source tracking. Tracking and segregation of an unknown number of moving sources are problems to be addressed in future work.

Our long-term goal is the development of a robust binaural system that benefits from but does not rely on spatial cues. While the proposed HMM framework is a step toward that goal and results indicate that this property has largely been achieved in terms of simultaneous organization, the system still fundamentally relies on azimuth to achieve sequential organization. An interesting direction for future work is thus developing a similar integration of monaural and spatial cues for sequential organization.

CHAPTER 7

CONTRIBUTIONS AND FUTURE WORK

7.1 Contributions

This dissertation addresses the problems of binaural localization and segregation in reverberant environments. Most existing binaural CASA systems perform localization-based grouping of T-F units to achieve segregation. Such systems typically treat localization as a first-stage subproblem so that grouping can be performed based on a limited set of estimated locations. While it is obvious that human performance benefits from a multitude of acoustic cues, localization-based grouping has become a dominant paradigm due to excellent performance in conditions with limited reverberation or background noise. In contrast to this approach, we integrate both monaural and binaural cues to achieve robust performance in real-world conditions.

In Chapter 3 we propose a novel system to integrate pitch and azimuth cues for localization and segregation of voiced speech. To this end we train a set of binaural models to extract azimuth-dependent cues and develop a probabilistic framework that decomposes segregation into separate simultaneous and sequential organization

processes. Using existing monaural CASA methods we illustrate that pitch cues are more effective than azimuth cues for simultaneous organization of voiced speech. We show that when azimuth cues can be integrated over large T-F regions, they can still be used reliably for sequential organization when sources are separated in space. We demonstrate that localization-based sequential organization outperforms an existing model-based approach. Finally, we show that integrating pitch cues also improves localization performance relative to binaural comparison methods, particularly for closely spaced sources or in reverberant environments.

In Chapter 4 we extend the framework from Chapter 3 to include additional monaural cues and develop an azimuth-dependent binaural model that both diminishes the training burden as compared to existing techniques and is adaptable to a new binaural setup (listener) and new environments. We perform extensive testing on multisource mixtures in reverberant and noisy conditions using both simulated and measured impulse responses. We show that monaural grouping improves localization of simultaneous sources relative to several binaural baselines and that the benefit of monaural grouping is most pronounced for distant sources and in low SNR conditions. We demonstrate that the proposed binaural model can achieve robust performance in real environments even when only anechoic measurements are available for training.

Chapter 5 seeks to identify a concrete computational objective for CASA-based segregation in reverberant environments. Motivated by psychoacoustics literature, we introduce the reflection boundary to the definition of the IBM so that target reverberation can be broken into useful and detrimental components. We perform a

series of experiments to identify IBM definitions that improve intelligibility of reverberant and noisy speech for normal hearing listeners. We conclude that in moderate to heavy reverberation, a commonly sought IBM that treats the fully reverberant target as the desired signal does not lead to improved intelligibility. In contrast, we show that IBM definitions with reflections boundary of 100 ms or less are capable of increasing speech intelligibility. We contend that although similar performance can be obtained with both an IBM based on the direct sound target and an IBM based on direct sound and early reflections, that the latter is more consistent with psychoacoustics and allows the use of a local SNR threshold within the range commonly used for anechoic signals.

Finally, in Chapter 6 we develop a binaural localization and segregation system that is more flexible and requires less prior knowledge than the systems presented in Chapters 3 and 4. We propose a novel HMM framework to perform simultaneous organization based jointly on pitch and azimuth cues. The framework incorporates MLPs trained on both feature types, which allows the relative contribution of pitch and azimuth to T-F grouping to be learned. This approach retains the benefit of pitch-based grouping, but allows for improved across-frequency grouping and grouping across local time intervals. Through a series of experiments we demonstrate that the proposed system outperforms state-of-the-art binaural segregation algorithms in both simulated and real environments, achieves more accurate detection and azimuth estimation relative to commonly used localization and tracking procedures, improves

simultaneous organization relative to using pitch alone, and outperforms a monaural approach to sequential organization.

7.2 Future Work

The system presented in Chapter 6 builds on those presented in earlier chapters and is thus a nice point of departure for future work. In spite of constraining the search space of the HMM using separate pitch and azimuth modules, computational complexity is still a concern. Particularly for online segregation and tracking, a more efficient algorithm for computation of the multisource posterior density would be necessary.

While we have observed good performance on multi-talker mixtures with reverberation and some additional diffuse background noise, one clear research goal is evaluation in a more diverse set of acoustic conditions. Understanding the effect of different types of interfering sounds as well as interferences with different spatial characteristics would lend further insight into the problem. Along these lines, another important step is to deal with moving sound sources (or a moving listener). While inclusion of a motion model for azimuth transitions would be straightforward to handle some motion locally in time, one might face a number of challenges. Most notably, azimuth-based sequential organization may become considerably more difficult in such circumstances. Further, source or listener movements may require localization and tracking to be performed in a continuous or more finely resolved azimuth space.

We have considered integration of spatial cues with monaural cues, but have primarily focused on pitch. As alluded to in Section 6.8, incorporating additional

monaural cues, particularly for sequential organization, should be beneficial. In this case, much like we saw when integrating pitch and spatial cues for simultaneous organization, the main challenge would be to combine cues for sequential organization so that each type of cue could be utilized when it provides reliable information. Speaker-dependent models [161, 187], speech-dependent models [134, 182, 185] or cepstral clustering [87] are promising monaural methods that could be incorporated.

In terms of localization, it would also be interesting to address three-dimensional localization. As discussed in Section 2.1, existing binaural systems for three-dimensional localization have dealt only with individual sound sources and have often focused on anechoic conditions. Monaural grouping could prove to be not just beneficial, but essential for localization in terms of elevation or distance for simultaneous sound sources. Further, since monaural grouping is based on detected source properties (e.g. pitch), information regarding local characteristics of the source could facilitate use of monaural spatial cues, which are necessary for localization in terms of elevation.

Finally, one of our main motivations is improving speech intelligibility for hearing impaired listeners. As such, it would be informative to perform subjective intelligibility experiments with the proposed approach to determine whether or not the system is capable of improving listening in difficult conditions.

BIBLIOGRAPHY

- [1] P. Aarabi. Self-localization dynamic microphone arrays. *IEEE Trans. Syst., Man, Cybern. C*, 32(2):474–484, 2002.
- [2] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *Proc. WASPAA*, 2001.
- [3] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.*, 65:943–950, 1979.
- [4] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney. Determination of the potential benefit of time-frequency gain manipulation. *Ear and Hearing*, 27(5):480–492, 2006.
- [5] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada. Blind separation of more speech than sensors with less distortion by combining source sparseness and ICA. In *Proc. IWAENC*, 2003.
- [6] T. Arbogast, C. R. Mason, and G. Kidd Jr. The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 117:2169–2180, 2005.
- [7] P. F. Assmann and Q. Summerfield. Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. *J. Acoust. Soc. Am.*, 88:680–697, 1990.
- [8] J. Benesty. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *J. Acoust. Soc. Am.*, 107(5):384–391, 2000.
- [9] V. Best, F. J. Gallun, S. Carlile, and B. G. Shinn-Cunningham. Binaural interference and auditory grouping. *J. Acoust. Soc. Am.*, 121(2):1070–1076, 2007.
- [10] J. Blauert. *Spatial Hearing - The Psychophysics of Human Sound Localization*. MIT Press, 1997.
- [11] M. Bodden. Modeling human sound-source localization and the cocktail-party-effect. *Acta Acoustica*, 1:43–55, 1993.

- [12] P. Boersma. Accurate short-time analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Inst. of Phonetic Sci.*, volume 17, pages 97–110, 1993.
- [13] J. S. Bradley, H. Sato, and M. Picard. On the importance of early reflections for speech in rooms. *J. Acoust. Soc. Am.*, 113:3233–3244, 2003.
- [14] M. Brandstein. Time-delay estimation of reverberated speech exploiting harmonic structure. *J. Acoust. Soc. Am.*, 105:2914–2919, 1999.
- [15] M. Brandstein and D. Ward, editors. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.
- [16] J. Breebaart, S. van de Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. *J. Acoust. Soc. Am.*, 110:1074–1088, 2001.
- [17] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990.
- [18] L. W. Brooks and I. S. Reed. Equivalence of the likelihood ratio processor, the maximum signal-to-noise ratio filter and the Wiener filter. *IEEE Trans. Aerosp. Electron. Sys.*, AES-8:690–692, 1972.
- [19] G. J. Brown and M. P. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8:297–336, 1994.
- [20] D. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang. Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask. *J. Acoust. Soc. Am.*, 120:4007–4018, 2006.
- [21] D. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang. Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers. *J. Acoust. Soc. Am.*, 125:4006–4022, 2009.
- [22] H. Buchner, R. Aichner, and W. Kellermann. Blind source separation for convolutive mixtures: a unified treatment. In *Audio Signal Processing*, chapter 10, pages 255–294. Kluwer Academic, 2004.
- [23] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann. Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering. In *Proc. ICASSP*, pages 97–100, 2005.
- [24] K. P. Burnham and D. R. Anderson. *Model Selection and multimodel inference: A practical information-theoretic approach*. Springer, 2002.
- [25] D. R. Campbell. The ROOMSIM user guide (v3.3), 2004.
- [26] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE*, 57:1408–1419, 1969.

- [27] J.-F. Cardoso. Blind signal separation: Statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.
- [28] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non gaussian signals. *Proc. IEEE*, 140:362–370, 1993.
- [29] J. Chen, J. Benesty, and Y. Huang. Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments. *EURASIP J. on App. Signal Proc.*, 2005(1):25–36, 2005.
- [30] P. C. Ching, Y. T. Chan, and K. C. Ho. Constrained adaptation for time delay estimation with multipath propagation. *IEEE Proc. Part F: Radar and Signal Proc.*, 138:453–458, 1991.
- [31] W.-S. Chou, K.-M. Cheong, and T.-S. Chi. A binaural algorithm for space and pitch detection. In *Proc. ICASSP*, 2011.
- [32] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker. A speech fragment approach to localising multiple speakers in reverberant environments. In *Proc. ICASSP*, pages 4593–4596, April 2009.
- [33] H. S. Colburn. Theory of binaural interaction based on auditory-nerve data. i. General strategy and preliminary results on interaural discrimination. *J. Acoust. Soc. Am.*, 54:1458–1470, 1973.
- [34] H. S. Colburn and A. Kulkarni. Models of sound localization. In A. N. Popper and R. R. Fay, editors, *Sound Source Localization*, pages 272–316. Springer New York, 2005.
- [35] H. S. Colburn, B. G. Shinn-Cunningham, G. Kidd Jr., and N. Durlach. The perceptual consequences of binaural hearing. *Int. J. Audiol.*, 45:S34–S44, 2006.
- [36] M. P. Cooke. *Modeling Auditory Processing and Organization*. Cambridge University Press, Cambridge, U. K., 1993.
- [37] M. P. Cooke and D. P. W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35:141–177, 2001.
- [38] M. P. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Commun.*, 34:267–285, 2001.
- [39] B. Cornelis, M. Moonen, and J. Wouters. A QRD-RLS based frequency domain multichannel wiener filter algorithm for noise reduction in hearing aids. In *Proc. EUSIPCO*, pages 1953–1857, 2010.

- [40] J. F. Culling, K. I. Hodder, and C. Y. Toh. Effects of reverberation on perceptual segregation of competing voices. *J. Acoust. Soc. Am.*, 114:2871–2876, 2003.
- [41] J. F. Culling and Q. S. Summerfield. Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *J. Acoust. Soc. Am.*, 98:785–797, 1995.
- [42] R. Cusani. Performance of fast time delay estimators. *IEEE Trans. Acoust., Speech, Signal Proc.*, 37(5):757–759, 1989.
- [43] C. J. Darwin. Spatial hearing and perceiving sources. In W. A. Yost, A. N. Popper, and R. R. Fay, editors, *Auditory Perception of Sound Sources*, pages 215–232. Springer, 2007.
- [44] C. J. Darwin and R. W. Hukin. Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *J. Acoust. Soc. Am.*, 102:2316–2324, 1997.
- [45] C. J. Darwin and R. W. Hukin. Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony. *J. Acoust. Soc. Am.*, 103:1080–1084, 1998.
- [46] C. J. Darwin and R. W. Hukin. Auditory objects of attention: The role of interaural time differences. *J. Exp. Psychol. Hum. Percept. Perform.*, 25:617–629, 1999.
- [47] C. J. Darwin and R. W. Hukin. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J. Acoust. Soc. Am.*, 107:970–977, 2000.
- [48] M. S. Datum, F. Palmieri, and A. Moiseff. An artificial neural network for sound localization using binaural cues. *J. Acoust. Soc. Am.*, 100:372–383, 1996.
- [49] A. de Cheveigne, S. McAdams, and C. M. H. Marin. Concurrent vowel identification. ii. Effects of phase, harmonicity, and task. *J. Acoust. Soc. Am.*, 101:2848–2856, 1997.
- [50] P. N. Denbigh and J. Zhao. Pitch extraction and separation of overlapping speech. *Speech Commun.*, 11:119–125, 1992.
- [51] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. Robust localization in reverberant rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, chapter 8, pages 157–180. Springer, 2001.

- [52] M. Dietz, S. D. Ewert, and V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Commun.*, 53:592–605, 2011.
- [53] H. Dillon. *Hearing Aids*. Thieme, 2001.
- [54] S. Doclo and M. Moonen. Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments. *EURASIP J. App. Signal Proc.*, 2003:1110–1124, 2003.
- [55] S. Doclo, A. Spriet, J. Wouters, and M. Moonen. Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction. *Speech Commun.*, 49:636–656, 2007.
- [56] W. R. Drennan, S. Gatehouse, and C. Lever. Perceptual segregation of competing speech sounds: the role of spatial location. *J. Acoust. Soc. Am.*, 114:2167–2177, 2003.
- [57] R. O. Duda and W. L. Martens. Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.*, 104(5):3048–3058, 1998.
- [58] N. Q. K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. Audio, Speech, Lang. Proc.*, 18:1830–1840, 2010.
- [59] N. I. Durlach. Equalization and cancellation theory of binaural masking level differences. *J. Acoust. Soc. Am.*, 35:1206–1218, 1963.
- [60] R. H. Dye. The combination of interaural information across frequencies: Lateralization on the basis of interaural delay. *J. Acoust. Soc. Am.*, 88:2159–2170, 1990.
- [61] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Proc.*, 32(6):1109–1121, 1984.
- [62] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Trans. Audio, Speech, Lang. Proc.*, 15:1741–1752, 2007.
- [63] C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am.*, 116 (5):3075–3089, 2004.
- [64] A. S. Feng and D. L. Jones. Location-based grouping. In D. L. Wang and G. J. Brown, editors, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, pages 187–208. Wiley/IEEE Press, 2006.

- [65] R. L. Freyman, U. Balakrishnan, and K. Helfer. Spatial release from informational masking in speech recognition. *J. Acoust. Soc. Am.*, 109:2112–2122, 2000.
- [66] O. L. Frost. An algorithm for linearly constrained adaptive array processing. *Proc. IEEE*, 60, 1972.
- [67] W. G. Gardner and K. D. Martin. HRTF measurements of a KEMAR. *J. Acoust. Soc. Am.*, 97:3907–3908, 1995.
- [68] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic phonetic continuous speech corpus, 1993.
- [69] E. L. J. George, S. T. Goverts, J. M. Festen, and T. Houtgast. Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners. *J. of Speech Lang. and Hearing Research*, 53:1429–1439, 2010.
- [70] B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138, 1990.
- [71] J. E. Greenberg, P. M. Peterson, and P. Zurek. Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. *J. Acoust. Soc. Am.*, 94:3009–3010, 1993.
- [72] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propagation.*, 30:27–34, 1981.
- [73] T. Habib and H. Romsdorfer. Comparison of SRP-PHAT and multiband-POPI algorithms for speaker localization using particle filters. In *Proc. DAFX*, 2010.
- [74] K. Han and D. L. Wang. An SVM based classification approach to speech separation. In *Proc. ICASSP*, pages 5212–5215, 2011.
- [75] D. J. Hand. Idiot’s Bayes—Not so stupid after all? *International Statistics Review*, 69(3):385–398, 2001.
- [76] S. Harding, J. Barker, and G. J. Brown. Mask estimation for missing data speech recognition based on statistics of binaural interaction. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):58–67, 2006.
- [77] M. L. Hawley, R. Y. Litovsky, and J. F. Culling. The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. *J. Acoust. Soc. Am.*, 115:833–843, 2004.
- [78] L. M. Heller and C. Trahiotis. Interference in detection of interaural delay in a sinusoidally amplitude-modulated tone produced by a second, spectrally remote sinusoidally amplitude-modulated tone. *J. Acoust. Soc. Am.*, 97:1808–1816, 1995.

- [79] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen. On optimal multi-channel mean-squared error estimators for speech enhancement. *IEEE Signal Processing Letters*, 16(10):885–888, October 2009.
- [80] T. Hidaka, L. L. Beranek, and T. Okano. Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls. *J. Acoust. Soc. Am.*, 98:988–1007, 1995.
- [81] N. I. Hill and C. J. Darwin. Lateralization of a perturbed harmonic: Effects of onset asynchrony and mistuning. *J. Acoust. Soc. Am.*, 100:2352–2364, 1996.
- [82] G. Hu and D. L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Netw.*, 15(5):1135–1150, 2004.
- [83] G. Hu and D. L. Wang. Auditory segmentation based on onset and offset analysis. *IEEE Trans. Acoust., Speech, Signal Proc.*, 15:396–405, 2007.
- [84] G. Hu and D. L. Wang. Segregation of unvoiced speech from nonspeech interference. *J. Acoust. Soc. Am.*, 124:1306–1319, 2008.
- [85] G. Hu and D. L. Wang. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Trans. Audio, Speech, Lang. Proc.*, 18:2067–2079, 2010.
- [86] J.-S. Hu and W.-H. Liu. Location classification of nonstationary sound sources using binaural room distribution patterns. *IEEE Trans. Audio, Speech, Lang. Proc.*, 17:682–692, 2009.
- [87] K. Hu and D. L. Wang. An approach to sequential grouping in cochannel speech. In *Proc. ICASSP*, 2011.
- [88] K. Hu and D. L. Wang. Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction. *IEEE Trans. Audio, Speech, Lang. Proc.*, 19:1600–1609, 2011.
- [89] R. W. Hukin and C. J. Darwin. Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification. *Percept. Psychophys.*, 57:191–196, 1995.
- [90] C. Hummersone, R. Mason, and T. Brookes. Dynamic precedence effect modeling for source separation in reverberant environments. *IEEE Trans. Audio, Speech, Lang. Proc.*, 18:1867–1871, 2010.
- [91] A. Hyvärinen, J. Karbunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, 2001.
- [92] IEEE. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.*, 17:227–246, 1969.

- [93] Y. Izumi, N. Ono, and S. Sagayama. Sparseness-based 2CH BSS using the EM algorithm in reverberant environment. In *Proc. WASPAA*, pages 147–150, Oct. 2007.
- [94] L. A. Jeffress. A place theory of sound localization. *Comparative Physiology and Psychology*, 41:35–39, 1948.
- [95] J. R. Jensen, M. G. Christensen, and S. H. Jensen. Joint DOA and fundamental frequency estimation methods based on 2-d filtering. In *Proc. EUSIPCO*, 2010.
- [96] Z. Jin and D. L. Wang. A supervised learning approach to monaural segregation of reverberant speech. *IEEE Trans. Audio, Speech, Lang. Proc.*, 17:625–638, 2009.
- [97] Z. Jin and D. L. Wang. HMM-based multipitch tracking for noisy and reverberant speech. *IEEE Trans. Audio, Speech, Lang. Proc.*, 19:1091–1102, 2011.
- [98] Z. Jin and D. L. Wang. Reverberant speech segregation based on multipitch tracking and classification. *IEEE Trans. Audio, Speech, Lang. Proc.*, 19:2328–2337, 2011.
- [99] M. Képesi, F. Pernkopf, and M. Wohlmayr. Joint position pitch tracking for 2-channel audio. In *Int. Workshop on Content based Multimedia Indexing*, 2007.
- [100] F. Keyrouz, W. Maier, and K. Diepold. Robotic binaural localization and separation of more than two concurrent sound sources. In *Proc. ISSPA*, 2007.
- [101] F. Keyrouz, Y. Naous, and K. Diepold. A new method for binaural 3-D localization based on HRTFs. In *Proc. ICASSP*, 2006.
- [102] G. Kidd Jr., C. R. Mason, A. Brughera, and W. M. Hartmann. The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acustica united with Acta Acustica*, 2005.
- [103] G. Kim, Y. Lu, Y. Hu, and P. Loizou. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *J. Acoust. Soc. Am.*, 126(3):1486–1494, 2009.
- [104] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. Wang. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *J. Acoust. Soc. Am.*, 126:1415–1426, 2009.
- [105] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust., Speech, Signal Proc.*, 24, no. 4:320–327, 1976.

- [106] Kuttruff. *Room Acoustics*. Taylor & Francis, 2000.
- [107] D. Li and S. E. Levinson. A Bayes-rule based hierarchical system for binaural sound source localization. In *Proc. ICASSP*, 2003.
- [108] N. Li and P. C. Loizou. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *J. Acoust. Soc. Am.*, 123, no. 3:1673–1682, 2008.
- [109] Y. Li, J. Woodruff, and D. L. Wang. Monaural musical sound separation using pitch and common amplitude modulation. *IEEE Trans. Audio, Speech, Lang. Proc.*, 17:1361–1371, 2009.
- [110] W. Lindemann. Extension of binaural cross-correlation model by contralateral inhibition. i: simulation of lateralization for stationary signals. *J. Acoust. Soc. Am.*, 80:1608–1622, 1986.
- [111] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *J. Acoust. Soc. Am.*, 106:1633–1654, 1999.
- [112] C. Liu, B. C. Wheeler, W. D. O’Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng. Localization of multiple sound sources with two microphones. *J. Acoust. Soc. Am.*, 108:1888–1905, 2000.
- [113] C. Liu, B. C. Wheeler, W. D. O’Brien, C. R. Lansing, R. C. Bilger, D. L. Jones, and A. S. Feng. A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers. *J. Acoust. Soc. Am.*, 110:3218, 2001.
- [114] J. P. A. Lochner and J. F. Burger. The influence of reflections on auditorium acoustics. *J. Sound Vib.*, 1:426–454, 1964.
- [115] P. Loizou, editor. *Speech enhancement: Theory and practice*. CRC Press, 2007.
- [116] A. Lombard, Y. Zheng, H. Buchner, and Kelle. TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis. *IEEE Trans. Audio, Speech, Lang. Proc.*, 19:1490–1503, 2011.
- [117] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP*, 2005:1110–1126, 2005.
- [118] Y.-C. Lu and M. Cooke. Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources. *IEEE Trans. Audio, Speech, Lang. Proc.*, 18:1793–1805, 2010.
- [119] R. F. Lyon. A computational model of binaural localization and separation. In *Proc. ICASSP*, 1983.

- [120] N. Ma, J. Barker, H. Christensen, and P. Green. Binaural cues for fragment-based speech recognition in reverberant multisource environments. In *Proc. INTERSPEECH*, 2011.
- [121] W.-K. Ma, B.-N. Vo, S. Singh, and A. Baddelay. Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach. *IEEE Trans. Signal Proc.*, 54:3291–3304, 2006.
- [122] R. P. S. Mahler. *Statistical multisource-multitarget information fusion*. Artech House, 2007.
- [123] M. I. Mandel and D. P. W. Ellis. EM localization and separation using interaural level and phase cues. In *Proc. WASPAA*, pages 275–278, Oct. 2007.
- [124] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis. Model-based expectation-maximization source separation and localization. *IEEE Trans. Audio, Speech, Lang. Proc.*, 18(2):382–394, February 2010.
- [125] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Proc.*, 13:845–856, 2005.
- [126] A. Masnadi-Shirazi and B. Rao. Separation and tracking of multiple speakers in a reverberant environment using a multiple model particle filtering glimpsing method. In *Proc. ICASSP*, 2011.
- [127] T. May, S. van de Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Trans. Audio, Speech, Lang. Proc.*, 19:1–13, 2011.
- [128] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, London, U.K., 5th edition, 2003.
- [129] A. K. Nabelek and P. K. Robinson. Monaural and binaural speech perception in reverberation for listeners of various ages. *J. Acoust. Soc. Am.*, 71:1242–1248, 1982.
- [130] T. Nakatani, M. Goto, and H. G. Okuno. Localization by harmonic structure and its application to harmonic sound stream segregation. In *Proc. ICASSP*, 1996.
- [131] L. Y. Ngan, Y. Wu, C. So, P. C. Ching, and S. W. Lee. Joint time delay and pitch estimation for speaker localization. In *Proc. of ICAS*, 2003.
- [132] M. Nilsson, S. Soli, and J. Sullivan. Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.*, 95:1085–1099, 1994.

- [133] J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *J. Acoust. Soc. Am.*, 119:463–479, 2006.
- [134] J. Nix and V. Hohmann. Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. *IEEE Trans. Audio, Speech, Lang. Proc.*, 15:995–1008, 2007.
- [135] J. Nix, M. Kleinschmidt, and Hohmann. Computational scene analysis of cocktail-party situations based on sequential monte carlo methods. In *Proc. CSSC*, 2003.
- [136] Özgür Yılmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.
- [137] K. J. Palomäki, G. J. Brown, and J. Barker. Missing data speech recognition in reverberant conditions. In *Proc. ICASSP*, 2002.
- [138] K. J. Palomäki, G. J. Brown, and D. L. Wang. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication*, 43:361–378, 2004.
- [139] L. Parra and C. Spence. Convolutional blind separation of non-stationary sources. *IEEE Trans. Speech Audio Proc.*, 8:320–327, 2000.
- [140] T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am.*, 60(4):911–918, 1976.
- [141] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice. An efficient auditory filterbank based on the gammatone function. Technical report, MRC App. Psych. Unit, Cambridge, 1988.
- [142] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems. Two-microphone separation of speech mixtures. *IEEE Trans. Neural Netw.*, 19:475–492, 2008.
- [143] M. Raspaud, H. Viste, and G. Evangelista. Binaural source localization by joint estimation of ILD and ITD. *IEEE Trans. Audio, Speech, Lang. Proc.*, 18(1):68–77, 2010.
- [144] K. Reindl, Y. Zheng, and W. Kellermann. Analysis of two generic wiener filtering concepts for binaural speech enhancement in hearing aids. In *Proc. EUSIPCO*, pages 988–993, 2010.
- [145] S. J. Rennie, P. Aarabi, T. Kristjansson, B. J. Frey, and K. Achan. Robust variational speech separation using fewer microphones than speakers. In *Proc. ICASSP*, 2003.

- [146] M. Reyes-Gómez, B. Raj, and D. P. W. Ellis. Multi-channel source separation by factorial HMMs. In *Proc. ICASSP*, 2003.
- [147] N. Roman, S. Srinivasan, and D. L. Wang. Binaural segregation in multisource reverberant environments. *J. Acoust. Soc. Am.*, 120(6):4040–4051, 2006.
- [148] N. Roman and D. L. Wang. Binaural tracking of multiple moving sources. *IEEE Trans. Audio, Speech, Lang. Proc.*, 16:728–739, 2008.
- [149] N. Roman, D. L. Wang, and G. Brown. Speech segregation based on sound localization. *J. Acoust. Soc. Am.*, 114:2236–2252, 2003.
- [150] N. Roman and J. Woodruff. Intelligibility of reverberant noisy speech with ideal binary masking. *J. Acoust. Soc. Am.*, 130:2153–2161, 2011.
- [151] S. T. Roweis. One microphone source separation. In *Neural Information Processing System 13*, pages 793–799, 2001.
- [152] H. Sawada, S. Araki, and S. Makino. A two-state frequency-domain blind source separation method for underdetermined convolutive mixtures. In *Proc. WASPAA*, pages 139–142, Oct. 2007.
- [153] H. Sawada, S. Araki, R. Mukai, and S. Makino. Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1592–1604, 2007.
- [154] B. M. Sayers and E. C. Cherry. Mechanism of binaural fusion in the hearing of speech. *J. Acoust. Soc. Am.*, 29:973–987, 1957.
- [155] M. Seltzer, B. Raj, and Stern. Speech recognizer-based microphone array processing for robust hands-free speech recognition. In *Proc. ICASSP*, 2002.
- [156] T. M. Shackleton and Meddis. The role of interaural time difference and fundamental frequency difference in the identification of concurrent vowel pairs. *J. Acoust. Soc. Am.*, 91:3579–3581, 1992.
- [157] T. M. Shackleton, R. Meddis, and Hew. The role of binaural and fundamental frequency difference cues in the identification of concurrently presented vowels. *J. Exp. Psychol.*, 47A:545–563, 1994.
- [158] S. A. Shamma, N. Shen, and P. Gopalaswamy. Stereausis: Binaural processing without neural delays. *J. Acoust. Soc. Am.*, 86:989–1006, 1989.
- [159] A. Shamsoddini and P. N. Denbigh. A sound segregation algorithm for reverberant conditions. *Speech Commun.*, 33:179–196, 2001.

- [160] Y. Shao and D. L. Wang. Model-based sequential organization in cochannel speech. *IEEE Trans. Audio, Speech, Lang. Proc.*, 14:289–298, 2006.
- [161] Y. Shao and D. L. Wang. Sequential organization of speech in computational auditory scene analysis. *Speech Commun.*, 51:657–667, 2009.
- [162] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.
- [163] K. U. Simmer, J. Bitzer, and C. Marro. Post-filtering techniques. In *Microphone Arrays: Signal processing techniques and applications*, pages 39–60. Springer, 2001.
- [164] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22:21–34, 1998.
- [165] M. Souden, J. Benesty, and S. Affes. On optimal frequency-domain multi-channel linear filtering for noise reduction. *IEEE Trans. Audio, Speech, Lang. Proc.*, 18:260–275, 2010.
- [166] A. Stéphenne and B. Champagne. A new cepstral prefiltering technique for estimating time delay under reverberant conditions. *Signal Proc.*, 59(3):253–266, 1997.
- [167] R. M. Stern, G. J. Brown, and D. L. Wang. Binaural sound localization. In D. L. Wang and G. J. Brown, editors, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, pages 147–185. Wiley, 2006.
- [168] R. M. Stern and H. S. Colburn. Lateral-position based models of interaural discrimination. *J. Acoust. Soc. Am.*, 77:753–755, 1985.
- [169] D. E. Sturim, M. S. Brandstein, and H. F. Silverman. Tracking multiple talkers using microphone-array measurements. In *Proc. ICASSP*, 1997.
- [170] C. Trahiotis and L. R. Bernstein. Detectability of interaural delays over select spectral regions: Effects of flanking noise. *J. Acoust. Soc. Am.*, 87:810–813, 1990.
- [171] J. Vermaak and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *Proc. ICASSP*, 2001.
- [172] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.
- [173] E. Vincent, R. Gribonval, and M. D. Plumbley. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87:1933–1950, 2007.

- [174] S. Vishnubhotla and C. Y. Epsy-Wilson. An algorithm for speech segregation of co-channel speech. In *Proc. ICASSP*, pages 109–112, April 2009.
- [175] H. Wallach, E. B. Newman, and M. R. Rosenzweig. The precedence effect in sound localization. *Am. J. Psychol.*, 52:315–336, 1949.
- [176] D. L. Wang. On ideal binary masks as the computational goal of auditory scene analysis. In P. Divenyi, editor, *Speech Separation by Humans and Machines*, pages 181–197. Kluwer Academic, Boston, MA, 2005.
- [177] D. L. Wang and G. J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, 10(3):684–697, 1999.
- [178] D. L. Wang and G. J. Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley/IEEE Press, Hoboken, NJ, 2006.
- [179] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner. Speech intelligibility in background noise with ideal binary time-frequency masking. *J. Acoust. Soc. Am.*, 125:2336–2347, 2009.
- [180] D. Ward, E. Lehmann, and R. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Trans. Speech Audio Proc.*, 11:826–836, 2003.
- [181] M. Weintraub. *A theory and computational model of auditory monaural sound separation*. PhD thesis, Stanford University, Department of Electrical Engineering, 1985.
- [182] R. Weiss, M. Mandel, and Ellis. Combining localization cues and source model constraints for binaural source separation. *Speech Commun.*, 53:606–621, 2011.
- [183] B. Widrow, P. Mantey, L. Griffiths, and B. Goode. Adaptive antenna systems. *Proc.*, 55:2143–2159, 1967.
- [184] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Koerner. A probabilistic model for binaural sound localization. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 36:982–994, 2006.
- [185] K. Wilson. Speech source separation by combining localization cues with mixture models of speech spectra. In *Proc. ICASSP*, 2007.
- [186] K. W. Wilson and T. Darrell. Learning a precedence effect-like weighting function for the generalized cross-correlation framework. *IEEE Trans. Audio, Speech, Lang. Proc.*, 14:2156–2164, 2006.

- [187] M. Wohlmayr, M. Stark, and Pern. A probabilistic interaction model for multipitch tracking with factorial hidden markov models. *IEEE Trans. Audio, Speech, Lang. Proc.*, 19:799–810, 2011.
- [188] J. Woodruff and D. L. Wang. Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization. Technical Report OSU-CISRC-5/08-TR27, The Ohio State University, 2008.
- [189] J. Woodruff and D. L. Wang. Integrating monaural and binaural analysis for localizing multiple reverberant sound sources. In *Proc. ICASSP*, pages 2706–2709, Mar. 2010.
- [190] J. Woodruff and D. L. Wang. Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization. *IEEE Trans. Acoust., Speech, Signal Proc.*, 18:1856–1866, 2010.
- [191] J. Woodruff and D. L. Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Trans. Audio, Speech, Lang. Proc.*, 20:1503–1512, 2012.
- [192] J. Woodruff and D. L. Wang. Binaural speech segregation based on pitch and azimuth tracking. In *Proc. ICASSP*, 2012.
- [193] W. S. Woods and H. S. Colburn. Test of a model of auditory object formation using intensity and interaural time difference discrimination. *J. Acoust. Soc. Am.*, 91:2894–2902, 1992.
- [194] W. S. Woods, M. Hansen, T. Wittkop, and B. Kollmeier. A simple architecture for using multiple cues in sound separation. In *Proc. ICSLP*, 1996.
- [195] S. N. Wrigley and G. J. Brown. Binaural speech separation using recurrent timing neural networks for joint F0-localisation estimation. In *Machine Learning for Multimodal Interaction*, pages 271–282. Springer Berlin / Heidelberg, 2008.
- [196] W. Zhang and B. D. Rao. A two microphone-based approach for source localization of multiple speech sources. *IEEE Trans. Audio, Speech, Lang. Proc.*, 18:1913–1928, 2010.
- [197] X. Zhong and J. R. Hopgood. Time-frequency masking based multiple acoustic sources tracking applying rao-blackwellised monte carlo data association. In *Proc. SSP*, 2009.