AUDITORY-BASED ALGORITHMS FOR SOUND SEGREGATION IN MULTISOURCE AND 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

Nicoleta Roman, M. S.

The Ohio State University

2005

Dissertation Committee:

Professor DeLiang Wang, Advisor

Professor Eric Fosler-Lussier

Professor Philip Schniter

Approved by

Advisor

Graduate Program in Department of Computer Science and Engineering
ABSTRACT

At a cocktail party, we can selectively attend to a single voice and filter out all the other acoustical interferences. This perceptual ability has motivated the emergence of a new field of study known as computational auditory scene analysis (CASA) which aims to build speech separation systems that incorporate principles of auditory organization. The psychological process of figure-ground segregation suggests that the target signal should be segregated as foreground while the remaining stimuli are treated as background. Accordingly, the computational goal of CASA should be to estimate an ideal time-frequency (T-F) binary mask, which selects the target if it is stronger than the interference in a local T-F unit. This dissertation investigates four aspects of CASA processing: location-based speech segregation in multisource environments, binaural tracking of multiple moving sources, binaural sound segregation in reverberant environments, and monaural segregation of reverberant speech.

The principal cues used by the auditory system to determine locations are the interaural time difference (ITD) and interaural intensity difference (IID) between the two ears. We observe that within a narrow frequency band, modifications to the relative strength of the target source with respect to the interference trigger systematic changes for ITD and IID. Moreover, for a fixed spatial configuration, this interaction produces a
characteristic clustering in the binaural feature space. Consequently, we propose a supervised learning approach to estimate the ideal binary mask using the estimated binaural features. A systematic evaluation in terms of signal-to-noise ratio (SNR) as well as automatic speech recognition (ASR) scores shows that the resulting system produces masks very close to the ideal binary ones in anechoic conditions. Furthermore, the model produces large speech intelligibility improvements with normal listeners.

In realistic environments, source motion requires consideration. Binaural cues are strongly correlated with source locations in T-F units dominated by only one source resulting in channel-dependent conditional probabilities. Consequently, we propose a multi-channel method that integrates these probabilities across reliable frequency channels in order to produce a likelihood function in the target space. Finally, a hidden Markov model (HMM) is employed for forming continuous tracks and automatically detecting the number of active sources.

While the above binaural systems perform optimally in anechoic conditions, reverberation affects the ITD and IID cues and therefore degrades their performance. For reverberant conditions, we propose a binaural segregation system that combines target cancellation through adaptive filtering and a binary decision rule to estimate the ideal binary mask. Specifically, we observe a correlation between the attenuation produced by the target cancellation stage and the relative strength between target and interference which is used subsequently to determine the target dominant T-F units. A major advantage of the proposed system is that, while requiring a fixed target location, it imposes no restrictions on the number, location or content of the
interfering sources. An extensive comparison using SNR as well as ASR results shows that our system outperforms standard two-microphone beamforming approaches.

While binaural cues provide important cues for auditory organization, psychoacoustic evidence suggests that monaural processing play a vital role in sound segregation. Monaural segregation utilizing periodicity has achieved considerable progress in handling additive noise in anechoic conditions. Reverberation, however, smears the harmonic structure of speech signals, and our evaluations show that the performance of pitch-based segregation deteriorates with the increase of room reverberation time. We propose a two-stage monaural separation system that combines the inverse filtering of the room impulse response corresponding to target location with a pitch-based segregation method. As a result of the first stage, the harmonicity of a signal arriving from target direction is partially restored while signals arriving from other locations are further smeared, and this leads to improved segregation. A systematic evaluation of the system shows that the proposed system results in considerable SNR gains across different conditions.
Dedicated to my husband, Cosmin, and to my two children, Ana and Andrei
ACKNOWLEDGMENTS

I express my gratitude towards my advisor Professor DeLiang Wang for guiding the research presented in this dissertation and for relentlessly keeping me on track throughout my graduate studies. Without his unabated trust and immense support this work would not have been completed. My future career will doubtlessly be marked by his energy, his passion for research and by his philosophy of always aiming high.

Many thanks go to Dr. Guy J. Brown for his close collaboration and his extremely helpful advice and suggestions on my research over the years.

I want to also extend my thanks to the faculty of Computer Science and Engineering Department as well as to the faculty of Electrical and Computer Engineering Department at The Ohio State University among which I would like to single out Professor Eric Fosler-Lussier, Professor James W. Davis, Professor Donna Byron, Professor Neelam Soundarajan, Professor Ashok Krishnamurthy and Professor Philip Schniter. The atmosphere in these departments has been very conducive to quality research and I feel lucky to have been a graduate student here.
Thanks to all the students who were part of the Perception and Neurodynamics Laboratory at The Ohio State University while I was there. They were great colleagues and friends! They all inspired me one way or the other: Guoning Hu by being a hard-working and passionate researcher, Soundararajan Srinivasan by being an infinite source of advice both on professional and personal matters, Yang Shao by offering his technical skills and his cheerful presence, and Yipeng Li by being enthusiastic and determined. Many thanks go also to the lab alumni: Dr. Xiuwen Liu, Dr. Mingyang Wu and Peter Chang as well as to the newest members: Lei Ding, Zhaozhang Jin and Saurabh Khanwalkar.

I want to thank my family and friends who provided me with love and support during these years. I could not have made it so far without them. Thanks also go to the Romanian graduate students' community and to the parish of St. Nicholas Orthodox Church in Columbus.

Last but not the least, I wish to acknowledge the financial support provided to me through two AFOSR grants (F49620-01-1-0027, FA9550-04-1-0117) and an NSF grant (IIS-0081058) to Prof. DeLiang Wang.
VITA

June 25, 1973 ............................................ Born in Galati, Romania

July, 1996 ............................................. B. S. Computer Science,
                                      Bucharest University, Bucharest, Romania

July, 1997 ............................................. M. S. Computer Science,
                                      Bucharest University, Bucharest, Romania

April, 2000 ............................................. M. S. Computer and Information Science,
                                     The Ohio State University

PUBLICATIONS

Journal Article


Conference Papers


**Technical Reports**


**FIELDS OF STUDY**

Major Field: Computer Science and Engineering
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Objectives</td>
<td>6</td>
</tr>
<tr>
<td>1.3</td>
<td>Organization of Dissertation</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Speech Segregation Based on Sound Localization</td>
<td>12</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Model Architecture</td>
<td>16</td>
</tr>
<tr>
<td>2.3</td>
<td>Auditory Periphery</td>
<td>19</td>
</tr>
<tr>
<td>2.4</td>
<td>Azimuth Localization</td>
<td>21</td>
</tr>
<tr>
<td>2.5</td>
<td>Ideal Mask Estimation</td>
<td>25</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Pure Tones</td>
<td>26</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Model</td>
<td>33</td>
</tr>
</tbody>
</table>

Abstract ........................................................................................................... ii
Dedication ........................................................................................................ v
Acknowledgments ............................................................................................... vi
Vita .................................................................................................................. vii
List of Tables ..................................................................................................... xiii
List of Figures .................................................................................................... xv
List of Acronyms .................................................................................................. xx
List of Symbols ................................................................................................. xxii
2.6 Evaluation and Comparison ............................................. 42
  2.6.1 SNR Evaluation .................................................. 43
  2.6.2 ASR Evaluation .................................................. 51
  2.6.3 Speech Intelligibility Evaluation ................................. 56
2.7 Discussion .................................................................... 60

3. Binaural Tracking of Multiple Moving Sources .................... 66
  3.1 Introduction ................................................................ 67
  3.2 Model Architecture .................................................. 70
  3.3 Modeling Auditory Motion ......................................... 73
  3.4 Statistical Tracking .................................................. 75
    3.4.1 Dynamics Model ............................................... 76
    3.4.2 Statistics of ITD and IID .................................... 79
    3.4.3 Likelihood Model ............................................. 84
    3.4.4 HMM-Based Source Tracking .............................. 88
  3.5 Results and Comparison ........................................... 89
  3.6 Discussion ................................................................ 96

4. Binaural Segregation in Multisource Reverberant Environments ... 99
  4.1 Introduction ................................................................ 100
  4.2 Proposed Model .................................................... 102
  4.3 Evaluation and Comparison ....................................... 108
    4.3.1 SNR Evaluation ............................................... 111
    4.3.2 ASR Evaluation ............................................. 116
  4.4 Discussion ................................................................ 126

5. Pitch-based Segregation of Reverberant Speech ................... 130
  5.1 Introduction ................................................................ 131
  5.2 Model Overview .................................................... 134
  5.3 Target Inverse Filtering .......................................... 138
  5.4 Pitch-Based Speech Segregation ................................. 142
    5.4.1 Auditory Periphery and Feature Extraction ............ 144
    5.4.2 Unit Labeling ............................................... 145
    5.4.3 Segregation .................................................. 149
  5.5 Evaluation .......................................................... 150
  5.6 Discussion ................................................................ 160

6. Contributions and Future Work ....................................... 164
  6.1 Contributions ....................................................... 164
  6.2 Insights Gained .................................................... 166
6.3 Future Work ................................................................. 168

Bibliography ........................................................................ 171
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Speech signals of the training set</td>
<td>34</td>
</tr>
<tr>
<td>2.2</td>
<td>Speech signals of the testing set</td>
<td>43</td>
</tr>
<tr>
<td>2.3</td>
<td>Noise signals of the testing set</td>
<td>44</td>
</tr>
<tr>
<td>3.1</td>
<td>Jump probabilities between subspaces with zero-, one-, two- and three-active sources</td>
<td>77</td>
</tr>
<tr>
<td>3.2</td>
<td>Estimated model parameters for one-source and two-source conditions</td>
<td>85</td>
</tr>
<tr>
<td>4.1</td>
<td>SNR evaluation for a one-source interference (Scene 1)</td>
<td>111</td>
</tr>
<tr>
<td>4.2</td>
<td>SNR evaluation for a two-speaker interference (Scene 2)</td>
<td>111</td>
</tr>
<tr>
<td>4.3</td>
<td>SNR evaluation for a four-speaker interference (Scene 3)</td>
<td>111</td>
</tr>
<tr>
<td>4.4</td>
<td>SNR evaluation for our model at different reverberation levels for a one-source interference and 0 dB input SNR</td>
<td>113</td>
</tr>
<tr>
<td>4.5</td>
<td>Comparison of the proposed system in terms of SNR with adaptive beamforming</td>
<td>116</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison in terms of speech recognition accuracy with the binaural system proposed by Palomäki et al. (2004)</td>
<td>126</td>
</tr>
<tr>
<td>4.7</td>
<td>Comparison in terms of speech recognition accuracy with the negative beamformer system proposed by Álvarez et al. (2002)</td>
<td>126</td>
</tr>
<tr>
<td>5.1</td>
<td>SNR results for target speech mixed with a female interference at three input SNR levels and different reverberation times</td>
<td>154</td>
</tr>
</tbody>
</table>
5.2 SNR results using estimated pitch tracks for target speech mixed with different noise types at three input SNR levels and $T_{60} = 0.35$ s. Target is at $0^\circ$ and interference at $45^\circ$ ................................................................. 156

5.3 SNR results using ideal pitch tracks for target speech mixed with different noise types at three input SNR levels and $T_{60} = 0.35$ s. Target is at $0^\circ$ and interference at $45^\circ$ ................................................................. 157

5.4 SNR results using ideal pitch tracks for target speech mixed with two type of noise at three input SNR levels and $T_{60} = 0.35$ s. Target and interference are both located at $0^\circ$ ................................................................. 159

5.5 Comparison between the proposed algorithm and spectral subtraction (SS). Results are obtained for target speech mixed with different noise types at three input SNR levels and $T_{60} = 0.35$ s. Target is at $0^\circ$ and interference at $45^\circ$ ................................................................. 160
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Schematic diagram of the model. Binaural signals are obtained by convolving input signals with measured head related impulse responses (HRIR) from a KEMAR dummy head. A model of the auditory periphery is employed. Azimuth localization for all the sources is based on a cross-correlation mechanism. ITD and IID are computed independently for different frequency channels. A pattern analysis block produces an estimation of an ideal binary mask, which enables the reconstruction of the target signal and the interfering sound ................................................. 16</td>
</tr>
<tr>
<td>2.2</td>
<td>Functions relating azimuth to ITD for three auditory channels with center frequencies of 500 Hz, 1 kHz, and 3 kHz ............................................................... 22</td>
</tr>
<tr>
<td>2.3</td>
<td>Azimuth localization for a mixture of male utterance at 0° and female utterance at 20°. The bottom plot in each panel shows a summation across all rows. (a) Cross-correlation functions for 128 frequency channels in the range 80 Hz – 5 kHz at time frame 40 (i.e. 400 ms after the start of the stimulus). For clarity, only every other channel is shown, resulting in 64 channels. (b) Skeleton cross-correlogram for the same time frame. The arrow indicates channels that contain roughly equal energy from both target and interference. (c) Pooled cross-correlogram for a stimulus of duration 1.5 seconds, shown every 20 ms .................................................. 24</td>
</tr>
<tr>
<td>2.4</td>
<td>Theoretical approximation for the mean ITD, ( \tau_{\text{max}} ), for two pure tones randomly distributed in a narrow band centered at 500 Hz. The y-axis corresponds to the relative strength ( R ). Two cases are shown: ( \beta = \pi / 4 ) (solid line) and ( \beta = 3\pi / 4 ) (dashed line) .......................................................... 29</td>
</tr>
</tbody>
</table>
2.5 The influence of filter bandwidth on the mean and variance of ITD and IID with respect to the relative strength $R$. The data is from simulations of two pure tones uniformly distributed in a narrow band. One tone is at 0° and the another is at 30°. The sampling frequency is 44.1 kHz. (a) Mean ITD as a function of $R$ for 500 Hz center frequency and four bandwidths between 0 Hz and 200 Hz. The auditory filter ERB here is 80 Hz. $M_1$ and $M_2$ correspond to the theoretical mean ITD as derived in Eq. 2.13 and Eq. 2.14, respectively. (b) ITD variance for the same condition as in (a). (c) Mean IID as a function of $R$ for a 2.5 kHz center frequency and five bandwidths between 0 Hz and 400 Hz. $M$ corresponds to the theoretical mean IID as derived in Eq. 2.15. The auditory filter ERB is 300 Hz. (d) IID variance for the same condition as in (c).......................... 32

2.6 Relationship between ITD/IID and the relative strength $R$ for a two-source configuration: target in the median plane and interference on the right side at 30°. (a) The scatter plot shows ITD and $R$ estimates from the training corpus for a channel with center frequency of 500 Hz. The solid curve shows the theoretical mean (see Eq. 2.14) and the dash curve shows the data mean. (b) Results for IID for a filter channel with center frequency 2.5 kHz. The solid curve shows the theoretical mean (see Eq. 2.15) and the dash curve shows the data mean ............................................. 35

2.7 Relationship between ITD/IID and the relative strength $R$ for a three-source configuration: target source in the median plane and interference at -30° and 30°. Statistics are obtained from the training corpus for a channel with center frequency close to 1.5 kHz. (a) Histogram of ITD and $R$ samples. (b) Histogram of IID and $R$ samples. (c) Clustering in the ITD-IID space ................................................................. 37

2.8 Discriminability comparison for the three binaural cues, ITD, IID and IED, the joint IED-IID space and the joint ITD-IID space. Error rates are displayed as a function of channel number (frequency) for a classification task for two spatial configurations. (a) Target source in the median plane and interference on the right side at 5°. (b) Target source in the median plane and interference on the right side at 30°. IED results are shown for frequencies above 1.5 kHz, i.e. above channel number 80 .......................... 40

2.9 Comparison between the estimated mask and the ideal binary mask for a two-source configuration. (a) Male target speech in the median plane. (b) Mixture of target speech and a female utterance on the right side at 30°. (c) The ideal binary mask. (d) The binary mask resulting from our model. The white regions indicate those T-F units dominated by target speech...... 41

xvi
2.10 Systematic results for two-source configuration with 5° azimuth separation. Black bars correspond to the SNR of the initial mixture, white bars indicate the SNR obtained using ideal binary mask, and gray bars show the SNR from our model. Results are obtained for speech mixed with ten types intrusions (Table 2.3) for different spatial configurations. (a) Target at 0°, interference at 5°. (b) Target at 40°, interference at 45°. (c) Target at 80°, interference at 85° ........................................ 45

2.11 Systematic results for two-source configuration with 30° azimuth separation. Black bars correspond to SNR of the initial mixture, white bars to the SNR obtained using an ideal binary mask, and gray bars to the SNR from our model. (a) Target at 0°, interference at 30°. (b) Target at 30°, interference at 60°. (c) Target at 60°, interference at 90° ......................... 47

2.12 Systematic results for two-source configuration using the Cooke corpus as the test corpus. Black bars correspond to SNR of the initial mixture, white bars to the SNR obtained using an ideal binary mask, and gray bars to the SNR from our model. (a) Target at 0°, interference at 5°. (b) Target at 0°, interference at 30° .......................................................... 48

2.13 Evaluation for a three-source configuration. The target is in the median plane and intrusions are at -30° and 30°. Black bars correspond to the SNR of the initial mixture, white bars to the SNR obtained using ideal binary mask, and gray bars to the SNR from our model ........................................ 49

2.14 SNR comparison between the Bodden model (white bars) and our model (gray bars) for a two-source configuration: target in the median plane and interference at 30°. The black bars corresponds to the SNR of the original mixture ................................................................. 50

2.15 Degradation of recognition score with deviations from an ideal binary mask evaluated for three SNR values: 5 dB (square), 0 dB (circle) and –5 dB (diamond). (a) Correctness score. (b) Accuracy score ...................... 54

2.16 Recognition performance at different SNR values for original mixture (dotted line), ideal binary mask (solid line) and estimated mask (dashed line). (a) Correctness score for a two-source case. (b) Accuracy score for a two-source case. (c) Correctness score for a three-source case. (d) Accuracy score for a three-source case ............................................. 56

2.17 Keyword intelligibility score (median values and interquartile ranges) before (white bars) and after processing (black bars) for a two-source condition (0° and 5°) at three SNR values : 0 dB, -5 dB and –10 dB ........ 59
2.18 Keyword intelligibility score (median values and interquartile ranges) before (white bars) and after processing (black bars) for a three-source condition (0°, 30° and -30°) at -10 dB SNR .............................................. 59

3.1 A schematic diagram of the multisource tracking system .................. 71

3.2 Schematic diagram of an HMM for forming continuous source tracks...... 76

3.3 IID reference functions for frequency in the range 80 Hz – 5000 Hz and azimuth in the range [-90°, 90°] ................................................................. 81

3.4 Histogram of estimated ITD and IID deviations from reference values for a channel with \( f_c = 1.5 \text{ kHz} \) in the two-source scenario ............... 83

3.5 Source tracking for two crossing sources with linear motion. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘*’ and ‘o’ tracks correspond to the estimated tracks........... 90

3.6 Source tracking for two crossing sources with nonlinear motion. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘*’ and ‘o’ tracks correspond to the estimated tracks......... 90

3.7 Source tracking for two sources with closely spaced motions. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘*’ and ‘o’ tracks correspond to the estimated tracks.......... 92

3.8 Tracking three non-stationary moving sources. (a) Left ear signal for the first speaker. (b) Left ear signal for the second speaker. (c) Left ear signal for the third speaker. (d) Continuous tracks obtained by applying our model. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘*’, ‘o’ and ‘+’ tracks correspond to the estimated tracks......................................................... 93

3.9 Tracking three non-stationary sources using a Kalman filter approach. (a) Summarized cross-correlation across time. (b) Continuous tracks using the Kalman filter approach. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘o’ tracks correspond to the estimated source locations ......................................................... 95

3.10 Source tracking for three stationary sources. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘*’ and ‘o’ tracks correspond to the estimated tracks........................................... 97
4.1 Schematic diagram of the proposed model. The input signal is a mixture of reverberant target sound and acoustic interference. At the core of the system is an adaptive filter for target cancellation. The output of the system is an estimate of the ideal binary mask ........................................ 103

4.2 Scatter plot of the output-to-input attenuation with respect to the relative strength for a frequency bin centered at 1 kHz. The white line corresponds to the –6 dB decision threshold used in the binary mask estimation .............. 106

4.3 Comparison between the estimated mask and the ideal binary mask for a five-source configuration. (a) Reverberant target speech. (b) Mixture of target speech presented at 0° and four interfering speakers at locations -135°, -45°, 45° and 135°. The SNR is 5 dB. (c) The mixture spectrogram overlaid by the estimated T-F binary mask. (d) The mixture spectrogram overlaid by the ideal binary mask. The recordings correspond to the left ear microphone .................................................................................................................. 109

4.4 Recognition performance for Scene 1 at different SNR values for the reverberant mixture (•), a fixed beamformer (▲), an adaptive beamformer (■), a system that combines target cancellation and spectral subtraction (●), our front end ASR using the estimated binary mask (●), our front-end ASR using the ideal binary mask (•). ......................................................... 121

4.5 Recognition performance for Scene 2 at different SNR values for the reverberant mixture (•), a fixed beamformer (▲), an adaptive beamformer (■), a system that combines target cancellation and spectral subtraction (●), our front end ASR using the estimated binary mask (●), our front-end ASR using the ideal binary mask (•). ......................................................... 122

4.6 Recognition performance for Scene 3 at different SNR values for the reverberant mixture (•), a fixed beamformer (▲), an adaptive beamformer (■), a system that combines target cancellation and spectral subtraction (●), our front end ASR using the estimated binary mask (●), our front-end ASR using the ideal binary mask (•). ......................................................... 123

5.1 Schematic diagram of the proposed two-stage model .................. 135

5.2 Effects of inverse filtering on room impulse responses. (a) A room impulse response for a target source presented in the median plane. (b) The effect of convolving the impulse response in (a) with an estimated inverse filter. (c) A room impulse response for one interfering source at 45° azimuth. (d) The effect of convolving the impulse response in (c) with the estimated inverse filter ................................................................. 141
5.3 Effects of reverberation and target inverse filtering on the harmonic structure of a voiced utterance. (a) Spectrogram of the anechoic signal. (b) Spectrogram of the reverberant signal corresponding to the impulse response in Fig. 5.2(a). (c) Spectrogram of the inverse-filtered signal corresponding to the equalized impulse response in Fig. 5.2(b). (d) Spectrogram of the reverberant signal corresponding to the room impulse response in Fig. 5.2(c). (e) Spectrogram of the inverse filtered signal corresponding to the impulse response in Fig. 5.2(d) .................................................. 143

5.4 Histograms of selected peaks in the high-frequency range (>800 Hz) for a male utterance. (a) Results for the clean signal. (b) Results for the reverberant signal. (c) Results for the inverse filtered signal. The solid lines are the corresponding pitch contours .................................................. 148

5.5 Comparison of pitch tracking in anechoic and reverberant conditions for a male voiced utterance. (a) Spectrogram of the anechoic signal. (b) Spectrogram of the reverberant signal corresponding to the impulse response in Fig. 5.2(a). (c) Pitch tracking results. The solid line indicates the anechoic pitch track. The ‘o’ track indicates the reverberant track ……… 151

5.6 Binary mask estimation for a mixture of target male utterance and interference female speech in reverberant and inverse-filtered conditions. (a) The estimated binary mask on the reverberant mixture. (b) The ideal binary mask for the reverberant condition. (c) The estimated binary mask on the filtered mixture. (d) The ideal binary mask for the inverse-filtered condition. The white regions indicate T-F units that equal 1 and the black regions indicate T-F units that equal 0.......................................................... 153
# LIST OF ACRONYMS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>Amplitude modulation</td>
</tr>
<tr>
<td>ASA</td>
<td>Auditory scene analysis</td>
</tr>
<tr>
<td>ASR</td>
<td>Automatic speech recognition</td>
</tr>
<tr>
<td>CASA</td>
<td>Computational auditory scene analysis</td>
</tr>
<tr>
<td>CMN</td>
<td>Cepstral mean normalization</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier transform</td>
</tr>
<tr>
<td>ERB</td>
<td>Equivalent rectangular bandwidth</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier transform</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov model</td>
</tr>
<tr>
<td>HRIR</td>
<td>Head related impulse response</td>
</tr>
<tr>
<td>HRTF</td>
<td>Head related transfer function</td>
</tr>
<tr>
<td>HTK</td>
<td>HMM toolkit</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent component analysis</td>
</tr>
<tr>
<td>IED</td>
<td>Interaural time difference between signal envelopes</td>
</tr>
<tr>
<td>IID</td>
<td>Interaural intensity difference</td>
</tr>
<tr>
<td>ITD</td>
<td>Interaural time difference</td>
</tr>
<tr>
<td>KEMAR</td>
<td>Knowles electronics manikin for acoustics research</td>
</tr>
<tr>
<td>LMS</td>
<td>Least mean square</td>
</tr>
<tr>
<td>LP</td>
<td>Linear prediction</td>
</tr>
<tr>
<td>LSCV</td>
<td>Least-squares cross-validation</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum a posteriori</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-frequency cepstral coefficient</td>
</tr>
<tr>
<td>MHT</td>
<td>Multiple hypothesis tracking</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood</td>
</tr>
<tr>
<td>OIR</td>
<td>Output-to-input energy ratio</td>
</tr>
<tr>
<td>RSR</td>
<td>Retained speech ratio</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-noise ratio</td>
</tr>
<tr>
<td>SRT</td>
<td>Speech reception threshold</td>
</tr>
<tr>
<td>TCM</td>
<td>Target cancellation module</td>
</tr>
<tr>
<td>T-F</td>
<td>Time-frequency</td>
</tr>
</tbody>
</table>
# LIST OF SYMBOLS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>Channel index</td>
</tr>
<tr>
<td>$f$</td>
<td>Frequency</td>
</tr>
<tr>
<td>$K$</td>
<td>Window length</td>
</tr>
<tr>
<td>$m$</td>
<td>Time frame</td>
</tr>
<tr>
<td>$R$</td>
<td>Relative strength between target source and interference</td>
</tr>
<tr>
<td>$t$</td>
<td>Time index</td>
</tr>
<tr>
<td>$T$</td>
<td>Time period</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Time lag</td>
</tr>
<tr>
<td>$\varphi$</td>
<td>Azimuth</td>
</tr>
<tr>
<td>$\omega$</td>
<td>Radian frequency</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Phase</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>$f_s$</td>
<td>Sampling frequency</td>
</tr>
<tr>
<td>$f_c$</td>
<td>Center frequency for channel $c$</td>
</tr>
<tr>
<td>$b_c$</td>
<td>Bandwidth of a gammatone filter $c$</td>
</tr>
<tr>
<td>$h_c$</td>
<td>Gammatone impulse response for a channel $c$</td>
</tr>
<tr>
<td>$l_c, \bar{l}_c$</td>
<td>The left ear auditory periphery response in a channel $c$ and its short-term average</td>
</tr>
<tr>
<td>$r_c, \bar{r}_c$</td>
<td>The right ear auditory periphery response in a channel $c$ and its short-term average</td>
</tr>
<tr>
<td>$C(c,m,\tau)$</td>
<td>Normalized cross-correlation for a channel $c$ at time frame $m$ and lag $\tau$</td>
</tr>
<tr>
<td>$C(c,m,\varphi)$</td>
<td>Cross-correlogram for a channel $c$ at time frame $m$ and azimuth $\varphi$</td>
</tr>
<tr>
<td>$S(c,m,\varphi)$</td>
<td>Skeleton cross-correlogram for a channel $c$ at time frame $m$ and azimuth $\varphi$</td>
</tr>
<tr>
<td>$p(m,\varphi)$</td>
<td>Pooled cross-correlogram at time frame $m$ and azimuth $\varphi$</td>
</tr>
<tr>
<td>$l(i), r(t)$</td>
<td>Left and right ear responses to a narrow band signal</td>
</tr>
<tr>
<td>$A$</td>
<td>Amplitude</td>
</tr>
<tr>
<td>$d_i$</td>
<td>Interaural time delay for source $i$</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>$H^{l,r}$</td>
<td>The left and the right ear HRTF</td>
</tr>
<tr>
<td>$\Delta\phi$</td>
<td>Sum of phase differences</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Cross-correlation function</td>
</tr>
<tr>
<td>$\tau_{\text{max}}$</td>
<td>Time lag corresponding to the peak in the cross-correlation function and its average</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Phase difference between two narrow band sources</td>
</tr>
<tr>
<td>$M$</td>
<td>Mean IID as derived in Eq. 2.15</td>
</tr>
<tr>
<td>$M_1$</td>
<td>Mean ITD as derived in Eq. 2.13</td>
</tr>
<tr>
<td>$M_2$</td>
<td>Mean ITD as derived in Eq. 2.14</td>
</tr>
<tr>
<td>$\Delta\omega$</td>
<td>Filter bandwidth</td>
</tr>
<tr>
<td>$h_j$</td>
<td>Smoothing parameters for the empirical distribution in Eq. 2.17</td>
</tr>
<tr>
<td>$s_c$</td>
<td>Auditory periphery response to target signal in channel $c$</td>
</tr>
<tr>
<td>$n_c$</td>
<td>Auditory periphery response to noise signal in channel $c$</td>
</tr>
<tr>
<td>$s_T$</td>
<td>Target signal resynthesized using the all-one mask</td>
</tr>
<tr>
<td>$s_E$</td>
<td>Target signal resynthesized using the estimated mask</td>
</tr>
<tr>
<td>$S_0$</td>
<td>Silence state space</td>
</tr>
<tr>
<td>$S_i^1$</td>
<td>State space for a single active source $i$</td>
</tr>
<tr>
<td>$S_{i,j}^2$</td>
<td>State space for two active sources $i$ and $j$</td>
</tr>
<tr>
<td>$S_3$</td>
<td>State space for three active sources</td>
</tr>
<tr>
<td>$s_m$</td>
<td>Hypothesized subspace at time frame $m$</td>
</tr>
<tr>
<td>$\theta(c)$</td>
<td>Frequency-dependent threshold on the peak height in the cross-correlation function for the selection of reliable channels</td>
</tr>
<tr>
<td>$\tau_{\text{ref}}$</td>
<td>Reference ITD value</td>
</tr>
<tr>
<td>$l_{\text{ref}}$</td>
<td>Reference IID value</td>
</tr>
<tr>
<td>$\iota_c$</td>
<td>Estimated IID in channel $c$</td>
</tr>
<tr>
<td>$T_c$</td>
<td>Set of all possible time lags in channel $c$</td>
</tr>
<tr>
<td>$\delta_t$</td>
<td>ITD deviation</td>
</tr>
<tr>
<td>$\delta_i$</td>
<td>IID deviation</td>
</tr>
<tr>
<td>$Q$</td>
<td>Noise level used in modeling the distribution of the ITD-IID space</td>
</tr>
<tr>
<td>$\Delta_t$</td>
<td>Maximum plausible ITD deviation</td>
</tr>
<tr>
<td>$\Delta_i$</td>
<td>Maximum plausible IID deviation</td>
</tr>
<tr>
<td>$L(\delta, \lambda)$</td>
<td>Laplacian distribution with parameter $\lambda$</td>
</tr>
<tr>
<td>$\lambda_c(c)$</td>
<td>Lambda parameter for $\delta_t$</td>
</tr>
<tr>
<td>$\lambda_i(c)$</td>
<td>Lambda parameter for $\delta_i$</td>
</tr>
<tr>
<td>$a_1, a_2, a_3, a_4$</td>
<td>Parameters used in modeling $\lambda_c(c)$ and $\lambda_i(c)$</td>
</tr>
<tr>
<td>$U$</td>
<td>Uniform distribution</td>
</tr>
<tr>
<td>$N(\mu, \sigma)$</td>
<td>Gaussian distribution with mean $\mu$ and variance $\sigma$</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>$N_b$</td>
<td>Smoothing factor for the observation probability in a time frame conditioned on the target space</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>Gate size for the channel gating mechanism</td>
</tr>
<tr>
<td>$\alpha_0, \alpha_2, \alpha_3$</td>
<td>Weighting parameters for the target subspaces with zero-, two- and three-sources</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>Normalization factor for the observation probability conditioned on the target space</td>
</tr>
<tr>
<td>$h_i$</td>
<td>Room impulse response corresponding to target source at the $i$'th microphone</td>
</tr>
<tr>
<td>$\theta_{\text{DIR}}(f)$</td>
<td>Threshold on the output-to-input energy ratio at frequency bin $f$</td>
</tr>
<tr>
<td>$n_E$</td>
<td>Noise signal resynthesized using the estimated mask</td>
</tr>
<tr>
<td>$s_{\text{IBM}}$</td>
<td>Target signal resynthesized using the ideal binary mask</td>
</tr>
<tr>
<td>$X_r, X_u$</td>
<td>The reliable and unreliable components of a spectral vector $X$</td>
</tr>
<tr>
<td>$h(t)$</td>
<td>Room impulse response corresponding to the target source</td>
</tr>
<tr>
<td>$\tilde{z}$</td>
<td>The inverse-filtered LP residual of reverberant speech</td>
</tr>
<tr>
<td>$q$</td>
<td>Inverse filter of the target room impulse response</td>
</tr>
<tr>
<td>$L$</td>
<td>Length of the inverse filter</td>
</tr>
<tr>
<td>$y_r$</td>
<td>LP residual of reverberant speech</td>
</tr>
<tr>
<td>$J$</td>
<td>Kurtosis function</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Update rate used in the inverse filtering algorithm</td>
</tr>
<tr>
<td>$E$</td>
<td>Teager energy operator</td>
</tr>
<tr>
<td>$p$</td>
<td>Target pitch in the pitch-based segregation stage</td>
</tr>
<tr>
<td>$l$</td>
<td>Time lag in the autocorrelation function</td>
</tr>
<tr>
<td>$A(c,m,\tau)$</td>
<td>Autocorrelation on the gammatone response at channel $c$, time frame $m$ and lag $\tau$</td>
</tr>
<tr>
<td>$A_E(c,m,\tau)$</td>
<td>Autocorrelation on the envelope of the gammatone response at channel $c$, time frame $m$ and lag $\tau$</td>
</tr>
<tr>
<td>$C(c,m)$</td>
<td>Cross-channel correlation on the gammatone response at channel $c$ and time frame $m$</td>
</tr>
<tr>
<td>$C_E(c,m)$</td>
<td>Cross-channel correlation on the envelope of the gammatone response at channel $c$ and time frame $m$</td>
</tr>
<tr>
<td>$\theta_L$</td>
<td>Threshold for discarding T-F units that disagree with target pitch at low frequencies</td>
</tr>
<tr>
<td>$\Delta_L$</td>
<td>Threshold for discarding T-F units that disagree with target pitch at high frequencies</td>
</tr>
<tr>
<td>$\theta_p$</td>
<td>Threshold on the peak ratio for selecting a T-F unit at low frequencies</td>
</tr>
<tr>
<td>$\theta_C$</td>
<td>Threshold on the cross-channel correlation used in segment formation</td>
</tr>
<tr>
<td>$F_0$</td>
<td>Fundamental frequency</td>
</tr>
<tr>
<td>$T_{60}$</td>
<td>Reverberation time measure; equal to the time required for the sound level to drop by 60 dB following the sound offset</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Motivation

For us it is relatively easy in most cases to filter out and comprehend a multitude of acoustic events that surrounds us in every moment. Imagine, for example, a cocktail party where we hear multiple voices, some background music and other environmental sounds at the same time. In this case, each acoustic source generates a vibration of the medium (air) and our hearing is determined by the superposition of all these vibrations impinging on the eardrums. As Helmholtz noted in 1863, the final waveform is “complicated beyond conception” (Helmholtz, 1863). Nonetheless, we are able to attend to and understand one particular talker in this situation. This perceptual ability is known as the “cocktail-party effect” – a term introduced by Cherry in 1953 (Cherry, 1953). Cherry’s original experiments have triggered research in widely different areas including speech perception in noise, selective attention, neural modeling, speech enhancement and source separation. The focus of this dissertation is on computational models for speech separation inspired by principles of auditory perception.
A solution to the general problem of sound separation in realistic environments is essential to many important applications including automatic speech and speaker recognition, hearing aid design and audio information retrieval. The field of speech and speaker recognition has seen much progress in recent years. However, the performance of current recognition systems degrades rapidly in the presence of noise and reverberation and the degradation is much faster compared to human performance in similar conditions (Lippmann, 1997; Huang et al., 2001). Hearing impairment affects about 10% of the total adult population and at least one third of the elderly population. To alleviate hearing loss, hearing aids are currently the primary choice. However, the user’s satisfaction with their hearing devices varies greatly with the situations in which they are used. According to a survey from 2000, while 90% of hearing aid users are satisfied in one-to-one conversations, only 25% of them are satisfied in larger groups (Kochkin, 2000). Noise reduction strategies that can deal with complex environments are therefore necessary in modern hearing aids (Dillon, 2001; Levitt, 2001). Finally, an important emerging application is information retrieval from archives of real audio recordings. Since the recordings usually contain mixtures of acoustic sources, separating these mixtures into the original sources can contribute to efficient segmentation, labeling and retrieval of audio information.

The sound separation problem has been investigated in the signal processing field for many years for both one-microphone recordings as well as multi-microphone ones (for recent reviews see Divenyi, 2005; Brandstein and Ward, 2001). One-microphone speech enhancement techniques include spectral subtraction (e.g., Martin, 2001), Kalman
filtering (Ma et al., 2004), subspace analysis (Ephraim and Trees, 1995) and autoregressive modeling (e.g., Balan et al., 1999). While having the advantage of requiring only one sensor, these algorithms make strong assumptions about interference and thus have difficulty in dealing with general acoustic mixtures. Microphone array algorithms are divided in two broad categories: beamforming and independent component analysis (ICA) (Brandstein and Ward, 2001). To separate multiple sound sources, beamforming takes advantage of their different directions of arrival while ICA relies on their statistical independence. A fixed beamformer, such as that of the delay-and-sum, constructs a spatial beam to enhance signals arriving from the target direction independent of the interfering sources. The primary limitations of a fixed beamformer are: 1) a poor spatial resolution at lower frequencies, i.e., the spatial response has a wide main lobe when the intermicrophone distance is smaller than the signal wavelength; and 2) spatial aliasing, i.e., multiple beams at higher frequencies when the intermicrophone distance is greater than the signal wavelength. To solve these problems a large number of microphones is required and constraints need to be introduced in order to impose a constant beam shape across the frequencies (Ward et al., 2001). Adaptive beamforming techniques, on the other hand, attempt to null out the interfering sources in the mixture (Griffiths and Jim, 1982; Widrow and Stearns, 1985; Van Compernolle, 1990). While they improve resolution significantly, the main disadvantage of such a beamformer is greater computation and adaptation time when the locations of interfering sources change. Note also that while an adaptive beamformer with two microphones is optimal for canceling a single directional interference, more microphones are required as the
number of noise sources increases (Weiss, 1987). A subband adaptive algorithm has been proposed by Liu et al. (2001) to address the multi-source problem. Their two-microphone system estimates the locations of all the interfering sources and uses them to steer independent nulls that suppress the strongest interference in each T-F unit. The underlying signal model is, however, anechoic and performance degrades in reverberant conditions. Similarly, the drawbacks of ICA techniques include the requirement that the number of microphones is greater than or equal to the number of sources and the poor performance in reverberant conditions (Hyvärinen et al., 2001). Some recent sparse representations attempt to relax this assumption (e.g., Zibulevsky et al., 2001) but the performance is limited.

As stated previously, this dissertation aims at a biologically relevant solution to the cocktail-party problem. Therefore, the scope of investigation will be limited to monaural (one-microphone) and binaural (two-microphone) recordings of the auditory scene. Extensive work has been undertaken to evaluate human speech intelligibility in noise for monaural conditions (Miller, 1947; Steeneken, 1992) as well as binaural conditions (Bronkhorst and Plomp, 1992). The results are quantified using the speech reception threshold which is the required signal-to-noise ratio (SNR) measured in decibels (dB) to yield a 50% intelligibility score. The results are highly dependent on the type of speech material as well as the type of noise material that has been used during testing. For broadband noises, the monaural SRT can be as low as -10 dB as, for example, in the case of a single-voice interference. The SRT rises quickly to -2 dB for a two-voice interference. In addition, when target and intrusions are presented at different spatial
locations binaural hearing can lead to a significant reduction in SRT of up to 10 dB (Bronkhorst and Plomp, 1992). This binaural advantage is the sum of two contributions: 1) the head shadow, i.e., the difference in SNR at the two ears; and 2) the ability to utilize the interaural differences at the two ears. The speech intelligibility experiments described above are conducted in anechoic conditions. Several studies explored the effect of reverberation on speech intelligibility in noisy conditions (Plomp, 1976; Culling et al., 2003). For a 0.4 s reverberation time, the study of Culling et al. (2003) has shown an increase of 5 dB for the monaural SRT compared to the anechoic conditions. However, note that the SRT reported in reverberation is below -10 dB. Taken together, these results clearly show the extraordinary ability of the auditory system to separate and recognize speech in realistic environments.

Inspired by this perceptual ability, research has been devoted to understand the mechanisms underlying the “cocktail-party problem”. A first coherent theory on the human ability to segregate signals from noisy mixtures was presented by Bregman in 1990. He argues that humans perform an auditory scene analysis (ASA) of the acoustic input in order to form perceptual representations of individual sources called streams. ASA takes place in two stages: the first stage decomposes the input into a collection of sensory elements while the second stage selectively groups the elements into streams that correspond to individual sound sources. According to Bregman, stream segregation is guided by a variety of grouping cues including proximity in frequency and time, pitch, onset/offset, and spatial location.
This ASA account has inspired a series of computational ASA (CASA) systems that have significantly advanced the state-of-the-art performance in monaural separation as well as binaural separation. A recent overview of CASA approaches can be found in Brown and Wang (2005). Mirroring the ASA processing described above, CASA systems generally employ two stages: segmentation (analysis) and grouping (synthesis). In segmentation, the acoustic input is decomposed into sensory segments (time-frequency regions), each of which supposedly originates from a single source. In grouping, the segments that are likely to come from the same source are put together. Monaural separation algorithms rely primarily on the pitch cue and therefore operate only on voiced speech. On the other hand, the binaural algorithms use the source location cues - time differences and intensity differences between the ears – which are independent of the signal content and thus can be used to track both voiced and unvoiced speech. Compared with the signal processing techniques described above, CASA systems make generally few assumptions about the acoustic properties of the interference and the environment.

1.2 Objectives

In the preceding section, we have described the “cocktail-party problem” and outlined our general approach to solve this problem. The final goal is to design a speech separation system that is able to extract a target speech signal from an acoustic mixture recorded in a real-world environment. We are interested in solutions inspired by the principles of human hearing and consequently investigate monaural and binaural models for ASA.
According to Marr’s framework for information processing (Marr, 1982), before delving into algorithmic and implementation aspects one needs to define the computational goal. At the core of many CASA systems is a time-frequency (T-F) mask. Specifically, the T-F units in the acoustic mixture are selectively weighted in order to enhance the desired signal. The weights can be binary or real (Srinivasan et al., 2004). The binary T-F masks are motivated by the masking phenomenon in human audition, in which a weaker signal is masked by a stronger one when they are presented in the same critical band (Moore, 2003). Additionally, from the speech segregation perspective, the notion of an ideal binary mask has been proposed as the computational goal of CASA (Roman et al., 2003; Wang, 2005). Such a mask can be constructed from a priori knowledge about target and interference; specifically a value of 1 in the mask indicates that the target is stronger than the interference in the corresponding T-F unit and 0 indicates otherwise. Speech reconstructed from ideal binary masks has been shown to be highly intelligible even when extracted from multi-source mixtures and also to produce substantial improvements in robust speech recognition (Cooke et al., 2001; Roman et al., 2003; Chang, 2004; Brungart et al., 2005). Hence, our computational goal is to estimate an ideal binary mask in order to extract the target speech from the acoustic mixture.

Multiple noise sources are simultaneously active in our acoustic environment. Hence, target separation from multisource interferences is indispensable for many speech applications operating in noisy environments. As seen in the previous section, binaural hearing provides an advantage in separating signals in multisource acoustic scenarios; this motivates our exploration of binaural segregation models. Note that a sound source
produces location-dependent interaural time differences (ITD) and interaural intensity
differences (IID) between the left and the right ears. These are the main cues that human
listeners use to determine the location of a sound source. How to efficiently utilize the
ITD and IID information when multiple sources are active at the same time is a
challenging problem. In addition, most binaural CASA systems assume that sound source
positions remain fixed for the duration of the recording. In real-world environments,
however, a flexible CASA system will require dealing with moving sound sources and/or
head movements.

Another challenging problem that CASA systems face in real environments is
reverberation which occurs almost everywhere except in an anechoic chamber. In an
enclosure, for example, an acoustic wave encountering a boundary loses part of the
energy and reflects the rest of the energy, a process that repeats itself until the energy of
the wave fades away. Consequently, a reverberated signal is the sum of time-delayed and
attenuated replicas of the same original signal. The reflections adding up at the eardrums
at a given time differ in their locations and their spectral content. As a result,
reverberation is smearing many of the acoustic cues used in segregation including
binaural cues and the harmonicity of the signals. These distortions affect both binaural
CASA systems as well as monaural ones.

Motivated by the challenges presented above, we address in this dissertation the
following aspects of CASA processing. First, we investigate binaural sound segregation
in multisource scenarios. Two objectives are established: to investigate the effect of a
broadband multisource interference on the binaural cues and to propose an algorithm that
operates efficiently in anechoic conditions with stationary sources. Second, we study the problem of tracking multiple moving sources with a binaural input. The goal is to search for a general framework for source localization and tracking and to apply this framework for the tracking of multiple speakers. Third, we extend our binaural investigation to reverberant scenarios. Here, the goal is to identify a computational model that can handle arbitrary noise sources in reverberation using only two microphones. However, speech intelligibility experiments have shown that monaural cues are more resilient to reverberation compared to the binaural ones (Culling et al., 2003). Hence our final investigation considers the problem of monaural speech segregation in reverberant environments. Our aim is to incorporate knowledge of target source location in a monaural framework thereby improving the performance of existing pitch-based CASA systems in realistic environments.

1.3 Organization of Dissertation

The rest of the dissertation is organized as follows. Chapter 2 presents our study on location-based speech segregation. We first provide a mathematical analysis of the ITD and IID cues in the context of pure tone interactions. The insights we gain from this analysis are then tested on the binaural cues extracted at the output of an auditory filterbank for a multiple speaker scenario. The main observation is that modifications to the relative strength of the target source with respect to the interference results in systematic changes for the ITD and IID cues in the narrow frequency bands. As a result,
the interaction between multiple sources produces a characteristic clustering in the ITD-IID space that can be used to estimate an ideal binary mask. We propose a pattern classification algorithm that generates masks very close to the ideal ones as determined using three different evaluation criteria: SNR, automatic speech recognition (ASR) accuracy and subjective listening tests.

In Chapter 3, we study the binaural tracking of multiple moving sources. We extend the observations on ITD and IID from the previous chapter and propose an HMM framework to detect and track individual sources across time. Central to the proposed system is the computation for the probability of the observed data conditioned on a source configuration which integrates, across different channels, the statistical distributions of ITD and IID.

In Chapter 4, we study sound segregation in reverberant environments with two-microphones. We propose a flexible figure-ground segregation framework that can deal with an arbitrary (possibly moving) number of noise sources. The system estimates an ideal binary mask at the output of a target cancellation module implemented using adaptive filtering. Systematic evaluations using both an SNR measure as well as ASR accuracy show large improvements over the baseline performance and better results over related two-microphone approaches.

In Chapter 5, we study monaural segregation in reverberant environments. We propose a two-stage model that combines inverse filtering with respect to target location and pitch-based speech segregation. A filter that aims at inverting a specific target location will smear room impulse responses corresponding to other locations. As a result,
the harmonicity of the target source is enhanced while interfering sources are further smeared. The proposed system shows considerable SNR improvements across different conditions.

Chapter 6 summarizes the contributions presented in this dissertation, discusses the insights gained from my doctoral research, and outlines future research directions.
CHAPTER 2

SPEECH SEGREGATION BASED ON SOUND LOCALIZATION

In this chapter we describe a novel supervised learning approach to speech segregation, in which a target speech signal is separated from interfering sounds using the spatial cues of ITD and IID. We observe that within a narrow frequency band, modifications to the relative strength of the target source with respect to the interference trigger systematic changes for estimated ITD and IID. For a given spatial configuration, this interaction produces characteristic clustering in the binaural feature space. Consequently, we perform pattern classification in order to estimate ideal binary masks. A systematic evaluation in terms of signal-to-noise ratio as well as automatic speech recognition performance shows that the resulting system produces masks very close to ideal binary ones. A quantitative comparison shows that our model yields significant improvement in performance over an existing approach. Furthermore, under certain conditions the model produces large speech intelligibility improvements with normal listeners. The work presented in this chapter has been published in the Journal of the Acoustical Society of America (Roman et al., 2003).
2.1 Introduction

It is widely acknowledged that for human audition, ITD is the main localization cue used at low frequencies (< 1.5 kHz), whereas in the high-frequency range both IID and interaural time differences between the envelopes of the signals (IED) are used (Blauert, 1997). The resolution of the binaural cues has implications for both localization and recognition tasks. Headphone experiments show that listeners can reliably detect 10-15 μs ITDs from the median plane, which correspond to a difference in azimuth of between 1 and 5 degrees. On the other hand, the smallest detectable change in IID by the human auditory system is about 0.5 dB to 1 dB at all frequencies. Resolution deteriorates as the reference ITD gets larger, and the difference limen can be as much as 10 degrees when the ITD corresponds to a source located far to the side of the head (Blauert, 1997).

Classical models for processing binaural cues compare the acoustic signals at the two ears, although they explain the binaural interaction through different mechanisms. These include extensions of the Jeffress coincidence model (Jeffress, 1948; Lindemann 1986; Gaik 1993), the equalization and cancellation theory (Durlach, 1972; Breebaart et al., 2001) and auditory nerve based models (Colburn, 1977; Stern and Colburn, 1978). The goal of this line of research is to explain experimental data for a number of psychoacoustical phenomena including lateralization, binaural masking levels, and the precedence effect (for a review see Stern and Trahiotis, 1995).

Increased speech intelligibility in binaural listening compared to the monaural case has also prompted research in designing cocktail-party processors based on
psychoacoustic principles (Lyon, 1983; Slatky, 1993; Bodden, 1993; Liu et al., 2001; Wittkop and Hohmann, 2003). Most cocktail-party-processor designs utilize the following observation: as the relative strength of the interference with respect to the target increases, certain attributes of the auditory event including location and spatial extent change systematically compared to the case of the target source alone. In particular, building on a previous cross-correlation model for sound localization, Bodden (1993) proposed a model that estimates optimal time-varying Wiener coefficients for all critical bands by comparing desired cross-correlation patterns to observed ones. Bodden’s model has shown that psychoacoustically motivated auditory mechanisms can produce substantial enhancement in speech intelligibility (Bodden, 1996).

In this chapter, we propose a sound segregation model using binaural cues extracted from the responses of a KEMAR dummy head that realistically simulates the filtering process of the head, torso and external ear (Burkhard and Sachs, 1975). Such a model can be applied to, among other things, enhancing speech recognition in noisy environments and improving binaural hearing aid design. As discussed in Chapter 1, our computational goal is an ideal binary mask. From a theoretical ASA perspective, an ideal binary mask gives a performance ceiling for all binary masks. Note that an ideal mask remains well-defined for situations when more than one target needs to be segregated. Moreover, such masks generate high quality reconstruction for a variety of signals, and have been shown to provide a highly effective front-end for robust speech recognition (Cooke et al., 2001). Furthermore, as will be shown later, deviations from ideal binary masks lead to gradual degradation in speech recognition performance. Hence, our model aims to estimate an
ideal binary mask using information about the spatial configuration of sound sources.

Statistics for the relationship between the relative strength of sources and the pattern of binaural cues are at the core of our system. We show for mixtures of multiple sound sources that there exists a strong correlation between the relative strength of target and interference and estimated ITD/IID, resulting in a characteristic clustering across frequency bands. Our aim is to maximize the performance of the system independently for different spatial configurations. Consequently, we employ a nonparametric classification method to determine decision regions in the joint ITD-IID feature space that corresponds to an optimal estimate for an ideal binary mask. An objective evaluation of the system with both SNR and ASR (automatic speech recognition) measures shows that the results of our system are comparable with those obtained using ideal binary masks. In addition, a speech intelligibility evaluation using normal listeners shows a large improvement under certain conditions.

The rest of the chapter is organized as follows. Section 2.2 contains an overview of the model. Section 2.3 describes the peripheral auditory model. Section 2.4 describes the azimuth localization algorithm. Section 2.5 is mainly devoted to the ideal binary mask estimation, which constitutes the core of the model. Section 2.6 presents the evaluation results of the system and a quantitative comparison with an existing binaural processor. In Section 2.7 we discuss related issues and conclude the chapter.
Figure 2.1. Schematic diagram of the model. Binaural signals are obtained by convolving input signals with measured head related impulse responses (HRIR) from a KEMAR dummy head. A model of the auditory periphery is employed. Azimuth localization for all the sources is based on a cross-correlation mechanism. ITD and IID are computed independently for different frequency channels. A pattern analysis block produces an estimation of an ideal binary mask, which enables the reconstruction of the target signal and the interfering sound.

2.2 Model Architecture

Our model consists of the following four stages: 1) a model of the auditory periphery; 2) binaural cue extraction and azimuth localization for both target and interference based on a cross-correlation mechanism; 3) estimation of an ideal binary mask; and 4) reconstruction of the target signal. Figure 2.1 illustrates the model architecture for the case of two sound sources.

The input to our model is a mixture of two or more signals at different, but fixed, locations: target speech and acoustic interference. Measurements of head-related transfer
functions (HRTF) are a standard method for realistic binaural synthesis. We utilize here a catalogue of HRTF measurements collected by Gardner and Martin (1994) from a KEMAR dummy head under anechoic conditions. The measurements consist of left/right KEMAR responses from a distance of 1.4 m in the horizontal plane, resulting in 128 point impulse responses at a sampling rate of 44.1 kHz. Binaural signals are obtained by filtering monaural signals with HRTFs corresponding to the direction of incidence. The responses to multiple sources are added at each ear. HRTFs introduce a natural combination of ITD and IID into the signals that is extracted by subsequent stages of our model.

The auditory periphery is simulated using a filterbank that models the cochlear filtering mechanism. In addition, the gains of the filters are adjusted to account for middle ear transfer, which is direction-independent. The output of each filter is processed using a simple model for hair-cell transduction, which performs half-wave rectification and square root compression. The output of the model gives a firing rate representation of auditory nerve activity.

Simulated auditory nerve responses from both ears are evaluated independently for all frequency bands in order to extract interaural differences. The most common method to determine ITD is cross-correlation of the corresponding left and right signals within individual frequency bands, which is calculated for time lags equally distributed in the plausible range. Our localization stage uses only ITD information. Consequently, the system cannot tell front from back. We restrict our model to the half-horizontal plane with azimuth in the range [-90°, 90°]. Due to some diffraction effects, a frequency
dependent nonlinear transformation from the time lag axis to the azimuth axis is necessary. The set of cross-correlations for all frequency bands and at all times results in a 3D structure called the “cross-correlogram”; where the coordinates are given by frequency, azimuth, time. A cross-correlogram is further evaluated to extract spatial information. Assuming fixed sources, the source locations are obtained as the positions of the maxima in a pooled cross-correlogram (Shackelton et al., 1992) – obtained by integrating the cross-correlogram across time and frequency. Further stages of our model use this spatial information: the number of sources, their locations and the target source location.

At the core of our system are decision rules that determine whether the target source is stronger than the interference in individual T-F units. The system is based on observed characteristic clustering of extracted ITD and IID features. The novelty of our approach lies in the introduction of supervised learning for different spatial configurations and across all frequency bands in a joint ITD-IID feature space. For a given frequency channel and a stimulus configuration, conditional probabilities are estimated from samples of ITD, IID and the corresponding relative strength based on a corpus of training data. Therefore, auditory grouping is implemented based on proximity in the ITD-IID space. The output of this pattern analysis is a time-frequency mask, which is an estimate of an ideal binary mask. The time-frequency resolution for the current implementation is 20-ms time frames with a 10-ms frame shift (see e.g. Wang and Brown, 1999), and 128 frequency channels that cover the range of 80 Hz to 5 kHz.
The last stage of the model is a reconstruction path, which allows the target signal to be recovered from the acoustic mixture by nullifying the T-F units dominated by interference. The method employed here is the same in principle to that described by Weintraub (1986) (see also Brown and Cooke, 1994). The target signal is reconstructed from the output of the gammatone filterbank. To remove across-channel phase differences, the output of a filter is time-reversed, passed through the gammatone filter and time-reversed again. Furthermore, the output for each filter is divided in 20-ms sections with 10-ms overlap that correspond to T-F units in the binary mask, and windowed with a raised cosine. Binary weights estimated in the previous stage are then applied to each section to remove the interference. This method achieves high quality reconstruction (Weintraub, 1986; Brown and Cooke, 1994; Wang and Brown, 1999).

2.3 Auditory Periphery

It is widely acknowledged that cochlear filtering can be modeled by a bank of bandpass filters. The filterbank employed here consists of 128 fourth-order gammatone filters (Patterson et al., 1988) following an implementation by Cooke (1993). The impulse response for filter channel $c$ has the following form:

$$h_c(t) = \begin{cases} t^{N-1} \exp(-2\pi b_c t) \cos(2\pi f_c t + \phi_c), & \text{if } t \geq 0 \\ 0, & \text{else} \end{cases}, \quad (2.1)$$
where the order of the filter is \( N = 4 \), \( b_c \) is the decay rate of the impulse response which is related to the bandwidth of the filter, \( f_c \) is the center frequency of the filter, and \( \phi_c \) is the phase (here we set \( \phi_c \) to zero).

The equivalent rectangular bandwidth (ERB) scale is a psychoacoustic measure of auditory filter bandwidth. The center frequencies \( f_c \) are equally distributed on the ERB scale between 80 Hz and 5 kHz, and specifically for each filter we set the bandwidth according to the following equations (Glasberg and Moore, 1990):

\[
ERB(f_c) = 24.7 \left( 4.37 f_c / 1000 + 1 \right), \quad (2.2)
\]

\[
b_c = 1.019 ERB(f_c). \quad (2.3)
\]

Since the HRTF reflects the filtering effects due to pinna and meatus but not the middle ear we adjust the gains of the gammatone filters in order to simulate the middle ear transfer function; such data is provided by Moore et al. (1997). We include this middle ear processing for the purpose of physiological plausibility. In the final step of the peripheral model, the output of each gammatone filter is half-wave rectified in order to simulate firing rates of the auditory nerve. Saturation effects are modeled by taking the square root of the rectified signal.

Psychophysical models for sound localization generally employ envelopes of the responses in the high-frequency range. This is supported by discrimination experiments using transposed stimuli, suggesting similar sensitivity to ITD for both low and high
frequency ranges (Bernstein and Trahiotis, 2001). Therefore, we additionally extract the envelopes using the Hilbert transform for channels with center frequencies above 1.5 kHz. Note that the envelope is not actually used in our current implementation; rather it is used in Section 2.5.2 as part of a comparison of the effectiveness of different interaural cues.

2.4 Azimuth Localization

Current models of azimuth localization almost invariably employ cross-correlation, which is functionally equivalent to the coincidence detection mechanism proposed by Jeffress (1948). Cross-correlation provides excellent time delay estimation for broadband stimuli, and for narrowband stimuli in the low-frequency range. However, for high-frequency narrowband signals it produces multiple ambiguous peaks. Here we use the normalized cross-correlation computed at lags equally distributed from –1 ms to 1 ms (−44 < τ < 44) using a rectangular integration window of 20 ms (corresponding to K = 880 samples below). The cross-correlation is computed for all frequency channels and updated every 10 ms, according to the following formula for frequency channel \( c \), time frame \( m \) and lag \( \tau \):

\[
C(c, m, \tau) = \frac{\sum_{k=0}^{K-1} (l_c(m-k) - \bar{l}_c)(r_c(m-k-\tau) - \bar{r}_c)}{\left(\sum_{k=0}^{K-1} (l_c(m-k) - \bar{l}_c)^2\right)^{1/2} \left(\sum_{k=0}^{K-1} (r_c(m-k-\tau) - \bar{r}_c)^2\right)^{1/2}},
\]  

(2.4)
Figure 2.2. Functions relating azimuth to ITD for three auditory channels with center frequencies of 500 Hz, 1 kHz, and 3 kHz.

where $l_c$, $r_c$ refer to the left and right auditory periphery output for channel $c$, and $\overline{l_c}$, $\overline{r_c}$ refer to their mean values estimated over the integration window.

For each frequency channel, ITD is estimated as the lag corresponding to the position of the maximum in the cross-correlation function. Diffraction effects introduce weak frequency dependences for ITD (MacPherson, 1991). As a result, we derive frequency-dependent nonlinear transformations to map the time-delay axis onto the azimuth axis, resulting in a cross-correlogram $C(c, m, \varphi)$ where $\varphi$ denotes azimuth. The mappings are obtained based on the cross-correlation output in response to white noise presented systematically at locations in the azimuth range $[-90^\circ, 90^\circ]$. Figure 2.2 shows three ITD-azimuth mappings, for channels with center frequencies of 500 Hz, 1 kHz, 3 kHz. The
functions are monotonic, being sigmoidal at low frequencies where diffraction effects are greater and increasingly linear at high frequencies.

In addition, a ‘skeleton’ \( S(c, m, \varphi) \) is formed by replacing the peaks in the cross-correlogram with Gaussians whose widths are narrower than the original peaks. That is, each local peak generates an impulse of the same height and then the obtained impulse train is convolved with a Gaussian. Here the width is linear with respect to the center frequency of the channel. This technique sharpens the cross-correlogram, an effect similar to a lateral inhibition mechanism (Arbib, 2003).

The cross-correlation method provides inconsistent results when two acoustic sources are present. Figure 2.3 shows the cross-correlation functions (Fig. 2.3(a)) and the skeleton cross-correlogram (Fig. 2.3(b)) for a mixture of male speech presented at 0° and female speech presented at 20°. Here, the width of the Gaussians in the skeleton cross-correlogram ranges from 4° at the low-frequency end to 2° at the high-frequency end. For frequency channels where one source is much stronger, activity is observed near the true location of that source. For T-F units where the two sources overlap the peak deviates, generally being closer to the more intense source. Peaks at both locations can occur in high-frequency channels – this ambiguity is due to the periodicity of the cross-correlation function. Hence, if little overlapping occurs for a sufficient number of channels a good estimate of the two source locations can be obtained at every time frame by pooling the cross-correlogram across all frequency channels. At time frame \( m \) and azimuth \( \varphi \), this yields the following pooled cross-correlogram:
Figure 2.3. Azimuth localization for a mixture of male utterance at 0° and female utterance at 20°. The bottom plot in each panel shows a summation across all rows. (a) Cross-correlation functions for 128 frequency channels in the range 80 Hz – 5 kHz at time frame 40 (i.e. 400 ms after the start of the stimulus). For clarity, only every other channel is shown, resulting in 64 channels. (b) Skeleton cross-correlogram for the same time frame. The arrow indicates channels that contain roughly equal energy from both target and interference. (c) Pooled cross-correlogram for a stimulus of duration 1.5 seconds, shown every 20 ms.
Improved localization results are obtained using the skeleton cross-correlogram proposed here over the standard cross-correlation. Summing across frequencies produces sharper peaks on the skeleton cross-correlogram; in the case of Fig. 2.3, the skeleton cross-correlogram gives a good estimate of source locations whereas the conventional cross-correlogram does not (compare the bottom plots in Fig. 2.3(a) and Fig. 2.3(b)). In Fig. 2.3(c) we display the pooled cross-correlogram for a signal of duration 150 frames (i.e., 1.5 seconds). Peaks in the pooled cross-correlogram indicate the locations of active sources at every frame. Assuming fixed sources, multiple locations can be reliably determined by further summating the pooled cross-correlogram across time as shown in the bottom plot of Fig. 2.3(c). This represents our method for azimuth localization.

2.5 Ideal Mask Estimation

The objective of this stage of the model is to develop an efficient mechanism for estimating an ideal binary mask, which selects the T-F units where the estimated signal energy is greater than the noise energy (i.e., greater than 0 dB SNR). Note that different SNR criteria are possible for defining an ideal binary mask (see Cooke et al., 2001). In the absence of evidence for a better SNR measure, we choose the 0 dB criterion for simplicity. We propose an estimation method based on the following observation regarding the auditory interaction of multiple sources. In a narrow band, the ITD and IID
corresponding to the target source exhibit azimuth-dependent characteristic values. As
the interference from additional sound sources increases, ITD and IID systematically shift
away from these values. Consequently, in a local T-F unit both binaural cues can be
potentially used to determine whether the target signal dominates.

In what follows, we analyze this phenomenon for the case of pure tones (see Slatky,
1993, for an extensive study of binaural cues with sinusoidal signals). Although in real-
world scenarios the conditions of this simplified model are generally not fulfilled, our
experimental results show that a similar trend holds for a variety of natural signals when
analyzed in narrow frequency bands. This analysis also serves to motivate the
introduction of our proposed algorithm for the general case in Section 2.5.2.

2.5.1 Pure Tones

We consider a simple model of two sources emitting pure tones in a narrow band. In
this case, the left-ear and the right-ear responses are given by:

\[
\begin{align*}
    l(t) &= |H_l^1(\omega)| A_1 \sin(\omega_1 t) + |H_l^2(\omega)| A_2 \sin(\omega_2 t + \Delta \phi) \\
    r(t) &= |H_r^1(\omega)| A_1 \sin(\omega_1 t + \omega_1 d_1) + |H_r^2(\omega)| A_2 \sin(\omega_2 t + \omega_2 d_2 + \Delta \phi)
\end{align*}
\] (2.6)

where \(A_i\) is the amplitude, \(\omega_i\) is the frequency, \(d_i\) corresponds to the interaural time
delay (equivalent to the phase difference between left and right HRTFs at frequency \(\omega_i\)),
and \(H_l^i(\omega)\) and \(H_r^i(\omega)\) represent respectively the right and left HRTF, for the \(i\)th
source. $\Delta \phi$ is the sum of phase differences between the initial signals and those due to the arrival times of the signals at the left ear.

To simplify, we neglect the magnitude of the HRTF response in analyzing ITD, which represents a reasonable assumption only in a narrowband low-frequency range. The cross-correlation function $\gamma(\tau)$ for infinite-duration signals is obtained by:

$$\gamma(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} r(t)l(t+\tau)dt.$$  \hspace{1cm} (2.7)

Observe that in approximating the cross-correlation function in a finite duration, there exists a tradeoff between the difference in frequency $|\omega_1 - \omega_2|$ and the total integration time. Therefore, we study the cross-correlation under the following two conditions:

**Case 1**: $\omega_1 = \omega_2 = \omega$

In this case, we have:

$$\gamma(\tau) = \frac{A_1^2}{2} \cos(\omega(\tau - d_1)) + \frac{A_2^2}{2} \cos(\omega(\tau - d_2)) +$$

$$+ A_1 A_2 \cos(\omega(\tau - \frac{d_1 + d_2}{2})) \cdot \cos(\Delta \varphi + \omega \frac{d_2 - d_1}{2}).$$  \hspace{1cm} (2.8)

Due to the periodicity of $\gamma(\tau)$, we study the cross-correlation function on a $2\pi$ interval centered at $\omega(d_1 + d_2)/2$. Without loss of generality, assume that the phase differences $\omega d_1, \omega d_2$ are in this interval; otherwise, simply shift the phases with
multiples of $2\pi$. To fix the discussion let $d_1 < d_2$. By observing the deviation of the peak location $\tau_{\text{max}}$ from the middle of the two sources, $(d_1 + d_2)/2$, we obtain the stronger source:

$$\tau_{\text{max}} > (d_1 + d_2)/2 \iff A_1 < A_2.$$  \hfill (2.9)

This result gives a threshold to decide which source is stronger based on ITD. Furthermore, we want to study how ITD changes with the relative strength $R = \frac{A_2}{A_1 + A_2} \in [0, 1]$. Hence, we derive the solution for $\tau_{\text{max}}$ as follows:

$$\tau_{\text{max}} = \frac{d_1 + d_2}{2} + \frac{1}{\omega} \left( \arctan \left[ \frac{(A_2^2 - A_1^2) \sin \beta}{(A_1^2 + A_2^2) \cos \beta + 2A_1A_2 \cos(\Delta \phi + \beta)} \right] + k \pi \right).$$  \hfill (2.10)

where $\beta = \omega \frac{d_2 - d_1}{2} \in [0, \pi]$ and $k$ is an integer. The relation obtained in Eq. 2.9 uniquely determines $k \in \{0, \pm 1\}$ for the $2\pi$ interval considered. More specifically, $\beta \leq \pi/2 \Rightarrow k = 0$ and $\beta > \pi/2 \Rightarrow k = 1$ when $A_1 < A_2$, and $k = -1$ when $A_1 > A_2$.

Furthermore, simulations and derivations show that a good approximation for the mean value $\bar{\tau}_{\text{max}}$ when $\Delta \phi$ varies uniformly in the range $[-\pi, \pi]$ is given by:

$$\bar{\tau}_{\text{max}} = \begin{cases} 
  d_1, & R < 0.5 \\
  \frac{d_1 + d_2}{2}, & R = 0.5 \\
  d_2, & R > 0.5 
\end{cases}.$$  \hfill (2.11)
Figure 2.4. Theoretical approximation for the mean ITD, \( \bar{\tau}_{\text{max}} \), for two pure tones randomly distributed in a narrow band centered at 500 Hz. The y-axis corresponds to the relative strength \( R \). Two cases are shown: \( \beta = \pi / 4 \) (solid line) and \( \beta = 3\pi / 4 \) (dashed line).

Case 2: \( \omega_1 \neq \omega_2 \)

In this case, due to the orthogonality of sine waves of different frequencies the cross-correlation function becomes:

\[
\gamma(\tau) = \frac{A_1^2}{2} \cos(\omega_1 (\tau - d_1)) + \frac{A_2^2}{2} \cos(\omega_2 (\tau - d_2)).
\]  \hspace{1cm} (2.12)

A closed form solution for the peak location in this case does not exist. Instead, we analyze the behavior of the peak location for relatively close angles, i.e.
In this interval, we apply a second-order Taylor expansion as an approximation for the cosine, resulting in a simple solution: \( \tau_{\text{max}} = \frac{A_1^2 \omega_1^2 d_1 + A_2^2 \omega_2^2 d_2}{A_1^2 \omega_1^2 + A_2^2 \omega_2^2} \).

Note that this is a monotonic function with respect to the relative strength \( R \).

For the general case, we observe that as the frequencies \( \omega_1 \) and \( \omega_2 \) vary uniformly in a narrowband centered at \( \omega \), a good approximation for the mean of \( \tau_{\text{max}} \) is given by:

\[
\tau_{\text{max}} = \frac{d_1 + d_2}{2} + \frac{1}{\omega} \left( \arctan \left( \frac{(A_2^2 - A_1^2)}{(A_1^2 + A_2^2) \tan \beta} \right) + k \pi \right), \quad k \in \{0, \pm 1\} \quad (2.13)
\]

which is the solution for the maximum position in Eq. 2.12 when \( \omega_1 = \omega_2 \). This function is monotonically increasing with respect to \( R \) when \( \beta < \pi/2 \) and decreasing when \( \beta > \pi/2 \). Fig. 2.4 shows the results when \( \omega = 500 \) Hz and \( \beta \) equals \( \pi/4 \) and \( 3\pi/4 \), respectively.

A systematic change in \( R \) also results in a corresponding shift in IID. A similar discussion applies here. That is, the frequency difference between the two tones affects the spread of IID distribution. We do not study the case \( \omega_1 = \omega_2 \) since the results for IID distribution are complex and not amenable to the analysis used here. In addition, IID is most reliable at high frequencies where filter bandwidths are large. Therefore, we consider the case \( \omega_1 \neq \omega_2 \). IID is approximated as the ratio of signal power at the two ears, resulting in the following expression:
IID = 10\log_{10} \frac{A^2_t \vert H^2_t(\omega_1) \vert^2 + A^2_r \vert H^2_r(\omega_2) \vert^2}{A^2_t \vert H^2_t(\omega_1) \vert^2 + A^2_r \vert H^2_r(\omega_2) \vert^2}, \quad (2.14)

where the power of a signal \( u(t) \) is \( \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} u^2(t) dt \). Note that IID is monotonic with respect to the relative strength \( R \).

The above analysis suggests that the distribution of the binaural cues in a given filter channel is directly influenced by the filter bandwidth \( \Delta \omega \). To test this, we simulate left and right signals using Eq. 2.6, where the relative strength is fixed, \( \Delta \varphi \) is uniformly distributed in the range \( [-\pi, \pi] \) and \( \omega_{L,2} \) in \( [\omega - \Delta \omega, \omega + \Delta \omega] \). Figs. 2.5(a) and 2.5(b) show the mean and the variance of ITD as a function of \( R \) for the condition of \( \omega = 500 \text{ Hz} \), 30° azimuth separation, 20-ms integration time and four \( \Delta \omega \) values in the range of 0 Hz to 200 Hz. In the figure, \( M_1 \) is the ITD mean as derived in Eq. 2.11 and it approximates well the case \( \Delta \omega = 0 \). \( M_2 \) is the ITD mean derived in Eq. 2.13 for the more general case \( \Delta \omega \neq 0 \). Similarly, Figs. 2.5(c) and 2.5(d) show results for IID when \( \omega = 2.5 \text{ kHz} \) and five \( \Delta \omega \) values in the range of 0 Hz to 400 Hz. Here, \( M \) is the IID mean as derived in Eq. 2.14. It is worth noting that the theoretical derivations of \( M_2 \) and \( M \) approximate well the simulation results when the bandwidth approaches the auditory filter ERB, which is 80 Hz for a 500 Hz center frequency and 300 Hz for 2.5 kHz. In addition, there is a systematic decrease in variance for both ITD and IID as \( \Delta \omega \) approaches the ERB. This behavior generalizes to other frequencies as well.

To conclude, our analysis shows that ITD and IID undergo systematic shifts from the ideal target values as the relative strength \( R \) of two sinusoidal sources is changed. A
Figure 2.5. The influence of filter bandwidth on the mean and variance of ITD and IID with respect to the relative strength $R$. The data is from simulations of two pure tones uniformly distributed in a narrow band. One tone is at 0° and the another is at 30°. The sampling frequency is 44.1 kHz. (a) Mean ITD as a function of $R$ for 500 Hz center frequency and four bandwidths between 0 Hz and 200 Hz. The auditory filter ERB here is 80 Hz. $M_1$ and $M_2$ correspond to the theoretical mean ITD as derived in Eq. 2.13 and Eq. 2.14, respectively. (b) ITD variance for the same condition as in (a). (c) Mean IID as a function of $R$ for a 2.5 kHz center frequency and five bandwidths between 0 Hz and 400 Hz. $M$ corresponds to the theoretical mean IID as derived in Eq. 2.15. The auditory filter ERB is 300 Hz. (d) IID variance for the same condition as in (c).

Comparison of the above theoretical derivations with the real data presented in the next subsection shows that the match is very close.
2.5.2 Model

The analysis of ITD and IID for pure tones shows relatively smooth changes with the relative strength $R$ in narrow frequency bands. In order to capture this relationship in the context of real signals, statistics are collected for individual spatial configurations during training. Binaural signals are obtained by convolving monaural signals with KEMAR HRTFs as explained in Section 2.2. We employ a training corpus consisting of 10 speech signals from the TIMIT database (Garofolo et al., 1993): 5 male utterances and 5 female utterances as presented in Table 2.1. The speaker ID in the table uniquely identifies the speaker in the TIMIT database where the first letter indicates the sex of the speaker. In the two-source case, we select S0-S4 to be the target and the rest interference. In the three-source case, we have S0-S3 as target signals and the 2 interfering sets are S4-S6 and S7-S9.

Estimates for ITD, IID and $R$ are extracted independently for all frequency channels. Since the cross-correlation function is periodic, resulting in multiple peaks for mid to high frequencies, we consider the following strategy for estimating ITD. We study deviations from the target ITD for individual frequency channels, which is obtained from the ITD-azimuth mappings presented in Section 2.4. Consequently, we compute $ITD_c$ as the peak location of the cross-correlation function in the range $f_c/f_s$ centered at the target ITD, where $f_c$ is the center frequency of channel $c$ and $f_s$ is the sampling frequency. $IID_c$ corresponds to the mean power ratio at the two ears, expressed in decibels:
<table>
<thead>
<tr>
<th>ID</th>
<th>Speaker ID</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0</td>
<td>MKLS0</td>
<td>“Primitive tribes have an upbeat attitude”</td>
</tr>
<tr>
<td>S1</td>
<td>FCKE0</td>
<td>“Only the best players enjoy popularity”</td>
</tr>
<tr>
<td>S2</td>
<td>MCDC0</td>
<td>“Our aim must be to learn as much as we teach”</td>
</tr>
<tr>
<td>S3</td>
<td>FEAR0</td>
<td>“Development requires a long-term approach”</td>
</tr>
<tr>
<td>S4</td>
<td>FDMS0</td>
<td>“Poets, moreover, dwell on human passions”</td>
</tr>
<tr>
<td>S5</td>
<td>FETB0</td>
<td>“Change involves the displacement of form”</td>
</tr>
<tr>
<td>S6</td>
<td>FCMM0</td>
<td>“The system works as an impersonal mechanism”</td>
</tr>
<tr>
<td>S7</td>
<td>MJWS0</td>
<td>“Most assuredly ideas are invaluable”</td>
</tr>
<tr>
<td>S8</td>
<td>MRVG0</td>
<td>“False ideas surfeit another sector of our life”</td>
</tr>
<tr>
<td>S9</td>
<td>MJRH0</td>
<td>“But in every period it has been humanism”</td>
</tr>
</tbody>
</table>

Table 2.1. Speech signals of the training set.

\[
IID_c = 20 \log_{10} \left( \frac{\sum c r^2_c(t)}{\sum c l^2_c(t)} \right), 
\]

(2.15)

where \( l_c \) and \( r_c \) refer to the left and right auditory periphery output of channel \( c \), respectively. Note that in computing \( IID_c \), we use 20 instead of 10 in order to compensate for the square root operation in the peripheral processing stage.

The relative amplitude is a measure of the relative strength between the target source and the acoustic interference, defined using root-mean-square values of the original signals at the “better ear” - the ear with higher SNR (see e.g. Shinn-Cunningham et al., 2001):
Figure 2.6. Relationship between ITD/IID and the relative strength $R$ for a two-source configuration: target in the median plane and interference on the right side at 30°. (a) The scatter plot shows ITD and R estimates from the training corpus for a channel with center frequency of 500 Hz. The solid curve shows the theoretical mean (see Eq. 2.14) and the dash curve shows the data mean. (b) Results for IID for a filter channel with center frequency 2.5 kHz. The solid curve shows the theoretical mean (see Eq. 2.15) and the dash curve shows the data mean.

$$R_c = \frac{\left(\sum_r s_c^2(t)\right)^{1/2}}{\left(\sum_r s_c^2(t) + \sqrt{\sum_r n_c^2(t)}\right)^{1/2}}$$

(2.16)

where $s_c$ refers to the response in channel $c$ to target signal and $n_c$ the response to the acoustic interference (noise).
Figure 2.6 shows empirical results obtained for a two-source configuration: target source in the median plane and interference at 30°. The scatter plot in Fig. 2.6(a) shows samples of $\text{ITD}_c$ and $R_c$ obtained for the channel with a center frequency of 500 Hz (about 7000 samples in total). In addition, we display the empirical mean of the samples and the theoretical one derived in Eq. 2.13. Similarly, Fig. 2.6(b) shows the results that describe the variation of $\text{IID}_c$ with $R_c$ for a channel with a center frequency of 2.5 kHz and compares the empirical mean with the one derived in Eq. 2.14. Note that $R_c$ incorporates the HRTF responses at the better ear. Therefore, the $R$ axis for the theoretical mean is converted accordingly. Figure 2.6 exhibits a systematic shift of the estimated ITD and IID with respect to $R$ for real signals. Moreover, the theoretical means obtained in the case of pure tones match the empirical ones very well. Similar matches are observed in other frequency channels and other spatial configurations.

The above observation extends to multiple-distracter scenarios. As an example, Fig. 2.7 displays smoothed histograms that show the relationship between $R_c$ and both $\text{ITD}_c$ (Fig. 2.7(a)) and $\text{IID}_c$ (Fig. 2.7(b)) for a three-source situation. Samples correspond to a frequency channel with a center frequency close to 1.5 kHz for target at 0° (median plane) and two interferences at −30° and 30°. Note that the interfering sources introduce systematic deviations of the binaural cues. Consider a particularly troubling case: the target is silent and two interferences have equal energy in a given T-F unit. This results in binaural cues indicating an auditory event at half of the distance between the two interference locations; for our setup, it is 0° - the target location. However, the data in
Figure 2.7. Relationship between ITD/IID and the relative strength $R$ for a three-source configuration: target source in the median plane and interference at -30° and 30°. Statistics are obtained from the training corpus for a channel with center frequency close to 1.5 kHz. (a) Histogram of ITD and $R$ samples. (b) Histogram of IID and $R$ samples. (c) Clustering in the ITD-IID space.

Fig. 2.7 suggest a low probability for this case. Figure 2.7 instead shows a clustering phenomenon, suggesting that in most cases only one source dominates a T-F unit.

By displaying the information in the joint ITD-IID space, we observe a location-based clustering of the binaural cues, which is clearly marked by strong peaks that correspond to distinct active sources as shown in Fig. 2.7(c). There exists a tradeoff between ITD and IID across frequencies, where ITD is most salient at low frequencies and IID at high frequencies. But a fixed cutoff frequency that separates the effective use of ITD and IID does not exist for different spatial configurations (see Fig. 2.8 later). This motivates our choice of a joint ITD-IID feature space that optimizes the system performance across different configurations. Differential training seems necessary for
different channels given that there exist variations of ITD and, especially, IID values with different center frequencies.

Since the goal is to estimate an ideal binary mask, we focus on detecting decision regions in the 2-dimensional ITD-IID feature space for individual frequency channels. Consequently, standard supervised learning techniques can be applied. For channel \( c \), we test the following two hypotheses. The first one is \( H_1 \): target is dominant or \( R_c > 0.5 \), and the second one is \( H_2 \): interference is dominant or \( R_c \leq 0.5 \). Based on estimates of the bivariate densities \( p(x \mid H_1) \) and \( p(x \mid H_2) \) the classification is done in accordance with the maximum a posteriori (MAP) decision rule: 

\[
p(H_1)p(x \mid H_1) > p(H_2)p(x \mid H_2)
\]

There exist a plethora of techniques for probability density estimation ranging from parametric techniques (e.g. mixture of Gaussians) to nonparametric ones (e.g. kernel density estimators). We initially tried the EM algorithm for learning Gaussian mixtures (Duda et al., 2001), but this did not prove to be as robust due to the following factors: (i) the true number of mixing components is usually unknown, and (ii) the algorithm tends to be sensitive to parameter initialization. Even for the two-source scenario, the method of computing ITD for mid- to high-frequencies can result in two-mode distribution for the \( H_2 \) hypothesis. In order to completely characterize the distribution of the data we use the kernel density estimation method independently for all frequency channels.

Kernel density estimation is well documented in the literature (Silverman, 1986), so we only summarize its essence here. Generally, the multidimensional kernel density estimate for \( n \) observations \( x_1, \ldots, x_n \) of dimensionality \( d \) is given by the following formula:
\[
\hat{f}(\mathbf{x}) = \frac{1}{\sum_{\mathbf{x}'} n h_{\mathbf{x}_1}...h_{\mathbf{x}_d}} \prod_{j=1}^{d} N \left( \frac{x_j - x_{ij}}{h_j} \right),
\]  

(2.17)

where \( \mathbf{x} = (x_1, ..., x_d) \) is a feature vector, \( x_{ij} \) is the \( j \)th element of \( \mathbf{x}_i \), \( N \) is a Gaussian function, and \( h_j \)'s are parameters called bandwidths that define the amount of smoothing for the empirical distribution. In our case, the ITD-IID feature space has dimensionality \( d = 2 \). The selection of the smoothing parameters is critical to the success of the estimation process: for too small values it approximates the data well but generalizes poorly and for too large values the structure of the data distribution disappear. One approach for finding optimal values is the least-squares cross-validation method (LSCV) (Silverman, 1986). We employ the LSCV method for high dimensions and the Gaussian kernel given by Sain et al. (1994) (p. 808). Optimal smoothing values are chosen as local minima in the range \([ n^{-1/6} \sigma, / 4, 3n^{-1/6} \sigma, / 2 \] \), where \( \sigma_i \) represents the standard deviation of the data set in the \( i \)th dimension and \( n \) is the size of sample data set.

One cue not employed in our model is IED. Auditory models generally use IED in the high-frequency range (see for example Bodden, 1993) since the auditory system becomes gradually insensitive to interaural phase differences above 1.5 kHz. In addition, the occurrence of multiple peaks at high frequencies in the cross-correlation function is much reduced for the IED cue. We have compared the individual performance of the three binaural cues: ITD, IID and IED, for a 1-dimensional classification task based on the kernel density estimation method presented above. An error is made whenever the estimated binary mask value for a T-F unit differs from the corresponding ideal value.
Figure 2.8. Discriminability comparison for the three binaural cues, ITD, IID and IED, the joint IED-IID space and the joint ITD-IID space. Error rates are displayed as a function of channel number (frequency) for a classification task for two spatial configurations. (a) Target source in the median plane and interference on the right side at 5°. (b) Target source in the median plane and interference on the right side at 30°. IED results are shown for frequencies above 1.5 kHz, i.e. above channel number 80.
Figure 2.9. Comparison between the estimated mask and the ideal binary mask for a two-source configuration. (a) Male target speech in the median plane. (b) Mixture of target speech and a female utterance on the right side at 30°. (c) The ideal binary mask. (d) The binary mask resulting from our model. The white regions indicate those T-F units dominated by target speech.

Figure 2.8 shows the error rates with respect to frequency channel using the Cooke corpus (see Section 2.6.1) as the test set, where we consider two cases: target source in the median plane and the acoustic interference at 5° (Fig. 2.8(a)) and 30° (Fig. 2.8(b)). IED results are given for the frequency range of interest - above 1.5 kHz (i.e. channel
number > 80). As the source separation increases, error rates for IED and IID improve. On the other hand, ITD loses discriminability for high-frequency channels where the multiple-peak problem results in the same ITD for both target and interference (Fig. 2.8(b)). Figure 2.8 also displays the corresponding error rates for the joint ITD-IID space and the joint IED-IID space, and it shows that the joint ITD-IID space yields the best overall performance across different spatial configurations. As indicated in Fig. 2.8, we have found no benefit for using IED after incorporating ITD and IID, and hence it is not utilized in our model.

2.6 Evaluation and Comparison

A binary mask produced by the model described in the last section approximates very well the corresponding ideal binary mask, which is obtained by comparing the energies of the original target and interference before mixing. As an example, Fig. 2.9 shows a comparison between the ideal binary mask and the estimated mask for a mixture of target male speech presented at 0° and interference female speech at 30° at the better ear. In the figure, a white pixel indicates a T-F unit in which the target dominates. The two masks are very similar, with an SNR difference of only 0.19 dB.

The performance of a segregation system can be assessed in different ways, depending on intended applications. To extensively evaluate our model, we use the following three criteria: 1) an SNR measure using the original target as signal; 2) ASR rates using our model as a front-end; and 3) human speech intelligibility tests. Results with each criterion are given below.
<table>
<thead>
<tr>
<th>ID</th>
<th>Speaker ID</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0</td>
<td>MWSB0</td>
<td>“Bright sunshine shimmers on the ocean”</td>
</tr>
<tr>
<td>S1</td>
<td>MDCD0</td>
<td>“Challenge each general’s intelligence”</td>
</tr>
<tr>
<td>S2</td>
<td>MDHS0</td>
<td>“The Thinker is a famous sculpture”</td>
</tr>
<tr>
<td>S3</td>
<td>MTAA0</td>
<td>“Only lawyers love millionaires”</td>
</tr>
<tr>
<td>S4</td>
<td>MRPC1</td>
<td>“Biblical scholars argue history”</td>
</tr>
<tr>
<td>S5</td>
<td>FPKT0</td>
<td>“They make us conformists look good”</td>
</tr>
<tr>
<td>S6</td>
<td>FJRE0</td>
<td>“Artificial intelligence is for real”</td>
</tr>
<tr>
<td>S7</td>
<td>FPAC0</td>
<td>“A good attitude is unbeatable”</td>
</tr>
<tr>
<td>S8</td>
<td>FREH0</td>
<td>“Too much curiosity can get you into trouble”</td>
</tr>
<tr>
<td>S9</td>
<td>FBCH0</td>
<td>“Clear pronunciation is appreciated”</td>
</tr>
</tbody>
</table>

Table 2.2. Target signals of the test set.

### 2.6.1 SNR Evaluation

To conduct an SNR evaluation, a segregated signal is reconstructed from a binary mask following the method described in Section 2.2. To quantitatively assess system performance, we measure in decibels the SNR using the original target speech before mixing as signal:

$$\text{SNR} = 10 \log_{10} \frac{\sum_{t} s_T^2(t)}{\sum_{t} (s_T(t) - s_E(t))^2},$$

where $s_T(t)$ represents the original target signal reconstructed using an all-one mask and $s_E(t)$ the estimated target reconstructed from the binary mask. Note that Eq. 2.18
<table>
<thead>
<tr>
<th>ID</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>N0</td>
<td>1 kHz tone</td>
</tr>
<tr>
<td>N1</td>
<td>Random noise</td>
</tr>
<tr>
<td>N2</td>
<td>Noise bursts</td>
</tr>
<tr>
<td>N3</td>
<td>“Cocktail” party noise</td>
</tr>
<tr>
<td>N4</td>
<td>Rock music</td>
</tr>
<tr>
<td>N5</td>
<td>Siren</td>
</tr>
<tr>
<td>N6</td>
<td>Telephone trill</td>
</tr>
<tr>
<td>N7</td>
<td>“Don’t ask me to carry an oily rag like…”</td>
</tr>
<tr>
<td>N8</td>
<td>“She had your dark suit in greasy wash…”</td>
</tr>
<tr>
<td>N9</td>
<td>“Why were we keen to use human…”</td>
</tr>
</tbody>
</table>

Table 2.3. Noise signals of the test set.

provides a measure that penalizes both retained noise and target distortion, and in the case of an all-one mask yields the original SNR. Our measure is more stringent than the conventional SNR measure; indeed our tests show that Eq. 2.18 gives systematically lower SNR values. To minimize the loss of target energy we take advantage of the higher initial SNR at the better ear. As a result, the reconstructed signal corresponding to the better ear contains more target energy. Therefore, all the following evaluations are performed at the better ear.

The system performance is measured on independent test corpora for different spatial configurations. For the two-source scenario, one test set is the corpus collected by Cooke (1993), chosen because it is commonly used in computational ASA studies (Brown and Cooke, 1994; Wang and Brown, 1999; Wu et al., 2003). The corpus contains 10 voiced
Figure 2.10. Systematic results for two-source configuration with 5° azimuth separation. Black bars correspond to the SNR of the initial mixture, white bars indicate the SNR obtained using ideal binary mask, and gray bars show the SNR from our model. Results are obtained for speech mixed with ten types intrusions (Table 2.3) for different spatial configurations. (a) Target at 0°, interference at 5°. (b) Target at 40°, interference at 45°. (c) Target at 80°, interference at 85°.

speech signals and 10 noise intrusions, encompassing a variety of common acoustic interferences such as telephone ringing, rock music, and other speech utterances. In
addition, we employ a second corpus containing 10 normal speech utterances from the TIMIT database (see Table 2.2) as target mixed with the 10 intrusions from the Cooke corpus (see Table 2.3). In the case of three sources, we use the Cooke corpus for testing: 5 speech signals form the target set and the other 5 form one interference source. The 10 intrusions then form the second interference source. Therefore, in this three-source corpus every mixture contains two utterances plus an additional intrusion.

For the two-source case, the model is systematically evaluated at the better ear for various combinations of azimuth angles. We compare the SNR gain obtained by our model against that obtained using an ideal binary mask. For the test corpus of Table 2.2, Fig. 2.10 shows the results for a spatial separation of 5° and target at azimuth 0°, 40° and 80°. Results are similar across mixtures in the same noise category; hence, we present the averaged result for each category. Very good results are obtained when the target is close to the median plane for an azimuth separation as small as 5°. Performance degrades when the target source is moved to the side of the head and this is a direct consequence of poorer resolution of the binaural cues at higher azimuth angles. When comparing with the SNR of the initial mixture, there is an average-SNR gain of 13.76 dB for the target in the median plane, and it reduces to 5.04 dB with the target at 80°. When the spatial separation increases, excellent results are obtained across all spatial configurations. Figure 2.11 shows results for target at 0°, 30° and 60° and interference at 30° to the right of target. Similar results are obtained for other spatial configurations. Figure 2.12 shows that the system performs equally well on the Cooke corpus. Figure 2.12(a) gives the
Figure 2.11. Systematic results for two-source configuration with 30° azimuth separation. Black bars correspond to SNR of the initial mixture, white bars to the SNR obtained using an ideal binary mask, and gray bars to the SNR from our model. (a) Target at 0°, interference at 30°. (b) Target at 30°, interference at 60°. (c) Target at 60°, interference at 90°.

...results for a 5° azimuth separation and the average improvement is 13.73 dB. Similarly, Fig. 2.12(b) gives the results for a 30° separation.
Our approach, like other location-based methods using cross-correlation, can be extended to cases with more than two sources. With given locations, our model performs target segregation in a similar manner, i.e. estimating an ideal binary mask following the method outlined in Section 2.5.2. Figure 2.13 illustrates the performance of the model in a three-source scenario with target located in the median plane and two interfering sources at $-30^\circ$ and $30^\circ$. Here, the 10 noise intrusions from the Cooke corpus are presented at $30^\circ$.
azimuth and the target is reconstructed based on the right ear mixture. As previously, results are mean values for the 10 types of noise intrusion. The performance degrades compared to the corresponding two-source situation, from an average SNR of about 12 dB to 4.10 dB. Still, the average SNR gain obtained is approximately 11.31 dB.

In order to draw a quantitative comparison with another binaural processing model, we have implemented the Bodden model (Bodden, 1993), which produces good-quality sound separation using source locations. The localization stage of this model uses an extended cross-correlation mechanism based on contralateral inhibition and it adapts to HRTFs. The separation stage of the model is based on estimation of the weights for a Wiener filter. Specifically, for a given T-F unit the weight is given by the ratio between a
desired excitation and an actual one. The actual excitation corresponds to the integration of the cross-correlation pattern across the azimuth axis and the ideal peak shape is used as a window to derive the desired excitation. The Bodden model differs from ours in several aspects. First, his sound localization stage builds on the previous models of Lindemann (1986) and Gaik (1993), which simulate aspects of the precedence effect for reverberant scenarios, whereas our localization stage is simpler and does not address the precedence effect. Second, his model requires only a target azimuth and no training is necessary as spatial configuration changes. Although these aspects add to the flexibility of his model, the estimation of Wiener filter weights appears less robust than our binary estimation of ideal masks. In addition, our configuration- and channel-specific training utilizes more
information provided by localization and makes an optimal use of frequency-dependent ITD and IID cues.

Bodden’s system uses a 24-channel filterbank intended to simulate critical bands. For a fair comparison, our implementation of the Bodden system uses the same time-frequency resolution employed in our system with a 128-channel gammatone filterbank; we also implemented the Bodden model with 24-channel critical bands and the results are not as good. We find that, when two sources are relatively close, the Bodden model is less robust than ours. Our comparison is based on the Cooke corpus and a spatial configuration of target at 0° and intrusion on the right side at 30°, an azimuth separation in the range where his model performs optimally. As displayed in Fig. 2.14, our model shows a considerable improvement over the Bodden system, producing 3.5 dB average improvement. The improvement is especially high for a few cases (e.g. N5 and N6) where our estimated masks result in large SNR improvements over the original mixtures.

2.6.2 ASR Evaluation

As discussed before, an ideal binary mask is defined a priori. Similar a priori masks have been shown to produce impressive performance when applied to the automatic recognition of noisy speech using a ‘missing data’ approach (Cooke et al., 2001). Given an observed speech spectral vector \( Y \), the classification problem in speech recognition involves the maximization of the posterior probability \( P(W|Y) \) where \( W \) is a valid word sequence. When parts of \( Y \) are masked by noise or other distortions, some components of \( Y \) are likely to be unreliable. The missing data approach addresses this problem by
partitioning $Y$ into its reliable and unreliable components, $Y_r$ and $Y_u$. In this approach, a continuous density hidden Markov model recognizer is modified such that only acoustic features indicated as reliable in the mask are used during decoding. There are different classification methods for missing-data recognition. Here, we use the bounded marginalization method (Cooke et al., 2001). Since our ideal binary masks are generated in a similar way to those used by Cooke et al., we would expect them to be an equally effective front-end to missing data ASR. We therefore use the missing-data technique proposed by Cooke et al. for our ASR evaluation. Specifically, we label the T-F units indicated as 1 in the binary mask as reliable and those indicated as 0 as unreliable.

Our motivation for ASR evaluation is two-fold. First, a practical system must estimate such a mask, and as a result deviations from an ideal mask must be considered. Hence we want to find how tolerant recognition performance is to deviations from an ideal mask. Second, we want to give a quantitative measure of the potential improvement on ASR performance using our speech segregation model as a front-end.

We have implemented the missing data algorithm with the same 128-channel gammatone filterbank as described in Section 2.3. Feature vectors are obtained using the instantaneous Hilbert envelope at the output of each gammatone filter. More specifically, each feature vector is extracted by smoothing the envelope using an 8-ms first-order filter, sampling at a frame-rate of 10 ms and finally log-compressing. As in the original study, the task domain is speaker independent recognition of connected digits. Thirteen (the number 1-9, a silence, very short pause between words, zero and oh) word-level models are trained using an HMM toolkit, HTK (Young et al., 2000). All except the short
pause model have 8 emitting states. The short pause model has a single emitting state, tied to the middle state of the silence model. The output distribution in each state is modeled as a mixture of 10 Gaussians. The grammar for this task allows for one or more repetitions of digits and all digits are equally probable. Finally, both training and testing are performed using the male speaker dataset in the TIDigits database (Leonard, 1984).

To study the sensitivity of an ideal mask to estimation error, our first test assesses the correctness score and the accuracy score (correctness minus word insertion errors) when a random deviation from an ideal binary mask is introduced. Here, we use for simplicity a monaural condition as in Cooke et al. (2001). Deviations are obtained by randomly flipping the same number of bits from 0’s and 1’s; the number is measured as percentage of the total number of 1’s in an ideal mask. The percentages tested are 0%, 5%, 10%, 20%, and 50%. Since the underlying acoustic energy associated with a T-F unit, or a bit, can vary in a large range, we further measure the energy deviation ratio as the ratio of the energy corresponding to flipped bits and the total energy corresponding to the ideal binary mask. The results for a male target speaker mixed with “car noise” (Cooke et al., 2001) are given in Fig. 2.15, where the abscissa indicates the energy deviation ratio. Three SNR levels for the mixture, i.e. –5 dB, 0dB and 5 dB are tested. Figure 2.15(a) gives the correctness score and Fig. 2.15(b) the accuracy score. Figure 2.15 shows that both correctness score and accuracy score decrease gradually and systematically as deviation ratio increases. This suggests that ideal binary masks are robust to estimation error. A comparison between Fig. 2.15(a) and Fig. 2.15(b) shows that the accuracy score degrades faster than the correctness score. This suggests that word insertions, which
Figure 2.15. Degradation of recognition score with deviations from an ideal binary mask evaluated for three SNR values: 5 dB (square), 0 dB (circle) and –5 dB (diamond). (a) Correctness score. (b) Accuracy score.

result from noise retention or word boundary blurring, are more sensitive to estimation error than recognition of present words.
The second test directly evaluates binary masks estimated by our system for binaural conditions with two and three sources. For all tests, the same male target speaker is located at 0°. Both training and testing of the system are performed on acoustic features from the left ear signal. Figure 2.16(a) and Fig. 2.16(b) show the correctness and accuracy scores for a two-source condition, where the interference is another male speaker at 30°. The performance of our model is compared against the ideal masks systematically for four SNR levels, i.e. 5 dB, 0 dB, –5 dB and –10 dB. Also shown in the figure is the baseline performance where the recognition is conducted on unprocessed mixtures from the left ear. Similarly, Fig. 2.16(c) and Fig. 2.16(d) show the results for the three-source case with an added female speaker at -30°. The results in Fig. 2.16 show that ideal binary mask exhibit only slight and gradual degradation in recognition performance with decreasing SNR and increasing number of sources. In the two-source case, the estimated masks perform equally well as the ideal masks. In the three-source case, the estimated masks do not perform as well, and this is to be expected since we know from Section 2.6.1 that the quality of ideal mask estimation for three sources is not as good as for two sources. Consistent with the observations from Fig. 2.15, performance degrades more quickly for the accuracy score than for the correctness score. Observe that large improvements over baseline performance are obtained across all conditions (to a lesser degree for the accuracy score in the three-speaker condition). This shows the strong potential of applying our model to robust speech recognition.
Figure 2.16. Recognition performance at different SNR values for original mixture (dotted line), ideal binary mask (solid line) and estimated mask (dashed line). (a) Correctness score for a two-source case. (b) Accuracy score for a two-source case. (c) Correctness score for a three-source case. (d) Accuracy score for a three-source case.

2.6.3 Speech Intelligibility Evaluation

Finally, we evaluate our model on speech intelligibility with human listeners. Before reporting the results, we should point out that human listeners have a remarkable ability
to perform ASA, and their superior ability to recognize speech in the presence of acoustic interference is the very motivation for our model design. Because of this, our tests focus on relatively low SNR conditions; otherwise scores will be indiscriminately high for both unprocessed mixtures and segregated speech.

We use the Bamford-Kowal-Bench sentence database that contains short semantically predictable sentences (Bench and Bamford, 1979) for intelligibility tests. The score is evaluated as the percentage of keywords correctly identified, ignoring minor errors such as tense and plurality (Stubbs and Summerfield, 1990). Two different spatial configurations are considered: a two-source configuration at 0° and 5°, and a three-source configuration at -30°, 0°, and 30°. To eliminate potential location-based priming effects (Maljkovic and Nakayama, 1996) we randomly swap the locations for target and interference for different trials. In the unprocessed condition, binaural signals are produced by convolving original signals with the corresponding HRTFs and the convolved signals are presented to a listener dichotically (see Bodden, 1993). In the processed condition, our algorithm is used to reconstruct the target signal at the better ear and results are presented diotically.

Twelve native English speakers with normal hearing, between 24-30 years old, participated in the experiments. The tests were conducted in a sound insulating booth (IAC Model 40a-9) and signals were presented over Sennheiser HD 256 headphones. At the beginning of a test, subjects were familiarized with the voice of a target male speaker and they were free to adjust the sound volume to a comfortable level. The task of a subject during each test run was to report what was comprehended and a human operator
marked the result. Each listener participated in a total of 8 conditions. Each condition contained 25 new, randomly chosen sentences, with the first 5 sentences used for practice only and their data discarded.

Figure 2.17 gives the keyword intelligibility score (median values and interquartile ranges) for the two-source configuration. Three SNR level are tested: 0 dB, -5 dB and –10 dB, where the SNR is computed at the better ear for each sentence. The interfering source used for this configuration is babble noise. The general finding is that our algorithm improves the intelligibility score for the tested conditions. The improvement becomes larger as the SNR decreases (61% at –10 dB), even though the algorithm introduces more target distortions at lower SNR levels. Our informal observations suggest, as expected, that the intelligibility score improves for unprocessed mixtures when two sources are more widely separated than 5°. Figure 2.18 shows the results for the three-source configuration, where our model yields a 40% improvement. Here the SNR is fixed at –10 dB at the better ear. The two interfering sources are one female speaker and a different male speaker. Note that, in this case, azimuth separation is high between the three sources. Though we have not formally tested in the three-source configuration, we would expect that a trend similar to the one in Fig. 2.17 occurs with respect to SNR levels; that is, the model improvement decreases as SNR increases.

We recognize that comprehensive human subject evaluations of a model would require a separate study (e.g. see Stubbs and Summerfield, 1990), and indeed this is a topic we intend to pursue in the future. Nonetheless, as far as we know, our system is the first binaural model that has been shown to produce a large speech intelligibility
Figure 2.17. Keyword intelligibility score (median values and interquartile ranges) before (white bars) and after processing (black bars) for a two-source condition (0° and 5°) at three SNR values: 0 dB, -5 dB and -10 dB.

Figure 2.18. Keyword intelligibility score (median values and interquartile ranges) before (white bars) and after processing (black bars) for a three-source condition (0°, 30° and -30°) at -10 dB SNR.
improvement for normal listeners (see Kollmeier and Koch, 1994; Shamsoddini and Denbigh, 2001). The configurations and SNR conditions under which improvement occurs will be systematically characterized in future investigation. Also, some musical noise occurs is signals resynthesized using binary T-F masks. Post-processing techniques have been can be subsequently applied to alleviate this problem and improve the listening quality (Araki et al., 2005).

2.7 Discussion

The human auditory system is capable of adapting to a variety of acoustical situations. A key feature of our model is the introduction of supervised learning for different spatial configurations, and training is conducted independently for different frequency channels. We assume that such training takes place before performing specific segregation tasks, and it would correspond to learning during the development stage. Supervised signals for a spatial configuration of target and intrusion could be supplied in a number of ways, including sound localization, signal estimation from a specific location, and even information extracted from a different modality (e.g. vision). It is worth emphasizing that, unlike a typical supervised learning situation, the training here does not need to capture the specific contents of training signals. As a result the model can be trained equally well using other natural sounds, and estimated distributions generalize in a broad range. In an earlier study (Roman et al., 2002), for example, we
employed a different training methodology and a different training corpus, but the system performance was very similar.

While satisfying the demands of an effective computational system, our model is motivated by physiological and psychoacoustical findings regarding the extraction of spatial features (Patterson et al., 1988). The peripheral processing is based on a gammatone filterbank, which has a foundation in physiology and psychoacoustics. Similarly, the cross-correlation mechanism for ITD extraction as well as the across-frequency integration for localization are supported by related physiological findings (Popper and Fay, 1992).

An open question concerns the role of spatial location in perceptual separation of competing sounds. The experiments by Culling and Summerfield (1995), using simulated vowels in which the formants were defined by noise bands, showed that simultaneous grouping across frequencies based on ITD is weak. Later experiments by Darwin and Hukin (1997; 1999) found that ITD plays a weak role in concurrent sound segregation, but a much stronger role in linking acoustic events from a common location over time. The recent experiments of Freyman et al. (2001) further showed a sizeable improvement in recognizing target speech in the presence of one or two competing speakers based on perceived spatial separation, which suggests a location-based grouping mechanism. Our computational results demonstrate that computed locations can play an effective role in across-frequency grouping. On the other hand, many monaural cues are also important for sound source segregation (see Chapter 1), and how to incorporate both monaural and binaural cues in a comprehensive system remains a challenge.
Our approach uses characteristic clustering of the joint ITD-IID space in order to accurately estimate an ideal binary mask. Related models for estimating target masks through clustering have been proposed previously (Tessier and Berthommier, 1997; Lehn, 1997; Glotin et al., 1999; Jourjine et al., 2000). Notably, the experimental results by Jourjine et al. (2000) suggest that speech signals in a multiple-speaker condition obey to a large extent disjoint orthogonality in time and frequency. That is, at most one source has a nonzero energy at a specific time and frequency. Such models, however, assume input directly from microphone recordings and head-related filtering is not considered. Simulation of human binaural hearing introduces different constraints as well as clues to the problem. First, both ITD and IID should be utilized since IID is more reliable for higher frequencies than ITD. Second, frequency-dependent combinations of ITD and IID arise naturally for a fixed spatial configuration. Consequently, channel-dependent training for each frequency band becomes necessary. Our tests with just ITD (as in Glotin et al.) or channel-independent classification (as in Jourjine et al.) yield considerably inferior performance.

As illustrated in Fig. 2.13, the proposed model can be used to extract target speech from an acoustic mixture that contains more than one intrusion. Although segregation results are expected to drop as the number of sources increases, this property of our model differs from blind source separation using independent component analysis (Hyvärinen et al., 2001) or spatial filtering using sensor arrays (Krim and Viberg, 1996); such techniques require that the number of sensors increases as the number of acoustic
sources increases. A main reason for this difference is that considerations of human audition play a large role in our model design.

Conventional two-microphone adaptive beamformers can develop one deep null which cancels almost perfectly one interference under optimal conditions (Greenberg and Zurek, 2001). The performance, however, degrades when the number of interfering sources increases and is largely affected by the relative SNR of the individual interferences in the reference channel. Weiss (1987) measured the attenuation of individual interferences in acoustical mixtures across different conditions. The experimental results in the anechoic case show attenuation up to 14.5 dB in the two-source case, when both target and interference are active during filter adaptation. For the three-source case, the performance degrades across all interferences by 4 dB and improvement can be as low as 0 dB. In comparison, our model works for a wide range of spatial configurations with two or more sources; for example, Fig. 2.13 shows that with three sources our model still obtains an average SNR gain of 11.3 dB. Conditions with high SNRs degrade the performance of adaptive beamforming. Our model, on the other hand, works especially well for high SNR scenarios. Sub-band versions of adaptive beamforming also exist (see for example Nordholm et al., 2003). In this case, the signal is analyzed independently in frequency bands and different directivity patterns are adaptively chosen in each band. This allows to cancel multiple interferences with nonoverlapping spectra (Cezanne and Pong, 1995). Conventional adaptive beamformers with slow adaptation rate are unable to track fast spectral changes in a multiple-speaker scenario, resulting in suboptimal performance. Using a frame-by-frame multi-source
localization algorithm, Liu et al. (2001) have proposed an equalization and cancellation system that has virtually zero adaptation time. Their two-microphone system exploits the location information in each frame and steers a different null in each frequency band, resulting in 6-7 dB gain in multi-speaker scenarios. Our model uses a similar strategy, by employing the localization cue independently in each T-F unit in order to cancel simultaneous interferences. Hence, binaural processing models including ours may have advantages over adaptive beamformers in a range of acoustical situations.

In terms of limitations, the model presented in this chapter does not address room reverberation or moving sound sources, which will be addressed in later chapters. Also, supervised training is required for different spatial configurations. This limits the flexibility of our system to cope with, say, diffuse background noise. In addition, the localization of many sources in reverberant conditions with just two sensors is a challenging topic. The situation becomes more complex when source motion is considered. Other auditory mechanisms, such as the precedence effect and forward/backward masking, could provide important cues to cope with reverberation in sound localization. Our model also does not address how to define a target in a multi-source situation; to address this issue would inevitably require some high-level processes such as attention and task specification.

To conclude, we have proposed a model for speech segregation based on spatial location. We have observed systematic deviations of the ITD and IID cues from the reference ones with respect to the relative strength between target and acoustic interference, and configuration-specific clustering in the joint ITD-IID feature space.
Consequently, supervised learning of binaural patterns is employed for individual frequency channels and different spatial configurations. Finally, the system estimates a binary mask in order to eliminate acoustic energy in time-frequency units where interference is stronger than target. Our model has been systematically evaluated using both SNR and ASR measures. Evaluation results show that the system estimates ideal binary masks very well and performance degradation is gradual with increasing number and intensity of interferences. In addition, when tested with normal listeners, the model produces large speech intelligibility improvements for two-source and three-source conditions.
CHAPTER 3

BINAURAL TRACKING OF MULTIPLE MOVING SOURCES

This chapter addresses the problem of tracking multiple moving sources based on binaural processing. Our study represents a first step in addressing auditory scene analysis with moving sound sources. As shown in the previous chapter, binaural cues are strongly correlated with source locations in time-frequency regions dominated by only one source. Hence, we propose a multi-channel algorithm that integrates probabilities across reliable frequency channels in order to produce a likelihood function in the target space. Finally, HMM is employed to form continuous tracks and automatically detect the number of active sources across time. Experimental results are presented for two- and three-source scenarios. A preliminary version of this work has been published in the Proceedings of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (Roman and Wang, 2003).
3.1 Introduction

The problem of tracking multiple moving targets arises in many engineering fields including surveillance, navigation, guidance, robotics and speech processing. In this chapter we are interested in the localization and tracking of multiple acoustic sources, such as the presence of concurrent speakers at a cocktail party. A solution to this problem is highly desirable in many speech processing applications including meeting segmentation and hands-free speech acquisition (Omologo et al., 1998; Ajmera et al., 2004).

Numerous multitarget tracking algorithms have been proposed, mostly developed for radar sensors (for a review see Stone, 2001). There are two main approaches to target tracking that utilize Bayesian inference: Multiple Hypothesis Tracking (MHT) and Bayesian filtering. The MHT attempts to optimally associate the noisy measurements over time to form multiple tracks. For a particular hypothesis, a Kalman filter is associated with each track and a MAP cost is computed using the Kalman filter innovation sequence and the \textit{a priori} track set probability. Finally, the estimated tracks are obtained by comparing all the hypothesized track sets using the MAP cost. Bayesian filtering, on the other hand, aims at the conditional mean estimation of the location state space. The conditional probability is recursively estimated by combining a model for the source motions and a likelihood for the state space given a set of noisy measurements. The Bayesian tracker has a closed-form solution only for a linear process with Gaussian noise which is equivalent to the Kalman filter in this case. In general, optimum MHT and
Bayesian solutions require an exponentially growing number of evaluations and therefore are deemed impractical. Hypothesis pruning and merging techniques have been proposed to reduce this computational burden including measurement gating (Read, 1979), probabilistic data association (Bar-Shalom and Tse, 1975) and Viterbi based algorithms (Buckley et al., 2000). An approximation to Bayesian filtering for nonlinear functions, non-Gaussian noises and multiple modal distributions is provided using sequential Monte-Carlo methods, also known as particle filtering (Gordon, 1997, Isard and Blake, 1998). When the number of active sources rapidly varies the above algorithms require complex birth/death rules to initiate and terminate individual tracks.

HMM has also been proposed for target tracking in sonar networks by employing the Markovian modeling of source dynamics in a discretized target space (Martinerie, 1997). It is important to note that this framework can handle multi-modal likelihood distributions. Due to the discrete Markov modeling, Viterbi decoding can be used to efficiently search for the most likely state sequences. The number of targets is, however, decided in this algorithm in a postprocessing step based on detection of local maxima in the likelihood distribution.

Several of the above techniques have been adapted and applied to the problem of speaker tracking using microphone arrays. To estimate the location of active source(s) in each time frame, these algorithms typically employ variants of the well-known generalized cross-correlation function (Knapp and Carter, 1976) or subspace-based methods (Krim and Viberg, 1996). The particle filtering theory, for example, has been extended to the tracking of one moving speaker in a reverberant environment (Vermaak
and Blake, 2001; Ward et al., 2003). For the tracking of multiple speakers, algorithms have been proposed that combine Kalman filtering with probabilistic data association techniques (Sturim et al., 1997; Potamitis et al., 2003). These multi-source tracking algorithms have been shown to provide good localization results using an array of microphones. However, when restricting the size of the array to only two sensors, the multi-source tracking problem becomes very challenging and little has been attained in this respect. As a solution, joint visual and auditory information is generally needed for the task, where audition helps mainly in resolving ambiguities during occlusions (Nakadai et al., 2001).

As seen in Chapter 2, location has been shown to be an effective cue for computational systems that attempt to separate individual talkers in noisy environments. The binaural cues of ITD and IID are strongly correlated with the source locations in T-F regions dominated by only one source. Hence, with accurate locations, the binaural cues can be used to segregate the original signals. However, in a realistic environment source motion or head movement have to be considered and the locations may have to be updated with each frame of data.

In this chapter, we study the tracking of multiple speakers based on the binaural response of a KEMAR dummy head. We propose an HMM framework where the change in the number of active tracks is modeled probabilistically. Specifically, the target space is modeled as a set of subspaces with jump probabilities between them. Each subspace models the tracking of a subset of possible active sources. Hence, unlike the methods presented above, the detection of tracks in the HMM is fully automatic and does not
require heuristic rules for track initialization and termination. Our approach extends an HMM-based model proposed by Wu et al. (2003) that has been successfully applied to the multi-pitch tracking problem. Due to the sparsity of speech signal distribution in the T-F domain, while some T-F units are heavily corrupted due to overlapping multiple sources, others are dominated by only one source and thus provide reliable information for localization. Because the binaural cues are strongly correlated with source locations in these regions, peaky statistical distributions describe the observations in the reliable channels. Hence, we propose to use a channel selection mechanism to determine reliable channels followed by a statistical integration of these channels in order to obtain the likelihood function for different target subspaces. In this chapter we report experimental results for the tracking of two and three simultaneous speakers.

The rest of the chapter is organized as follows: the next section gives an overview of the system. Section 3.3 describes auditory motion modeling. Section 3.4 contains details of the proposed statistical model. Section 3.5 gives simulation results and the last section concludes the chapter.

3.2 Model Architecture

Our multisource tracking system consists of the following four stages: 1) a model of the auditory periphery and binaural cue estimation; 2) a channel selection mechanism that identifies reliable channels in each time frame; 3) a multichannel statistical integration method that produces the likelihood function for target subspaces; and 4) a continuous
Figure 3.1. A schematic diagram of the multisource tracking system.

HMM model for multisource tracking. Figure 3.1 illustrates the model architecture for the case of two moving sources.

The input to our model is a simulated binaural response of a KEMAR dummy head to an acoustic scene with multiple moving sources. We utilize here the catalog of HRTF measurements for anechoic conditions at fixed source locations on a sphere around the KEMAR (see Chapter 2). Interpolation is then used to obtain HRTF filters for arbitrary positions on the sphere. Here we restrict the motion of individual sources to the half horizontal plane with azimuth in the range [-90°, 90°] as used in Chapter 2. The system is, however, extensible to cover the entire azimuth range since ITD and IID used jointly
can potentially differentiate between front and back. Hence, for each moving source left and right ear signals are obtained by filtering with time-varying HRTFs that correspond to the source trajectory on the frontal semicircle. The responses to multiple sources are added at the two ears and form the binaural input to our system.

In the first stage, the resulting left and right ear mixtures are analyzed using the auditory periphery model described in Chapter 2. In short, the signals are passed through an auditory filterbank of 128 fourth-order gammatone filters with channel center frequencies equally distributed on the ERB scale between 80 Hz and 5 kHz. A simple model of hair cell transduction consisting of half-wave rectification and square root operation is then applied to the filtered signals. Finally, for each frequency channel, normalized cross-correlation functions are computed in consecutive time frames using Eq. 2.4. The lag of a peak in the cross-correlation function is a candidate for ITD estimation. At high frequencies where multiple peaks are present the set of all possible time lags is considered, and this creates ambiguity in localization. We resolve this ambiguity by using IID information, which is computed using Eq. 2.15 as the ratio of signal power at the two ears.

Channel selection comprises the second stage of our system. This stage attempts to select the reliable channels defined as being primarily dominated by only one source while removing the more corrupted ones. Here, we use the height of the peak in the cross-correlation function as a measure of the channel reliability. The third stage is the multichannel integration of location information. As seen in Chapter 2, the conventional approach is to summate the cross-correlation functions across all frequency channels. A
peak in the summary cross-correlation suggests an active source while the height of the peak gives its likelihood. This approach, however, under-utilizes the location information in each frequency channel. In our system, we consider the statistical distribution of the ITD-IID estimates. Given a configuration hypothesis, we first formulate the probability of each channel supporting the hypothesis and then we employ an integration method to produce the likelihood of observing the configuration. For configurations with more than one active source a gating mechanism is used to associate the observations with one of the sources.

The last stage of the algorithm is to form continuous azimuth tracks using a continuous HMM framework. We propose an HMM model that allows jumping between subspaces within each of which only a subset of the total number of sources is active. The framework combines the likelihood model from the previous stage, a model for the dynamics of source motion and jump probabilities between the individual subspaces. Finally, optimal azimuth tracks are obtained using the Viterbi algorithm.

3.3 Modeling Auditory Motion

For human audition, sound source localization is primarily achieved with the binaural cues of ITD and IID. For a moving sound, there are changes in ITD and IID that may provide velocity information and enable the listener to perceive and track the changing source location. The transmission path between the acoustic source and the receiver contains many subsystems, i.e. the loudspeaker, the microphone and the ear canal. Here,
we use the diffuse-field equalized HRTFs for which all the factors that are not location-dependent are eliminated. The HRTF catalog (Gardner and Martin, 1994) provides 256 point impulse responses for a fixed number of locations residing on a 1.4 m radius sphere around the KEMAR head. In particular, the resolution in the horizontal plane is 5° azimuth. The sampling rate is fixed at 44.1 kHz.

An attractive property of HRTFs is that they are almost minimum-phase (Mehrgardt and Mellert, 1977). Therefore, a standard way of modeling HRTFs is to decompose the system into a cascade of a minimum-phase filter and a pure delay line (Begault, 1994). The motivation is that minimum-phase systems behave better than the raw measurements for interpolation both in the phase and the magnitude response. In addition, a minimum-phase reconstruction of HRTF does not have perceptual alterations (Kistler and Wightman, 1992). Here, we reconstruct the minimum-phase part through appropriate windowing in the cepstral domain. Specifically, the negative cepstral coefficients are set to 0 and a minimum-phase filter is then obtained by inverting the truncated cepstrum (Gold and Morgan, 2000). The time delay part is estimated as the mean of the group delay in the range of interest from 80 Hz to 5 kHz.

To simulate a continuous motion, the impulse response of an arbitrary direction of incidence is obtained by interpolating separately the minimum-phase filters and the time delays corresponding to neighboring entries in the HRTF catalog. Since we are simulating motions in the horizontal plane, a simple two-way linear interpolation is applied. The impulse response is then reconstructed from the cascade of the resulting minimum-phase filter and the time delay. Finally, to synthesis the binaural response of
the KEMAR dummy head to one moving source a monaural signal is upsampled at 44.1 kHz and filtered with the corresponding time-varying left and right impulse responses. The synthesized multiple sources are added at the two ears and fed to the tracking system.

3.4 Statistical Tracking

Our problem of tracking the azimuths of multiple acoustic sources is formulated here in an HMM framework. An HMM is a doubly stochastic process where an underlying stochastic (Markovian) process that is not directly observable (i.e. “hidden”) is observed through another stochastic process that produces a sequence of observations (Rabiner and Juang, 1993). An HMM is completely defined by the following: 1) the possible target state space; 2) the transition probabilities which reflect the evolution of the target states across time; and 3) the observation probabilities conditioned on the target states, also known as the observation likelihood. Figure 3.2 illustrates our proposed HMM framework. A state in the target space specifies what the active sources are as well as their azimuth information at a particular time frame. The target space is decomposed into subspaces; each subspace corresponds to a subset of active sources. Hence, the transition probability between states in neighboring time frames takes into account both the jump probability between subspaces and the temporal evolution within individual subspaces. Finally, a statistical model that integrates ITD and IID observations in different frequency channels is used to construct the observation likelihood in the target space. To increase
the robustness of the system only frequency channels that are dominated by a single source and thus deemed reliable are considered in our statistical integration.

3.4.1 Dynamics Model

In a practical multi-source tracking situation, the number of active sources at a particular time is generally unknown. In this work, we assume a maximum of three sources and aim to assign separate tracks to each of the sources. Hence, we define the target state space as the union of eight possible subspaces as follows:
Table 3.1. Jump probabilities between subspaces with zero-, one-, two- and three-active sources.

<table>
<thead>
<tr>
<th></th>
<th>$\rightarrow S_0$</th>
<th>$\rightarrow S_1^i$</th>
<th>$\rightarrow S_1^j$</th>
<th>$\rightarrow S_1^{i,j}$</th>
<th>$\rightarrow S_2^{1,3}$</th>
<th>$\rightarrow S_2^{2,3}$</th>
<th>$\rightarrow S_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_0$</td>
<td>0.9663</td>
<td>0.0112</td>
<td>0.0112</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$S_1^i$</td>
<td>0.0692</td>
<td>0.6590</td>
<td>0</td>
<td>0.1359</td>
<td>0.1359</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$S_2^{1,2}$</td>
<td>0.0692</td>
<td>0</td>
<td>0.6590</td>
<td>0</td>
<td>0.1359</td>
<td>0.1359</td>
<td>0</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.0692</td>
<td>0</td>
<td>0.6590</td>
<td>0</td>
<td>0.1359</td>
<td>0.1359</td>
<td>0</td>
</tr>
<tr>
<td>$S_2^{1,3}$</td>
<td>0.0347</td>
<td>0.0347</td>
<td>0</td>
<td>0.7077</td>
<td>0</td>
<td>0</td>
<td>0.2230</td>
</tr>
<tr>
<td>$S_2^{2,3}$</td>
<td>0.0347</td>
<td>0</td>
<td>0.0347</td>
<td>0</td>
<td>0.7077</td>
<td>0</td>
<td>0.2230</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.0448</td>
<td>0.0448</td>
<td>0.0448</td>
<td>0.8655</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1. Jump probabilities between subspaces with zero-, one-, two- and three-active sources.

\[
S = S_0 \cup S_1^i \cup S_1^j \cup S_1^{i,j} \cup S_2^{1,2} \cup S_2^{1,3} \cup S_2^{2,3} \cup S_3, \quad (3.1)
\]

where $S_0$ is the silence space, $S_1^i$ is the state space for a single active source $i$, $S_2^{i,j}$ is the state space for two simultaneously active sources $i$ and $j$, and $S_3$ is the state space for all three active sources. A state is represented as a 3-dimensional vector $x = (\phi^i, \phi^j, \phi^k)$, where each dimension $\phi^i$ gives the azimuth for the $i$'th source or indicates that the source is silent.
State transitions in a Markov model provide a standard statistical framework for dealing with multiple dynamic models (e.g., Koch, 2001). Suppose that the state of the system at frame $m$, $\mathbf{x}_m = (\mathbf{\phi}_m^1, \mathbf{\phi}_m^2, \mathbf{\phi}_m^3)$, is in the subspace $s_m$ and the sources are independent of each other. Then the state transitions are described by:

$$p(\mathbf{x}_m, s_m | \mathbf{x}_{m-1}, s_{m-1}) = p(s_m | s_{m-1}) \prod_{i \in I} p(\mathbf{\phi}_m^i | \mathbf{\phi}_{m-1}^i), \quad (3.2)$$

where $p(s_m | s_{m-1})$ is the jump probability between subspaces, $I$ is the set of active sources at time frame $m$, and $p(\mathbf{\phi}_m^i | \mathbf{\phi}_{m-1}^i)$ gives the temporal evolution of the $i$’th source.

The jump probabilities between state spaces of zero-, one-, two- and three-sources in consecutive time frames are estimated using mixtures of three speech utterances from the TIMIT database. For this, speech activity detection is performed separately on each individual utterance by using a threshold on the signal energy. This enables the detection of the number of active sources at each time frame in the mixture. We assume that at most one source can be turned on or off during one time frame. Also, one-source and two-source subspaces are considered equally probable. The resulting jump probabilities between the eight subspaces are reported in Table 3.1.

We assume that an active source moves slowly following a linear trajectory model with additive Gaussian noise. Also, when a source transitions from silence to activity we assume a uniform distribution in the azimuth space. Therefore the dynamics of the $i$’th source is described by:
where $\text{nil}$ stands for silence, $N(\varphi, \sigma)$ is the Gaussian distribution with mean $\varphi$ and standard deviation $\sigma$ (set to a small value), and $U$ is the uniform distribution in the azimuth range $[-90^\circ, 90^\circ]$.

### 3.4.2 Statistics of ITD and IID

For a particular T-F unit, the normalized cross-correlation function (see Eq. 2.4) has a maximum of 1 when the left and right signals are identical except for a time delay and an intensity difference. This condition is satisfied when only one source is active in the corresponding T-F unit. The computed ITD and IID reflect in this case the actual source location. However, when sources from different locations are all strong in a T-F unit, the left and right mixture signals do not satisfy this condition anymore and the maximum in the normalized cross-correlation function decreases. Moreover, as seen in Chapter 2, ITD and IID deviate from the actual source locations and can indicate phantom sources. Hence, we utilize the peak height of the cross-correlation function as a measure of reliability in individual T-F units. Therefore, a T-F unit is considered reliable (i.e., dominated by only one source) and thus selected if its peak height exceeds a threshold $\theta(c)$. The thresholds $\theta(c)$ are estimated so that 80% of all noisy T-F units are rejected, where a unit is considered noisy if the relative strength $R$ of target with respect to...
interference is less than 0.2. We observe that $\theta(c)$ is a linearly decreasing function with respect to channel index $c$.

For each selected T-F unit, the estimated ITD and IID signal a specific source location. By studying the deviation of the estimated ITD and IID values from the reference values, we can derive the probability of one selected channel supporting a location hypothesis. For each frequency channel, the reference values are obtained from simulated white noise signals at locations in the azimuth range $[-90^\circ, 90^\circ]$. The ITD reference values have been already reported in Section 2.4. As seen in Fig. 2.2, ITD is monotonic with respect to azimuth but has a slight dependency on channel center frequency. Figure 3.3 shows IID reference values for all frequency channels. As seen in the figure, IID is highly dependent on both channel frequency and azimuth.

Consider channel $c$ and azimuth $\phi$ for which the ITD and IID reference values are $\tau_{\text{ref}}(c, \phi)$ and $t_{\text{ref}}(c, \phi)$. For a given T-F unit, we define the ITD and IID deviations as:

$$\delta_t = \tau - \tau_{\text{ref}}(c, \phi), \quad (3.4a)$$

$$\delta_i = t - t_{\text{ref}}(c, \phi), \quad (3.4b)$$

where $\tau$ is the lag of the closest peak in the cross-correlation function to $\tau_{\text{ref}}(c, \phi)$ and $t$ is the estimated IID.

Statistics of the deviations $\delta_t$ and $\delta_i$ are collected separately for each frequency channel across different time frames. Figure 3.4 shows the results of these deviations for
Figure 3.3. IID reference functions for frequency in the range 80 Hz – 5000 Hz and azimuth in the range [-90°, 90°].

a channel with center frequency $f_c$ of 1.5 kHz. The ITD and IID deviations are obtained for the one-source scenario using a small set of 10 utterances from the TIMIT database and various linear motion patterns. As seen in the figure, both histograms are sharply centered at zero. Consequently, we model the joint distribution of ITD and IID deviations in channel $c$ as a combination of a Laplacian distribution and a uniform distribution which models the background noise:
\begin{equation}
\begin{aligned}
p_c(\delta_r, \delta_i) &= (1-q)L(\delta_r, \lambda_r(c))L(\delta_i, \lambda_i(c)) + qU_c(\Delta_r, \Delta_i),
\end{aligned}
\end{equation}

where \(0 < q < 1\) is the noise level. \(U_c(\Delta_r, \Delta_i)\) is the uniform distribution in the plausible range for \(\delta_r \in [-\Delta_r, \Delta_r]\) in lag step and \(\delta_i \in [-\Delta_i, \Delta_i]\) in dB. \(\Delta_r = 20\) and \(\Delta_i = \max(\frac{f_s}{2f_c}, 44)\), where \(f_s\) is the sampling frequency and 44 steps correspond to a delay of 1 ms (see Chapter 2). \(L(\delta, \lambda)\) is the Laplacian distribution with parameter \(\lambda\) defined by:

\begin{equation}
L(\delta, \lambda) = \frac{1}{2\lambda} \exp\left(-\frac{|\delta|}{\lambda}\right).
\end{equation}

We observe that the parameters \(\lambda_r(c), \lambda_i(c)\) are channel dependent: \(\lambda_r(c)\) decreases abruptly with increasing \(c\) (or center frequency) whereas \(\lambda_i(c)\) is slowly increasing. To obtain smooth parameters across channels we use the following approximation:

\begin{equation}
\begin{aligned}
\lambda_r(c) &= a_1 + a_2 / f_c, \\
\lambda_i(c) &= a_3 + a_4 \cdot c.
\end{aligned}
\end{equation}

Similarly, ITD and IID statistics are extracted for multi-source scenarios with two and three active sources. We employ a set of 10 binaural mixtures using the same utterances as before and various linear motion patterns. For a selected T-F unit, the dominant source is obtained by comparing the energies of the individual sources and the ITD and IID.
Figure 3.4. Histogram of estimated ITD and IID deviations from reference values for a channel with $f_c = 1.5$ kHz in the multi-source scenario.

deviations are computed relative to the dominant source. While the deviations exhibit the same peaky distributions as in the one-source scenario, their variance increases due to the interaction between the sources.

The maximum likelihood (ML) method is then used to estimate the parameters $a_1$, $a_2$, $a_3$, and $a_4$ for the one-source and the multi-source scenarios assuming a fixed noise level $q$ across all conditions and frequency channels. This ensures that the background noise and the unreliable channels do not influence the comparison between one-source and
multi-source scenarios. ML estimation gives $q=0.03$. The parameters $a_1$, $a_2$, $a_3$, and $a_4$ are reported in Table 3.2.

### 3.4.3 Likelihood Model

In this subsection we derive the conditional probability density $p(T_c, t_c | x)$, often referred to as the likelihood, which statistically describes what a single frame of ITD and IID observations relate to the joint state $x$ of the source locations to be tracked. Here, $T_c$ is the set of time lags $\tau_c$ corresponding to the local peaks in the cross-correlation function and $t_c$ is the estimated IID for channel $c$. The braces denote all frequency channels.

First, we consider the conditional probability $p(T_c, t_c | x)$ for the one-source subspaces, i.e. $x \in S_1^s \cup S_2^s \cup S_3^s$. For channel $c$, we compute the deviations $\delta_x, \delta_i$ as described in Eq. 3.4 using as reference values $\tau_{ref} (c, \varphi)$ and $t_{ref} (c, \varphi)$ where $\varphi$ refers to the azimuth of the hypothesized active source. Then, the conditional probability of the observations in channel $c$ with respect to the one-source state $x$ is given by:

$$p(T_c, t_c | x) = \begin{cases} p_c (\delta_x, \delta_i), & \text{if channel } c \text{ is selected} \\ qU_c (\Delta_x, \Delta_i), & \text{else} \end{cases},$$

where the symbols are as described in Eq. 3.5 and Eq. 3.7 and the parameters are estimated for the one-source scenario. Note that the uniform background noise is assigned to an unreliable channel.
<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-source</td>
<td>0.1328</td>
<td>59.0497</td>
<td>0.3666</td>
<td>0.0026</td>
</tr>
<tr>
<td>Multi-source</td>
<td>0.1293</td>
<td>500.000</td>
<td>1.2306</td>
<td>0.0071</td>
</tr>
</tbody>
</table>

Table 3.2. Estimated model parameters for one-source and multi-source conditions.

By assuming independence between observations in different channels, the conditional probability in a frame can be easily obtained by multiplying the conditional probabilities in individual channels. However, the observations are usually correlated due to the wideband nature of speech signals and this results in “spiky” distributions. This is known as the probability overshoot phenomenon. To alleviate this problem, the observation probability in the current time frame conditioned on the one-source state $x$ is therefore smoothed using a root operation (Hand and Yu, 2001):

$$p(\{T_c, t_c\} | x) = \kappa^{N_b} \prod_c \sqrt{p(T_c, t_c | x)} ,$$

(3.9)

where $N_b=20$ is the smoothing factor and $\kappa$ is a normalization factor.

Next, we consider the conditional probability $p(\{T_c, t_c\} | x)$ for the two-source case, i.e. $x \in S_2^{1.2} \cup S_2^{1.3} \cup S_2^{2.3}$. Similar to the one-source case, we compute the deviations $\delta_i^x$
and \(\delta_i^k\) with respect to the \(k'\)th hypothesized source, where \(k = 1, 2\). Observe that a selected channel should signal only one source under the assumption that only one speaker dominates a reliable T-F unit. Moreover, all channels whose ITD and IID deviations with respect to the same source are relatively small should support the same source hypothesis. Consequently, we employ a gating technique to associate channels with the hypothesized sources. Specifically, we label channel \(c\) as belonging to the \(k'\)th source if the corresponding deviations satisfy 
\[
|\delta_i^c| < \varepsilon \lambda_i(c) \quad \text{and} \quad |\delta_t^c| < \varepsilon \lambda_t(c)
\]
where \(\varepsilon = 5\) is the gate size. Assume that the \(k'\)th source is the stronger among the two. Then the conditional probability for channel \(c\) under this assumption is given by:

\[
p(T_c, t_c | x, k) = \begin{cases} 
q U_c(\Delta_t, \Delta_r), & \text{if channel } c \text{ not selected} \\
p_c(\delta_t^k, \delta_i^k), & \text{if channel } c \text{ belongs to source } k, \\
\max[p_c(\delta_t^i, \delta_i^i), p_c(\delta_t^2, \delta_i^2)], & \text{else}
\end{cases}
\] (3.10)

where all the parameters are derived for the multi-source case. We apply integration of the individual probabilities across all channels as done in Eq. 3.9 to give the conditional probability \(p(T_c, t_c | x, k)\) for the current time frame under the assumption that the \(k'\)th hypothesized source is the strongest. Finally, the conditional probability \(p(T_c, t_c | x)\) for the current time frame is the larger of assuming either the first or the second hypothesized source to be the stronger source:

\[
p(\{T_c, t_c\} | x) = \alpha_i \max[p(\{T_c, t_c\} | x, 1), p(\{T_c, t_c\} | x, 2)],
\] (3.11)
where \( \alpha_2 \) is used to adjust the relative strength of the two-source space.

Note that without the gating mechanism, Eqs. 3.10 and 3.11 simplify to a simple max operation in the selected channels. However, this operation attempts to overfit the data with a two-source model by assigning the noisy observations produced by one source to two closely spaced sources. The gating mechanism is one way to penalize the overfitting due to noise.

Similar to the two-source case, we consider the conditional probability \( p(T_c, t_c | x) \) for the three-source case, i.e. \( x \in S_3 \). Eqs. 3.10 and 3.11 are easily extensible to three sources by considering all the three-source permutations and utilizing an additional parameter \( \alpha_3 \) to adjust the relative strength of the \( S_3 \) subspace.

After training we fix \( \alpha_n \) as follows: \( \alpha_2 = 1 \) and \( \alpha_3 = e^{-4.25} \). Finally, we fix the probability of the current time frame conditioned on the silence state, i.e. \( x \in S_0 \):

\[
p(\{T_c, t_c\} | x) = \kappa \alpha_0, \quad \text{where} \quad \alpha_0 = e^{-60}.
\] (3.12)

The above \( \alpha_n \) parameters provide different weights for the individual subspaces. In addition to the actual active sources, a few unreliable channels may indicate the presence of a spurious source. The differential weights are needed to avoid this spurious source occurrence.
3.4.4 HMM-Based Source Tracking

For the HMM framework described above, the state space and the time axis are discretized and the standard Viterbi algorithm is employed in order to identify the optimal sequence of states (Jelinek, 1997). The algorithm attempts to reconstruct the initial tracks of the most probable sound sources in the scene. Consequently, the decision of the system at every time frame includes the number of currently active sources and their estimated locations.

The computational cost of our HMM framework is mainly due to the large target space which increases with the maximum number of sources considered. This cost can be reduced significantly by employing several efficient implementation techniques. First, the computations are performed in the log domain thus reducing the number of multiplication and root operations. Second, pruning is used to reduce the number of states to be searched for deciding the current candidate states. Since the original tracks are slowly moving in time, the difference of azimuth in consecutive time frames, hence search, can be restricted considerably. Specifically, we allow an azimuth range of \([-3\sigma, 3\sigma]\) where \(\sigma = 2^\circ\) is the standard deviation in the motion model of individual sources. Finally, beam search is employed to reduce the state space considered in the evaluation of the current time frame. In each time frame, beam searching is performed so that any state whose maximum log probability falls more than 20 below the maximum of all states is not considered.
3.5 Results and Comparison

The performance of the HMM tracking system presented in Section 3.4 is illustrated in this section for two-source and three-source scenarios. As described in Section 3.3, binaural synthesis is used to simulate moving sources in the auditory space of a KEMAR dummy head. Given a binaural mixture as input, the system aims at identifying the number of active speakers at a particular time and constructing continuous trajectories for each of the sources.

Figure 3.5 shows the result of tracking two simultaneous speakers: one male and one female for a duration of 2.5 s. The original speech utterances are equalized to have the same energy level before binaural synthesis. As seen in the figure, the speakers follow a linear motion with respect to the azimuth on the frontal semicircle. The first speaker moves from 40°, which is on the right side of the KEMAR, to -40° on the left side while the second speaker starts at -40° and ends at 40°. Hence, the two trajectories intersect each other in the middle. The system is able to indicate when a source is active and track the two sources across time as long as it is not entirely masked by the interference. Two types of gaps are detected by the system: when the source is silent and when the source is masked across all frequency channels by the other source.

Our system is not restricted to linear motions. Figure 3.6 shows the result of tracking one female and one male speaker moving on two cosine azimuth trajectories that also cross each other in the middle. As above, the two speech utterances are equalized before
Figure 3.5. Source tracking for two crossing sources with linear motion. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘*’ and ‘o’ tracks correspond to the estimated tracks.

Figure 3.6. Source tracking for two crossing sources with nonlinear motion. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘*’ and ‘o’ tracks correspond to the estimated tracks.
binaural synthesis. Note that while the two source locations are correctly identified across time, the system switches the trajectories at the intersection point. However, as seen in Fig. 3.5 our system could disambiguate between two tracks at a crossing point when the likelihood is dominated by a single continuous source. In this example, the source corresponding to the ‘○’ track is dominated by the source corresponding to the ‘*’ track at the crossing point which facilitates the tracking of the latter one. This helps the disambiguation of the two tracks.

Figure 3.7 highlights the robustness of the system to close trajectories. Two male speakers are simulated following nonlinear trajectories with respect to azimuth. The speech utterances are equalized as described above. The two trajectories are symmetric with respect to the median plane. The first speaker oscillates on the right side of the KEMAR while the second trajectory oscillates on the left side. Note that the distance between the two trajectories can be as small as 10° when both speakers approach the median plane. As seen in the figure, the system makes associations and reconstructs the two trajectories. In some cases, a strong source may mask the presence of other sources. Consequently, as seen in Fig. 3.7, this results in gaps in the estimated tracks.

Figure 3.8 shows results for a challenging scenario with three speakers following nonlinear motions. Two males and one female speech utterances are used to obtain the three binaural signals. These signal energies are equalized as described above and the left ear signal for each speaker is displayed in Fig. 3.8(a), Fig. 3.8(b) and Fig. 3.8(c) respectively. As seen in the figure, the system is able to detect the pauses between words in the utterances. Such word level accuracy is required in real speech applications where
Figure 3.7. Source tracking for two sources with closely spaced motions. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘*’ and ‘o’ tracks correspond to the estimated tracks.

the talkers may utter only a few words for the duration of a particular recording. Since we assume that at most only one source can be turned on or off during one time frame, there are no transitions between the 1-source space and the three-source space. This causes the switching of tracks corresponding to the first and the third speakers as seen in Fig. 3.8(d).

Finally we compare our approach with a combination of Kalman filtering and data association techniques as proposed in Sturim et al. (1997) for the tracking of multiple speakers using microphone array measurements. Figure 3.9 shows the extracted tracks using this Kalman filtering approach for the same three source configuration as used in Fig. 3.8. For azimuth estimation, we employ the skeleton cross-correlogram described
Figure 3.8. Tracking three non-stationary moving sources. (a) Left ear signal for the first speaker. (b) Left ear signal for the second speaker. (c) Left ear signal for the third speaker. (d) Continuous tracks obtained by applying our model. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘∗’, ‘○’ and ‘□’ tracks correspond to the estimated tracks.
in Section 2.4 which is similar in principle to the generalized cross-correlation method. First, the time-delay axis for the normalized cross-correlations is mapped to the azimuth axis using the reference ITD values. Next, each peak in the cross-correlation function is replaced with a narrow width Gaussian and all the individual channels are summed together. The results for the summary cross-correlation across time are shown in Fig 3.8a. Here the brighter regions correspond to stronger activities. For an anechoic situation, strong peaks are usually well correlated with the active sources. Hence, at each time frame we select all the azimuths corresponding to the prominent peaks in the summary cross-correlation function. As seen in Fig. 3.9(a), this representation exhibits spurious as well as missing peaks for a considerable number of frames. Smoothing these observations using Kalman filtering can improve the location estimation. In Sturim et al. (1997), the Kalman filter is used for the tracking of single source tracks. In particular, we assume here a second order auto regressive model for the source motion. In addition, a data association algorithm is used to initialize and terminate tracks. The new observations are associated with individual tracks using acceptance regions that take into account the variance of measurement noise and the possible motion of target (Sturim et al., 1997). Observations that cannot be associated with any of the active tracks are used in the initialization of a new track. Note that in their approach there is no correspondence between estimated tracks across time. This differs from our system which uses the continuity of the tracks at the boundaries between the one-, two- and three-source subspaces to reconstruct the individual tracks across time. Note also that our statistical model performs substantially better in estimating the individual source locations.
Figure 3.9. Tracking three non-stationary sources using a Kalman filter approach. (a) Summarized cross-correlation across time. (b) Continuous tracks using the Kalman filter approach. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘o’ tracks correspond to the estimated source locations.
3.5 Discussion

We have proposed a new approach for tracking multiple moving sound sources. The performance of our system is determined by two factors. First is an across-frequency statistical integration method for localization. Second, an HMM framework that imposes continuity constraints across time for individual tracks is combined with a switching mechanism for transition between subspaces consisting of different number of active sources. As a result, the system is able to automatically detect the number of active sources at a given time and to estimate their locations. Such a behavior is highly desirable in speech applications where speakers spontaneously change locations and utter only words in a sporadic way. The power of the system is contingent on the combination of the two factors. It would be interesting to analyze the relative contribution of these two factors to the overall accuracy of the system. In particular, if the first factor is found to be dominant it may be possible to reduce the computational load of the HMM.

Our system can also be applied to the multi-source localization of stationary sources. Figure 3.10 shows such an example with three stationary sources: one female speaker at -30°, one male speaker at 0° and another female speaker at 30°. The signals for the three sources are equalized to have the same average energy at the two ears. To demonstrate the system capability to jump between the subspaces with zero-, one-, two- and three-sources, we let the three speech utterances start and end at different times. As shown in the figure, the system correctly detects the number of sources for the majority of time
Figure 3.10. Source tracking for three stationary sources. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The ‘*’, ‘o’ and ‘□’ tracks correspond to the estimated tracks.

frames. Moreover, the source locations are estimated to within $5^\circ$ of true azimuths. This demonstrates the robust performance of our system in localizing stationary sources. Recall that in Chapter 2 the localization of stationary sources is obtained by summing the cross-correlations across both frequency and time. Each prominent peak in the resulting summary cross-correlation indicates an active source. However, by pooling the cross-correlogram only across frequency in each time frame, spurious peaks and missing peaks will result in significant tracking errors. Hence, tracking of individual sources across time as well as detection of the number of sources at a given time is needed for improved accuracy.
Although the current system does not consider reverberation, our framework is promising for reverberant conditions also. Under reverberation, ITD and IID cues become noisy due to the multiple reflections of a sound source. However, the acoustic onsets are generally unaffected by the reflections and thus could provide reliable ITD and IID information. Therefore, an onset detector can be incorporated in our channel selection stage in order to improve the localization of sound sources.

Although in our simulations we have considered a maximum of three sources, our tracking framework is extensible to an arbitrary number of sources. With increased number of sources, the number of reliable channels decreases and hence the dynamics part of the model should play a more dominant role. To improve the robustness of the likelihood model MAP decoding can also be considered as an alternative to the Viterbi algorithm used in the present study. However, the state space grows exponentially with the number of sources and thus efficient pruning strategies will become increasingly necessary. Also, the system needs to incorporate additional information in order to robustly identify possible direction changes at the crossing points, such as spectral and pitch continuity.
CHAPTER 4

BINAURAL SEGREGATION IN MULTISOURCE REVERBERANT ENVIRONMENTS

In a natural environment, speech signals are degraded by both reverberation and concurrent noise sources. While humans can operate under these conditions using only two ears, current two-microphone algorithms perform poorly when reverberation and multiple noise sources are both present. In this chapter, we propose a binaural CASA system which segregates reverberant target signal from multisource reverberant mixtures by utilizing only the location information of target source. The proposed system combines target cancellation through adaptive filtering followed by a binary decision rule to estimate the ideal T-F binary mask. The key observation is that the attenuation due to target cancellation in a T-F unit is systematically correlated with the relative strength between target and interference. A comprehensive evaluation shows that the proposed system results in large SNR gains across all conditions. In addition, comparisons using SNR as well as automatic speech recognition results show that our system outperforms standard two-
Microphone beamforming approaches as well as a recent binaural processor. A preliminary version of this work has been published in the *Proceedings of 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Roman and Wang, 2004).

4.1 Introduction

As shown in Chapter 2, location-based segregation algorithms can produce very good results for multi-talker scenarios under anechoic conditions. The underlying assumption in those systems is that ITD and IID show location-based characteristic clustering and thus provide reliable cues for sound segregation. In reverberant conditions, however, this anechoic model is inadequate. For each sound source, reverberation introduces potentially an infinite number of sources due to reflections from hard surfaces; each reflection having a different location. As a result, the estimation of ITD and IID in individual T-F units becomes unreliable with the increase of reverberation and the performance of location-based segregation systems degrades under these conditions. A notable exception is the binaural system proposed by Palomäki et al. (2004) which includes an inhibition mechanism that emphasizes the onset portions of the signal and groups them according to common location. The system shows improved speech recognition results across a range of reverberation times. Evaluations under reverberation have also been reported for two-microphone algorithms that combine pitch information...
with binaural cues or other signal-processing techniques (Luo and Denbigh, 1994; Shamsoddini and Denbigh, 2001; Barros et al., 2002).

As discussed in previous chapters, other approaches to sound separation include microphone array beamforming techniques. The fixed beamformer such as that of the delay-and-sum, constructs a spatial beam to enhance signals arriving from the target direction independent of the interfering sources. While being robust to reverberation, this approach requires a large number of microphones in order to construct a narrow beam and achieve good noise cancellation. Adaptive beamforming techniques, on the other hand, attempt to cancel the interference by placing a spatial null at each interfering location. This type of processing is optimal for one interference in anechoic conditions using only two microphones. The performance, however, degrades rapidly when the number of sources or the reverberation level increase.

The model proposed here is motivated by our pursuit of a binaural solution to target segregation in real world conditions with both reverberation and the presence of multiple concurrent sound sources. We achieve sound segregation on the basis of target cancellation through adaptive filtering by observing a correlation between the amount of cancellation produced in individual T-F units and the relative strength between target and interference. Consequently, the input-to-output attenuation level is employed to estimate the ideal binary mask. Since the system depends only on the location of the target, it works for a variety of interfering sources including moving intrusions and impulsive ones. Álvarez et al. (2002) proposed a related system that combines a first-order differential beamformer to cancel the target and obtain a noise estimate and spectral
subtraction to enhance the target source, but their results are not satisfactory in reverberant conditions.

Although the speech reconstructed directly from the ideal binary mask is highly intelligible, ASR systems are sensitive to the small distortions produced during resynthesis and hence do not perform well on the reconstructed signals. Two methods have been proposed to alleviate this problem: the missing data recognizer as used in Chapter 2 and a reconstruction method for the unreliable features proposed by Raj et al. (2004) that is followed by a standard ASR system. While the first method constraints the ASR to operate on spectral features, the second method reconstructs the spectrograms in the front-end and the ASR can capitalize on the advantage of cepstral features over the spectral ones. For evaluation purposes, we use spectrogram reconstruction similar to the one proposed by Raj et al. (2004) and show substantial speech recognition improvements over baseline and other related two-microphone approaches.

The rest of the chapter is organized as follows. The next section defines the problem and describes the model. Section 4.3 gives an extensive evaluation of our system as well as comparison with related models. The last section concludes the chapter.

4.2 Proposed Model

The proposed model consists of two stages as shown in Fig. 4.1. In the first stage, the system performs target cancellation through adaptive filtering. In the second stage, the
Figure 4.1. Schematic diagram of the proposed model. The input signal is a mixture of reverberant target sound and acoustic interference. At the core of the system is an adaptive filter for target cancellation. The output of the system is an estimate of the ideal binary mask.

The input signal shown in Fig. 4.1 can be modeled as, assumes that a desired speech source $s$ has been produced in a reverberant enclosure and recorded by two microphones to produce the signal pair $(x_1, x_2)$. We assume that the transmission path from the target location to the microphones is a linear system and can be modeled as:

$$Y_1 = H_1 S + N_1$$

$$Y_2 = H_2 S + N_2$$

system labels as 1 those T-F units that have been largely attenuated in the first stage since those units are likely to have originated from the target source.
where $h_i(t)$ corresponds to the room impulse response for the $i$th microphone at time $t$. In this problem formulation, the challenge arises when an unwanted interference pair $(n_1, n_2)$ is also added at the input of the microphones. The interference here is a combination of multiple reverberant sources and additional background noise. The target is assumed fixed but no restrictions are imposed on the number, location, or content of the interfering sources. In realistic conditions, the interference can suddenly change its location and may also contain impulsive sounds. Under these conditions, it is hard to localize each individual source in the scene. The goal is therefore to remove or attenuate the noisy background and recover the reverberant target speech based only on the target source location.

Our objective here is to develop an effective mechanism to estimate an ideal binary mask, which selects the T-F units where the local SNR exceeds a threshold of 0 dB. The relative strength between target signal and interference for a T-F unit is defined as:

$$R(f, t) = \frac{|X_1(f, t)|}{|X_1(f, t)| + |N_1(f, t)|},$$  \hspace{1cm} (4.2)$$

where $X_1(f, t)$ and $N_1(f, t)$ are the corresponding Fourier transforms of the reverberant target signal and the noise signal at frequency $f$ and time $t$ for microphone 1 (primary microphone). The noise signal includes all the interfering sources. Thus, a T-F unit is set
to 1 in the ideal binary mask if $R(f,t)$ exceeds 0.5, otherwise it is set to 0. A similar measure has been used in Chapter 2 (see Eq. 2.16) for additive noise conditions.

In the classical adaptive beamforming approach (Griffith and Jim, 1982), the filter learns to identify the differential acoustic transfer function of a particular noise source and thus perfectly cancels only one directional noise source. Systems of this type are unable to cope well with multiple noise sources or diffuse background noise. As an alternative, we propose to use the adaptive filter only for target cancellation and then process the noise reference obtained using a nonlinear method described below in order to obtain an estimate of the ideal binary mask (see also Roman and Wang, 2004). This approach offers a potential solution to the problem of multiple interfering sources in the background.

In the experiments reported here, we assume a fixed target location and the filter in the target cancellation module (TCM) is trained in the absence of interference. A white noise sequence of 10 s duration is used to calibrate the filter. We implement the adaptation using the Fast-Block Least Mean Square (LMS) algorithm with an impulse response of 375 ms length (6000 samples at 16 kHz sampling rate) (Haykin, 2002). After this training phase, the filter parameters are fixed and the system is allowed to operate in the presence of interference. Both the TCM output $Z(f, t)$ and the noisy mixture at the primary microphone $Y_1(f, t)$ are analyzed using a short time-frequency analysis. The time-frequency resolution is 20-ms time frames with a 10-ms frame shift and 257 discrete Fourier transform (DFT) coefficients. Frames are extracted by applying a running Hamming window to the signal.
As a measure of signal suppression at the output of the TCM unit, we define the output-to-input energy ratio as follows:

\[
OIR(f,t) = \frac{|Z(f,t)|^2}{|Y_i(f,t)|^2},
\]

(4.3)
Consider a T-F unit in which noise is zero. Ideally, the TCM module cancels perfectly the target source resulting in zero output and therefore $OIR(f, t) \rightarrow 0$. On the other hand, T-F units dominated by noise are not suppressed by the TCM and thus $OIR(f, t) \gg 0$. Hence, a simple binary decision can be implemented by imposing a decision threshold on the estimated output-to-input energy ratio. The estimated binary mask is 1 in those T-F units where $OIR(f, t) > \theta_{\text{thr}}(f)$ and 0 in all the other units.

Figure 4.2 shows a scatter plot of $R$ and $OIR$ obtained for individual T-F units corresponding to a frequency bin at 1 kHz. The results are extracted from 100 mixtures of reverberant target speech fixed at 0° azimuth mixed with four interfering speakers at -135°, -45°, 45° and 135° azimuths. The room reverberation time, $T_{60}$, is 0.3 s; $T_{60}$ is the time required for the sound level to drop by 60 dB following the sound offset. The input SNR considering reverberant target as signal is 5 dB. Observe that there exists a correlation between the amount of cancellation in the individual T-F units and the relative strength between target and interference. In order to simplify the estimation of the ideal binary mask we have throughout this chapter used a frequency-independent threshold of -6 dB for the output-to-input energy ratio. The -6 dB threshold is obtained from Eq. 4.2 when the reverberant target signal and the noise signal have equal energy. As seen in the figure, the binary masks estimated using this threshold remove most of the noise at the expense of some target speech energy loss. However, informal listening experiments show that the reconstructed signals remain highly intelligible.
4.3 Evaluation and Comparison

We have evaluated our system on binaural stimuli, generated using the room acoustic model described in Palomäki et al. (2004). The reflection paths of a particular sound source are obtained using the image reverberation model for a small rectangular room (6m×4m×3m) (Allen and Berkley, 1979). The resulting impulse response is convolved with the measured head related impulse responses (HRIR) of a KEMAR dummy head in order to produce the two binaural inputs to our system. Specific room reverberation times are obtained by varying the absorption characteristics of room boundaries. The position of the listener was fixed asymmetrically at (2.5m×2.5m×2m) to avoid obtaining near identical impulse responses at the two microphones when the source is in the median plane. All sound sources are presented at different angles at a distance of 1.5 m from the listener. For all our tests, target is fixed at 0° azimuth unless otherwise specified. To test the robustness of the system to various noise configurations we have performed the following tests: 1) an interference of rock music at 45° (Scene 1); 2) two concurrent speakers (one female and one male utterance) at azimuth angles of -45° and 45° (Scene 2); and 3) four concurrent speakers (two female and two male utterances) at azimuth angles of -135°, -45°, 45° and 135° (Scene 3). The initial and the last speech pauses in the interfering utterances have been deleted in conditions Scene 2 and Scene 3 making them more comparable with condition Scene 1. The signals are upsampled to 44.1 kHz and convolved with the corresponding left and right ear HRIRs to simulate the individual sources for the above three testing conditions (Scene 1-Scene 3). Finally, the
Figure 4.3. Comparison between the estimated mask and the ideal binary mask for a five-source configuration. (a) Reverberant target speech. (b) Mixture of target speech presented at $0^\circ$ and four interfering speakers at locations $-135^\circ$, $-45^\circ$, $45^\circ$, and $135^\circ$. The SNR is 5 dB. (c) The mixture spectrogram overlaid by the estimated T-F binary mask. (d) The mixture spectrogram overlaid by the ideal binary mask. The recordings correspond to the left ear microphone.
reverberated signals at each ear are summed and then downsampled to 16 kHz. In all our evaluations, the input SNR is calculated at the left ear using reverberant target speech as signal. While in Scene 2 and Scene 3 the SNR at the two ears is comparable, the left ear is the ‘better ear’ – the ear with higher SNR - in the Scene 1 condition. In the case of multiple interferences, the interfering signals are scaled to have equal energy at the left ear.

The binaural input is processed by our system as described in Section 4.2 in order to estimate the ideal T-F binary mask which is defined with respect to reverberant target energy and using 0- db local SNR criterion. In all our results, the binary mask is computed and the signal is resynthesized at the better ear (left ear). Figure 4.3 demonstrates the performance of our system for Scene 3 where target utterance is the male utterance “Bright sunshine shimmers on the ocean”. The room conditions are T\text{60}=0.3 s and 5 dB input SNR. Observe that the estimated mask is able to estimate well the ideal binary mask especially in the high target energy T-F regions.

To systematically evaluate our segregation system, we use the following performance measures: 1) SNR evaluation using reverberant target speech as signal as well as the ideal binary mask output; and 2) ASR accuracy using our model as front-end. Quantitative comparisons with related approaches are also provided.
Table 4.1. SNR evaluation for a one-source interference (Scene 1).

<table>
<thead>
<tr>
<th>Input SNR</th>
<th>-5 dB</th>
<th>0 dB</th>
<th>5 dB</th>
<th>10 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output SNR</td>
<td>6.36</td>
<td>11.55</td>
<td>15.87</td>
<td>19.69</td>
</tr>
<tr>
<td>RSR</td>
<td>59</td>
<td>74</td>
<td>84</td>
<td>91</td>
</tr>
</tbody>
</table>

Table 4.2. SNR evaluation for a two-speaker interference (Scene 2).

<table>
<thead>
<tr>
<th>Input SNR</th>
<th>-5 dB</th>
<th>0 dB</th>
<th>5 dB</th>
<th>10 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output SNR</td>
<td>4.82</td>
<td>10.18</td>
<td>14.68</td>
<td>18.54</td>
</tr>
<tr>
<td>RSR</td>
<td>58</td>
<td>73</td>
<td>83</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 4.3. SNR evaluation for a four-speaker interference (Scene 3).

<table>
<thead>
<tr>
<th>Input SNR</th>
<th>-5 dB</th>
<th>0 dB</th>
<th>5 dB</th>
<th>10 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output SNR</td>
<td>3.41</td>
<td>8.94</td>
<td>13.68</td>
<td>17.79</td>
</tr>
<tr>
<td>RSR</td>
<td>52</td>
<td>66</td>
<td>79</td>
<td>89</td>
</tr>
</tbody>
</table>

4.3.1 SNR Evaluation

We perform SNR evaluations for the three conditions described above using 10 speech signals from the TIMIT database as target: five female utterances and five male
utterances. Results are given in Table 4.1, Table 4.2 and Table 4.3. The room reverberation time is 0.3 s in all conditions and the system is evaluated for the following four input SNR values: -5 dB, 0 dB, 5 dB and 10 dB. The room reverberation time is 0.3 s in all conditions and the system is evaluated for the following four input SNR values: -5 dB, 0 dB, 5 dB and 10 dB. In order to assess the system performance, output SNR and retained speech ratio (RSR) are computed as follows (see Wang and Brown, 1999):

\[
\text{Output SNR} = 10 \log_{10} \left( \frac{\sum_t s_E^2(t)}{\sum_t n_E^2(t)} \right),
\]

(4.4)

\[
\text{RSR} = \sum_t s_E^2(t) \left/ \sum_t s_T^2(t) \right.,
\]

(4.5)

where \( s_T(t) \) is the reverberant target signal resynthesized through an all-one mask, \( s_E(t) \) is the reverberant target signal resynthesized through the estimated mask, and \( n_E(t) \) is the noise signal resynthesized through the same mask. While the output SNR measures the level of noise that remains in the reconstructed signal, the RSR measures the percentage of target energy loss. The results show SNR improvements in the range of 8-11 dB while preserving much of the target energy (~70-90%) for input SNR levels greater than 0 dB. Observe that the system performance degrades at lower SNR values because of increased overlap between target and interference. The RSR may be improved by imposing a higher threshold on the output-to-input attenuation level at the expense of increasing the residual noise in the output signal.
Table 4.4 shows the performance of our system for six reverberation times between 0.0 s (anechoic) and 0.5 s (e.g., large living rooms and classrooms) which are obtained by simulating room impulse responses with different room absorption characteristics. Results are reported for Scene 1 and 0 dB input SNR. For each room configuration, the filter in the TCM module is adapted using 10 s of white noise simulated at target location as mentioned earlier. Overall, the system performance degrades by 8 dB output SNR when $T_{60}$ is 0.2 s compared to the anechoic case while preserving the same RSR. This is partly due to the smearing of the spectrogram of individual sources as reverberation time increases which results in increased overlap between target and interference.

We compare our algorithm with the standard delay-and-sum beamformer which is computationally simple and requires no knowledge about the interfering sources. As discussed in the introduction, while fixed beamformers are computationally simple and
require only the target direction, they require a large number of microphones to obtain a good resolution. For our two-microphone configuration, the delay-and-sum beamformer produces only an average of 1.2 dB SNR gain across all three conditions.

To compare our model with adaptive beamforming techniques, we implement the two-stage adaptive filtering strategy described in Van Compernolle (1990) that improves the classic Griffith-Jim model under reverberation. The first stage is identical to our target cancellation module and is used to obtain a good noise reference. The second stage uses another adaptive filter to model the difference between the noise reference and the noise portion in the primary microphone in order to extract the target signal. Here, training for the second filter is done independently for each noise condition (Scene 1 - Scene 3) in the absence of target signal using 10 s white noise sequences presented at each location in the tested configuration. The length of the filter is the same as the one used in the TCM (375 ms). Note that this approach requires adaptation for any change in both target source location as well as any interfering source location. As expected, the adaptive beamformer is optimal for canceling out one interfering source and hence gives in the Scene 1 condition an SNR gain of 13.61 dB. However, the second adaptive filter is not able to adapt to the noise configuration when multiple interferences are active since each source has a different differential path between the microphones. The adaptive beamformer thus produces an SNR gain of 3.63 dB in the Scene 2 condition and only 2.74 dB in the Scene 3 condition. The advantage for both the fixed beamformer as well as the adaptive one is that target signal distortions are minimal in the output when the filters are calibrated. By comparison, our system introduces some target energy loss. However,
note that in the Scene 3 condition our system produces an SNR gain of 8 dB while losing less than 30% energy in the target signal for input SNR levels greater than 0 dB.

Given our computational objective of estimating the ideal binary mask, we also employ an SNR evaluation that uses the signal reconstructed from the ideal binary mask as ground truth (see Hu and Wang, 2004):

$$SNR_{IBM} = 10 \log_{10} \frac{\sum_{t} s_{IBM}^2(t)}{\sum_{t} (s_{IBM}(t) - s_E(t))^2}$$, \hspace{1cm} (4.6)

where $s_{IBM}(t)$ represents the target signal reconstructed using the ideal binary mask and $s_E(t)$ is the estimated target reconstructed from the binary mask produced by our model. The main advantage of Eq. 4.6 is that it provides a single measure that accounts for both retained noise as well as target distortion. Table 4.5 provides a comparison between our proposed system and the adaptive beamformer using this SNR measure. In order to extend the evaluation to the adaptive beamformer, the waveform at the beamformer output needs to be converted into a binary mask representation. Assuming target energy and noise energy are uncorrelated in individual T-F units, we can construct a binary mask as follows. For each T-F unit, if the energy ratio between the beamformer output and the input mixture is greater than 0.5 we label the unit as 1, otherwise we label the unit as 0. The signal resynthesized by applying this mask to the output waveform is used in Eq. 4.6 as the estimated target. As seen in Table 4.5, our system provides consistent improvements over the adaptive beamformer in low input SNR scenarios with multiple
Table 4.5. Comparison of the proposed system in terms of SNR with adaptive beamforming.

interferences (Scene 2 and Scene 3).

A combination of target cancellation using a first-order differential beamformer and a spectral subtraction technique has been proposed previously by Álvarez et al. (2002). Since the first stage of our system produces a noise estimate, alternatively we can combine our adaptive filtering stage with spectral subtraction to enhance the reverberant target signal.

4.3.2 ASR Evaluation

We also evaluate the performance of our system as a front-end to a robust ASR system. The task domain is speaker independent recognition of connected digits and the
word level HMMs are the same as used in Chapter 2. Training is performed using the 4235 clean signals from the male speaker dataset in the TIDigits database downsampled to 16 kHz to be consistent with our model. Testing is performed on a subset of the testing set containing 229 utterances from 3 speakers which is similar to the test used in Palomäki et al. (2004). The test speakers are different from the speakers in the training set. The test signals are convolved with the corresponding left and right ear target impulse responses and noise is added as described above to simulate the three conditions Scene 1-Scene 3.

Recall that in Chapter 2 we have evaluated our estimated binary mask using a missing-data recognizer that operates with spectral features. While this approach has shown promising results with additive noise in anechoic conditions, extension to reverberant conditions has proved to be problematic (for details see Palomäki et al., 2004). Therefore, we adapt the spectrogram reconstruction method proposed by Raj et al. (2004) to reverberant environments as described below, which shows improved performance over the missing-data recognizer.

We have trained the HMMs with clean utterances from the training data using feature vectors consisting of the 13 mel-frequency cepstral coefficients (MFCC) including the zeroth order cepstral coefficient, $C_0$, as the energy term together with their first and second order temporal derivatives. MFCCs are used as feature vectors as they are most commonly used in state-of-the-art recognizers (Rabiner and Juang, 1993). Cepstral mean normalization (CMN) is applied to the cepstral features in order to improve the robustness of the system under reverberant conditions (Shire, 2000). Frames are extracted
using 20 ms windows with 10 ms overlap. A first-order preemphasis coefficient of 0.97 is applied to the signal. The recognition result using clean test utterances is 99% accuracy. Using the reverberated test utterances, performance degrades to 94% accuracy.

The CMN applied on the MFCC features provides a relatively robust front-end for our task domain under the moderate reverberant conditions considered here. Hence, a reasonable approach is to remove the noise component from our acoustic mixture in the front-end processor and to feed an estimate of the reverberant target to the MFCC-based ASR. Although subjective listening tests have shown that the signal reconstructed from the ideal binary mask is highly intelligible (Roman et al., 2003; Chang, 2004; Brungart et al., 2005), the extraction of MFCC features from such a mask does not have high quality due to the mismatch arising from the T-F units labeled 0, which smears the entire cepstrum through the cepstral transform (Cooke et al., 2001). A similar problem occurs when our second stage is replaced by spectral subtraction since spectral subtraction performs poorly in the T-F regions dominated by interference where over-subtraction or under-subtraction occurs. One way to handle this problem is by estimating the original spectral values in those T-F units using a prior speech model. This approach has been suggested by Raj et al. (2004) in the context of additive noise. In this approach, a noisy spectral vector $Y$ at a particular frame is partitioned in its reliable $Y_r$ and its unreliable and $Y_u$ components. The task is to reconstruct the underlying true spectral vector $X$. Assuming that the reliable features $Y_r$ are approximating well the true ones $X_r$, a Bayesian decision is then employed to estimate the remaining $X_u$ given only the reliable component. Hence, this approach works seamlessly with the T-F binary mask that our speech segregation
system produces. Here, the reliable features are the T-F units labeled 1 in the mask while the unreliable features are the ones labeled 0. Although the reliable data in our system contains some reverberation, we train the prior speech model only on the clean data. This actually avoids the trouble of obtaining a prior for each deployment condition, and is desirable for robust speech recognition.

The speech prior is modeled empirically as a mixture of Gaussians and trained with the same clean utterances used in ASR training:

\[
p(X) = \sum_{k=1}^{M} p(k) p(X | k),
\]

where \(M=1024\) is the number of mixtures, \(k\) is the mixture index, \(p(k)\) is the mixture weight and \(p(X | k) = N(X; \mu_k, \Sigma_k)\).

Previous studies (Cooke et al., 2001; Raj et al., 2004) have shown that a good estimate of \(X_u\) is its mean conditioned on \(X_r\):

\[
E_{X_u | X_r, 0 \leq X_u \leq Y_u} (X_u) = \sum_{k=1}^{M} p(k | X_r) \int_{0}^{Y_u} X_u p(X_u | k, 0 \leq X_u \leq Y_u) dX_u,
\]

where \(p(k | X_r)\) is the \textit{a posteriori} probability of the \(k\)'th Gaussian given the reliable data and the integral denotes the expectation \(\bar{X}_u\) corresponding to the \(k\)'th mixture. Note that under the additive noise condition, the unreliable parts may be constrained as \(0 \leq X_u \leq Y_u\) (Cooke et al., 2001). In our implementation, we have assumed that the prior can be modeled using a mixture of Gaussians with diagonal covariance. Theoretically,
this is a good approximation if an adequate number of mixtures are used. Additionally, our empirical evaluations have shown that for the case of $M=1024$ this approximation results in an insignificant degradation in recognition performance while the computational cost is greatly reduced. Hence, the expected value can now be computed as:

$$\hat{X}_u = \begin{cases} 
\mu_{u,k}, & 0 \leq \mu_{u,k} \leq Y_u \\
Y_u, & \mu_{u,k} > Y_u \\
0, & \mu_{u,k} < 0
\end{cases}$$  \hspace{1cm} (4.9)

The \textit{a posteriori} probability of the $k$'th mixture given the reliable data is estimated using the Bayesian rule from the simplified marginal distribution $p(X_r | k) = N(X_r; \mu_{r,k}, \sigma_{r,k})$ obtained without utilizing any bounds on $\hat{X}_u$. While this simplification results in a small decrease in accuracy, it gives substantially faster computation of the marginal.

Speech recognition results for the three conditions Scene 1, Scene 2 and Scene 3 are reported separately in Fig. 4.4, Fig. 4.5 and Fig. 4.6 at five SNR levels: –5 dB, 0 dB, 5 dB, 10 dB and 20 dB. Results are obtained using the same MFCC-based ASR as the back-end for the following approaches: fixed beamforming, adaptive beamforming, target cancellation through adaptive filtering followed by spectral subtraction, our proposed front-end ASR using the estimated mask and finally our proposed front-end ASR using the ideal binary mask. The baseline results correspond to the unprocessed left ear signal. Observe that our system achieves large improvements over the baseline performance across all conditions. Note that the ASR performance depends on the interference type.
Figure 4.4. Recognition performance for Scene 1 at different SNR values for the reverberant mixture (⋆), a fixed beamformer (▼), an adaptive beamformer (▲), a system that combines target cancellation and spectral subtraction (■), our front end ASR using the estimated binary mask (●), our front-end ASR using the ideal binary mask (◆).

and we obtain the best accuracy score in the two speaker interference. As seen also in the SNR evaluation, the adaptive beamformer outperforms all the other algorithms in the case of a single interference (Scene 1). However, as the number of interferences increases, the performance of the adaptive beamformer degrades rapidly and approaches
Figure 4.5. Recognition performance for Scene 2 at different SNR values for the reverberant mixture (*), a fixed beamformer (▼), an adaptive beamformer (▲), a system that combines target cancellation and spectral subtraction (■), our front end ASR using the estimated binary mask (●), our front-end ASR using the ideal binary mask (♦).

the performance of the fixed beamformer in the Scene 3 condition. As described in the previous subsection, we can combine our adaptive filtering stage with spectral subtraction to cancel the interference. As illustrated by the recognition results in Fig. 4.5 and Fig. 4.6, this approach outperforms the adaptive beamformer in the case of multiple concurrent
interferences. While spectral subtraction improves the SNR gain in target-dominant T-F units, it does not produce a good target signal estimate in noise-dominant regions. Note that our front-end ASR employs a better estimation of the spectrum in these unreliable T-F units and therefore results in large improvements over the spectral subtraction method. Although the results using our front-end ASR show substantial performance gains,
further improvement can be achieved as can be seen in the results reported for the ideal binary mask.

We compare our system with the binaural system proposed by Palomäki et al. (2004) which produces substantial recognition improvements on the same digit recognition task as used here. Their system combines binaural localization with precedence effect processing in order to detect reliable spectral regions that are not contaminated by interfering noise or echoes. Recognition is then performed in the log spectral domain by employing the missing data ASR system. In order to account for the reverberant environment, spectral energy normalization is employed. While our system can handle a variety of interfering sources, the binaural system of Palomäki et al. was developed for only one-interference scenarios. Table 4.6 compares the two systems for the case of one interfering source of rock music, which was used in Palomäki et al. The recognition results for the Palomäki et al. system as reported by the authors while the results for our system have been produced using the same configuration setup. Listener is located in the middle of the room while target and interfering sources are located at 20° and -20° respectively. $T_{60}$ is 0.3 s and the input SNR is fixed before the binaural presentation of the signals at three SNR levels: 0 dB, 10 dB and 20 dB. Note that we obtain a marked improvement over the system of Palomäki et al., in the low SNR conditions. By utilizing ITD and IID information only during acoustic onsets, the mask obtained by their system has a limited number of reliable units. This limits the amount of information available to the missing data recognizer for the decoding (Srinivasan et al., 2004). In our system, on
the other hand, a novel encoding of the target source location leads to the recovery of more target dominant regions and this results in a more robust front-end for ASR.

We further compare our system with the negative beamforming approach proposed by Álvarez et al. (2002) and the results are reported in Table 4.7. In order to compare with this approach we simulate the input for a two-microphone array with 5 cm intermicrophone distance using the image reverberation model (Allen and Berkley, 1979). We use the same room configuration, the same interfering signals and the same spatial configuration as in the Scene 3 condition described above. The system proposed by Álvarez et al. uses a first-order differential beamformer to cancel the direct path of the target signal. Since target is fixed at 0°, the adaptation parameter in the differential beamformer is fixed to 0.5 across all frequencies. The output of the differential beamformer contains both the reverberant path of the target signal as well as an estimate of the additional interfering sources. An additional frequency-equalizing curve is applied on this output since the amount of attenuation performed by this beamformer varies with the frequency of the signal as well as its location. This equalizing-curve is trained using white noise at the corresponding interfering locations. The estimated noise spectrum is finally subtracted from the spectrum of one of the two microphone mixtures (the left one) and the results are fed to the MFCC-based ASR. Our system is trained on the new configuration to obtain the TCM adaptive filter using the parameters described in Section 4.2. The T-F mask produced by our system is then used to reconstruct the spectrogram using the prior speech model. As Table 4.7 shows, our system significantly outperforms the system of Álvarez et al. (2002) for a range of SNRs.
Table 4.6. Comparison in terms of speech recognition accuracy with the binaural system proposed by Palomäki et al. (2004).

<table>
<thead>
<tr>
<th>Input SNR</th>
<th>0 dB</th>
<th>10 dB</th>
<th>20 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (MFCC+CMN)</td>
<td>13.04</td>
<td>43.01</td>
<td>81.85</td>
</tr>
<tr>
<td>Palomäki et al. system</td>
<td>32.7</td>
<td>78.8</td>
<td>91.9</td>
</tr>
<tr>
<td>Proposed system</td>
<td>47.58</td>
<td>81.59</td>
<td>91.80</td>
</tr>
</tbody>
</table>

Table 4.7. Comparison in terms of speech recognition accuracy with the negative beamformer system proposed by Álvarez et al. (2002).

<table>
<thead>
<tr>
<th>Input SNR</th>
<th>0 dB</th>
<th>10 dB</th>
<th>20 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (MFCC+CMN)</td>
<td>11.69</td>
<td>40.99</td>
<td>82.80</td>
</tr>
<tr>
<td>Álvarez et al. system</td>
<td>24.14</td>
<td>51.61</td>
<td>73.39</td>
</tr>
<tr>
<td>Proposed system</td>
<td>31.59</td>
<td>75.00</td>
<td>91.94</td>
</tr>
</tbody>
</table>

4.4 Discussion

In natural settings, reverberation alters many of the acoustical properties of a sound source reaching our ears, including smearing out the binaural cues due to the presence of multiple reflections. This is especially detrimental when multiple sound sources are present in the acoustic scene since the acoustic cues are now required to distinguish
between the competing sources. Location based algorithms that rely on the anechoic assumption of time delayed and attenuated mixtures are therefore prone to failure in reverberant scenarios. An adaptive filter can therefore be used to better characterize the target location in a reverberant room. We have presented in this chapter a novel two-microphone sound segregation system that performs well under such realistic conditions. Our approach is based on target cancellation through adaptive filtering followed by an analysis of the output-to-input attenuation level in individual T-F units. The output of the system is an estimate of an ideal binary mask which labels the T-F components of the acoustic scene dominated by the target sound.

Classical two-microphone noise cancellation strategies process the input using linear adaptive filters and while being optimal in the one-interference condition, they are unable to cope with multiple interferences. By using a nonlinear strategy in the second stage, our system is able to cancel an arbitrary number of interferences using only two microphones. As shown in our SNR evaluation, the system is able to outperform existing beamforming techniques across a range of input SNRs. Note that while our processing produces some target signal distortion, we preserve most of the target energy (> 70%) at input SNRs greater than 0 dB. The balance between noise cancellation and target distortion can be controlled in our system by varying the output-to-input attenuation threshold. In high SNR conditions, for example, a more relaxed threshold will ensure less target distortion at the expense of some background noise. Further, informal listening tests show that the reconstructed signals are highly intelligible.
Since the first stage of our system provides a noise estimate, an alternative nonlinear strategy for the second stage is spectral subtraction. A combination of target cancellation through differential beamforming and spectral subtraction has been proposed previously by Álvarez et al. (2002). Informal listening tests have shown that the signals obtained using spectral subtraction in the second stage of your system have a similar quality to the ones resynthesized from the estimated binary masks. An SNR evaluation using the reverberant target as signal shows a slight improvement using the spectral subtraction method. However, as seen in the ASR evaluation, the binary masks provide a complimentary advantage when coupled with missing data techniques and can provide sizeable ASR improvements. Further improvement can be obtained by using spectral subtraction in combination with our binary mask estimation. For example, we observe additional improvements (3%-5% absolute percentages) when using spectral subtraction to clean the reliable regions prior to the spectrogram reconstruction.

In terms of application to real-world scenarios our adaptive filtering strategy has several drawbacks. First, the adaptation of the inverse filter requires data on the order of a few seconds and thus any fast change in target location (e.g. walking) will have an adverse impact on the system. Second, the system needs to identify signal intervals that contain no interference to allow for the filter to adapt to a new target position. On the other hand, note that our system requires training only with respect to target location and is therefore insensitive to changes in the locations of interfering sources, unlike adaptive beamforming whose training is conditioned on the positions of all sound sources in the environment.
We use the approach proposed by Raj et al. (2004) to reconstruct the clean target signal in the unreliable T-F units. This allows for our system to be utilized as a front-end to a standard speech recognition system operating using cepstral features. In a systematic comparison, our system shows substantial performance gains over baseline and significant improvements over related approaches. Note that our prior and ASR models are trained on clean speech and hence our algorithm is applicable when recognition in changing reverberant environments is desired. However, if samples of reverberant target are available or dereverberation techniques become applicable we can utilize these to further improve ASR performance.
CHAPTER 5

MONAURAL SEGREGATION OF REVERBERANT SPEECH

While previous chapters have addressed speech segregation in binaural conditions, psychoacoustic evidence suggests that monaural processing plays a dominant role in human speech perception. The auditory system is capable of segregating sources that originate from nearby or same location; in these conditions spatial filtering does not apply and segregation must rely on monaural features. Speech segregation based on periodicity has achieved considerable progress in handling additive noise. However, little research in monaural segregation has been devoted to reverberant scenarios. Reverberation smears the harmonic structure of speech signals, and our evaluations using a pitch-based segregation algorithm show that an increase in the room reverberation time causes a degradation in performance due to the loss in periodicity for the target signal. In this chapter, we propose a two-stage monaural separation system that combines the inverse filtering of the room impulse response corresponding to target location with a pitch-based speech
segregation method. As a result of the first stage, the harmonicity of a signal arriving from target direction is partially restored while signals arriving from other locations are further smeared, and this leads to improved segregation in the second stage. By using the harmonicity of target source as the segregation cue, the system is able to handle various scenarios including where interfering sources originate at target location. In this situation, both target as well as interfering sources are dereverberated. This again facilitates the successful application of our pitch-based segregation stage. A systematic evaluation of the system shows that the proposed system results in considerable signal-to-noise ratio gains across different conditions. A preliminary version of this work will appear in the *Proceedings of 2005 INTERSPEECH* (Roman and Wang, 2005).

5.1 Introduction

In a natural environment, a desired speech signal often occurs simultaneously with other interfering sounds such as echoes and background noise. In this chapter, we study the monaural separation of reverberant speech. Our monaural study is motivated by the following two considerations. First, an effective one-microphone solution to sound separation is more desirable in applications such as automatic speech recognition and hearing prosthesis. Second, although binaural listening improves the intelligibility of target speech under anechoic conditions (Bronkhorst, 2000), this binaural advantage disappears when the interfering sources come from the same direction or nearby
directions. Also, it appears that the binaural advantage is diminished under reverberation (Plomp, 1976; Culling et al., 2003). All these emphasize the dominant role of monaural hearing in realistic conditions.

Perceptually, one of the most effective segregation cues used by the auditory system is the fundamental frequency (F0) (Darwin and Carlyon, 1995). Accordingly, much work has been devoted to build computational systems that exploit the F0 of a desired source to segregate its harmonics from the interference (e.g., Weintraub, 1986; Cooke, 1993; Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004). In particular, the system proposed by Hu and Wang (2004) exploits a differential strategy to segregate resolved and unresolved harmonics. More specifically, periodicities detected in the response of a cochlear filterbank are used at low frequencies to segregate resolved harmonics. In the high-frequency range, however, the cochlear filters have wider bandwidths and a number of harmonics interact within the same filter, causing amplitude modulation (AM). In this case, their system exploits periodicities in the response envelope to group unresolved harmonics. In this chapter, we propose a pitch-based speech segregation method that follows the same principles while simplifying the calculations required for extracting periodicities. The system shows good performance when tested with a variety of noise intrusions under anechoic conditions. However, when the pitch varies with time in a reverberant environment, reflected waves with different F0s arrive simultaneously with the direct sound at the ear. This multipath situation causes smearing of the signal in the sense that harmonic structure is less clear in the signal
Due to the loss of harmonicity, the performance of pitch-based segregation degrades in reverberant conditions.

One method for removing the reverberation effect is to pass the reverberant signal through a filter that inverts the reverberation process and hence reconstructs the original signal. Due to the non minimum-phase nature of a typical room impulse response, a perfect one-microphone reconstruction requires a noncausal infinite impulse response filter with a large delay (Neely and Allen, 1979). Exact inverse filtering using causal finite impulse responses can be obtained using multiple microphones by assuming no common zeros among the different room impulse responses (Miyoshi and Kaneda, 1988). Inverse filtering techniques which partially dereverberate the reverberant signal have also been studied (Gillespie and Atlas, 2002). However, these algorithms assume a priori knowledge of the room impulse responses, which is often impractical. Several strategies have been proposed to estimate the inverse filter in unknown acoustical conditions (Furuya and Kaneda, 1997; Gillespie et al., 2001; Nakatani and Miyoshi, 2003). In particular, the system developed by Gillespie et al. (2001) estimates the inverse filter from an array of microphones using an adaptive gradient-descent algorithm that maximizes the kurtosis of linear prediction (LP) residuals. The restoration of LP residuals results in both a reduction of perceived reverberation as well as an improvement of spectral fidelity in terms of harmonicity. In this chapter, we employ a one-microphone adaptation of this strategy proposed by Wu (2003; Wu and Wang, 2005).

The dereverberation algorithms described above are designed to enhance a single reverberant source. Here, we investigate the effect of inverse filtering as pre-processing.
for a pitch-based speech segregation system in order to improve its robustness in a reverberant environment. The key idea is to estimate the filter that inverts the room impulse response corresponding to the target source. The effect of applying this inverse filter on the reverberant mixture is two-fold: it improves the harmonic structure of target signal while smearing those signals originating at other locations. Using an SNR evaluation, we show that the inverse filtering stage improves the separation performance of the proposed pitch-based system. To our knowledge, the proposed system is the first study that addresses monaural speech segregation with room reverberation.

The rest of the chapter is organized as follows. The next section defines the problem domain and presents a model overview. Section 5.3 gives a detailed description of the dereverberation stage employed in this chapter. Section 5.4 gives a detailed description of the proposed pitch-based segregation stage. Section 5.5 presents systematic results on pitch-based segregation both in reverberant and inverse filtered conditions. We also make a comparison with the spectral subtraction method. Section 5.6 concludes the chapter.

5.2 Model Overview

The speech received at one ear in a reverberant enclosure undergoes both convolutive and additive distortions (see also (4.1)):

\[ y(t) = h(t) * s(t) + n(t) , \]  

(5.1)
Figure 5.1. Schematic diagram of the proposed two-stage model.

where ‘∗’ indicates convolution. \( s(t) \) is the clean speech signal to be recovered, \( h(t) \) models the acoustic transfer function from target speaker location to the ear, and \( n(t) \) is the reverberant background noise which usually contains interfering sources at other locations. As explained in the introduction, the problem of monaural speech segregation has been studied extensively in the additive condition by employing the periodicity of target speech. However, room reverberation poses an additional challenge by smearing the spectrum and weakening the harmonic structure. Consequently, we propose a two-stage speech segregation model: 1) inverse filtering with respect to target location in
order to enhance the periodicity of target signal; 2) pitch-based speech segregation. Figure 5.1 illustrates the architecture of the proposed model.

The input to our model is the left ear response of a KEMAR dummy head to two or more sound sources in a small reverberant room, simulated using the same configuration as used in Chapter 4.3. Note that two different positions in the room produce impulse responses that differ greatly in their structure. The mixture signals are finally sampled at 16 kHz.

In the first stage, a finite impulse response filter is estimated that inverts the target room impulse response \( h(t) \). Adaptive filtering strategies for estimating this filter are sensitive to background noise (Haykin, 2002). For simplicity, here we perform this estimation during an initial training stage in the absence of noise. We employ the inverse filtering strategy proposed by Gillespie et al. (2001), which is a practical system using a relatively small amount of training data. This method exploits the fact that the signal to be recovered is speech by employing an LP-based metric and produces improved harmonicity for the target source. The inverse filter is applied to the entire mixture and the result is fed to the next stage.

In the second stage, a pitch-based segregation system is employed to separate the inverse-filtered target signal from other interfering sounds. The signal is analyzed using a gammatone auditory filterbank in consecutive time frames to produce a time-frequency decomposition. A standard mechanism for periodicity extraction employs a correlogram which is a collection of autocorrelation functions computed at individual filters (Licklider, 1951; Slaney and Lyon, 1993). For a particular T-F unit in the low-frequency
range, the autocorrelation faithfully encodes its periodicity information. In the high-frequency range, the filters have a wide bandwidth and multiple harmonics activate the same filter, thus creating beats at a rate corresponding to the fundamental period (Helmholtz, 1863). Such amplitude modulation can be detected using the envelope-based autocorrelation. Our system employs a peak selection mechanism to reveal likely periodicities in the autocorrelation functions of individual T-F units. Further, the system decides whether the underlying target is stronger than the combined interference by comparing these periodicities with a given target pitch.

However, labeling at the T-F unit level is a very local decision and prone to noise. Following Bregman’s conceptual model, previous CASA systems employ an initial segmentation stage followed by a grouping stage in which segments likely to originate from the same source are grouped together (see e.g. Wang and Brown, 1999). By definition, a segment is composed of spatially contiguous units dominated by a single sound source. Hence, grouping at the segment level improves the system robustness compared to the simple T-F labeling. Here, we combine the unit labeling described above with the segmentation framework proposed by Hu and Wang (2004). First, segments in the low-frequency range are generated using cross-channel correlation and temporal continuity. These segments are grouped into a target stream and a background stream according to the labeling of their T-F components. Similarly, segments are added to the target stream in the high-frequency range using envelope-based cross-channel correlation. The result of this process is a binary mask that assigns 1 to all the T-F units in the target
stream and 0 otherwise. Finally, a speech waveform is resynthesized from the resulting binary mask using the method described in Section 2.2.

5.3 Target Inverse Filtering

As described in section 5.1, inverse filtering is a standard method used for deriving the original target signal. We employ the method proposed by Gillespie et al. (2001) which attempts to blindly estimate the inverse filter from reverberant speech data. Based on the observation that peaks in the LP residual of speech are smeared under reverberation, an online adaptive algorithm estimates the inverse filter by maximizing the kurtosis of the inverse-filtered LP residual of reverberant speech $\tilde{z}(t)$:

$$\tilde{z}(t) = q^T y_r(t),$$

(5.2)

where $y_r(t) = [y_r(t-1), ..., y_r(t-L+1), y_r(t)]$ and $y_r(t)$ is the LP residual of the reverberant speech from the target source, and $q$ is an inverse filter of length $L$. The inverse filter is derived by maximizing the kurtosis of $\tilde{z}(t)$, which is defined as:

$$J = \frac{E[\tilde{z}^4(t)]}{E^2[\tilde{z}^2(t)]} - 3.$$  

(5.3)

The gradient of the kurtosis with respect to the inverse filter $q$ can be approximated as follows (Gillespie et al., 2001):
\[
\frac{\partial J}{\partial \hat{q}} \approx \left\{ \frac{4 \left( E[\hat{z}^2(t)]\hat{z}^3(t) - E[\hat{z}^4(t)]\hat{z}(t) \right)}{E[\hat{z}^2(t)]} \right\} y_r(t). \tag{5.4}
\]

Consequently, the optimization process in the time-domain is given by the following update equation:

\[
\hat{q}(t+1) = \hat{q}(t) + \mu f(t)\hat{y}_r(t), \tag{5.5}
\]

where \( \hat{q}(t) \) is the estimate of the inverse filter at time \( t \), \( \mu \) denotes the update rate and \( f(t) \) denotes the term inside the braces of Eq. 5.4.

In general, the LP model attempts to fit the spectrum of the reverberant speech signal. In a particular frame, it may be hard to distinguish between the contributions due to the room impulse response and those due to the LP speech spectrum. However, while speech is nonstationary and consequently has an LP spectrum which varies across time frames, the room impulse response remains constant. Hence, due to the averaging across training samples the inverse filter learns only the characteristics of the room impulse response.

A direct time-domain implementation of the above update equation is not desirable since the requirement for a long filter results in a significant increase in the computational complexity of the algorithm (Haykin, 2002). Frequency-domain adaptive filters provide a solution to this problem. Here, we use the fast-block LMS implementation for one microphone signals described by Wu and Wang (2005). The signal is processed block by block using a size \( L \) for both filter length and block length using the following update equations:
\[ Q'(n+1) = Q(n) + \mu \sum_{k=1}^{M} F(k) Y^*_r(k), \quad (5.6) \]

\[ Q(n+1) = \frac{Q'(n+1)}{|Q'(n+1)|}, \quad (5.7) \]

where \( F(k) \) and \( Y_r(k) \) represent the fast Fourier transform (FFT) of \( f(t) \) and \( y_r(t) \) for the \( k \)'th block, and \( Q(n) \) represents the estimate for the FFT of inverse filter \( q \) at iteration \( n \). \( M \) represents the number of blocks and the superscript * is the complex conjugation. Equation 5.7 ensures that the estimate of the inverse filter is normalized.

The system is trained on reverberant speech from the target source sampled at 16 kHz and presented alone. We employ a training corpus consisting of ten speech signals from the TIMIT database: five female utterances and five male utterances. An inverse filter of length \( L=1024 \) is adapted for 500 iterations on the training data.

Figure 5.2 shows the outcome of convolving an estimated inverse filter with both the target room impulse response as well as the room impulse response at a different source location. The room reverberation time, \( T_{60} \), is 0.35 s. The two source locations are 0° (target) and 45°. As can be seen in Fig. 5.2(b), the equalized response for the target source is far more impulse-like compared to the room impulse response in Fig. 5.2(a). On the other hand, the impulse response corresponding to the interfering source is further smeared by the inverse filtering process, as seen in Fig. 5.2(d). Fig. 5.3 illustrates the effect of reverberation as well as that of inverse filtering on the harmonic structure of a voiced utterance. The filters in Fig. 5.2 are convolved with a clean signal to generate the
Figure 5.2. Effects of inverse filtering on room impulse responses. (a) A room impulse response for a target source presented in the median plane. (b) The effect of convolving the impulse response in (a) with an estimated inverse filter. (c) A room impulse response for one interfering source at 45° azimuth. (d) The effect of convolving the impulse response in (c) with the estimated inverse filter.

signals in Fig. 5.3. For a constant pitch contour, reverberation produces elongated tails but largely preserves the harmonicity. However, once the pitch changes reverberation
smears the harmonic structure. For a given change in pitch frequency, higher harmonics vary their frequencies more rapidly compared to lower ones. Consequently, higher harmonics are more susceptible to reverberation as can be seen in Fig. 5.3(b). Figure 5.3(c) shows that an inverse filter is able to recover some of the harmonic components in the signal. To exemplify the smearing effect on the spectrum of an interfering source, we show the convolution of the same utterance with the filters corresponding to Fig. 5.2(c) and Fig. 5.2(d) and the results are given in Fig. 5.3(d) and Fig. 5.3(e), respectively.

Finally, the target inverse filter is applied on the reverberant mixture composed of both target speech and interference and the resulting signal feeds to the second stage of our model.

5.4 Pitch-based Speech Segregation

The proposed pitch-based speech segregation system uses a given target pitch contour to group harmonically related components from the target source. Our system follows the principles of segmentation and grouping from the system of Hu and Wang (2004). However, we simplify their algorithm by extracting periodicities directly from the correlogram. Also, compared to the sinusoidal modeling approach used for computing AM rates in Hu and Wang (2004), our simplified implementation is more robust to intrusions in the high frequency range resulting in more efficient T-F unit labeling. A detailed description of the model is given below.
Figure 5.3. Effects of reverberation and target inverse filtering on the harmonic structure of a voiced utterance. (a) Spectrogram of the anechoic signal. (b) Spectrogram of the reverberant signal corresponding to the impulse response in Fig. 5.2(a). (c) Spectrogram of the inverse-filtered signal corresponding to the equalized impulse response in Fig. 5.2(b). (d) Spectrogram of the reverberant signal corresponding to the room impulse response in Fig. 5.2(c). (e) Spectrogram of the inverse filtered signal corresponding to the impulse response in Fig. 5.2(d).
5.4.1 Auditory Periphery and Feature Extraction

The signal is filtered through a bank of 128 fourth-order gammatone filters with center frequencies aligned on the ERB scale between 80 and 5000 Hz (see Chapter 2). In addition, envelopes are extracted for channels with center frequencies higher than 800 Hz as follows (Rouat et al., 1997). A Teager energy operator is applied to the filtered signals in each frequency channel. This is defined as \( E(t) = x^2(t) - x(t+1)x(t-1) \) for a signal \( x(t) \). Then, the signals are low-pass filtered at 800 Hz using a third-order Butterworth filter and high-pass filtered at 64 Hz.

The correlogram \( A(c, m, \tau) \) for channel \( c \), time-frame \( m \), and lag \( \tau \) is computed by the following autocorrelation using a window of 20 ms (\( K = 320 \)):

\[
A(c, m, \tau) = \frac{\sum_{k=0}^{K} g_c(m-k)g_c(m-k-\tau)}{\sqrt{\sum_{k=0}^{K} g_c^2(m-k)} \sqrt{\sum_{k=0}^{K} g_c^2(m-k-\tau)}},
\]

where \( g_c \) is the gammatone filter output and the correlogram is updated every 10 ms. The range for \( \tau \) corresponding to the plausible pitch range of 80 Hz to 500 Hz is from 32 to 200 time lags. At high frequencies, the autocorrelation based on response envelopes reveals the amplitude modulation rate that coincides with the fundamental frequency for one periodic source. Hence, an additional envelope-based correlogram \( A_e(c, m, \tau) \) is computed for channels in the high-frequency range (> 800 Hz) by replacing the filter
output \( g \) in Eq. 5.8 with its extracted envelope. This correlogram representation of the acoustic signal has been successfully used in Wu et al. (2003) for multi-pitch analysis.

Finally, the cross-channel correlation between normalized autocorrelations in adjacent channels is computed in each T-F unit as:

\[
C(c, m) = \sum_{\tau=0}^{N-1} A(c, m, \tau) A(c + 1, m, \tau),
\]

(5.9)

where \( N=200 \) corresponds to the minimum pitch frequency of 80 Hz. Since adjacent channels activated by the same source tend to have similar autocorrelation responses, the cross-channel correlation has been used as an effective feature in previous segmentation studies (see e.g. Wang and Brown, 1999). Similarly, envelope-based cross-channel correlation \( C_e(c, m) \) is computed for channels in the high-frequency range (>800 Hz) to capture the amplitude modulation rate.

### 5.4.2 Unit Labeling

A pitch-based segregation system requires a robust pitch detection algorithm. We employ here the multi-pitch tracking algorithm proposed by Wu et al. (2003) that produces up to two pitch contours and has shown good performance for a variety of intrusions. The system combines correlogram-based pitch and channel selection mechanisms within a statistical framework in order to form multiple tracks that correspond to the active sources in the acoustic scene. However, the assignment of the
overlapping pitch contours is needed when the interference also has harmonic structure. For this, the ‘ideal’ pitch contour is extracted using Praat (Boersma and Weenink, 2002) from the target signals and used as the ground truth for the sole purpose of deciding which of two overlapping pitch contours belongs to the target utterance. The resulting estimated pitch track is used for identifying individual T-F units that belong to target as described below.

The labeling of an individual T-F unit is carried out by comparing the lag $p$ corresponding to target pitch with the periodicity of the normalized correlogram. In the low-frequency range, the system selects the time lag $l$ that corresponds to the closest peak in $A(c,m,\tau)$ from the pitch lag $p$. For a particular channel, the distribution of the selected time lags is sharply centered around the pitch lag and its variance decreases as the channel center frequency increases. Here, a T-F unit is discarded if the distance between the two lags $|p - l|$ exceeds a threshold $\theta_L$. We have found empirically that a value of $\theta_L = 0.15(f_s / f_c)$ results in a good performance, where $f_s$ is the sampling frequency and $f_c$ is the center frequency of channel $c$. Finally, the unit is labeled 1 if $A(c,m,l)$ is close to the maximum of $A(c,m,\tau)$ in the plausible pitch range:

$$\frac{A(c,m,l)}{\max_{\tau \in [32,200]} A(c,m,\tau)} > \theta_p,$$

where $\theta_p$ is fixed to 0.85. The unit is labeled 0 otherwise.
In the high-frequency range, we adapt the peak selection mechanism developed by Wu et al. (2003). First, the envelope correlogram $A_e(c, m, \tau)$ of a periodic signal exhibits a peak both at the pitch lag and at the double of the pitch lag. Thus, the system selects all the peaks that satisfy the following condition: A peak with time lag $l$ must have a corresponding peak that falls within the 5% interval around the double of $l$. If no peaks are selected, the T-F unit is labeled 0. Second, a harmonic interference introduces peaks at lags around the multiples of its pitch lag. Therefore, our system selects the first peak that is higher than half of the maximum peak in $A_e(c, m, \tau)$ for $\tau \in [32, 200]$. The T-F unit is labeled then 1 if the distance between the time lag of the selected peak and the target pitch lag does not exceed a threshold $\Delta_L = 15$, the unit is labeled 0 otherwise. All the above parameters were optimized by using a small training set and found to generalize well over a test set.

The distortions on harmonic structure due to room reverberation are generally more salient in the high-frequency range. Figure 5.4 illustrates the effect of reverberation as well as inverse filtering in frequency channels above 800 Hz for a single male utterance. The filters in Fig. 5.2(a) and Fig. 5.2(b) are used to simulate the reverberant signal and the inverse-filtered signal, respectively. At each time frame, we display the histogram of time lags corresponding to selected peaks. As can be seen from the figure, inverse filtering results in sharper peak distributions and improved harmonicity in comparison with the reverberant condition. The corresponding pitch contours are extracted using Praat (Boersma and Weenink, 2002) for each separate condition. From a different measure, the channel selection mechanism retains 79 percent of the total signal energy by
Figure 5.4. Histograms of selected peaks in the high-frequency range (>800 Hz) for a male utterance. (a) Results for the clean signal. (b) Results for the reverberant signal. (c) Results for the inverse filtered signal. The solid lines are the corresponding pitch contours.
applying inverse-filtering as compared to 58 percent without inverse filtering. As a reference, the system retains 94 percent signal energy in the anechoic condition.

5.4.3 Segregation

The final segregation of the acoustic mixture into a target and a background stream is based on combined segmentation and grouping. The main objective is to improve on the pitch-based T-F unit labeling described above using segment-level features. The following steps follow the general segregation strategy from the Hu and Wang model (2004).

In the first step, segments are formed using temporal continuity and the gammatone-based cross-channel correlation. Specifically, neighboring T-F units are iteratively merged into segments if their corresponding cross-channel correlation $C(c,m)$ exceeds a threshold $\theta_c=0.985$ (Hu and Wang, 2004). The segments formed at this stage are primarily located in the low-frequency range. A segment agrees with the target pitch at a given time frame if more than half of its T-F units are labeled 1. A segment that agrees with the target pitch for more than half of its length is grouped into the target stream; otherwise it is grouped in the background stream.

The second step primarily deals with potential segments in the high-frequency range. Segments are formed by iteratively merging T-F units that are labeled 1 but not selected in the first step for which the envelope cross-channel correlation $C_e(c,m)$ exceeds the
threshold $\theta_c$. Segments shorter than 50 ms are removed (Hu and Wang, 2004). All these segments are grouped to the target stream.

This final step performs an adjustment of the target stream so that all T-F units in a segment bear the same label and no segments shorter than 50 ms are present. Furthermore, the target stream is iteratively expanded to include neighboring units that do not belong to either stream but are labeled 1.

With the T-F units belonging to the target stream labeled 1 and the other units labeled 0, the segregated target speech waveform can then be resynthesized from the resulting binary T-F mask for systematic performance evaluation, to be discussed in the next section.

5.5 Evaluation

Two types of ASA cues that can potentially help a listener to segregate one talker in noisy conditions are: localization and pitch. Darwin and Hukin (2000) compared the effects of reverberation on spatial, prosodic and vocal-tract size cues for a sequential organization task where the listener’s ability to track a particular voice over time is examined. They found that while location cues are seriously impaired by reverberation, the F0 contour and vocal-tract length are more resistant cues. In our experiments, we also observe that pitch tracking is robust to moderate levels of reverberation. To illustrate this, Figure 5.5 compares the results of a pitch tracking algorithm (Wu et al., 2003) on a single male utterance in anechoic and reverberant conditions where $T_{60} = 0.35$ s. The only
Figure 5.5. Comparison of pitch tracking in anechoic and reverberant conditions for a male voiced utterance. (a) Spectrogram of the anechoic signal. (b) Spectrogram of the reverberant signal corresponding to the impulse response in Fig. 5.2(a). (c) Pitch tracking results. The solid line indicates the anechoic pitch track. The ‘o’ track indicates the reverberant track.
distortions observed in the reverberant pitch track compared to the anechoic one are elongated tails and some deletions in time frames where pitch changes rapidly.

Culling et al. (2003) have shown that while listeners are able to exploit the information conveyed by the F0 contour to separate a desired talker, the smearing of individual harmonics in reverberation degrades this capability. However, compared to location cues, the pitch cue degrades gradually with increasing reverberation and remains effective for speech separation (Culling et al., 2003). In addition, as illustrated in Fig. 5.4, inverse filtering with respect to target location improves signal harmonicity. We therefore assess the performance of two viable pitch-based strategies: 1) segregating the reverberant target from the reverberant mixture and 2) segregating the inverse-filtered target from the inverse-filtered mixture. Consequently, the speech segregation system described in Section 5.4 is applied separately on the reverberant mixture and the inverse-filtered mixture. As described in Section 5.2, we have evaluated the system on the left-ear response of a KEMAR dummy head, using a room acoustic model implemented by Palomäki et al. (2004). In addition, the inverse filter of the target room impulse response is obtained from training data as explained in Section 5.3 and applied on the whole reverberant mixture to obtain the inverse filtered mixture.

Figure 5.6 shows the binary masks obtained for a mixture of target male speech presented at 0° and interference female speech at 45°. Reverberant signals as well as inverse-filtered signals for both target and interference are produced by convolving the original anechoic utterances with the filters from Fig. 5.2. The signals are mixed to give an overall 0 dB SNR in both conditions. The ideal binary mask is constructed from the
Figure 5.6. Binary mask estimation for a mixture of target male utterance and interference female speech in reverberant and inverse-filtered conditions. (a) The estimated binary mask on the reverberant mixture. (b) The ideal binary mask for the reverberant condition. (c) The estimated binary mask on the filtered mixture. (d) The ideal binary mask for the inverse-filtered condition. The white regions indicate T-F units that equal 1 and the black regions indicate T-F units that equal 0.

premixing target and intrusion as follows: a T-F unit in the mask is assigned 1 if the target energy in the unit is greater than the intrusion energy and 0 otherwise. This corresponds to a 0 dB local SNR criteria for ideal mask generation (see Chang, 2004). The figure shows an improved segregation capacity in the high frequency range in the inverse-filtered case (Fig. 5.6 (c)) as compared to the reverberant case (Fig. 5.6 (a)).
<table>
<thead>
<tr>
<th>Reverberation Time</th>
<th>-5 dB</th>
<th>0 dB</th>
<th>5 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anechoic</td>
<td>8.78</td>
<td>11.61</td>
<td>13.93</td>
</tr>
<tr>
<td>$T_{60}=0.05$ s</td>
<td>7.25</td>
<td>8.54</td>
<td>10.65</td>
</tr>
<tr>
<td>$T_{60}=0.10$ s</td>
<td>7.35</td>
<td>8.16</td>
<td>9.46</td>
</tr>
<tr>
<td>$T_{60}=0.15$ s</td>
<td>6.37</td>
<td>7.09</td>
<td>8.24</td>
</tr>
<tr>
<td>$T_{60}=0.20$ s</td>
<td>5.59</td>
<td>6.52</td>
<td>7.39</td>
</tr>
<tr>
<td>$T_{60}=0.25$ s</td>
<td>4.74</td>
<td>6.06</td>
<td>6.79</td>
</tr>
<tr>
<td>$T_{60}=0.30$ s</td>
<td>4.47</td>
<td>5.57</td>
<td>6.22</td>
</tr>
<tr>
<td>$T_{60}=0.35$ s</td>
<td>4.55</td>
<td>5.36</td>
<td>6.13</td>
</tr>
</tbody>
</table>

Table 5.1. SNR results for target speech mixed with a female interference at three input SNR levels and different reverberation times.

To conduct a systematic SNR evaluation, a segregated signal is reconstructed from a binary mask following the method described in Section 5.2. Given our computational objective of estimating the ideal binary mask, we use the signal reconstructed from the ideal binary mask as the ground truth in our SNR evaluation (see Chapter 4).

We perform the SNR evaluations using as target the set of 10 voiced male sentences collected by Cooke (1993) for the purpose of evaluating voiced speech segregation systems. The following 5 noise intrusions are used: white noise, babble noise, a male utterance, music, and a female utterance. These intrusions represent typical acoustical interferences occurring in real environments. In all cases, target is fixed at 0°. The babble
noise is obtained by presenting natural speech utterances from the TIMIT database at the following 8 separated positions around the target source: \( \pm 20^\circ, \pm 45^\circ, \pm 60^\circ, \pm 135^\circ \). For the other intrusions, the interfering source is located at 45°, unless otherwise specified. Also, the reverberation time for the experiments described below equals 0.35 s, unless otherwise specified. This reverberation time falls in the typical range for living rooms and office environments. When comparing the results between the two strategies the target signal in each case is scaled to yield a desired input SNR. Each value in the following tables represents the average output SNR of one particular intrusion mixed with the 10 target sentences.

We first analyze how pitch-based speech segregation is affected by reverberation. Table 5.1 shows the performance of our pitch-based segregation system applied directly on reverberant mixtures when \( T_{60} \) increases from 0.05 s to 0.35 s. The mixtures are obtained using the female speech utterance as interference and three levels of input SNR: -5 dB, 0 dB, 5 dB. The ideal pitch contours are used here to generate the results. As expected, the system performance degrades gradually with increasing reverberation. The individual harmonics are increasingly smeared and this results in a gradual loss in energy especially in the high frequency range as illustrated also in Fig. 5.6. The decrease in performance for \( T_{60} = 0.35 \) s compared to the anechoic condition ranges from 4.23 dB at -5 dB input SNR to 7.80 dB at 5 dB input SNR. Overall, however, the segregation algorithm provides consistent gains across a range of reverberation times, showing the robustness of the pitch cue. Observe that a sizeable gain of 9.55 dB is obtained for the – 5 dB input SNR even when \( T_{60} = 0.35 \) s.
Table 5.2. SNR results using estimated pitch tracks for target speech mixed with different noise types at three input SNR levels and $T_{60} = 0.35$ s. Target is at 0° and interference at 45°.

Now we analyze how inverse-filtering impacts the overall performance of our speech segregation system. The results in Table 5.2 are given for both the reverberant case (Reverb) and inverse-filtered case (Inverse) at three SNR levels: -5 dB, 0 dB and 5 dB. The results are obtained using the estimated pitch tracks provided by the multi-pitch tracking algorithm of Wu et al. (2003) as explained in Section 5.4.2. The performance depends on input SNR and type of interference. A maximum improvement of 12.46 dB is obtained for the female interference at -5 dB input SNR. The proposed system (Inverse) has an average gain of 10.11 dB at -5 dB, 6.45 dB at 0 dB and only 2.55 dB at 5 dB.
<table>
<thead>
<tr>
<th>Input SNR</th>
<th>-5 dB</th>
<th>0 dB</th>
<th>5 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reverb</td>
<td>Inverse</td>
<td>Reverb</td>
</tr>
<tr>
<td>White noise</td>
<td>5.94</td>
<td>5.38</td>
<td>6.19</td>
</tr>
<tr>
<td>Babble noise</td>
<td>3.25</td>
<td>4.23</td>
<td>5.14</td>
</tr>
<tr>
<td>Male</td>
<td>1.90</td>
<td>5.08</td>
<td>4.49</td>
</tr>
<tr>
<td>Music</td>
<td>3.89</td>
<td>6.25</td>
<td>5.73</td>
</tr>
<tr>
<td>Female</td>
<td>4.55</td>
<td>7.23</td>
<td>5.36</td>
</tr>
<tr>
<td>Average</td>
<td>3.90</td>
<td>5.63</td>
<td>5.38</td>
</tr>
</tbody>
</table>

Table 5.3. SNR results using ideal pitch tracks for target speech mixed with different noise types at three input SNR levels and T<sub>60</sub> = 0.35 s. Target is at 0° and interference at 45°.

When compared to the reverberant condition a 2-3 dB improvement is observed for the male and female intrusions at all SNR conditions. Almost no improvement is observed for white noise or babble noise. Moreover, inverse filtering decreases the system performance in the case of white noise at low SNRs by attempting to over-group T-F units in the high frequency range. For comparison, results using the ideal pitch tracks are presented in Table 5.3. The improvement obtained by using ideal pitch tracks is small and shows that the pitch estimation method is accurate.

As seen in the results presented above, the major advantage of the inverse-filtering stage occurs for harmonic interference. In all the cases presented above the interfering
source is located at $45^\circ$, and the inverse filtering stage further smears its harmonic structure. However, if the interfering source is located at a location near the target source the inverse filter will dereverberate the interference also. Table 5.4 shows SNR results for both white noise as well as female speech intrusions when the interference location is fixed at $0^\circ$, the same as the target location. As expected, in the white noise case, the results are similar to the ones presented in Table 5.3. However, the relative improvement obtained using inverse filtering compared to the reverberant condition is largely attenuated to the range of 0.5-1 dB. This shows that smearing the harmonic structure of the interfering source plays an important role in boosting the segregation performance in the inverse-filtered condition.

As mentioned in Section 5.1, our system is probably the first study on monaural segregation of reverberant speech. As a result, we cannot quantitatively compare with an existing system. In an attempt to put our performance in perspective, we show a comparison with the spectral subtraction method, which is a standard speech enhancement technique (O’Shaughnessy, 2000). To apply spectral subtraction in practice requires robust estimation of interference spectrum. To put spectral subtraction in a favorable light, the average noise power spectrum is computed \textit{a priori} within the silent periods of the target signal for each reverberant mixture. This average is used as the estimate of intrusion and is subtracted from the mixture. The SNR results are given in Table 5.5, where the reverberant target signal is used as ground truth for the spectral subtraction algorithm and the inverse-filtered target signal is used as ground truth for our algorithm. As shown in the table, the spectral subtraction method performs significantly
Table 5.4. SNR results using ideal pitch tracks for target speech mixed with two type of noise at three input SNR levels and $T_{60} = 0.35$ s. Target and interference are both located at 0°.

<table>
<thead>
<tr>
<th>Input SNR</th>
<th>-5 dB</th>
<th>0 dB</th>
<th>5 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reverb</td>
<td>Inverse</td>
<td>Reverb</td>
</tr>
<tr>
<td>White noise</td>
<td>6.37</td>
<td>6.76</td>
<td>6.30</td>
</tr>
<tr>
<td>Female</td>
<td>4.82</td>
<td>5.51</td>
<td>5.74</td>
</tr>
</tbody>
</table>

worse than our system, especially at low levels of input SNR. This is because of its well known deficiency in dealing with non-stationary interferences. At 5 dB input SNR the spectral subtraction outperforms our system when the interference is white noise, babble noise or music. In those cases with relatively steady intrusion, the spectral subtraction algorithm tends to subtract little intrusion but it introduces little distortion to the target signal. By comparison, our system is a target-centered algorithm that attempts to reconstruct the target signal on the basis of periodicity. Target components made inharmonic by reverberation are therefore removed by our algorithm, thus introducing more distortion to the target signal. It is worth noting that the ceiling performance of our algorithm without any interference is 8.89 dB.
Table 5.5. Comparison between the proposed algorithm and spectral subtraction (SS). Results are obtained for target speech mixed with different noise types at three input SNR levels and $T_{60} = 0.35$ s. Target is at $0^\circ$ and interference at $45^\circ$.

<table>
<thead>
<tr>
<th>Input SNR</th>
<th>-5 dB</th>
<th>0 dB</th>
<th>5 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
<td>Proposed</td>
<td>SS</td>
</tr>
<tr>
<td>White noise</td>
<td>2.40</td>
<td>3.36</td>
<td>6.54</td>
</tr>
<tr>
<td>Babble noise</td>
<td>-2.76</td>
<td>2.74</td>
<td>1.98</td>
</tr>
<tr>
<td>Male</td>
<td>-4.05</td>
<td>4.11</td>
<td>0.77</td>
</tr>
<tr>
<td>Music</td>
<td>-1.37</td>
<td>4.45</td>
<td>3.22</td>
</tr>
<tr>
<td>Female</td>
<td>-3.31</td>
<td>5.40</td>
<td>1.46</td>
</tr>
<tr>
<td>Average</td>
<td>-1.81</td>
<td>4.01</td>
<td>2.79</td>
</tr>
</tbody>
</table>

5.6 Discussion

In natural settings, reverberation alters many of the acoustical properties of a sound source reaching our ears, including smearing out its harmonic and temporal structures. Despite these alterations, moderate reverberant speech remains highly intelligible for normal-hearing listeners (Nabelek and Robinson, 1982). When multiple sound sources are active, however, reverberation adds another level of complexity to the acoustic scene. Not only does each interfering source constitute an additional masker for the desired source, but also does reverberation blur many of the cues that aid in source segregation. The recent results of Culling et al. (2003) suggest that reverberation degrades human
ability to exploit differences in F0 between competing voices, producing a 5 dB increase in speech reception threshold for normal intonated sentences in monaural conditions.

We have investigated pitch-based monaural segregation in room reverberation and report the first systematic results on this challenging problem. We observe that pitch detection is relatively robust in moderate reverberation. However, the segregation capacity is reduced due to the smearing of the harmonic structure resulting in a gradual degradation in performance as the room reverberation time increases. As seen in Table 5.1, compared to anechoic conditions there is an average decrement of 5.33 dB for a two-talker situation with $T_{60} = 0.35$ s. Observe that this decrement is consistent with the 5 dB increase in speech reception threshold reported by Culling et al. (2003).

To reduce the smearing effects on the target speech, we have proposed a pre-processing stage which equalizes the room impulse response that corresponds to target location. This pre-processing results in both improved harmonicity for signals arriving from the target direction as well as smearing of competing sources at other locations, and thus provides a better input signal for the pitch-based segregation system. The extensive evaluations show that our system yields substantial SNR gains across a variety of noise conditions.

The improvement in speech segregation obtained in the inverse filtering case is limited by the accuracy of the estimated inverse filter. In our study, we have employed a practical algorithm that estimates the inverse filter directly from reverberant speech data. When the room impulse response is known, better inverse filtering methods exist, e.g. the linear least square equalizer proposed by Gillespie and Atlas (2002). This type of pre-
processing leads to increased target signal fidelity and thus produces large improvements in speech segregation. In terms of applications to real-world scenarios our inverse-filtering faces several drawbacks. First, the adaptation of the inverse filter requires data on the order of a few seconds and thus any fast change in the environment (e.g. head movements, walking) will have an adverse impact on the inverse-filtering stage. Second, the stage needs to identify signal intervals that contain no interference to allow for the filter adaptation. On the other hand, our pitch-based segregation stage can function without training and is robust to a variety of environmental changes. Hence, whenever the adaptation of the inverse filter is infeasible, one can use our pitch-based segregation algorithm directly on the reverberant mixture.

Speech segregation in high input SNR conditions presents a challenge to our system. We employ a figure-ground segregation strategy that attempts to reconstruct the target signal by grouping harmonic components. Consequently, inharmonic target components are removed by our approach even in the absence of interference. While this problem is common in both anechoic and reverberant conditions, it worsens in reverberation due to the smearing of harmonicity. To address this issue probably requires examining the inharmonicity induced by reverberation and distinguishing such inharmonicity from that caused by additive noise. This is a topic requiring further investigation.

In the segregation stage, our system utilizes only pitch cues and thus is limited to the segregation of voiced speech. Other ASA cues such as onsets, offsets and acoustic-phonetic properties of speech are also important for monaural separation (Bregman, 1990). Recent research has shown that these cues can be used to separate unvoiced
speech (Hu and Wang, 2003; 2005). Future work will need to address unvoiced separation in reverberant conditions.
6.1 Contributions

In this dissertation, we have investigated four important problems in auditory-based source segregation: location-based speech segregation, binaural tracking of moving sources, binaural segregation in multisource reverberant environments, and monaural speech segregation in reverberant environments. Our computational goal has been the estimation of an ideal binary mask which selects only the target dominant T-F units from an acoustic mixture and cancels the others.

In Chapter 2, we have studied speech segregation based on sound localization in multisource environments. The principal cues that our system uses are ITD and IID. Following a T-F decomposition of the left and right ear mixtures using the auditory filterbank, ITD and IID cues are estimated in each T-F unit. We observe systematic deviations of the ITD and IID cues which results in a configuration-specific clustering in the joint ITD-IID feature space. Our main contribution is the introduction of supervised learning for different spatial configurations and frequency bands on the observed ITD-
IID space. This approach minimizes the discrepancies between the estimated masks and the ideal binary ones in each scenario. Evaluation results using both SNR and ASR accuracy show that the system estimates ideal binary masks almost perfectly. In addition, when tested with normal listeners, the model produces large speech intelligibility improvements for two-source and three-source conditions. For example, at -10 dB input SNR the system improves the intelligibility score from 20% to 80%. These tests were the first to show that computational models of this kind can produce speech intelligibility improvements for human listeners (Roman et al., 2003).

In Chapter 3, an HMM framework is proposed for tracking multiple active speakers. Due to the sparsity of speech signals, each active source likely triggers a different subset of T-F units. The first contribution of this study is a statistical integration of the ITD and IID cues across reliable frequency channels at each time frame which results in a robust likelihood over the target space. The second contribution is the novel introduction of the HMM framework that can track individual sources and automatically switch between hypotheses with different numbers of active sources.

In Chapter 4, we have studied segregation in multisource reverberant environments using two microphones. Standard two-microphone noise cancellation strategies process the input using linear adaptive filters and can be used to efficiently cancel a single interference. However, these algorithms cannot deal with multiple interfering sources. Our contribution is to couple target cancellation through adaptive filtering with a binary decision rule that estimates the target dominant T-F units. This is a figure-ground segregation approach that requires knowledge about the target source but imposes no
restrictions on the number or positions of interfering sources. In addition, we have adapted the approach proposed by Raj et al. (2004) to reconstruct the target spectrogram in the unreliable T-F units. This allows our system to be utilized as a front-end to a standard speech recognition system operating using cepstral features. In a systematic comparison, our system shows substantial performance gains over baseline and other related approaches.

In Chapter 5, we have studied monaural, pitch-based speech segregation in reverberation. To reduce the smearing effects on the target speech, we have proposed a pre-processing stage which equalizes the room impulse response that corresponds to target location. This pre-processing results in both improved harmonicity for signals arriving from the target direction as well as smearing of competing sources at other locations, and thus provides a better input signal for pitch-based segregation. Evaluations show that our system yields substantial SNR gains across a variety of noise conditions.

6.2 Insights Gained

At the beginning of this dissertation, we set out to investigate several aspects of CASA based source segregation. A number of insights have emerged during my research and they are summarized below. A key insight in this dissertation is that ideal T-F binary masks represent a very effective goal for speech segregation systems. We have demonstrated this by showing dramatic improvements in speech intelligibility and speech recognition performance. Part of my success stems from a conscious, explicit estimation
of the ideal binary mask. In a sense, my dissertation reaffirms the importance of clarifying the computational goal in perceptual information processing (Marr, 1982; Wang, 2005).

In the study of segregation based on sound localization, we started our investigation by analyzing how the ITD and IID cues were affected by the presence of simultaneous acoustic sources. We found that for narrow frequency bands there are systematic changes for both ITD and IID when modifying the relative strength of a target source with respect to the acoustic interference. Hence, supervised learning using ITD and IID deviations became apparent and has since been used to estimate ideal binary masks by clustering T-F units dominated by a common location. By employing the localization cue independently in each T-F unit, a binaural CASA system is able to efficiently handle multiple interferences. This represents a major advance when compared to standard signal processing techniques where the required number of microphones increases with increasing number of sources.

Due to the wide range of speech energy in the T-F domain, while some T-F units are heavily corrupted due to the overlap of multiple sources, others are dominated by only one source. Therefore, a statistical integration that takes into account the differential distributions of ITD and IID between reliable and unreliable channels provides an optimal approach to sound localization. For multiple moving sources, the locations can be predicted and tracked across time as shown in Chapter 3. We find that the HMM framework introduced by Wu et al. (2003) can be effectively extended to tracking moving sound sources.
In a reverberant environment, the energy from a sound source reaches the eardrums after undergoing multiple reflections. As a result, ITD and IID become smeared and lose their characteristics. Hence, one needs a different approach to encode the location information for use as a segregation cue. A filter can be learned to better exhibit the characteristics of a location in reverberation. Realistic environments contain, however, multiple sources and background noise and localizing all active sources is hard if not impossible. Hence, the use of a figure-ground segregation framework offers a general strategy to segregation in the presence of an arbitrary (possibly moving) number of interfering sources.

Reverberation also contributes to the smearing of harmonic structure in sound, presenting a challenge to pitch-based segregation. One solution to this problem is the application of a dereverberation algorithm prior to pitch-based segregation. A key insight in Chapter 5 is the use of the estimated, inverted room impulse response corresponding to target source. This has the effect of enhancing the harmonicity of the target signal while deteriorating the harmonicity of the interfering ones. Since our segregation extracts the harmonic components of the target, further smearing of interference does not cause much of a problem. Consequently, this strategy results in improved pitch-based segregation.

6.3 Future Work

A number of issues remain to be addressed in future research. In Chapter 2 we have proposed a binaural approach to sound segregation that performs optimally for fixed
source locations. In an effort to relax this constraint, we have since proposed a binaural tracking algorithm in Chapter 3. A natural extension would therefore be to combine the two algorithms and provide a more general solution for binaural segregation in anechoic environments.

Several concerns exist when we extend our multisource tracking algorithm from Chapter 3 to reverberant conditions. As discussed in Section 6.2, reverberation smears ITD and IID cues and thus is expected to degrade the performance of the proposed algorithm. A number of improvements can be envisaged. For example, a stricter channel selection mechanism that utilizes the precedence effect could help enhance localization. Also, by employing supervised learning techniques, the ITD and IID cues can be differentially weighted depending on their relative reliability.

In our target-based monaural and binaural reverberant segregation systems, the target location is fixed currently. However, recall that we allow for movement of interfering sources. For slowly moving interferences, a standard technique in noise cancellation involves the identification of signal intervals that contain no target followed by filter adaptation. The algorithms proposed for Chapters 4 and 5 can employ a similar strategy to handle a slow target motion.

As outlined in Bregman’s ASA account of auditory perception, monaural and binaural cues provide complementary cues for grouping (Bregman, 1990). Chapter 5 represents one such attempt that utilizes the location information of the target in conjunction with a monaural segregation technique. Psychoacoustic evidence (Darwin and Hukin, 1997; Darwin and Hukin, 1999) suggests that localization likely follows
monaural grouping but subsequently aids in the organization of individual sources across time – known as sequential organization. Perhaps an ultimate solution to the “cocktail-party” problem lies in a CASA system in which monaural and binaural cues are extracted interactively and work in tandem.


