Graz University of Technology

Doctoral Thesis

Auditory Inspired Methods for Multiple Speaker Localization and Tracking Using a Circular Microphone Array

Tania Habib

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria

First Examiner:
Prof. Dr. Gernot Kubin,
Graz University of Technology, Austria

Second Examiner:
Prof. Dr. Walter Kellermann,
University of Erlangen-Nuremberg, Germany

Co-Advisor:
Dr. Harald Romsdorfer
Graz University of Technology, Austria

Graz, July 2011

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

....

date (signature)

Kurzfassung

Die vorliegende Dissertation beschreibt ein neues Verfahren für die Lokalisierung und Verfolgung von mehreren akustischen Quellen mit Hilfe eines Mikrofon-Arrays.

Die Verwendung von Mikrofon-Arrays bietet eine Verbesserung des Sprachsignals bei Aufnahmen in Besprechungs- und Büroräumen. Eine gebräuchliche Lösung zur Sprachsignalverbesserung in realistischen Umgebungen mit Umgebungslärm und Mehrwegeausbreitung ist die Verwendung sogenannter "Beamforming"-Techniken, die Signale aus der gewünschten Richtung durch konstruktive Interferenz verstärken und Signale aus anderen Richtungen durch destruktive Interferenz abschwächen. Diese Beamforming-Algorithmen benötigen als Vorwissen die Position der Quelle. Deswegen sind Algorithmen zur Lokalisierung und Verfolgung von akustischen Quellen eine wesentliche Komponente eines solchen Systems. Konventionelle Lokalisierungsalgorithmen verschlechtern sich jedoch in realen Aufnahmesituationen, sobald mehrere Sprecher gleichzeitig sprechen.

Im Gegensatz zu konventionellen Lokalisierungsalgorithmen verwendet der in dieser Dissertation vorgestellte Algorithmus zusätzlich zur Positionsinformation die Grundfrequenz bzw. Tonhöhe (auf Englisch "Pitch") des Sprachsignals. Dieser sogenannte "Position-Pitch"-Algorithmus verwendet zur Vorverarbeitung der Sprachsignale eine Multi-Band Gamma-Tone-Filterbank, die in ihrer Funktion vom menschlichen Gehör inspiriert ist. Die Funktion dieser Gamma-Tone-Filterbank wird im Detail analysiert. Diese Methode verwendet unter anderem ein Frequenzselektionskriterium, welches beim menschlichen neuronalen System beobachtet wurde und zu einer robusteren Lokalisierung mehrerer, gleichzeitiger Sprachquellen beiträgt. Im Rahmen dieser Arbeit werden zwei Algorithmen zur Quellenverfolgung untersucht, welche die Lokalisierung von mehreren Sprechern weiter verbessern: der Erste basiert auf der Gruppierung von spektralen und temporalen Regionen, die ähnliche Grundfrequenzcharakteristiken aufweisen. Der Zweite verwendet Partikel-Filter (auch sequentielle Monte-Carlo-Methode genannt) basierend auf den Positionsschätzungen des "Position-Pitch"-Algorithmus. Abschließend wird ein neuer

Partikel-Filter basierter Lokalisierungs- und Verfolgungsalgorithmus präsentiert und es werden mehrere Lösungen für Probleme mit Partikel-Filter basierten Algorithmen vorgestellt. Unter anderem eine Methode zur Verbesserung des Wahrscheinlichkeitsmodels basierend auf Informationen über die Quellaktivität.

Alle in dieser Arbeit vorgestellten Lokalisierungs- und Verfolgungsalgorithmen wurden anhand von Sprachaufnahmen mit einem zirkulären 24-kanaligen Mikrofon-Array in verschiedenen realen, akustischen Umgebungen getestet. Die Sprachquellen wurden dafür entweder direkt mit Sprechern bzw. mit über Lautsprechern abgespielten Orignalaufnahmen realisiert. Die in dieser Dissertation entwickelten Methoden erzielen im Durchschnitt eine Verbesserung von etwa 20% gegenüber dem als Stand der Technik geltenden SRP-PHAT Algorithmus.

Abstract

This thesis presents a new approach to the problem of localizing and tracking multiple acoustic sources using a microphone array.

The use of microphone arrays offers enhancements of speech signals recorded in meeting rooms and office spaces. A common solution for speech enhancement in realistic environments with ambient noise and multi-path propagation is the application of so-called beamforming techniques, that enhance signals at the desired angle, using constructive interference, while attenuating signals coming from other directions, by destructive interference. Such beamforming algorithms require as prior knowledge the source location. Therefore, source localization and tracking algorithms are an integral part of such a system. However, conventional localization algorithms deteriorate in realistic scenarios with multiple concurrent speakers.

In contrast to conventional localization algorithms, the localization algorithm presented in this thesis makes use of fundamental frequency or pitch information of speech signals in addition to the location information. This "position-pitch"-based algorithm pre-processes the speech signals by a multiband gammatone filterbank that is inspired from the auditory model of the human inner ear. The role of this gammatone filterbank is analyzed and discussed in details. For a robust localization of multiple concurrent speakers, a frequency-selective criterion is explored that is based on a study of the human neural system's use of correlations between adjacent sub-band frequencies. This frequency-selective criterion leads to more robust localization and pitch cues. In the following, two different kinds of tracking algorithms that further improve localization accuracy of an arbitrary number of speakers are presented: the first one is based on grouping of spectro-temporal regions formed by fundamental frequency cues. The second one applies sequential Monte Carlo methods or particle filters using the location cues provided by the multiband position-pitch algorithm. Finally, a novel particle filter-based joint position and pitch tracking algorithm is presented. Various solutions are proposed for the existing problems faced by the particle filter-based trackers, including an improvement in the likelihood model on information of source activity.

All proposed speaker localization and tracking algorithms are tested using real-world recordings made with a 24-channel uniform circular microphone array using loudspeakers and human speakers under various acoustic environments. The proposed techniques give on average 20% more accurate results than the state-of-the-art SRP-PHAT algorithm.

Acknowledgment

I would like to express my gratitude to the people who helped and assisted me in achieving this milestone.

I am indebted to my supervisor Prof. Gernot Kubin for giving me the opportunity to work with him. His intellect combined with the kind heartedness makes him a truly special person. His comments and suggestions throughout my work helped in better understanding of the problem.

I am grateful to Prof. Walter Kellermann for hosting me at his institute. I learned a lot about my field during my stay at his lab. My sincere thanks for his detailed and useful comments during the preparation of this thesis.

I would like to thank Dr. Marián Képesi for introducing me to this field and helping me build the interest to pursue the problem. I am grateful for the support of my co-advisor Dr. Harald Romsdorfer during the later years of my work. His critical thinking abilities have been of great value to me.

I am thankful to the Higher Education Commission of Pakistan for funding the major part of my research.

I am also thankful to my colleagues at SPSC lab for providing a friendly atmosphere for work and willingness to help whenever I was stuck with some technical problem. I am particularly grateful to Barbara Schuppler and Christina Leitner for proofreading my thesis. I would like to thank Barbara for writing the German abstract of my thesis and for her suggestions regarding the structure of the thesis.

I am grateful to my parents for instilling the confidence and the courage to go to a foreign land to fulfill my dreams. This beautiful journey unknowingly also helped me find my significant other, Salman. Who celebrated with me the successes and gave me constant support and encouragement during the difficult time.

Finally, I am thankful to Allah for blessing me with the ability to undertake this task.

Graz, 2011 Tania Habib

To my parents, Sabuhi Naz and Habib Alam

Contents

List of Acronyms **xvii**

List of Symbols **xviii**

1 Introduction **1**

1.1 Review of the Selected Doctoral Theses 3
1.1.1 Summary of Main Goals and Results 3
1.1.2 Comparison with the Goals of the Selected Theses 7
1.2 Outline of the Thesis 9
1.3 Work Contributions 11

2 Background **13**

2.1 Signal Models 13
2.2 Acoustic Source Localization 16
2.3 Generalized Cross Correlation based Localization Methods 19
2.3.1 Maximum Likelihood Weighting 20
2.3.2 Phase Transform Weighting 20
2.3.3 Bandpass Weighting 21
2.3.4 Cross Power Spectrum Phase 21
2.4 Steered Response Power Methods 23
2.5 Spectral-Estimation Based Localization Methods 27
2.6 Adaptive Eigenvalue Decomposition Algorithm 28
2.7 Blind Source Separation Based Time Delay Estimation Methods . . . 29
2.8 Pitch Based Localization Algorithms 29
2.8.1 The Position-Pitch Algorithm 30

2.8.2 Joint Time-Delay and Pitch Based ML Estimator 30
2.8.3 Subspace Method for Time-Delay and Frequency Estimation . 32
2.8.4 State-space Approach for Time-Delay and Pitch Estimation . 32
2.8.5 A Pitch-Based GCC Weighting Function 33
2.8.6 Excitation Source-Based Time-Delay Estimation 34
2.8.7 Time Delay and Pitch Estimation Using a Neural Network . . 35
2.8.8 Correlogram-Based Joint Time Delay and Pitch Estimation . . 35
2.9 Summary 36

3 Data Acquisition and Corpus Building 37
3.1 Available Microphone Array Databases 37
3.2 Reasons for a New Database 38
3.3 Microphone Array Design 39
3.4 Speech Database and Speaker Setup 45
3.5 Speech Database Reference Segmentation and Labeling 46
3.6 Room Acoustics and Background Noise 49
3.7 Evaluation Metrics 52
3.8 Summary 53

4 Joint Position-Pitch Estimation Based Source Localization 55
4.1 The Position Pitch Algorithm 55
4.1.1 The PoPi Plane 58
4.1.2 Multi-Microphone Position Pitch Algorithm 59
4.1.3 Cepstrum Based Weighting Function 62
4.2 The Multiband Position-Pitch Algorithm 64
4.3 Frequency Selection Based MPoPi Method 75
4.4 Spectro-Temporal Fragment Based MPoPi Method 79
4.5 Experimental Evaluations 84
4.5.1 Controlled Experiments 85
4.5.2 Real Speakers Experiments 94
4.6 Discussion 101
4.7 Conclusions 104

5 Acoustic Source Tracking 107
5.1 Background 107
5.2 Particle Filter Based Source Tracking 109

5.3 Dynamic Model . . . 114
5.4 Multiband Position-Pitch Estimation Based Likelihood Function . . . 115
5.5 Voice Activity Detection . . . 115
5.6 Particle Filter with Integrated Voice Activity Detection . . . 116
5.7 Proposed Modification . . . 116
5.8 Experimental Evaluations . . . 126
5.8.1 Controlled Experiments . . . 126
5.8.2 Real Speakers Experiments . . . 136
5.9 Discussion . . . 138
5.10 Conclusions . . . 142

6 Conclusions and Future Work **143**

A Relationship Between SRP-PHAT and MPoPi Approaches **149**

B Computational Complexity **151**
B.1 Example . . . 152

C Other Work **153**

Bibliography **161**

List of Acronyms

1D,2D,3D	One, Two and Three-dimensional
ACF	Auto Correlation Function
ACG	Auto-Correlogram
AEDA	Adaptive Eigenvalue Decomposition Algorithm
ASL	Acoustic Source Localization
ASLT	Acoustic Source Localization and Tracking
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CCF	Cross Correlation Function
DFT	Discrete Fourier Transform
DoA	Direction of Arrival
ERB	Equivalent Rectangular Bandwidth (of auditory filters)
FFT	Fast Fourier Transform
GCC	Generalized Cross Correlation (function)
GCF	Global Coherence Field
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICA	Independent Component Analysis
IFFT	Inverse Fast Fourier Transform
KF	Kalman Filters
MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single-Output
MPoPi	Multiband Position-Pitch Algorithm

MMSE	Minimum Mean Square Error
MSE	Mean Square Error
MSTD	Mean Standard Deviation
PDF	Probability Density Function
PHAT	Phase Transform
PoPi	Position-Pitch Algorithm
RIR	Room Impulse Response
RMSE	Root Mean Square Error
RT60	60dB Reverberation Time
SBF	Steered Beamformer Function
SIMO	Single-Input Multiple-Output
SIR	Sequential Importance Function
SIS	Sequential Importance Sampling
SISO	Single-Input Single-Output
SMC	Sequential Monte Carlo
SNR	Signal-to-Noise Ratio
SRP	Steered Response Power (of the SBF)
SRR	Signal to Reverberant Ratio
STFT	Short-Time Fourier Transform
TDE	Time Delay Estimation
TDoA	Time Difference of Arrival
UCA	Uniform Circular Array
ULA	Uniform Linear Array

List of Symbols

Standard Operators, Functions and Symbols

$(\ \cdot\ )^*$	Complex conjugate
$[\ \cdot\ ]^T$	Transpose operator
cos	Cosine function
exp	Exponential function
∞	Infinity
$\|\ \cdot\ \|$	Vector 2-norm
$\lceil\ \cdot\ \rceil$	Ceiling function
$\lfloor\ \cdot\ \rfloor$	Floor function
$\log_{10}$	Logarithm to the base 10
$\|\ \cdot\ \|$	Absolute value (scalar) or determinant (matrix)
$\mathbb{E}\{\cdot\}$	Statistical Expected Value
$\mathbb{R}^+$	Real-positive function
max	Max operator
min	Min operator
$\propto$	Proportionality
sin	Sine function
$p(\cdot)$	Probability
$*$	Convolution

$\forall$	For all
$\triangleq$	Generic correspondence or definition

Indexing Variables and Units

J_p	Number of retained particles
M	Number of microphones
M_p	Number of microphone pairs
N_p	Number of particles
N_s	Number of sources
i	Particle number
m	Microphone number
m_p	Microphone pair number
s	Source label
$[\cdot]^\circ$, deg	Angle in Degrees
°C	Degree Celsius
Hz	Hertz
Ohms	Impedance
dB	Decibel
m	Meter
m/sec	Velocity
sec	Second

Specific Functions and Variables

A	Auto-correlation
C	Correlation coefficient
CC	Cross-correlations of gammatone filterbank
F_0	Fundamental Frequency
F_c	Center frequency of gammatone filter

F_s	Sampling frequency
K	Number of cross-correlation peaks
L	Number of frequency channels and time frames in a fragment
N	Number of frames
N_{eff}	Effective number of particles
$O(\varphi_0)$	Correlation lag corresponding to DoA
P	Power of a signal
R	Cross-correlation function
W	Weighting function of GCC
$\boldsymbol{\Theta}$	DoA state vector
$\boldsymbol{\mathcal{Y}}, \mathcal{Y}$	Observation vector and scalar observation variable
$\mathcal{F}_p$	Speech fragment
$\epsilon_{\text{quantization}}$	Spatial quantization error
$\mathbf{F}$	Fundamental frequency state vector
$\mathbf{I}$	Unity matrix
$\mathbf{Y}$	Raw data matrix
$\mathcal{K}$	Total number of discrete frequencies
$\mathcal{L}(x; l, \lambda)$	Laplace distribution with location parameter l and scale parameter λ
$\mathcal{N}(x; \mu, \Sigma)$	Gaussian distribution with mean vector μ and covariance matrix $\sum$
$\mathcal{P}$	Pseudo-likelihood function
$\mathcal{T}$	Raw data transform
$\mathcal{U}$	Uniform random variable
$\mathcal{Z}$	Normalization constant
a	Attenuation parameter
b	Normalization factor
c	Speed of sound in air

d	Distance between a pair of sensors
f	Frequency index
h	Impulse response
j	Imaginary number:$\sqrt{-1}$
k	Frame index
$n(t)$	Process noise
$q(\cdot)$	Importance density function
$s(t)$	Speech signal
t	Continuous time variable
$u(t)$	Noise-like sound source
$v(t)$	Measurement noise
w	Particle weight
$x(t)$	Continuous-time signal
$x[n]$	Discrete-time signal
Δ	Error threshold
β	Attenuation factor
$\boldsymbol{\alpha}$	Joint DoA-Fundamental frequency state vector
$\boldsymbol{\gamma}$	Position vector
δ	Dirac impulse
ϵ	Small value
η	Distance between source and microphone
$\hat{\varphi}$	Source DoA estimate
κ	Discrete temporal frequency
ω	Angular frequency
ϕ	Elevation
π	Pi

ψ	Interaural coherence function
ρ	Joint position-fundamental frequency function
σ	Standard deviation
τ	(relative) Time-delay
θ	Azimuth
ξ	Threshold constant
ζ	Direction of propagation

Set and Vector Notation

(φ_0, F_0)	Pair of elements
$[x\ y\ z]$	Line vector
$\mathcal{Y}_{1:k}$	Set of elements:$\{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_k\}$
$\{1, \ldots, N\}$	Interval of discrete values
$\{w^{(i)}\}_{i=1}^{N_p}$	Set of discrete values:$\{w_1, w_2, \ldots, w_{N_p}\}$

Chapter 1

Introduction

In today's world, hands-free communication has become an essential part of day-to-day activities. It exists as an acoustic front end of telephony and speech dialog systems to name a few. In practice, these systems are placed in adverse acoustic environments with ambient noise. Moreover, the distance between the speaker and microphones decreases the power level of recorded speech signal resulting in poor quality signal acquisition.

The emergence of array signal processing techniques is offering improved system performance for multiple input systems. A comprehensive overview about the field can be found in [1, 2, 3]. The multi-channel system allows to solve problems, such as source localization and tracking, which is difficult with single-channel systems. The problem of source localization has been analyzed and various solutions are proposed keeping different situations in mind in fields of radar, sonar, seismology, geophysics, ultrasonics, and global positioning systems. These applications differ considerably from the speech localization problem addressed here in many aspects. Primarily, the time-delay estimates for the above mentioned fields are evaluated relative to an absolute time-scale or a single reference sensor. This strategy is inappropriate because of the radiation pattern of speech signals is not ideal in realistic environments. Moreover, the accurate estimation of Time Difference of Arrival (TDoA) relative to a single reference sensor may not be possible due to signal incoherence among the sensors. This is why speech source localization methods rely on pairwise Time Delay Estimation (TDE), where the signals received at two (or more) spatially separated microphones are compared for estimating the source location.

This thesis focuses on the subject of time-delay estimation using a circular microphone array. The TDE task is possible with any given placement of sensors or microphones but the strategic positioning of the microphones can yield optimal spatial resolution. Furthermore, a microphone array system can enhance the desired source signal or attenuate noise source signal coming from different directions by using sources' location information.

The problem of speech localization of multiple speakers using a microphone array is depicted in Fig. 1.1. One of the main factors contributing to the complexity of the problem is the acoustic environment where the array has been placed. In this case, the microphones not only pick up the speech signals but the reverberated signals together with the environmental noise. Therefore, the aim is to robustly localize and track one or more concurrent speakers using only the data acquired by the microphone array.

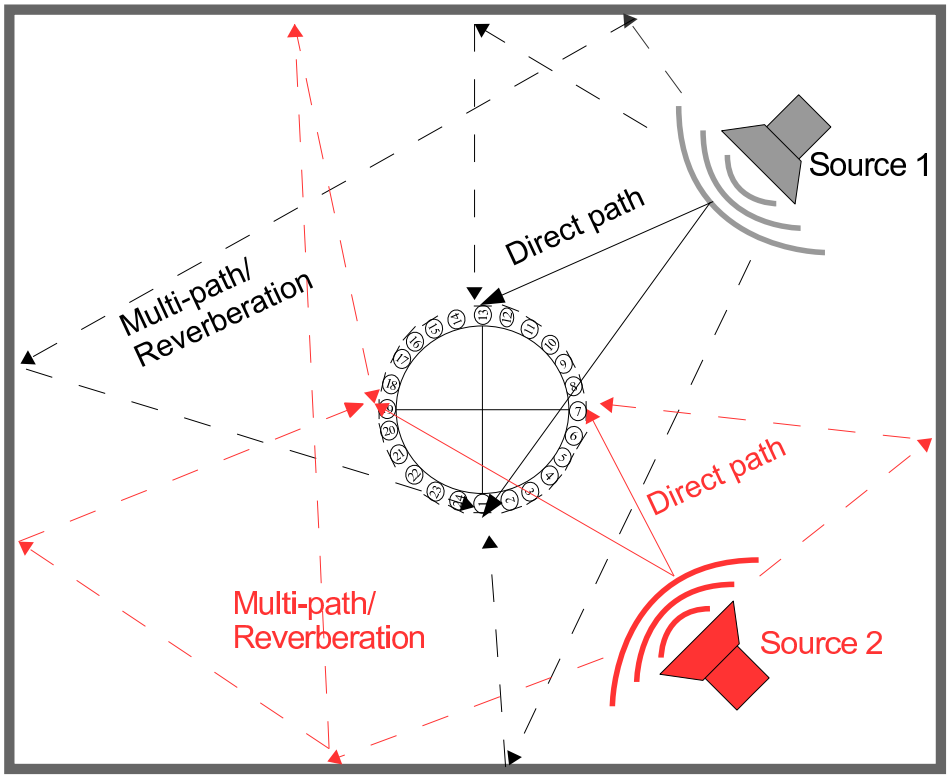


Figure 1.1: Illustration of the speech localization problem in a reverberant environment. The microphones are placed on a uniform circular ring. The positions of the microphones are fixed and known. With the given setup, the aim is to localize and track one or more active sources using only the data acquired from the microphone array.

Recently, meeting rooms equipped with different kinds of sensors have become popular. These are referred to as smart meeting rooms, where a microphone array is used to record meetings with multiple participants. Such a setup facilitates remote meetings. Furthermore, the recorded data can be used for automatic structuring and indexing of the meetings. It is not limited to audio and speech processing but it is also widely studied in computer vision, human computer interaction, and information retrieval. An accurate detection, localization, and tracking of speakers is also essential for media processing tasks: for steering a video camera towards an active speaker [4], for conference telephony systems [5], for speech enhancement of the active stream using the microphone array beamforming for distant speech recognition [6], and to provide accumulated information for speaker identification. All these tasks are crucial for increasing the interactive experience between the participants of a meeting. The meeting room environment, however, poses a number of challenges such as multiple concurrent speakers, short utterances, and background noise sources. A number of recordings emulating some of the above mentioned scenarios have been made to test the performance of the novel algorithms introduced in this thesis.

In the next section, a review of selected relevant doctoral theses in the area of acoustic source localization and tracking is presented. Moreover, a comparison with their respective goals is carried out to highlight the different issues that this thesis is addressing.

1.1 Review of the Selected Doctoral Theses

1.1.1 Summary of Main Goals and Results

A Framework for Speech Source Localization using Sensor Arrays [7]

[7] presents the first few comprehensive works done for the problem of acoustic source localization using multiple microphones. This thesis is authored by Michael Shapiro Brandstein, at the Brown University, U.S.A. The key idea behind the work is to provide a complete framework for speech source localization starting with the theoretical foundation of the problem. The author introduced some error criteria for

location estimates along with various source detection and estimate-error prediction methods. Furthermore, the author presents a novel closed-form locator, named the *Linear Intersection* (LI) method. In order to estimate the source position using LI method, the concept of sensor-pair geometry is introduced, where, for a pair of microphones, the candidate positions of the speakers are half-hyperboloids with the center defined in the middle of the microphones. The hyperbola-to-cone approximation gives the source bearing lines. For multiple pairs, the intersection of these bearing lines indicates the likely source position. A frequency-domain time-delay estimator intended for the speech source environment was also proposed. This method is based on the cross-power spectrum between a pair of microphones and requires minimal computational resources to produce precise TDoA estimates. The developed algorithms have been first tested on a pair of sensors, and later on data recorded with a 10-element bilinear array system.

A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays [8]

Also the work by Joseph Hector Dibiase [8] was carried out at the Brown University. The basic idea behind this work is to improve the localization accuracy by introducing a low-latency technique. The author analyzed the performance of the conventional cross-correlation methods for speaker localization using real-world recordings. The results show that these methods require longer segments ($\backsim$ 200 msec) to achieve high accuracy. The use of such long analysis lengths is not permissible in real-time applications of source localization. Furthermore, the conventional beam-steering method known as the Steered Response Power (SRP) method is applied for the speaker localization. It is then combined with one of the weighting functions of Generalized Cross-Correlation (GCC) methods known as the PHAse Transform (PHAT) to create a new filter-and-sum technique called as "SRP-PHAT" algorithm. The author proved that the summation of microphone signals by exploiting microphone redundancy provides more accurate location estimates in comparison to the combination of TDoA estimates. Furthermore, it was shown in the thesis that for such methods, a segment length of 20 msec can be used to generate accurate position estimates. The performance of SRP-PHAT was compared to SRP and GCC-PHAT in a set of experiments on data collected from both small aperture and large aperture arrays. Moreover, a theoretical relationship is

developed between SRP-PHAT and GCC-PHAT, where the author shows that the SRP-PHAT response can be determined by summations of GCC-PHAT functions over multiple microphones. Although SRP-PHAT is computationally expensive, it provides accurate location estimates in realistic environments. There are a number of solutions presented in the literature for the computational complexity problem. A summary of those methods is presented later in Chapter 2.

Spatio-Temporal Analysis of Spontaneous Speech with Microphone Arrays [9]

[9] presents the work by Guillaume Lathoud at EPFL Lausanne, Switzerland. The underlying concept behind this work is to develop a robust and accurate post-processing stage for the localization system. It combines the location cues with spectral features to create the observation vectors used to solve the problem of speaker identification. Besides the speaker identification task, the author also proposed modifications to the source localization methods, i.e., a Phase Domain Metric (PDM) which is used for detection and localization of multiple speaker in the thesis. For every assumed source location, the theoretical and observed phase values are compared and the actual source position is determined by minimization of the resulting PDM. For the two 8-channel circular microphone arrays, a sector wise scheme for the microphone arrays is used for speaker detection. Furthermore, the acoustic power in each sector is modeled using unsupervised probabilistic modeling to perform the speaker detection-localization task.

The author proposed a short-term clustering scheme for speaker identification, which performs Speech/Non-Speech (SNS) decisions for each cluster to remove non-speech clusters. His short-term clustering method rejects the non-speech noise sources and detects the beginning and end times of each speech utterance. To address the question of who spoke a given utterance, he investigated how to determine the speaker identity of each speech utterance when using distant microphones. The author used an agglomerative clustering method, where speech utterances from the same speaker are progressively grouped together into a single long-term cluster. Moreover, a distant microphones based speaker clustering scheme was proposed, where, the Bayesian Information Criterion (BIC) was changed to combine features such as Mel-Frequency Cepstral Coefficients (MFCCs) and location estimates.

Particle Filtering Methods for Acoustic Source Localization and Tracking [10]

One of the first successful applications of the particles filtering approach to the problem of acoustic source localization and tracking was presented by Eric André Lehmann in [10] at the Australian National University. The thesis presents a general framework for acoustic source localization using Sequential Monte Carlo (SMC) methods referred to as particle filters. The author proposed four different algorithms, which combined the GCC and Steered Beamforming (SBF) methods with Gaussian and pseudo-likelihood functions within the particle filtering framework. The proposed methods are focused on the localization and tracking of a single source using a frame length of 64 msec. The concept of sequential importance sampling is used to revise the acoustic source localization algorithms. With the use of importance sampling, a valuable property of re-initialization is integrated at a low algorithm level. Furthermore, the author showed that importance sampling based methods are able to recover from complete track losses, to detect new targets entering the acoustic scene, and to switch between alternating speakers. The author suggested that the particle filters based on the importance sampling principle are better suited for practical applications as compared to previously developed filters.

A distributed array setup of eight microphones was considered in the thesis. The microphones were arranged in a pairwise manner on each wall of the enclosure. The performance of the proposed methods was evaluated using both synthetic and real-life samples of audio data. In order to avoid the unpredictable outcome of random initializations, the particles were always initialized at the actual speaker positions. Therefore, this thesis only demonstrates the tracking ability of the algorithms. The proposed sequential estimation approaches are shown to outperform conventional localization methods. In addition to that, the theoretical performance analysis of the methods developed is carried out by deriving a modified version of the Cramér-Rao bound.

Acoustic Source Tracking using Sequential Monte Carlo [11]

[11] presents the work carried out by Maurice Fallon at the University of Cambridge, U.K. The thesis proposes several novel techniques to solve the acoustic source tracking problem. The algorithm uses the signal-to-signal correlation between a pair

of microphones to determine speaker's orientation. This concept is based on the observation that the microphone pairs in front of the speaker will exhibit higher correlation compared to those behind the speaker. The steered beamformer based likelihood function proposed in [10] was modified for the "Track-Before-Detect" framework. A discretization of the SBF surface was proposed to efficiently utilizes the available computing power. This approach avoids the position measurements assignment to a particular source and makes a simple extension of the algorithm to track two or more concurrent sources. Moreover, the thesis presents an extension of the multi-target tracking algorithm to track intermittent speakers. The developed algorithm basically is a variable dimension particle filter which constantly monitors the surveillance region for changes in activity to propose new source positions. The likelihood weights of the particles are determined by an importance weighting scheme. The proposed scheme probabilistically combined hypothesized prior behavioral information, the previous particle positions, and the current measurement data to estimate particles' weights. This particle distribution can then be used to infer the number and location of the active speech sources. All the proposed algorithms were tested using real audio data recorded using a distributed setup of 12 microphones.

1.1.2 Comparison with the Goals of the Selected Theses

In this section, the goals of this thesis are compared with the work summarized in the previous section. This thesis focuses on the problem of "Acoustic Source Localization and Tracking in Meeting Rooms/Office Space using a Circular Microphone Array" (here the term acoustic source means the speech source). Therefore, the aim is to develop robust and practical multi-source localization and tracking algorithms performing well in realistic environments with various background noise and strong reverberations. In this work, I explore an interesting and most often neglected area in the context of the speech localization problem, which deals with combining a speech related feature known as fundamental frequency or pitch with the time-delay estimation task.

The work presented in [8, 7] uses the acoustic properties of the environment along with a geometric placement of the sensors to localize the speakers. In the present thesis, a GCC based joint position-pitch (PoPi) algorithm [12] is used as a baseline algorithm. The PoPi method exploits the periodicity present in voiced speech by carrying out a parameterized sampling of the cross-correlation between two spatially

separated microphones. The summation of a certain number of cross-correlation peaks results in a 2D plane where one dimension represents the position of the source, and the other represents the pitch values. Different modifications for the PoPi algorithm will be presented in the present thesis, including some auditory inspired pre-processing techniques [13]. These auditory techniques make use of the periodicity information, and help with the pre-grouping of cues for the location and the pitch of different speakers. This pre-grouping then helps the joint position-pitch decomposition in concurrent speaker scenarios. The proposed methods will be compared with the SRP-PHAT method [8] on real-world recordings using a 24-channel circular microphone array. The proposed methods show superior performance to the SRP-PHAT under strong reverberant conditions with high levels of background noise.

The work in [9] provides the spatio-temporal analysis of speech signals acquired by a microphone array. The main goal of the work was to merge location cues with spectral cues for the speaker identification task. To extract the location cues, the author used the traditional SRP-PHAT method with some modifications summarized in Section 1.1.1. The speaker localization methods proposed in this thesis can offer improved location accuracy over SRP-PHAT algorithm, especially for multiparty speech, where short, fast-changing speaker turns are commonly occurring. Finding solutions for the speaker identification task, however, would be beyond the scope of this work.

In most practical scenarios, localization methods suffer from the increasing number of anomalies as the acoustic conditions become more challenging. The work in [10] provides algorithms for successful tracking of a single source in realistic scenarios. The problem of tracking multiple concurrent speakers, however, is not addressed in [10]. Furthermore, the work in [10] lacks an elaborate statistical representation of the speech signal in the observation model of the particle filter framework. The inclusion of voice activity along with acoustic features like pitch into the observation variable can enhance the tracking performance. Contrary to the work in [10], this thesis combines the fundamental frequency and auditory inspired pre-processing techniques into a conventional particle filtering framework. The other main difference is the experimental setup. The authors in [10, 11] used a distributed microphone array setup, where the microphones were fixed on the walls of the room. This thesis presents results from experiments with a circular microphone array placed in the center of the room to test the speaker tracking performance of the proposed algo-

rithms. Moreover, the author in [11] proposed particle filtering based methods for tracking more than one speaker using the *track-before-detect* methodology. This is a good strategy to circumvent the data association problem in multi-target tracking. The methodology deviates from the main goal of the current work, which is to provide robust speaker detection/localization algorithms. Therefore, this strategy is not explored in this thesis. These methodologies require the actual speaker positions to be known *a priori*. To avoid this condition, the author in [11] presented a variable dimension particle filtering algorithm, where the birth/death rules for the particles are defined heuristically. These rules are somewhat difficult to generalize for different experimental setups. On the one hand, this thesis explores the Markov Chain Monte Carlo (MCMC) sampling techniques within the particle filtering framework. The importance sampling technique is combined with the basic bootstrapping approach to form a new sampling method with some easily tunable parameters. These methods do not require the actual position of speakers to be known *a priori* for particles' initialization.

1.2 Outline of the Thesis

The formulation of auditory inspired cross-correlation-based methods for source localization and tracking using a circular microphone array requires some important steps, which are presented in the following chapters.

Basics of the Acoustic Source Localization Problem

Chapter 2 presents the fundamental ideas, which are explored in this thesis starting with the choice of the signal model, which is crucial for the underlying algorithms. Hence, it needs to be selected carefully. An overview of the state-of-the-art methods is presented. This overview highlights a lacking area of research in the context of speech localization, where little effort has been made to extensively explore the use of fundamental frequency for speech localization problem. Moreover, a summary of the localization algorithms utilizing the periodicity in voiced speech (or pitch) information is presented. In these methods, the pitch information is either combined in, or aiding the process of source localization. Finally, the main hypothesis of the current work is formulated.

Corpus Building

The detailed analysis of the algorithms is only possible when a large database covering different realistic scenarios is used. Chapter 3 outlines some of the well-known microphone array databases along with the reasons to create the new corpus. First, the details of the experimental setup are presented starting with the multi-channel system used for the recordings. Second, the design parameters of the in-house microphone array are further discussed. Then, different speaker configurations recorded for the corpus are outlined. Another important task for experimental evaluations is the generation of ground truth values. The speech segmentation task used to produce speaker activity label files is discussed in detail. The evaluation metrics for different algorithm comparisons and a number of acoustic measurements made in the recording room are also listed.

Auditory Inspired Cross-Correlation Methods for Source Localization

Chapter 4 presents a detailed analysis of the joint position-pitch algorithm through illustrative examples. Furthermore, the modifications to the PoPi algorithm are outlined along with their respective limitations in different scenarios. An investigation of the human auditory system models (the interesting aspects of how humans localize and segregate multiple speakers in complex environments) used in Computational Auditory Scene Analysis (CASA) field lead to a new algorithm referred to as Multiband Position Pitch (MPoPi) algorithm. This method uses the auditory filterbank (which models the human cochlea) as a preprocessing step before the PoPi decomposition. Moreover, different CASA techniques are used to group the location and pitch cues which proves beneficial for concurrent speaker scenarios. These techniques are combined at the low algorithm level. The resulting algorithms are then combined for the multi-channel system and extensively evaluated over a large database. The details of this corpus are presented in Chapter 3. The proposed algorithms outperform the well-known SRP-PHAT algorithm in noisy conditions, and concurrent speaker scenarios.

Tracking of Active Speakers using Particle Filters

In Chapter 5, the Acoustic Source Localization (ASL) methods proposed in Chapter 4 are combined in the particle filtering framework. The reasons behind using a

tracker are discussed and a summary of notable work done for the given problem will be presented. A new likelihood function is proposed using the MPoPi method, which also incorporates voice activity information. The CASA based methods presented in Chapter 4 are modified to be used as importance sampling functions. The Markov Chain Monte Carlo (MCMC) sampling techniques are combined to present novel particle filtering algorithms. Moreover, a novel position and pitch tracking algorithm is developed for tracking multiple speakers. The Acoustic Source Localization and Tracking (ASLT) algorithms introduced in the chapter are evaluated repeating some of the experiments outlined in Chapter 4 and including new experiments especially aimed for the tracking algorithms. The combination of CASA and SMC techniques yield more robust results than traditional methods such as the SRP-PHAT.

Discussion of the Results and Conclusions

Chapter 6 discusses the proposed approaches for the speaker localization and tracking task draws conclusions regarding the proposed methods. Moreover, the experimental results of the doctoral theses presented in Section 1.1.1 will be compared with the results of this thesis. The thesis closes with an outline of open issues and interesting new directions for future research.

1.3 Work Contributions

The scientific contributions that appear in this thesis are extended versions of the published work listed below:

In the beginning of work, the joint position-pitch algorithm was analyzed in detail by extensive evaluations. The traditional approaches for pitch and time-delay estimation were combined with the PoPi method and the results of these modifications were presented at SAM'08 [14].

An auditory system modeling the human cochlea model was used as a preprocessing step to the PoPi algorithm. The resulting method is referred as Multiband Position Pitch (MPoPi) algorithm. The underlying mechanism along with experimental evaluation and comparison with the SRP-PHAT was presented at

HSCMA'08 [15], and INTERSPEECH'08 [16]. Furthermore, the pitch information was combined at low algorithm level to generate coherent spectro-temporal regions for the multichannel system. The enhancement dubbed as MPoPi-STF algorithm was tested on two concurrent speaker scenarios, and presented at INTERSPEECH'10 [17]

The proposed localization methods were combined with Sequential Monte-Carlo (SMC) techniques known as particle filters. The speech-related information was also combined in the likelihood models, which were further refined to deal with multiple speakers tracking using multiple sensors. These methods were presented at DAFx'10 [18], and SAM'10 [19].

Moreover, the combination of the frequency-selective criterion with a new particle filtering algorithm is accepted for publication at INTERSPEECH'11 [20].

At the moment, a journal paper consisting of new contributions made in this thesis is being prepared for the EURASIP Journal on Advances in Signal Processing (Special Issue: Sparse Signal Processing).

Chapter 2

Background

This chapter presents an overview of the fundamental ideas used for the work presented in this thesis. The first section presents an overview of different signal models widely used in the literature to solve the problem of acoustic source localization. In the next section, the problem of source localization is developed using a circular array geometry. The advantage of this geometry is emphasized by exploring the spatial diversity achievable by such a design. This is followed by an overview of the existing localization methods. The performance of the outlined algorithms is investigated against the factors affecting the accuracy of source localization [21]: the Signal to Noise Ratio (SNR), and the reverberation time RT_{60}, which is defined as the time taken by a sound to die away 60 dB below the signal initial power. Other contributing factors taken into consideration are the number of sources, spatial diversity, source motion, signal statistics, number of sensors, and the array geometry. In the last part of the chapter, a special class of algorithms is presented, which exploit the periodicity present in voiced speech (also known as pitch) along with time-delay estimation. This lays out the foundation of the main hypothesis tested in this work, which is that making use of the quasi-periodicity of speech signals improves the cross-correlation based time-delay estimation methods under low SNR and reverberant conditions, and in particular for multiple speaker scenarios.

2.1 Signal Models

Consider M microphones placed inside a room with fixed dimensions, the medium in the room is assumed to be homogeneous, non-dispersive and lossless. The Doppler

effect is negligible as no matter how fast the speakers move, they cannot move faster than the speed of sound in air ($\backsim$ 345 m/sec). The signal $x_m(t)$ received at any microphone is represented as

$$x_m(t) = h_m(t) * s(t) + v_m(t), \tag{2.1}$$

where $m \in \{1, \cdots, M\}$, $s(t)$ is the source signal, and $h_m(t)$ is the Room Impulse Response (RIR) between the source and the microphone m. $v_m(t)$ is the measurement noise recorded at the microphone, which is assumed to be uncorrelated with the speech signal and other microphones.

The simplest signal model is the single source free space model, which assumes no multipath propagation effects [22]. Furthermore, it considers a point source emitting at far-field conditions, where the distance of the source is much larger than the arrays' dimensions or aperture. In such scenarios, the curvature of the spherical wave is much smaller than the aperture size, thus justifying the planar wave assumption for the impinging wave [22]. Therefore, the sound waves propagate along a straight line from the source to the microphone. This is known as *direction of propagation*. The direction opposite to that is defined as the Direction of Arrival (DoA) [22]. The assumption simplifies (2.1), where only the direct path is taken into account ignoring any multi-path propagations, and results in

$$x_m(t) = a_m s(t - \tau_m) + v_m(t), \tag{2.2}$$

where a_m is the attenuation parameter dependent on the distance between the source and microphone, and τ_m is the propagation delay (relative time-delay) between the source and the microphones. In general τ_m is unknown, but with the given geometry of the array, the determination of the source direction is well defined mathematically. Fig. 2.1 illustrates the far-field DoA estimation. A plane wave is impinging on a pair of microphones at distance d with an incident angle θ (i.e., the normal to the wave front). By taking one microphone as the reference, the signal needs to travel an additional distance of $d\cos\theta$ to reach the other microphone. Hence, the TDoA between the two sensors is $\tau = \dfrac{d\cos\theta}{c}$, where c is the speed of sound in air. There is a unique relationship between DoA and TDoA in far-field, where measuring the one is essentially the same as measuring the other.

Approaches have been developed in the literature which takes the multipath effects into account and which provide better modeling of real environments. If the

impulse response between the source and the microphone is known, the localization problem can be solved by combining two or more of the direct path components of the responses. In these models, the impulse response $h_m(t)$ is approximated by a Finite Impulse Response (FIR) filter. The overall situation can be modeled as a Single Input Multiple Output (SIMO) system for single source localization as shown in Fig. 2.2(a). In this scheme, adaptive filters are used to estimate the impulse responses through a constrained Least Mean Square (LMS) algorithm [23]. This technique is known as the Adaptive Eigenvalue Decomposition Algorithm (AEDA), which is discussed in Section 2.6 of this chapter. The Multiple Inputs Multiple Outputs (MIMO) system is used for multi-speaker localization, which is illustrated for two sources in Fig. 2.2(b). The impulse responses in MIMO system are calculated using the Independent Component Analysis (ICA) technique [24]. The MIMO system architecture has been exploited by Blind Source Separation (BSS) systems proposed in [25, 26, 27] for multiple source localization. These BSS algorithms are based on the ICA technique to estimate the impulse responses. The details of these algorithms are discussed in Section 2.7 of this chapter.

In this thesis, the free space signal model is used to solve the problem of speaker localization. The reason behind this decision is that the free space model is conceptually simple. Moreover in common meeting rooms and office space scenarios, the direct propagation paths are accessible to the microphone array to determine the time-delay estimates. On top of that periodicity information combined with

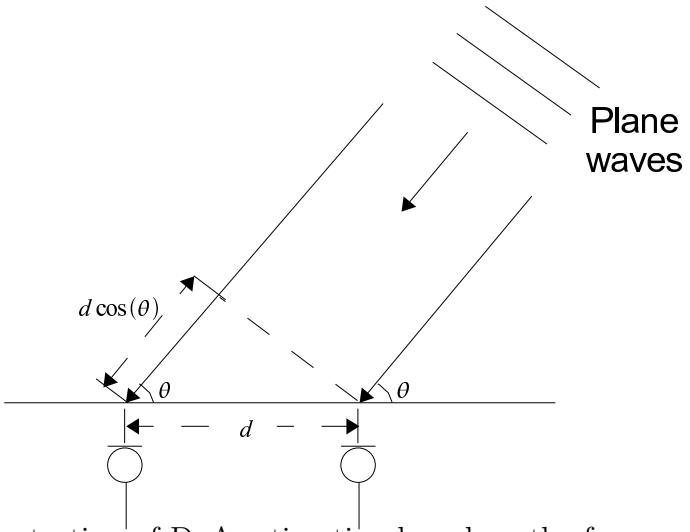


Figure 2.1: Illustration of DoA estimation based on the free space signal model for a pair of microphones. The source is placed at far-field and the resulting planer waves are impinging at an incident angle θ normal to the wave front, and d is the distance between the microphones (as defined in [3]).

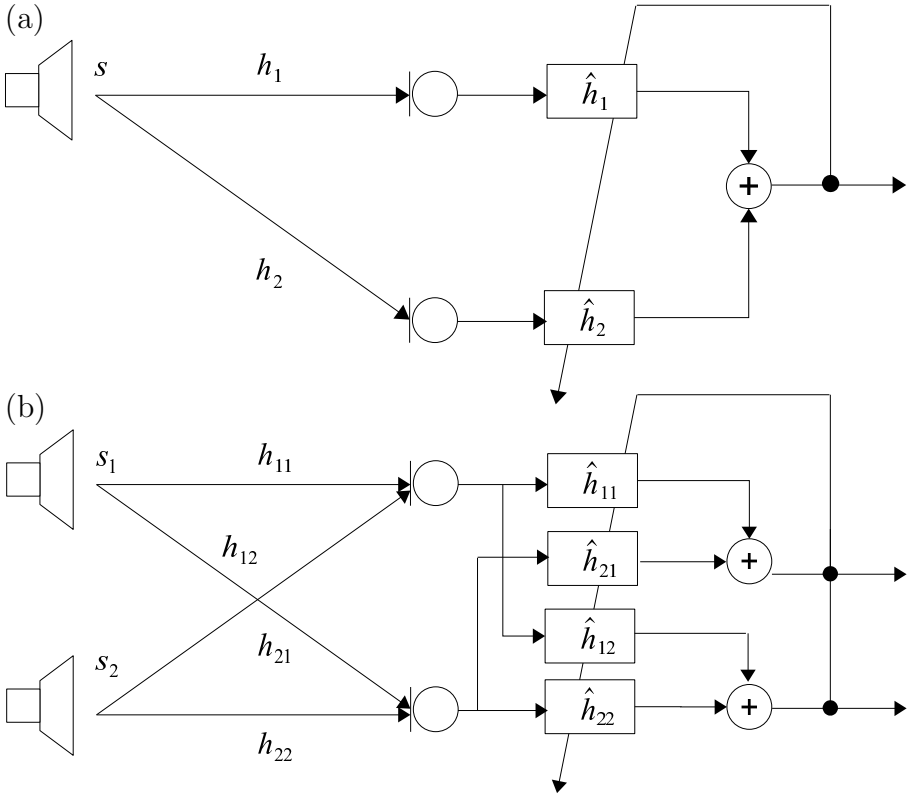


Figure 2.2: Sound source localization using a single source (a) SIMO system, and a two source (b) MIMO system. In the SIMO systems, adaptive filters are used to estimate the impulse responses through a constrained Least Mean Square (LMS) algorithm. The impulse responses in MIMO system are calculated using the Independent Component Analysis (ICA) technique. Based on the given architecture, the direction estimates are determined by combining two or more of the direct path components of the impulse responses.

auditory inspired pre-processing and tracking techniques can yield robust location estimates in real reverberant and noisy environments. The Direct to Reverberation Ratio (DRR) of 1.51 dB was measured in the room, where the distance between source and microphone was 2 m, the details of the acoustic measurements are presented in Section 3.7.

2.2 Acoustic Source Localization

With the known geometry of an array and TDoA measurements, the localization problem can be defined as in [28]. Let $\boldsymbol{\gamma}_m$ be the position of microphones in Cartesian

coordinates forming a uniform circular ring as shown in Fig. 2.3 given as

$$\boldsymbol{\gamma}_m = [x_m \quad y_m \quad z_m]^T, \quad m = 1, \cdots, M. \tag{2.3}$$

The center of the array is considered as a reference and positioned at the origin of the coordinate system, $\boldsymbol{\gamma}_0 = [0 \quad 0 \quad 0]^T$. The acoustic source is placed at $\boldsymbol{\gamma}_s = [x_s \quad y_s \quad z_s]^T$. The distance between the source and the m^{th} microphone is denoted by

$$\eta_m \triangleq \|\boldsymbol{\gamma}_m - \boldsymbol{\gamma}_s\| = \sqrt{(x_m - x_s)^2 + (y_m - y_s)^2 + (z_m - z_s)^2}. \tag{2.4}$$

The difference between the microphone m and j from the source is known as the

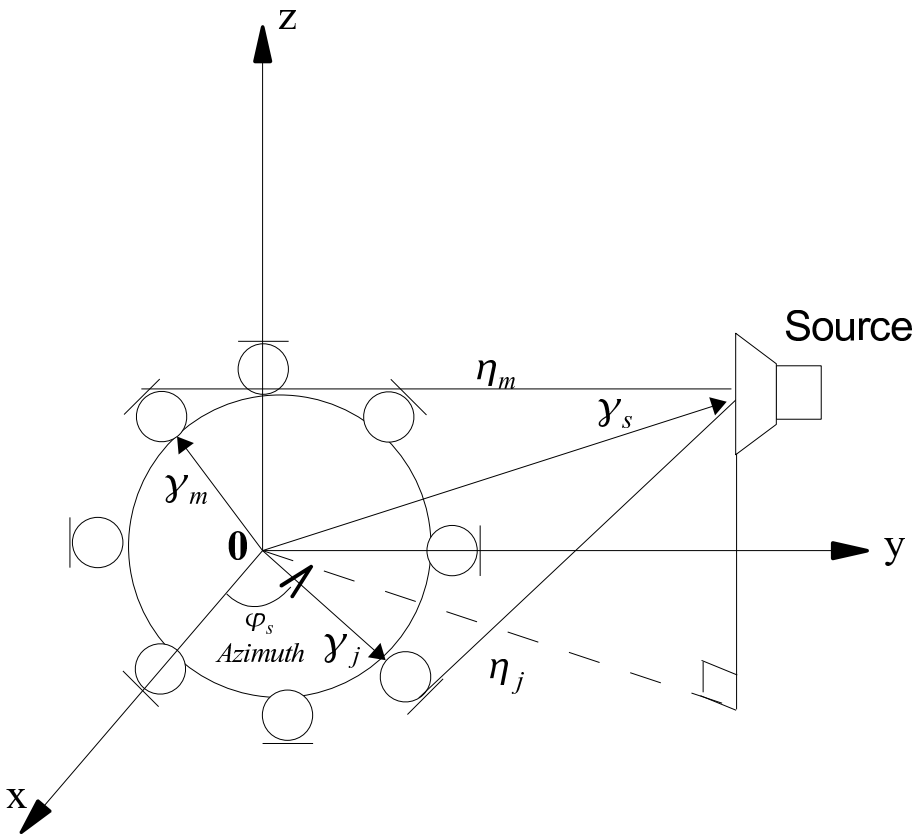


Figure 2.3: Illustration of the source localization problem using a uniform circular array. $\boldsymbol{\gamma}_m$ denotes the position of the m^{th} microphone. η_m is the distance between the source and microphone m. For a pair of microphones m and j, the range difference $d_{mj} = \eta_m - \eta_j$ is proportional to the relative time-delay τ_{mj}. The azimuth DoA φ_s of the source is defined with respect to the 0° reference direction which is axis of the circular array (as defined by [28]).

range difference d_{mj} given by

$$d_{mj} \triangleq \eta_m - \eta_j. \tag{2.5}$$

Hence the range difference is proportional to the relative time-delay τ_{mj} between microphone m and j is given as

$$d_{mj} = c \cdot \tau_{mj}, \tag{2.6}$$

where c is the speed of sound (in m/sec), which can be estimated from the air temperature $\mathrm{t_{air}}$ as $c \approx 331 + 0.610 \cdot \mathrm{t_{air}}$ ($\mathrm{t_{air}}$ in °C). In this thesis, the DoA is defined with respect to the 0° reference direction which is the axis of the circular array. Generally, the 0° reference direction can be assigned along any pair of the array. In this thesis, for a 8-channel array, the reference pair is formed by microphones 1 and 5 (for 24-channel array, it will be microphones 1 and 13).

In [7], the author showed that for a pair of microphones the locus of potential source positions corresponds to one-half of a hyperboloid of two sheets (for illustration, see [7, p.14]), where the microphone position is the foci. For multiple pairs of microphones, the cross-sections of different hyperboloids are used to estimate source position. In the far-field case, the hyperboloid-to-cone approximation presented in [7] leads to simplification of the localization problem. Under such scenarios, the range coordinates are ignored and only the DoA information is taken into account.

Fig. 2.4 presents the spatial locations (hyperbolae) related to TDoA values in time samples using 4 pairs of microphones in a circular geometry. For the given geometry only oppositely placed pairs are used. The range of the TDoA estimates is constrained by the distance between the microphones ($d = 0.55\,\mathrm{m}$) and the speed of sound in the air (fixed at 345 m/sec). In order to estimate the DoA of the source, the hyperboloid-to-cone approximation for the far-field case should be applied for each microphone pair. So that a unique relationship between the TDoA value and DoA can be established (i.e., $\theta = \cos^{-1}(\frac{c \cdot \tau}{d})$). Furthermore, the number of microphones was restricted to 8 instead of 24 for the sake of clarity in the plot. Moreover, it is shown in Section 4.5.1 that a minimum number of 4 pairs of microphones are required to achieve good spatial resolution. In the actual array, 24 microphones form a circular geometry, and 12 pairs of microphones are considered for the localization task.

The methods to solve the acoustic source localization problem can be broadly classified into three categories, that is the ones based on TDoA information, methods

based upon maximizing the output of the steered beamformer, and the techniques using high-resolution spectral estimation concepts. A detailed analysis of the state-of-the-art techniques were presented in [29] (parts of which will be used in this chapter as well). The following sections will present a brief overview of these techniques. In addition, the techniques based on different propagation models and analysis tools closely associated with blind channel identification and source separation concepts are also discussed.

2.3 Generalized Cross Correlation based Localization Methods

The most common pairwise TDoA method is the Generalized Cross-Correlation (GCC) algorithm and its variants [30]. A detailed overview these techniques can

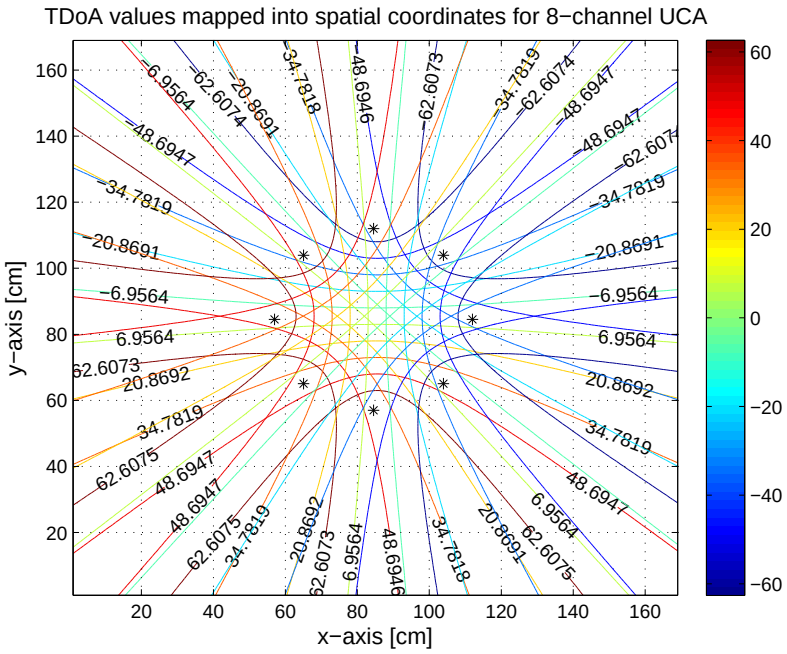


Figure 2.4: Illustration of the TDoA mapping into spatial coordinates. Spatial locations for a uniform circular array of 8 microphones with a diameter of 0.55 m marked by "$*$" is shown. The mapping provides the top-view of the simulated room with dimensions of 1.69×1.69 m along the horizontal and vertical axes, whereas the scale in the plot represents the dimension in cm. The colorbar of the plot shows the range of plotted TDoA values starting from the smallest to the highest.

be found in [30, 31, 32]. The GCC-based methods assume the single-source free space model to estimate the TDoA of propagating waves between two microphones. The TDoA estimates from multiple microphone pairs are then be used to estimate the location of a sound source. The GCC function $R_{x_1x_2}(t,\tau)$ at time instant t and for a given time lag τ is calculated as the inverse Fourier transform of the received signal cross-spectrum $X_1(t,\omega)X_2^*(t,\omega)$, where $X_1(t,\omega)$ is the Fourier transform of the windowed signal $x_1(t)$ and $X_2^*(t,\omega)$ is the complex conjugate of the Fourier transform of the windowed signal $x_2(t)$, which is weighted by a weighting function, $W(t,\omega)$ defined at time instant t is given as

$$R_{x_1x_2}(t,\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} W(t,\omega)\, X_1(t,\omega)X_2^*(t,\omega)\, \exp(j\omega\tau)\, d\omega. \tag{2.7}$$

The type of filtering or weighting functions used with the GCC function is crucial to the performance. Some of the well-known weighting functions proposed in the literature are outlined below.

2.3.1 Maximum Likelihood Weighting

The Maximum Likelihood (ML) weighting function is derived from the magnitude-squared response of the coherence function using signal and noise information. The ML weighting emphasizes different frequencies according to signal to noise considerations. In practice, the coherence function is not known *a priori* and needs to be estimated from the given data. The approximated ML weighting $\hat{W}_{\mathrm{ML}}(t,\omega)$, which is to be used in place of $W(t,\omega)$ in (4.2) for a frame of observed data is given as

$$\hat{W}_{\mathrm{ML}}(t,\omega) = \frac{|X_1(t,\omega)||X_2(t,\omega)|}{|N_1(t,\omega)|^2|X_2(t,\omega)|^2 + |N_2(t,\omega)|^2|X_1(t,\omega)|^2}. \tag{2.8}$$

Here $X_1(t,\omega)$ and $X_2(t,\omega)$ are the received microphone spectra, and $N_1(t,\omega)$ and $N_2(t,\omega)$ represent the noise components that are assumed to be estimated over silent periods. Although it is theoretically optimal for stationary scenarios when there is single-path propagation in the presence of uncorrelated noise, this process can become more difficult in presence of non-stationary signals.

2.3.2 Phase Transform Weighting

The combination of GCC as defined in (4.2) and PHAse Transform (PHAT) weighting known as GCC-PHAT [30]. has been shown to perform well in realistic en-

vironments [33]. The GCC-PHAT whitens the microphone signals, this results in a cross-spectrum function retaining only the phase information. Hence, the phase transform employing the weighting function $W_{\text{PHAT}}(t,\omega)$ is given by

$$W_{\text{PHAT}}(t,\omega) = |X_1(t,\omega)X_2^*(t,\omega)|^{-1}. \tag{2.9}$$

It eliminates the influence of the spectral magnitudes and produces a GCC function that depends entirely on the phase of the cross-spectrum.

2.3.3 Bandpass Weighting

The simplest weighting function is one that attenuates frequencies outside of the band of interest. For speech, this band is typically 300 Hz - 6 kHz. It is advantageous to suppress frequency components below 300 Hz because much of the power in this range is from low-frequency noise. Furthermore, for the long wavelengths of low-frequency propagating waves, it is difficult to determine their direction of propagation using a small-aperture array. The bandpass weighting $W_{\text{BP}}(t,\omega)$ can be used with other weighting functions as its primary role is to emphasize the speech signal energy band, which is given as

$$W_{\text{BP}}(t,\omega) = \begin{cases} 1, & 2\pi 300 \text{ Hz} \leq \omega \leq 2\pi 6000 \text{ Hz}; \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

2.3.4 Cross Power Spectrum Phase

The frequency domain formulation of the GCC-PHAT principle was presented in [34, 35, 36], which is known as the Cross-power Spectrum Phase (CSP) method. It is based on the GCC method, where the weighting function is chosen such that the resulting cross-power spectrum between the two sensors is normalized; thus keeping only the phase difference information. Therefore, the normalized CSP $\mathcal{S}(t,\omega)$ is given as

$$\mathcal{S}(t,\omega) = \frac{X_1(t,\omega)X_2^*(t,\omega)}{|X_1(t,\omega)||X_2^*(t,\omega)|}. \tag{2.11}$$

The inverse Fourier transform of $\mathcal{S}(t,\omega)$ results in the cross-correlation $R_{x_1x_2}(t,\tau)$. A visual representation referred to as Coherence Measure (CM) was proposed in [36].

It is expected to have a prominent peak at the respective time-delay between the two sensors. The CM corresponding to $R_{x_1x_2}(t,\tau)$ is given as

$$C_{12}(n,l) = R_{x_1x_2}(n,l), \tag{2.12}$$

where n indicates the time instant in samples, and l represents the number of samples corresponding to the lag τ and $R_{x_1x_2}(n,l)$ denotes the digital representation of $R_{x_1x_2}(t,\tau)$. The maximum lag estimate for time samples n $(1 \leq n \leq N)$ can be derived as

$$\hat{l}_{12} = \underset{l}{\operatorname{argmax}} \left[\sum_{n=1}^{N} \mathbf{C}_{12}(n,l) \right]. \tag{2.13}$$

The CM can be calculated for every pair of microphones by determining the maximum lag estimate of each pair. These delays can be combined depending on the geometry of the array to determine the position of the source.

The cross-correlation based TDE techniques require longer segment lengths to improve performance. The dynamic environments of various acoustic source localization applications, however, require high update rates (which means short data segments). There are various studies conducted to evaluate the performance of the GCC techniques in realistic environments. The study in [37] evaluated the cross-correlation based localization technique in simulated reverberant environment. The authors reported that, when using a pair of microphones, the anomalies in time-delay estimation jumped from 0 to 90 percent as the reflection coefficient increases from 0.6 to 0.8. The source was placed at a distance of 4 m from the microphones.

This outcome suggests that cross-correlation techniques should be further investigated and whether the use of auditory-based methods inspired from psychoacoustics domain can improve the location accuracy of cross-correlation-based methods. The details of this study is presented in Chapter 4 of this thesis.

In addition to the contributions of the present work, there have been some promising methods utilizing the GCC-based approach to localize multiple speakers in reverberant environments. [38] presents the extension of GCC-based methods to localize multiple speakers by a TDoA disambiguation scheme in reverberant environments. To resolve the problem of TDoA ambiguity among multiple microphones, the authors exploited the raster condition where the direct paths in cross-correlations can be identified by combining the extremum positions of cross-correlations and the auto-correlations. The other condition based on the redundancy of TDoAs was termed

zero cyclic sum condition. In the first step, the echo path TDoAs are identified in the cross-correlation and removed by using the raster matching with the corresponding auto-correlations. To exploit the zero cyclic sum conditions, a concept of consistent graphs was introduced in [38] where each node in the graph represents a microphone and the edges represents the TDoA between two microphones. The synthesis of consistent graphs rather than analysis lead to consistent triplet concept where the sum of actual TDoAs in the consistent triplet must be equal to zero. An efficient synthesis algorithm was developed in [38] to combine and extend the consistent triplets to larger TDoA graphs using some simple rules. This scheme results in consistent TDoA graphs each containing the TDoA estimates belonging to a single speaker. The proposed scheme was tested for two speaker data recorded using eight microphone placed at different locations in a room with reverberation time of 300 msec. The proposed algorithm outperformed the well-known SRP-PHAT by yielding smaller TDoA errors and successfully localizing both speakers positions. Another advantage of the TDoA disambiguation scheme is that it can be applied to TDoA estimates derived using any localization algorithm.

2.4 Steered Response Power Methods

An array with multiple microphones can act as a beamformer [22]. To focus on a signal coming from a particular direction, the steering parameters of the array are adjusted, which makes the beamformer to steer towards the given direction. The output power of a beamformer in such cases is known as the Steered Response Power (SRP) expressed according to [8] as:

$$\mathbf{P}(\Delta_1 \dots \Delta_M) = \int_{-\infty}^{\infty} Y(\omega, \Delta_1 \dots \Delta_M)\; Y^*(\omega, \Delta_1 \dots \Delta_M) d\omega, \tag{2.14}$$

where $Y(\omega, \Delta_1 \dots \Delta_M)$ is the output of the filter-and-sum beamformer (for illustration, see Fig.6.1, [8, p.76]) with a set of M steering delays $\Delta_1 \dots \Delta_M$, and $Y^*(\omega, \Delta_1 \dots \Delta_M)$ is the complex conjugate of the output response. The methods which estimate the source location by finding the maximum in the output power or in the SRP of the beamformer are called direct methods. The SRP has a maximum when the steering direction of the beamformer matches with the actual location of the source. There is an extensive discussion in [8] regarding the usage of the SRP method for source localization and beamforming. The author observed that

the steering delays, $\hat{\Delta}_1 \dots \hat{\Delta}_M$ which maximize (2.14) correspond to the TDoA estimates among microphones [8]. This behavior is similar to the GCC function for two microphones, which peaks when the time-delay τ corresponds to the TDoA of sound waves between the two microphones. For the m^{th} and n^{th} microphone signals, the TDoA $\hat{\tau}_{mn}$ will be the difference between delays that maximize the SRP given as

$$\hat{\tau}_{mn} = \hat{\Delta}_m - \hat{\Delta}_n. \tag{2.15}$$

These steering delays can be computed from the SRP by scanning over a a predefined region and then using the estimated steering delays to steer the beamformer [8]. The estimate of the source location is determined when the steered power peaks or a global maximum is achieved for a specific spatial point. Hence, SRP is a function of the candidate location of the source $\boldsymbol{\gamma}$. Therefore, (2.14) can be rewritten as

$$\mathbf{P}(\boldsymbol{\gamma}) = \mathbf{P}(\Delta_1 \dots \Delta_M), \qquad \Delta_m = \tau_0 - \frac{|\boldsymbol{\gamma}_m - \boldsymbol{\gamma}|}{c}. \tag{2.16}$$

For the far-field assumption, the propagation delays can be expressed in terms of the assumed direction of propagation, $\boldsymbol{\zeta}_0$, as follows

$$\Delta_m = -\frac{\boldsymbol{\zeta}_0 \cdot \boldsymbol{\gamma}_m}{c}. \tag{2.17}$$

Here the negative sign shows that the look vector points in direction opposite to the direction of propagation. It can be represented in terms of azimuth and elevation angles θ and ϕ, respectively, which is given as

$$\boldsymbol{\zeta}_0 = \begin{bmatrix} \cos\phi \sin\theta \\ \cos\phi \cos\theta \\ \sin\phi \end{bmatrix}. \tag{2.18}$$

Here the angles θ and ϕ represent the look direction, relative to the array's origin. The discussion above leads to defining SRP in terms of GCC [8], which can be represented for a pair of microphones resulting in the m^{th} and n^{th} signals, as:

$$R_{mn}(t,\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} W_{mn}(t,\omega)\ X_m(t,\omega) X_n^*(t,\omega)\ \exp(j\omega\tau) d\omega, \tag{2.19}$$

where the time lag τ can be represented as $\Delta_{mn} = \Delta_m - \Delta_n$. Hence, the SRP response for M number of microphones can be computed by summing all possible

pairwise GCC crossings, which are time shifted by the differences in the steering delays given as [8]:

$$\mathbf{P}(\Delta_1 \dots \Delta_M) = 2\pi \sum_{k=1}^{M} \sum_{q=1}^{M} R_{kq}(\Delta_q - \Delta_k). \tag{2.20}$$

The Steered Response Power Phase Transform

[8] proposed a PHAT weighting from the GCC techniques to be used into the cross-correlation function in (2.19), the resulting algorithm is known as Steered Response Power Phase Transform (SRP-PHAT) [8]. In practice, the SRP-PHAT function $\tilde{\mathbf{P}}_b^{\text{PHAT}}(\Delta_1 \dots \Delta_M)$ is determined as

$$\tilde{\mathbf{P}_b}^{\text{PHAT}}(\Delta_1 \dots \Delta_M) = \sum_{\kappa=1}^{\mathcal{K}} \tilde{Y}_b^{\text{PHAT}}(\kappa, \Delta_1 \dots \Delta_M) \tilde{Y}_b^{*\text{PHAT}}(\kappa, \Delta_1 \dots \Delta_M), \tag{2.21}$$

where the SRP response using the PHAT weighted function $\tilde{Y}_b^{\text{PHAT}}(\kappa, \Delta_1 \dots \Delta_M)$ is determined as

$$\tilde{Y}_b^{\text{PHAT}}(\kappa, \Delta_1 \dots \Delta_M) = \sum_{m=1}^{M} \frac{X_{m,b}(\kappa)}{|X_{m,b}(\kappa)|} \exp(-j\omega\Delta_m), \qquad \kappa = 1, \cdots, \mathcal{K} \tag{2.22}$$

where $X_{m,b}(\kappa)$ is the DFT of the b^{th} block of the m^{th} microphone signal and the output is a function of the discrete temporal frequency κ and a set of m continuous steering delays. The summation is taken over $\mathcal{K}$ discrete frequencies to obtain the SRP.

The author showed in [8] that the accuracy of location estimation can be improved by exploiting the microphone redundancy. This is achieved by combining the microphones' signals rather than by combining the multiple TDoA estimates. The work presented in [8] focuses on highlighting the key issues in performance degradation of various source localization algorithms such as reverberation, which even in a mild case can severely impact the performance of short-time GCC-based localization techniques. The performance of SRP-PHAT was reported to outperform SRP and GCC-PHAT using real microphone array data that were collected in rooms having 200 to 400 msec of reverberation time [8]. Furthermore the author showed that although SRP-PHAT requires much more computation, it can provide accurate results using smaller data segments in comparison to GCC-based localization

techniques such as GCC-PHAT, which makes SRP-PHAT more useful for moving sources.

The SRP-PHAT algorithm generates an acoustical energy map in the search space to find the position of the active source. The grid-based search approach makes SRP-PHAT computationally expensive [8]. Another problem is the occurrence of local maxima in SRP space, which makes the conventional search algorithms such as simplex or a gradient based techniques unfeasible. Moreover, in SRP based methods, the search space can be large (e.g., the whole room such as a meeting room). In order to decreases the search space, the authors in [39] presented a Stochastic Region Contraction (SRC) based SRP-PHAT algorithm, which decreases the computational complexity by factor of three. There is no assurance, however, that the maximum found by the algorithm belongs to the desired source because it can also be due to a brief dominant noise peak caused by reverberation or non-speech source. The authors later presented two enhanced SRP-PHAT methods for multiple concurrent speakers. One technique consists of Gaussian Mixture Models (GMM) trained by the Expectation Maximization (EM) algorithm to estimate clusters from the SRP-PHAT function. The other technique is a Region Zeroing (RZ) method which applies thresholding on the SRP-PHAT response to find the maximum points in the plane. Then a 50 cm minimal distance rule between two speakers is applied to omit the SRP-PHAT values around the maximum points. The remaining maximal points are then clustered to estimate source location [40]. These techniques are promising, but they are tested on a limited amount of data, which does not validate their true significance.

In addition to optimization based techniques to reduce the search volume, an approach is proposed in [41] that discretizes the search space into active volumes or sectors and performs the localization only in those sectors. Another interesting modification to SRP-PHAT is presented in [42], where an inverse mapping of the relative delays to possible source positions is presented. The proposed method reduces the computational load while keeping the localization process accurate. Moreover, the authors in [43] evaluated the use of interpolation functions to increase the time resolution (hence increasing the spatial resolution which decreases the effects of spatial quantization) of the cross-correlations used in the SRP-PHAT algorithm. Using real data recorded in a concert hall, the authors showed the effectiveness of these interpolation techniques.

A theoretic comparison of the baseline joint position-pitch method with the SRP-

PHAT technique has been developed and presented in Appendix A. The performance of the proposed methods presented in this thesis has been compared with the SRP-PHAT method.

2.5 High-Resolution Spectral-Estimation Based Localization Methods

The subspace methods were originally designed for narrowband signals in the fields of radar and communications. Some of the well-known methods are MUSIC [44] and ESPRIT [45] algorithms. One of the popular variants of the MUSIC algorithm is Root-Music [46]. These approaches are "based on the eigen-decomposition of the spatio-temporal covariance matrix of the microphone signals" [47, p.240]. The covariance matrix can be decomposed into orthogonal eigenvectors. These eigenvectors corresponding to the largest eigenvalues form an orthogonal basis of a subspace known as signal subspace, the rest forms the noise subspace. One main advantage of these techniques is that they can localize multiple sources. There have been various efforts made to make these narrowband techniques useful for broadband signals such as speech. One of the well-known modification for speech signal using a circular microphone array was presented in [48], which is known as Eigen-Beam ESPRIT (EB-ESPRIT) algorithm. The EB-ESPRIT algorithm utilizes wave-field decomposition technique as a basis for ESPRIT algorithm. In practice, the covariance matrix is generally unknown and needs to be estimated, which requires that the sources should be fixed and that the signals should be stationary to perform temporal averaging. This is hardly the case for speech signals and the acoustic environments. Multi-path propagation still seems to be a problem with these methods, since reverberations have to be modeled explicitly. Recently, a new technique using EB-ESPRIT method with spherical microphone array was proposed in [49] for multiple speaker localization in reverberant environment.

2.6 Adaptive Eigenvalue Decomposition Algorithm

The Adaptive Eigenvalue Decomposition Algorithm (AEDA) [23] is a well-known localization method based on the SIMO system shown in Fig. 2.2(a). AEDA belongs to a group of methods working under the principle that the relative time-delay between two microphone signals can be determined by the temporal difference between the maxima of the room impulse responses. These maxima are associated with the direct path between the source and the microphone. The effects of reverberations are negligible on direct paths, as they appear later in the impulse response. This leads to the blind estimation of impulse responses. The AEDA is based on the convolutive reverberant model for sound propagation, which means that AEDA estimates the direct delay path without assuming free space or delay only signal model. According to the SIMO system presented in Fig. 2.2(a), the author in [23] showed that the two recorded signals x_1 and x_2 of a single source s at time frame k can be combined with their respective impulse responses h_1 and h_2 given as

$$x_1 * h_2 = s * h_1 * h_2 = x_2 * h_1, \tag{2.23}$$

which in the noiseless case gives the equality

$$x_1 * h_2 - x_2 * h_1 = 0. \tag{2.24}$$

This relationship is then used by the author in [23] to develop an LMS algorithm which is adaptive and has the capability to jointly estimate the impulse responses by using the signal covariance matrix. The underlying assumption is that the combined impulse response is the eigenvector of the signal covariance matrix whose eigenvalue equals 0. The AEDA outperforms the conventional localization algorithm as it takes into account the effects of reverberations. However one limitation of AEDA is that it requires the covariance matrix to be full rank. However, it was reported in [50] that for periodic signals this condition is hardly met and harmonic signals have ill-conditioned covariance matrices. For smaller data segments of around 20 msec, there might be only voiced or unvoiced speech in the segment but rarely a mixture of both. Moreover, the AEDA gives a single TDoA estimate, whereas the cross-correlation-based techniques provides a localization function [10]. It is difficult to form pseudo-likelihood functions presented in Section 5.4 of this thesis with the AEDA as its localization function is essentially a delta function. Therefore, a comparison with

this algorithm is omitted in this thesis.

2.7 Blind Source Separation Based Time Delay Estimation Methods

The extension of the SIMO model to multiple sources is known as blind MIMO channel identification [51]. The MIMO model for two sources is shown in Fig. 2.2(b). This class of methods does not require geometrical knowledge of the microphone array. These approaches were shown in [25, 26] to be closely linked to Blind Source Separation (BSS) techniques. Thus, the solution of blind MIMO channel identification problem results in determining the complete model of the meeting room, which allows joint localization and separation of active sources in the acoustic scene. It was reported in [21] that the broadband BSS can also be regarded as a generalized subspace approach based on the block-diagonalization of the signal correlation matrix using second order statistics. Furthermore, it was illustrated in [21] that the broadband MIMO filtering approach generalizes and unifies both the traditional subspace methods and the Blind System Identification (BSI) methods. The BSS method described in [51] referred to as TRINICON (TRIple-N Independent component analysis for CONvolutive mixtures) framework is one of the most recent and powerful TDoA estimation techniques, where the systematic incorporation of time lags into the correlation matrix can handle room reverberation. To address various issues of the BSS techniques such as the spatial ambiguity in case of multiple arrays, localization of more than two sources, new methods and modifications are proposed in [52, 53, 27].

2.8 Pitch Based Localization Algorithms

The localization methods discussed above only use geometry dependent features, for example TDoAs, and do not exploit any speech related feature. One such feature is the periodicity present in voiced speech, which is also known as pitch. The pitch is one of the characteristics of human speech, created by the vocal chords of a speaker during voiced portions of speech. It appears in time-domain speech signals as a repeating waveform. The period between the adjacent peaks of the

repeating waveform determines the pitch (for more details on speech production, see Ch. 3 in [54]). The term pitch and fundamental frequency are interchangeably used throughout this thesis. The work presented in this thesis hypothesizes that making use of the quasi-periodicity of speech signals helps the cross-correlation based time-delay estimation methods under low SNR conditions. The summary of the methods presented in the literature using similar ideas is given below.

2.8.1 The Position-Pitch Algorithm

In [12], a two-channel based joint position and pitch extraction method known as PoPi algorithm was presented. In the PoPi algorithm, the acoustic sources are illustrated by their position and fundamental frequency in a 2D plane known as the PoPi plane. The underlying concept of the method is based on two well-known features of the cross-correlation. The first one is that the fundamental frequency of the source is encoded in form of a lag between each correlation peak, which is the inverse of the fundamental frequency. The second feature is the correlation lag corresponding to the cross-channel delay (or time-delay), which is present in the joint shift of all the cross-correlation peaks. The frequency domain formulation of [12] is presented in [55], where the basic properties of the representation are further explored and improvements are proposed by applying multiple microphones with circular arrangement. Both techniques were tested on synthetic voiced signals without any multipath propagation and background noise effects. Under these ideal conditions, these methods show promising results but they fail to localize speech sources in a realistic environment. The PoPi algorithm is used as a baseline method for this thesis. In Chapter 4, the PoPi method is extended to a multichannel system and new weighting and pre-processing techniques are presented to improve the algorithm's performance in realistic environments.

2.8.2 Joint Time-Delay and Pitch Based Maximum Likelihood Estimator

The study conducted by [56] was the first one to propose a maximum likelihood estimator to determine the time-delay and pitch of the speech signals assuming that the signals are received at a pair of microphones. The authors [56] introduced a

scoring function which depends on postulated pitch period P and time-delay D by taking two microphones. A new signal is first computed by averaging a microphone signal for each postulated period and its multiples. The theoretic time-delays from the other microphones in the array are further added by circular shifting the signals (modulo P), and averaging over all the microphones. A second signal is constructed in similar manner but averaged over the squared samples of the microphone signals. Then the scoring function is formed by taking the difference between these two new signals and finally a peak detector finds the delays and period where the maximum score is attained.

Through the use of ML method, the authors showed that the proposed technique for pitch estimation uses the data efficiently by taking into account $\frac{1}{2}N(N-1)$ ($N = \lfloor \frac{T}{P} \rfloor$ is defined as number of complete periods contained in the data frame T) cross-correlations than the correlation method which uses only $N-1$ cross-correlation between adjacent length P data segments. Similarly, for the time-delay estimation, the proposed method uses $\binom{M}{2} \cdot N^2$ cross-correlations between length P data segments than correlation method which uses $\binom{M}{2} \cdot N$ combinations.

The experiments were conducted on a synthetic data by shifting the speech signal by few samples to emulate the speech recorded at a second sensor. A white noise was further added to the speech signals to test the algorithm for different SNR levels ranging from -20 dB to 5 dB. The Root Mean Square Error (RMSE) is reported to decrease by 25% in comparison to the correlation method at -15 dB SNR. This setup is somewhat restrictive and does not validate the use of such scheme for real recorded signals from a microphone array. The idea of the scoring function is similar to the baseline PoPi method where instead of microphone signals; the cross-correlation between a pair of microphones is used to determine the position-pitch relations. In [56], one microphone is considered as a reference and to account for the delays of other microphones, a circular shift is applied to the reference signal. In practice, the other channels in the microphone array also contain effects of multi-path propagation and background noise which cannot be compensated by a simple delay. Moreover, the choice of a reference microphone channel is critical to the performance of the method. The poor selection can lead to erroneous results. The PoPi algorithm does not make that selection step rather different pairs of microphones are combined to determine pitch and DoA estimates. All the pairs in the UCA sample the spatial region differently. Hence the pairwise combination leads to robust location estimates than a reference microphone-based technique.

2.8.3 Subspace Algorithm for Joint Time-Delay and Frequency Estimation of Multiple Sinusoids

The joint estimation of time-delay and frequency of sinusoidal signals has been reported in [57]. According to this technique the source signal can be represented as a sum of complex sinusoids with unknown amplitudes and distinct normalized radian frequencies. The authors in [57] presented a subspace method to jointly estimate time-delay and frequency using covariance matrix of received signals. This technique derives the frequencies from the eigenvectors. Moreover, the speech signal used for evaluation is simulated by using a complex sinusoidal model [58]. The output at the second sensor is generated by shifting the original signal by a fixed delay. The proposed method outperformed the GCC algorithm for different SNR levels. The time-delay and frequency estimation is inferior to the Cramér Rao Lower Bound (CRLB) by only a few dB. Here again the absence of real speech recorded in a realistic acoustic condition is a problem, therefore one cannot guarantee similar performance for real-world recordings. Even though the subspace approaches are known to accommodate multiple speakers, they cannot handle multi-path propagations. Therefore, much effort is needed to make the given approach workable in real environments.

2.8.4 State-space Approach for Joint Time-Delay and Pitch Estimation for Speaker Localization

In [59], a method based on state-space realization of a subspace method [57] was proposed. The frequencies of the constituent components are determined directly from the eigenvalues of the state transition matrix, where the time-delay is determined using the observation matrix and the pitch is calculated from a set of estimated frequencies which are harmonically related. To demonstrate the performance of the method, computer simulations were carried out for complex sinusoids, synthetic speech, and real speech. For different SNR levels, the estimation performance of the time-delay and the frequency is tested for different SNR levels. The proposed method is reported to perform well for SNR $\geq -10\,\mathrm{dB}$. There was no effort made to record the signals with real microphones but an artificial time-delay was created between the two signals by shifting one signal by some samples to generate the second signal.

2.8.5 A Pitch-Based GCC Weighting Function

A technique using a pitch based GCC weighting function was tested for time-delay estimation process on data recorded at a pair of microphones in [60]. It makes use of the observation that segments of periodic speech will maintain a degree of periodicity when subjected to the effects of noise and reverberation. This feature can be incorporated in the GCC function in the form of an extended weighting function. The author suggested that all source relevant DoA information is located at the corresponding harmonics, as these can be assumed to be of considerably higher energy than environmental noise and to be in general well separated from other sources' harmonics. The authors proposed to model the speech spectrum $X(\omega)$ as a product of spectral envelope $H(\omega)$ and an excitation spectrum $E(\omega)$. An error criterion is generated by dividing the spectrum into frequency bands centered around the harmonics of the fundamental frequency, therefore the error ε_n associated with the n^{th} harmonic is given as

$$\varepsilon_n = \frac{1}{2\pi} \int_{a_n}^{b_n} |X(\omega) - A_n E(\omega)|^2 d\omega, \tag{2.25}$$

where the interval $[a_n, b_n]$ is centered around the n^{th} harmonic and A_n is the corresponding complex spectral amplitude, which is calculated to minimize the error function. All these error functions are summed to generate the error criterion for a given fundamental frequency. As this process is exhaustive in terms of computation, any traditional time-domain pitch estimation procedure can be used to determine a coarse pitch estimate, which can be refined later by frequency-domain analysis. The author proposed a normalized error E_n for each n^{th} harmonic, which calculates the degree of accuracy with which the periodic excitation matches the observed spectral region given by

$$E_n = \frac{\varepsilon_n}{\frac{1}{2\pi} \int_{a_n}^{b_n} |X(\omega)|^2 d\omega}. \tag{2.26}$$

This acts as a measure of harmonicity for every signal received from a pair of microphones where a value close to 0 indicates strong harmonic interval and a value close to 1 indicates noise of unvoiced interval. Hence, a new weighting function $W_P(\omega)$ for GCC is formulated by incorporating these harmonic voicing mixture error for each microphone pair and is given by

$$W_P(\omega) = \frac{(1 - \max(E_{n1}, E_{n2}))^{\varsigma}}{|X_1(\omega) X_2^*(\omega)|}, \quad \omega \in [a_n, b_n] \tag{2.27}$$

where the value of ς is selected between 1 and 2. This was shown to be an effective value based upon a limited number of experiments. The performance of the method was compared with the Phase Transform estimator (Section 2.3.2) and the ML weighted estimator (Section 2.3.1). The recordings at different positions were simulated themselves by convolving speech signals with the impulse responses which were simulated by the image method [61]. Furthermore, different acoustic conditions were artificially created such as the room reverberation time varying from 0 to 0.3 sec and artificial noise was added to the signals at different SNR levels ranging from 0 to 40 dB. For comparison among different weighting functions, statistical parameters such as bias, variance, and RMSE were calculated. The pitch-based weighting function outperformed the PHAT-weighting function under noisy conditions. It was reported to be less vulnerable to the effects of reverberation in comparison to the ML weighting function. Although it is speculated in [60] that the pitch-based weighting can be advantageous for multiple source scenarios, the extension of the technique to multiple speakers does not seem straightforward.

2.8.6 Excitation Source-Based Time-Delay Estimation

In [62], a speech enhancement method was proposed based on the observations that the spectral processing techniques performance degrade in presence of other speech sources, background noise and reverberations. It was further shown in [62] that the impulse-like excitation present in voiced speech is robust in the sense that the relative position of the peaks in the excitation signal or epochs remains unchanged in the direct sound signals only shifted by a fixed delay due to spatial locations of the microphones. The excitation signal was generated by applying Hilbert Envelope (HE) to the Linear Prediction (LP) residual signal. The locations of the estimated epochs were then used to determine the time-delay estimate between a pair of microphones. The knowledge of time-delay was then used to enhance the speech signal, which was done by aligning the HEs of the LP residual signals according to the estimated time-delay and then exciting the modified LP residual signals with a time-varying all-pole filter. The proposed time-delay estimation method for a single speaker was compared with GCC method using a 50 msec frame with 10 msec frame shift. The proposed method showed smaller sample deviation in comparison to the GCC method for different pairs of microphones. Recently, in [63], similar idea was used by the authors to estimate the instantaneous frequencies of two speakers

from the mixed signal recorded with a pair of microphones in a reverberant environment. In this scenario, the time-delays for both speakers are estimated using the excitation source based TDE technique presented in [62]. In order to estimate the instantaneous frequencies of the two speakers, a zero frequency filtering algorithm was applied on the delay compensated HEs of the LP residual signals.

2.8.7 Joint Time Delay and Pitch Estimation Using a Neural Network

The joint time-delay and fundamental frequency estimation in form of a neural network was explored in [64]. The authors proposed an extension to Recurrent Timing Neural Networks (RTNNs) and suggested its usage in estimating joint Interaural Time Difference-Fundamental Frequency (ITD-F_0) cues. This extension forms a 2D RTNN by adding a second layer of coincidence detectors, which results in one axis of RTNN representing F_0 and the other ITD. The sources can be distinguished based on their separation in ITD-F_0 space. The grouping and the segregation are carried out within individual frequency channels without recourse to across-channel estimates of F_0 and ITD. The system is evaluated on spatialized speech created using Head Related Transfer Functions (HRTF) measured from a KEMAR artificial dummy head [65] in an anechoic environment. To create concurrent speaker scenarios, the speech signals were convolved with different measured HRTFs and then added at a relative SNR of 0 dB. In this approach, no reverberation effects were considered and the SNR of combined speakers was not varied in order to see the effects of loudness mismatches. This method requires a complex setup of RTNN-layers and to analyze different ITD and pitch combinations. Moreover, the performance of proposed RTNNs using real-world data is missing.

2.8.8 The Correlogram-Based Joint Time Delay and Pitch Estimation

Another technique proposed in [66] makes use of joint ITD and F_0 cues to localize multiple speakers in reverberant environments. This method uses binaural data recorded for different speaker configurations using a dummy head placed in a real environment. The multi-speaker data are filtered through an overlapping bandpass

filterbank known as "gammatone filterbank", whose center frequencies are uniformly spaced on an Equivalent Rectangular Bandwidth (ERB) scale (details presented in (4.9)). The ITDs are estimated using the cross-correlation of each frequency band, where the cross-correlations are determined between the filtered left and right ear signals. The fundamental frequency is estimated from the auto-correlogram of the averaged left and right ear signals. The method was tested by employing four baseline systems: two incorporate the spectro-temporal extent of the speech segment and the other two are based on localization cues only. Evaluations using real world recordings showed that the proposed technique allows robust location estimates from a noisy cross-correlogram ITD cues by integrating the spectro-temporal regions which are dominated by a single source. In [67], the method was further extended by weighting each spectro-temporal fragment and using a weighted mean to combine the fragments to gather location estimates. These weighting criteria are based on models such as the perceptual precedence effect, the measure of interaural coherence between the two ears' signals, and the data driven approach trained in matched acoustic conditions. This scheme has shown to perform reasonably well in reverberant environments.

2.9 Summary

This chapter presented different signal models based on the physical properties of wave signals and acoustic environments. The state-of-the-art localization methods exploiting these signal models and signal characteristics are summarized. Furthermore, a group of algorithms exploring the advantage of combining a speech related feature or namely the fundamental frequency, with the time-delay estimation task is discussed. Based on similar lines, a recent technique first presented in [12] becomes the main starting point of my work which is explored in detail in Chapter 4. The shortcomings of the original algorithm are discussed in detail and modifications are proposed to improve its performance in realistic environments.

Chapter 3

Data Acquisition and Corpus Building

This chapter gives an overview of existing microphone array databases, and outlines the reasons for building an in-house microphone array. Furthermore, the issues related to various existing corpora are discussed, which motivates the recording of an extensive multi-channel database addressing the outlined problems in existing corpora. In the next section, the details of the recording setup including the specification of the microphone array are presented. The design characteristics of the array are specified in the following section by using different specifications, such as the spatial resolution of the array, the beam pattern, and the maximum operating frequency. Then details about the speaker setup and reference speech database used for the recordings are given. The process of speech segmentation and speaker labeling is discussed in detail through illustrative examples. In the last part of the chapter, some basic acoustic measurements of the recording room are presented followed by a discussion of the distortion metrics used for calculation of background noise and reverberation levels.

3.1 Available Microphone Array Databases

During the last decade, the increase in multi-channel system applications for hands-free speech acquisition and enhancement brought a number of corpora created by different research groups all over the world. These databases are focused on different microphone array applications. One of the largest and most notable corpus is the Automatic Multi-Party Interaction (AMI) meeting corpus [68]. It is a multi-modal database containing 100 hours of meeting recordings including close-talking

microphones, microphone arrays, individual, and room-view cameras with electronic output from a slide projector and whiteboard. Smart rooms equipped with all the above mentioned sensors were designed at three different locations of the participating research groups. One of them was designed at Idiap, Switzerland; the other was at the University of Edinburgh; and the third was at TNO Human Factors Research Institute, The Netherlands. The design elements for all smart rooms prepared for the recordings were slightly different at each location. My main interest was to explore only the microphone array data used in the corpus. Two kinds of geometries were used in the corpus, a linear and a circular array. The diameter for the circular array was kept at 0.2 m diameter in all smart meeting rooms. A detailed annotation of the recordings was made publicly available by the end of 2007 on the project website [69].

Another well-known microphone array database was developed for the EU funded Computers in the Human Interaction Loop (CHIL) project [70]. In this project, a smart room was designed with a distributed microphone network in the form of seven T-shaped arrays, each consisting of 4 microphones. The arrays were placed on the walls of the room. The room was also equipped with a 64-channel NIST MARK III [71] linear microphone array fixed on one of the walls. All arrays were used for the multi-channel recordings.

A recent microphone array database was prepared for another EU funded project, the Distant-talking Interfaces for Control of Interactive TV (DICIT) project [72]. The project aim was to develop and analyze advanced audio/speech techniques using multi-microphone devices, which were used as add-on features in interactive TV systems. A 15 element nested linear microphone array was built to carry out multi-channel tasks such as source localization, beamforming, acoustic echo cancellation, and distant speech recognition. It is a special array consisting of four linear sub-arrays, of which three consist of five microphones and the fourth consists of seven microphones. The recordings have been recently made public on the project's website [72].

3.2 Reasons for a New Database

The above mentioned corpora show the increasing trend of publicly available microphone array databases. This trend promotes the research of advanced techniques in

multi-channel audio/speech processing for different kinds of real-world applications. For this work, an effort has been made in a similar direction to build an in-house microphone array, which is used in the meeting room ("cocktail party room") of the Signal Processing and Speech Communication (SPSC) Lab at Graz University of Technology, Austria. There were multiple reasons to build an independent setup, one of the reasons is the need for reference files, which are used as ground truth values for both localization and pitch estimation tasks. All corpora mentioned above deal with a scenario where only a single speaker is active in one frame. The current work explores the performance of the proposed algorithms for concurrent speaker scenarios, where multiple speakers are assumed to be active at the same time. This challenging task requires the playback files to be pre-processed before the multi-speaker recordings, as the tasks of concurrent speakers' Voice Activity Detection (VAD) is itself subject of ongoing research. To the best of my knowledge, no robust off-the-shelf VAD algorithm is available for automatic annotations of the recorded multi-channel data for concurrent speaker scenarios. Therefore, a new process of generating speech segmentation for concurrent speaker scenarios is presented later in this chapter.

The other reason for the development of an intensive database is the control of different acoustic environment parameters, which includes various background noise conditions: electronic projector, environmental noise coming from open windows, small events like opening or closing of the meeting room door, and a spatial noise source. For this purpose, the cocktail party room is used to create and record these specific, well-controlled acoustic conditions.

Fig. 3.1 presents the SPSC's Uniform Circular Array (UCA). Even though the current setup is immobile, it serves as a benchmark for smaller and portable arrays. The flexibility of the array as shown in Fig. 3.1 gives an opportunity to try different array diameters, which were absent in the corpora mentioned above. The analysis of the proposed algorithms with different array configurations helps to understand the effects of the number of sensors and the array diameter on the performance of the algorithms.

3.3 Microphone Array Design

The SPSC's UCA consists of 24 microphones positioned equidistantly with a maximum diameter of 0.55 m on a circular ring. The specifications of the microphones

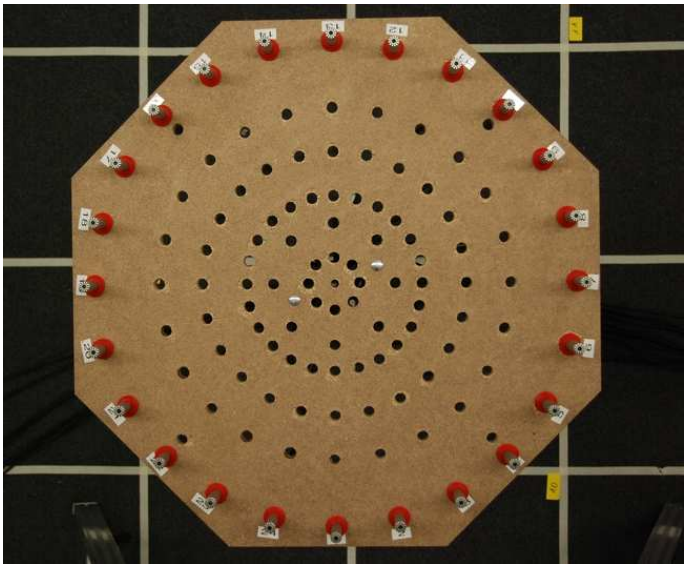


Figure 3.1: SPSC microphone array in UCA shape consisting of 24 microphones with variable diameter ranging from 0.2 m to 0.55 m. The microphones can be placed in any of the boreholes of th planar fixture made out of plywood.

are as follows:

- Model No: *Behringer* ECM 8000 [73]
- Transducer Type: Electret condenser
- Frequency Response: 15 Hz to 20 kHz
- Pick-up Pattern: Omnidirectional
- Sensitivity: up to - 60 dB
- Impedance: 600 Ohms

The RME Fireface 800 audio interface [74] is used for multi-channel recordings. The two ADAT digital I/O on the RME Fireface 800 allows to connect two external converters. Two *Behringer* ADA 8000 [75] A/D and D/A converters are used with the RME Fireface 800 to support recordings of up to 24 channels. The digitized signals are passed to a laptop via a FireWire 400 cable. The specifications of the RME Fireface 800 are as follows:

- 8 x Analog line I/O (used for recording of 8 of the 24 channels)
- 2 x ADAT Digital I/O (the two *Behringer* ADA 8000 connected to these two ports were used to record the remaining 16 of the 24 channels)

- Up to 192 kHz sampling rate
- Up to 24 bit sample resolution
- 2 Microphone/Line Inputs with Preamps (unused)
- 2 Instrument/Line Inputs (unused)
- 1 x stereo headphone output (unused)
- 2 x MIDI I/O (unused)

The multi-channel recordings were stored on an external hard-drive connected to the laptop at a sampling rate of 48 kHz, and 16 bits amplitude resolution. The reason to use 16 bits resolution was because the location accuracy of the proposed algorithm did not improve by using 24 bits resolution [76]. One other concern was the memory aspect, as the 24 bits wave-files require one-third more memory space than the 16 bits wave-files.

The array has four different diameter rings as shown in Fig. 3.1 in the range of $d = [0.2, 0.3, 0.4, 0.55]$ m. These four rings allows the use of 24 microphones plus two additional smaller rings allows the use of only 8 microphones. Using this flexibility multiple data sets with different array diameters were recorded to test the performance of the proposed algorithms. The number of microphone pairs used to generate direction estimates is restricted to the pairs formed by oppositely placed microphones. This results in 12 pairs in case of a 24-channel microphone array. This arrangement simplifies the direction estimation process as all the microphone pairs share a common center.

There are several physical specifications of the microphone array affecting the source localization problem discussed in [7], e.g., the sampling frequency, the distance between the microphones d, and the speed of sound c. Therefore, it is important to illustrate the possible range of discrete time-delay values with the given geometry. The DoA θ can be calculated from the delay τ by the relation $\cos\theta = \frac{c \cdot \tau}{F_s \cdot d}$. Based on the physical design of the SPSC array (at $d = 0.55$ m), and the speed of sound given by $c = 346.43$ m/sec (calculated for $t_{air} = 25°C$), the number of possible time-delays is in the range of $[-76, 76]$ samples. Using the defined relation, Table 3.1 shows the DoA θ calculated for all the positive time-delay τ values along with the corresponding spatial quantization error $\epsilon_{\text{quantization}}$. The results are symmetric for negative time-delay values. A best resolution of approximately 0.75° or corresponding smallest quantization error of ±0.3760° is achieved for an angle of 0°

("broadside"). This resolution is better than what was reported in [77], where the main difference was the smaller distance between microphone pairs (in their case, $d = 0.3\,\mathrm{m}$), and $F_s = 44.1\,\mathrm{kHz}$. The error remains within a factor of 2 of the best value within a range of 66 samples (60°), whereas in [77], it stays small within a range of 30 samples (51.5°).

Some preliminary analysis of the array design is made by calculating the delay-and-sum beam pattern in linear and polar coordinates. The bandwidth of interest for speech signals lies in the range of 100 Hz to 8000 Hz and the signal is arriving from $\phi = \theta = \pi/4$ or 45°. Fig. 3.2 shows the delay-and-sum beam pattern for the UCA of $d = 0.55\,\mathrm{m}$ diameter consisting of $M = 24$ microphones, which is calculated using the following relation [78]:

$$B(\mathbf{k}, \mathbf{k}_L) = \frac{1}{M} \sum_{m=0}^{M} \exp\left[j \frac{\pi d}{\lambda} \left\{ \sin\theta \cos\left(\phi - \frac{2\pi m}{M} \right) - \sin\theta_L \cos\left(\phi_L - \frac{2\pi m}{M} \right) \right\} \right], \tag{3.1}$$

where $\mathbf{k}$ is the wavenumber, $\mathbf{k}_L$ is the wavenumber of the plane wave source with spherical coordinates $(\theta_L = \pi/4, \phi_L = \pi/4)$, and λ denotes the wavelength of plane wave. The beam pattern of circular array does not suffer from grating lobes prominent in case of linear arrays; however, this plot is based on a simulation assuming anechoic environment which does not validate similar performance in multi-path environments. The beam pattern in polar coordinates at different frequencies are shown in Fig. 3.3.

In theory, the design of the microphone array offers a limited range of frequencies where it can localize speakers. The distance between the microphones has similar effect in spatial-domain as the time distance has on the signal frequencies sampled in the time-domain. This effect is called spatial aliasing [79]. To avoid the effects of spatial aliasing, a relationship to calculate the maximum operating frequency $f_{\max}$ for a given array configuration was presented in [79]:

$$f_{\max} = \frac{c}{2 \cdot d_\theta}, \tag{3.2}$$

where c is the speed of sound in air and d_theta is the minimum distance of array. For a $M = 24$ channels UCA of $d = 0.55\,\mathrm{m}$ diameter, where $d_\theta = d \cdot \sin(\frac{\pi}{M})$ is defined as the minimum distance of the array (this relationship was presented for

Table 3.1: Quantization error for the SPSC microphone array in shape of a Uniform Circular Array (UCA) consisting of 24 microphones with a diameter of 0.55 m. τ is the time-delay in integer multiple of the sampling interval, θ is the corresponding DoA angle in degrees, and $\epsilon_{\text{quantization}}$ is the maximum quantization error in degrees.

τ	θ^{o}	$\varepsilon_{\text{quantization}}$	τ	θ^{o}	$\varepsilon_{\text{quantization}}$	τ	θ^{o}	$\varepsilon_{\text{quantization}}$
0	0	0.3759	26	19.9487	0.4010	52	43.0282	0.5187
1	0.7519	0.3760	27	20.7506	0.4031	53	44.0656	0.5279
2	1.5039	0.3761	28	21.5568	0.4053	54	45.1214	0.5379
3	2.2561	0.3763	29	22.3675	0.4077	55	46.1971	0.5486
4	3.0088	0.3766	30	23.1830	0.4102	56	47.2944	0.5602
5	3.7620	0.3769	31	24.0034	0.4129	57	48.4148	0.5729
6	4.5158	0.3773	32	24.8291	0.4156	58	49.5606	0.5866
7	5.2704	0.3778	33	25.6603	0.4185	59	50.7339	0.6017
8	6.0259	0.3783	34	26.4974	0.4216	60	51.9373	0.6183
9	6.7825	0.3789	35	27.3406	0.4248	61	53.1740	0.6367
10	7.5403	0.3795	36	28.1903	0.4282	62	54.4473	0.6571
11	8.2994	0.3803	37	29.0468	0.4318	63	55.7616	0.6801
12	9.0599	0.3811	38	29.9105	0.4356	64	57.1218	0.7060
13	9.8221	0.3820	39	30.7817	0.4396	65	58.5338	0.7357
14	10.5860	0.3829	40	31.6609	0.4438	66	60.0052	0.7700
15	11.3519	0.3840	41	32.5485	0.4482	67	61.5452	0.8102
16	12.1198	0.3851	42	33.4450	0.4529	68	63.1656	0.8583
17	12.8899	0.3863	43	34.3508	0.4579	69	64.8823	0.9171
18	13.6624	0.3875	44	35.2665	0.4631	70	66.7164	0.9911
19	14.4375	0.3889	45	36.1927	0.4686	71	68.6985	1.0881
20	15.2152	0.3903	46	37.1300	0.4745	72	70.8747	1.2230
21	15.9959	0.3918	47	38.0790	0.4808	73	73.3208	1.4294
22	16.7796	0.3935	48	39.0405	0.4874	74	76.1795	1.8063
23	17.5665	0.3952	49	40.0153	0.4945	75	79.7921	2.9957
24	18.3569	0.3970	50	41.0042	0.5020	76	85.7835	2.9957
25	19.1509	0.3989	51	42.0082	0.5100	-	-	-

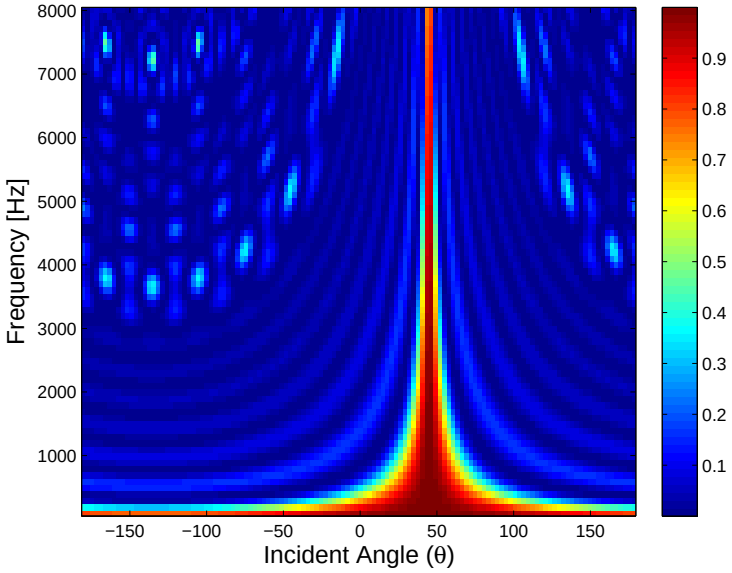


Figure 3.2: Delay-and-sum beam pattern of SPSC microphone array consisting of 24 microphones with diameter of 0.55 m. The steering direction is fixed at 45° azimuth. The colorbar represents the amplitude of the beam pattern.

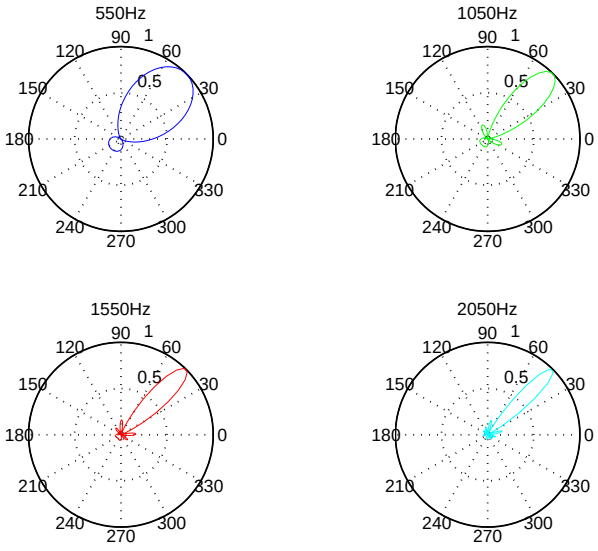


Figure 3.3: Delay-and-sum beam pattern in polar coordinates at different frequencies of the SPSC microphone array. The steering direction is fixed at 45°. The radial dimension show the amplitude of the beam pattern.

Table 3.2: $f_{\max}$ values for different sensors' spacing.

d [m]	$f_{\max}$ [kHz]
0.55	2.4
0.4	3.8
0.3	4.38
0.2	6.57

circular array in [80, p.32]). The maximum frequency $f_{\max}$ below which the array can localize the speech source is $f_{\max} = 2.4\,\text{kHz}$ for $c = 343\,\text{m/sec}$. Following similar calculations for the other three diameter configurations by keeping the number of microphones and the speed of sound constant, the trend is shown in Table 3.2. It is important to note that the given results are just approximations to give an idea about the characteristics of the array. The upper limit for the frequency increases with decreasing sensor spacing. On the other hand, the quantization error for the location estimates will increase for a small aperture array. Hence, there is a tradeoff between selection of physical parameters of the microphone array, which should be chosen carefully according to the desired application in mind. In the present work, the higher spatial resolution achieved by 0.55 m diameter array is preferred over the frequency range as most of the useful spectral information of speech signals such as the first and second formants (for male, female and young children) are usually below 2.4 kHz. Furthermore, the distant speech recognition experiments carried out in [80] were successfully conducted for a closer range of 2.24 kHz. Therefore, the current setup not only improves location accuracy, it can be used for speech enhancement algorithms such as beamforming.

3.4 Speech Database and Speaker Setup

The speakers or speech sources used for the recordings are taken from the Grid corpus [81]. This database consists of single-channel high-quality recordings for 36 speakers with 500 sentences for each speaker. The duration of each sentence is 3 sec. The Grid corpus provides the opportunity to generate large datasets using various speaker combinations. For the playback, the wave-files of the Grid corpus were sampled at 48 kHz. The resampled wave-files were played back through a Yamaha MSP5A loudspeaker mounted on a stand at a fixed distance to the array in the same horizontal plane. In all the recordings the distance of speakers remain fixed

and only the directions are varied. Concurrent speaker scenarios were recorded by simultaneously playing back the speech files by loudspeakers placed at different positions. This procedure facilitates the repetition of experiments and generation of speaker label files prior to the recordings.

Different speaker combinations are explored ranging from a single speaker to multiple speakers. In this thesis, a maximum number of four concurrent speakers is considered. In all cases, different speakers from the Grid corpus were used to emulate such scenarios. The details about the number of speaker combinations for each scenario are listed in the Section 4.5.1. This part of the recorded database is termed "Controlled Experiments", as exact positions and voicing information of all concurrent speakers are available for a minimum frame-size of 20 msec. Another set of experiments using human speakers was recorded in the cocktail-party room with a group of four participants, two males and two females, who enacted two real-world scenarios. One scenario recorded in the SPSC meeting room was a presentation scenario, where one speaker was standing close to the white board or projector screen, and the rest were the audience sitting in fixed positions. The other case was a meeting scenario taking place among a group of participants sitting closely and/or in front of each other. In both cases, the array was placed in the center of the room. These recordings are termed "Real Speakers Experiments", where manual labeling is done by carrying out listening tests on one of the input channels of the array. The annotation was further improved by examining the waveform and spectrogram of the recorded speech file. A small subset of the multi-channel database was recorded using human speakers enacting some moving speaker scenarios. These experiments are termed "Mobile Speakers", where either one or two speakers moved in a stepwise manner creating different scenarios presented in Section 4.5.2. The speaker segmentation and labeling of this task was most difficult, because there was no head tracking system available in the meeting room. The annotations were done by video recording a view of the scenarios, where the audio streams from the video and the microphone array were synchronized and labeled manually.

3.5 Speech Database Reference Segmentation and Labeling

The position-pitch speaker localization algorithm relies on voiced segments of speech, as only these segments exhibit the necessary pitch information. Therefore, all source

speech signals that were used for recording every single and multi-speaker scenario throughout this thesis were segmented and annotated into voiced, unvoiced, and silent segments. After recording, the appropriate time-delays for each microphone input, latency of the recording software, delays due to playback were added to the segment time stamps and the segments of all speakers were overlaid. The Pure Data (PD) [82] software directly access the multi-channel soundcard so there are no laptop operating system latency issues. This overlaid multi-speaker segmentation and annotation was specified using a new label format.

Description of Label Format

The label format is based on the HTK label file format [83], which is illustrated by an example later in the section. The label file was a text file which was assigned the same base name as the audio file. The speech labels were defined as follows:

Speech Labels:

vcd:	voiced
uvcd:	unvoiced
sil:	silence

To disambiguate multiple speakers, the speaker tags were assigned with every voiced, unvoiced and silent label. These tags were based on the order number of the speaker in the Grid Corpus, which has 36 speakers. In the given application, the label files are used to evaluate the estimation accuracy of the localization algorithms. Therefore, the speaker position in terms of DoA was also assigned to every speaker tag and speech label. Moreover, the distance of the speaker from the array was appended to each label as well. The list of assigned speaker labels is as follows

Speaker ID: <SPKR><Speech Labels><ANGL><DIST>

<SPKR>:	sp01, sp02, sp03, sp04
<Speech Labels>:	vcd, unvcd, sil
<ANGL>:	integer value indicating speaker direction in degrees relative to $\theta = 0^\circ$
<DIST>:	value indicating distance between speaker and array center in meters up to two decimal points

Fig. 3.4 shows an example of a reference file for a two concurrent speakers scenario. The first two columns show the start and end samples of the speech segments. The

Example:

<start sample>	<end sample>	<segment information>
1200000	7320000	sp03-sil-142-2,00
1200000	5870000	sp04-sil-310-2,00
5870000	6180000	sp04-unvcd-310-2,00
6180000	7390000	sp04-vcd-310-2,00
7320000	9730000	sp03-vcd-142-2,00
7390000	8230000	sp04-unvcd-310-2,00
8230000	8650000	sp04-sil-310-2,00
8650000	8750000	sp04-unvcd-310-2,00
8750000	10090000	sp04-vcd-310-2,00
9730000	9840000	sp03-sil-142-2,00
9840000	12590000	sp03-vcd-142-2,00
10090000	10280000	sp04-unvcd-310-2,00
10280000	10640000	sp04-sil-310-2,00
⋮	⋮	⋮

Figure 3.4: An example of a reference file for the two concurrent speaker scenario.

third column corresponds to the segment information, where all the labels defined above are combined for each speaker. This is the label file for a male and female recording, where "sp03" was a male speaker file taken from the Grid corpus and was reproduced at 142° during the recording. The "sp04" was a female speaker file taken from the Grid corpus and reproduced at 310° during the recordings. Therefore the segment labels append this information to the speech labels. The distance of both speakers from the array was 2 m. A small script was created to read this label file and make the speaker activity table for every recording, where each column of the table denotes the recorded speaker for the given scenario. The resulting table is used as a reference to evaluate the proposed algorithms' performance.

For the segmentation and annotation of voiced, unvoiced, and silent segments of speech a tool provided by the SYNVO GmbH was employed, cf. [84]. This tool is based on a two-stage neural network for classification into voiced, unvoiced, and silent segments of speech. The first stage resembles a speaker-independent model for estimation of voiced, unvoiced, and silent posteriors on an individual analysis frame of the speech signal. The analysis frame has a length of 5 msec and is shifted by 1 msec. Given these posteriors, the second stage classifies the current speech frame into voiced, unvoiced, or silent taking a context of 25 frames before and after the current frame into consideration. Hence, this segmentation provides an accurate

framewise speaker labels which aid the evaluation process.

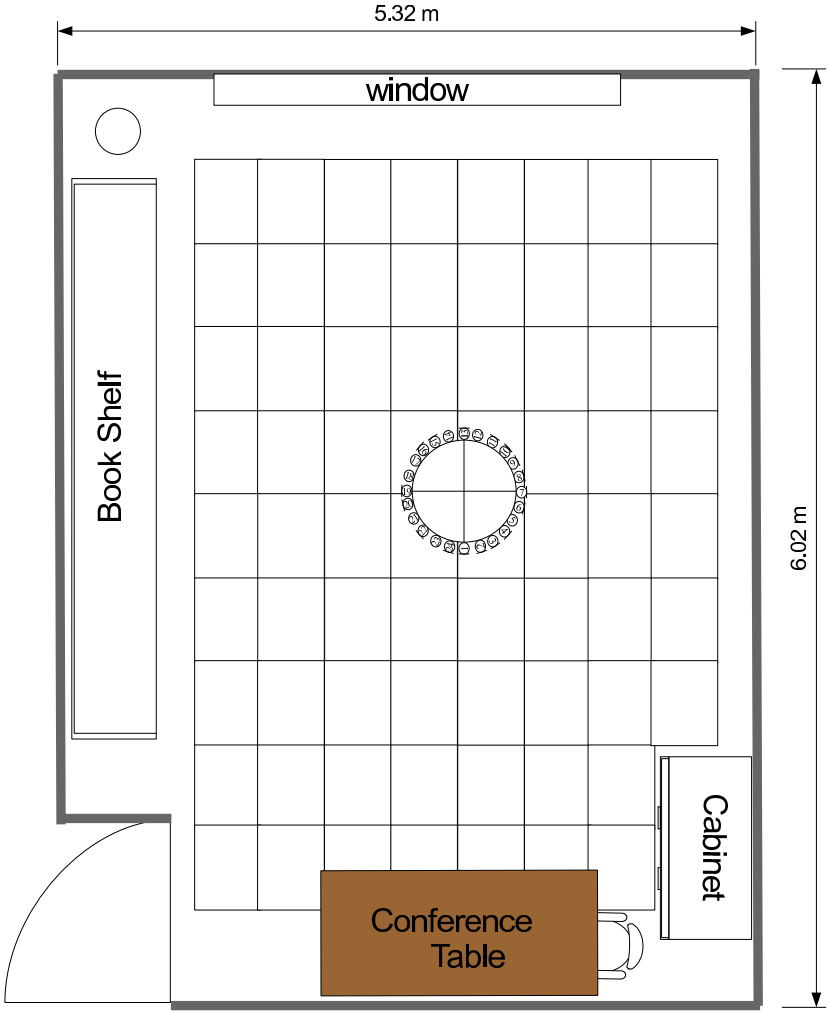


Figure 3.5: SPSC cocktail party room layout.

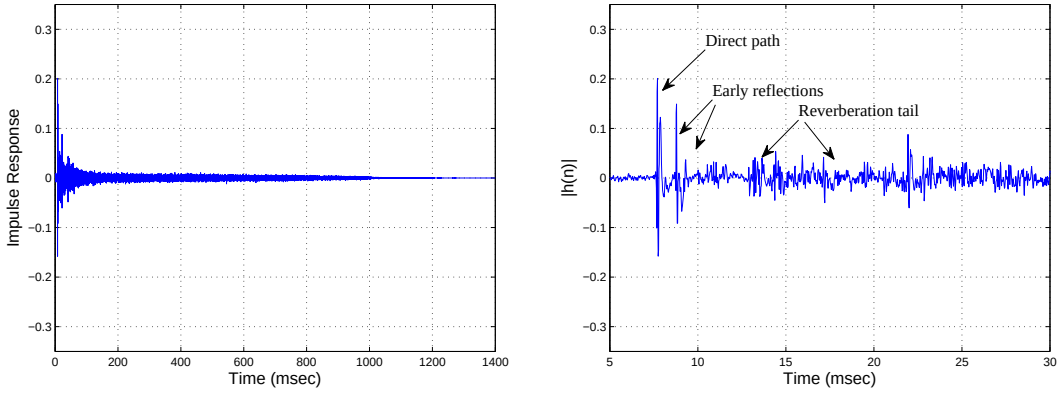
3.6 Room Acoustics and Background Noise

A series of recording sessions consisting of the speaker setups discussed in Section 3.4 took place in the cocktail party room of the SPSC Lab. Fig. 3.5 presents the room layout. This meeting room has the dimensions $6.02 \times 5.32 \times 3\,\mathrm{m}$ and one of the walls of the room has a large window partly covered by blinds that were set open during the recordings. The floor is covered with standard carpet. No particular effort was made to reduce the reverberation in the room. The room was divided

into 72 sections, each measuring $0.5 \times 0.5\,\mathrm{m}^2$. The division supports an accurate measurement of speaker positions relative to the array.

Furthermore, some basic acoustic measurements were carried out in the cocktail party room, for example, the Room Impulse Response (RIR) was measured using a Maximum Length Sequence (MLS). The measurement microphone was placed at a distance of 2 m from the loudspeaker. The impulse response can be described as a sequence of delayed delta impulses. "The delays are associated with the geometrical length of the related propagation paths. The amplitude of the impulse response depends on the reflection coefficients of the boundaries and on the inverse path lengths" [47, p.232]. The measured impulse response is shown in Fig. 3.6(a), where a close-up of a 25-millisecond segment of the RIR is shown in Fig. 3.6(b). The direct-path component and some strong reflected components are highlighted. To calculate the reverberation time RT_{60}, the Schroeder formula is used [85]. Fig. 3.7 shows the plot of the Schroeder function of the impulse response from Fig. 3.6(a). A reverberation time $RT_{60} \approx 500\,\mathrm{msec}$ was measured by linear extrapolation of the curve from the level of $-5\,\mathrm{dB}$ to $-12\,\mathrm{dB}$ (for details about echo and reverberation characteristics, see [6, p.48]).

The microphone array consists of 24 omnidirectional microphones placed in a uniform circular ring with a diameter of 0.55 m. The diameter of the array can be varied to four different settings as shown in Fig. 3.1. All diameters were used for



(a) (b)

Figure 3.6: (a) Measured Room Impulse Response (RIR) using Maximum Length Sequence (MLS), (b) A 25-millisecond segment showing the close-up of the RIR including direct path and early reflections.

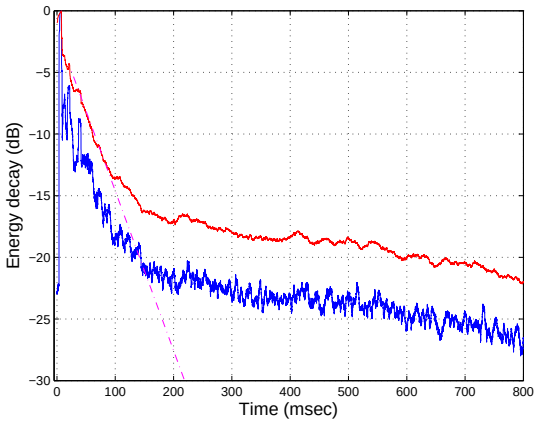


Figure 3.7: The upper curve is the energy decay curve and the bottom curve shows the impulse response envelope. The energy decay curve is estimated by the Schroeder function using the RIR from Fig.3.6(a). The "dashed-dot-line" is the linear regression curve used for extrapolation of the reverberation time RT_{60}.

the recordings by keeping the number of microphones constant in all cases. The recordings were carried out in two phases. In the first phase, 18 speakers: 8 males and 10 females chosen from the Grid Corpus were played back through Yamaha MSP5A loudspeakers at different positions relative to the array. The array was placed in center of the room. The loudspeakers were positioned at a height of 1.39 m maintaining a constant distance of 2 m from the array. A set of speech files containing male and female utterances were mixed into longer segments modeling different speaker interaction behaviors in a spatialized multi-speaker scenario. Furthermore, different speaker combinations were played back simultaneously for up to 4 speakers. The reference label files were generated as explained in Section 3.5, which allows to carry out detailed performance analysis of all localization algorithms. In the second phase, the real speaker scenarios introduced in Section 3.4 were recorded. The annotations for real speakers were done manually. The playback and recording process was controlled by a software on a single laptop, and the captured audio was saved directly to the hard disk of the laptop with 16 bit resolution and a sampling rate of 48 kHz. The playback and the recordings were controlled by a patch made in PD software.

3.7 Evaluation Metrics

All the localization algorithms are evaluated on a framewise basis, for which a metric, denoted as ACC is used. This measure has also been used in [86] for the localization estimates. The measure is based on the normalized number of frames where the estimated localization angle $\hat{\varphi}_n$ is close enough to the true angle φ_0 given as

$$\mathrm{ACC} = \frac{1}{N}\sum_{n=1}^{N}\delta^*(\varphi_0, \hat{\varphi}_n) \times 100\%, \tag{3.3}$$

where N is the number of frames. δ^* is defined as

$$\delta^*(a,b) = \begin{cases} 1 & \text{if } |a-b| \leq \Delta \\ 0 & \text{otherwise,} \end{cases} \tag{3.4}$$

where Δ is a grace boundary or error threshold around the true angle within which the estimated angle $\hat{\varphi}_n$ is considered to be correct. The true angle may belong to any speaker. This evaluation metric is used throughout the thesis to compare different algorithms. In addition to a fixed threshold, the accuracy results can be plotted as a Cumulative Distribution Function (CDF) versus error threshold Δ. This metric is used to emphasize and compare the estimation accuracy and variance of the estimates for all the algorithms. In some cases, the value of grace boundary was fixed at 5°. This corresponds to the minimal inter-speaker distance of 35 cm in 2 m distance from the array.

The database is recorded with the different acoustic conditions discussed in Section 3.2. Therefore, it is important to introduce different distortion measures used to evaluate the algorithms. A frequently used distortion measure for additive noise is the SNR value, which can be defined as the ratio of the power of the desired signal to the power of the noise in the distorted signal measured in the logarithmic *decibel* scale given as [6]:

$$\mathrm{SNR} \triangleq 10\log_{10}\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}, \tag{3.5}$$

where P is the average power measured over the system bandwidth. Different kinds of background noise present in a typical office or meeting room are considered in this work. The kinds of noise against which all algorithms are tested are as follows:

- Beamer Noise

- Environmental noise coming from an open window
- Opening and closing of door
- Machine Noise coming from a computer

Moreover, a special set of recordings was carried out, where a loudspeaker was placed on the floor at the same distance from the array as the speech sources. A white noise signal was played back through the loudspeaker, and the speech sources were placed at different positions. The setup was recorded for scenarios with single and two speech sources at SNR values ranging from $-5\,\mathrm{dB}$ to $20\,\mathrm{dB}$.

As the SNR does no account for the reverberation of the room, a channel based measure known as the Direct to Reverberation Ratio (DRR) [87] is used, which is defined as the ratio between the energy of the direct path (computed from the first part of the impulse response, i.e., a few milliseconds around the dominant peak), and the energy of the remaining reflection paths given as

$$\mathrm{DRR} = 20\log_{10}\Big(\frac{||\mathbf{h}_d||_2}{||\mathbf{h}-\mathbf{h}_d||_2}\Big)\,\mathrm{dB}. \tag{3.6}$$

A DRR of $1.51\,\mathrm{dB}$ was measured in the cocktail party room, where the source was placed $2\,\mathrm{m}$ away from the array. A signal-based reverberant measure known as the Signal to Reverberation Ratio (SRR) [87] is given as

$$\mathrm{SRR} = 20\log_{10}\Big(\frac{||\mathbf{s}_d||_2}{||\hat{\mathbf{s}}-\mathbf{s}_d||_2}\Big)\,\mathrm{dB}, \tag{3.7}$$

where $\mathbf{s}$ is the original speech signal, $\mathbf{s}_d = \mathbf{h_d}^T\mathbf{s}$ is the delayed and scaled version of the speech signal without reflections, and $\hat{\mathbf{s}}$ is the reverberant signal. An SRR of $1.1\,\mathrm{dB}$ was measured with the same settings as defined above.

3.8 Summary

In this chapter, well-known microphone array corpora are summarized along with the reasons to build a new microphone array and to record a large database. The characteristics of the array are explored by theoretical measures such as the beam pattern, and maximum operating frequencies achieved by different diameters of the array. The details of the recorded corpus for different speaker setups are discussed.

Furthermore, a detailed discussion of the speech segmentation and labeling process is presented, which makes it possible to evaluate the algorithms for concurrent speaker scenarios. Moreover, the details of the recording room are listed along with basic room acoustic measurements. In the end, a list of evaluation metrics for measuring the accuracy of the algorithms and acoustic conditions such as background noise and reverberation is presented. All the illustrative examples and evaluations carried out in Chapter 4, and Chapter 5 are based on this corpus.

Chapter 4

Joint Position-Pitch Estimation Based Source Localization

This chapter presents various auditory inspired cross-correlation-based methods for Acoustic Source Localization (ASL). These methods are based on the joint position-pitch decomposition of the cross-correlation between a pair of microphones. The baseline algorithm known as the Position Pitch (PoPi) algorithm is analyzed in detail. This method is extended to a multi-channel system consisting of a 24-channel UCA. The auditory inspired preprocessing is combined with the PoPi algorithm to form a new algorithm known as the Multi-band Position Pitch (MPoPi) algorithm. A new frequency-selective criterion is combined within the MPoPi algorithm, which groups the frequency channels belonging to the same speaker. This algorithm is referred to as MPoPi-FS. Moreover, the grouped channels are temporally linked by using a pitch tracker to form extended spectro-temporal regions or fragments. These spectro-temporal regions are combined within the MPoPi algorithm. The modified algorithm is called MPoPi-STF. Finally, the proposed ASL algorithms are tested on real-world recordings under various acoustic scenarios.

4.1 The Position Pitch Algorithm

The joint position and pitch extraction (PoPi) algorithm was summarized in Section 2.8.1. The PoPi algorithm provides a feature set consisting of relevant cues required for segmenting sources in multiple source scenarios.

The PoPi algorithm performs a parameterized sampling process on the cross-correlation function R_t at time instant t between a pair of microphone signals to extract the common periodicities together with the corresponding DoA φ_0 value given by

$$\rho_t(\varphi_0, F_0) = b \cdot \sum_{k=-K}^{K} R_t(\lfloor k \cdot L(F_0) + O(\varphi_0) \rfloor), \tag{4.1}$$

where $\rho_t(\varphi_0, F_0)$ is the PoPi plane at time t and the cross-correlation function $R_t(\tau)$ for a given time lag is calculated as the inverse Fourier transform of the received signal cross-spectrum $X_1(t,\omega)X_2^*(t,\omega)$, where $X_1(t,\omega)$ is the Fourier transform of the windowed signal $x_1(t)$ and $X_2^*(t,\omega)$ is the complex conjugate of the Fourier transform of the windowed signal $x_2(t)$ given as

$$R_t(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_1(t,\omega)X_2^*(t,\omega)\, \exp(j\omega\tau)\, d\omega. \tag{4.2}$$

In the formulation defined in (4.1), b denotes a normalization factor which is discussed later in this section, K defines the cross-correlation interval used for summation of samples. $L(F_0)$ is the first time-lag, which depends on the pitch parameter F_0 according to the relation $L(F_0) = \frac{F_s}{F_0}$, and F_s is the sampling frequency of the recorded signals. The time-lag value is rounded using the floor function to convert real numbers that can result for the time-lag values to close integers.

The normalization factor b can be set equal to 1 or can be set to the reciprocal of the number of correlation peaks considered. The summation runs over a symmetric interval from $-K$ to K but it can also be ran from a specific correlation peak $-K_1$ to K_2. In this thesis, b is set to $\frac{1}{2K+1}$. The value of K is chosen to be 3 because according to the range of F_0 (from 80 Hz to 400 Hz) defined in Section 4.1.1 leads to pitch periods at $F_s = 48$ kHz to be from 600 samples to 120 samples. For a selected frame length of 2048 samples used for the evaluations of the experiments in Section 4.5, only up to third harmonic of the maximum pitch period can be evaluated. The sampling function $\rho_t(\varphi_0, F_0)$ generates the DoA value φ_0 using the relation $O(\varphi_0) = \frac{d.cos(\varphi_0).F_s}{c}$ (ref. Fig. 2.1), where τ denotes the correlation lag corresponding to the DoA of interest, d is the distance between the microphones, and c is the speed of sound in air. Therefore, the main characteristic of this technique is that it represent the speech signals received by two sensors in a two dimensional Position-Pitch (PoPi) plane.

The upper summation line in Fig. 4.1 illustrates the relationship defined in (4.1),

where the cross-correlation between a pair of microphones is presented. The main peak of the cross-correlation is shifted by the number of samples, which encodes the time-delay between the received signals by a pair of microphones. The position of this peak is searched over a certain range depending upon the distance between the two microphones and the sampling frequency F_s (the range for the SPSC UCA with $d = 0.55\,\mathrm{m}$ and $F_s = 48\,\mathrm{kHz}$ equals ± 76 samples). The second feature, the fundamental frequency, is related to the spacing $L(F_0)$ between the attenuated multiples of the main cross-correlation peak. The main peak is marked by "$\circ$" and the multiples by "$\square$" exhibiting a uniform distance related to the pitch F_0. The second summation line will be discussed later in this section.

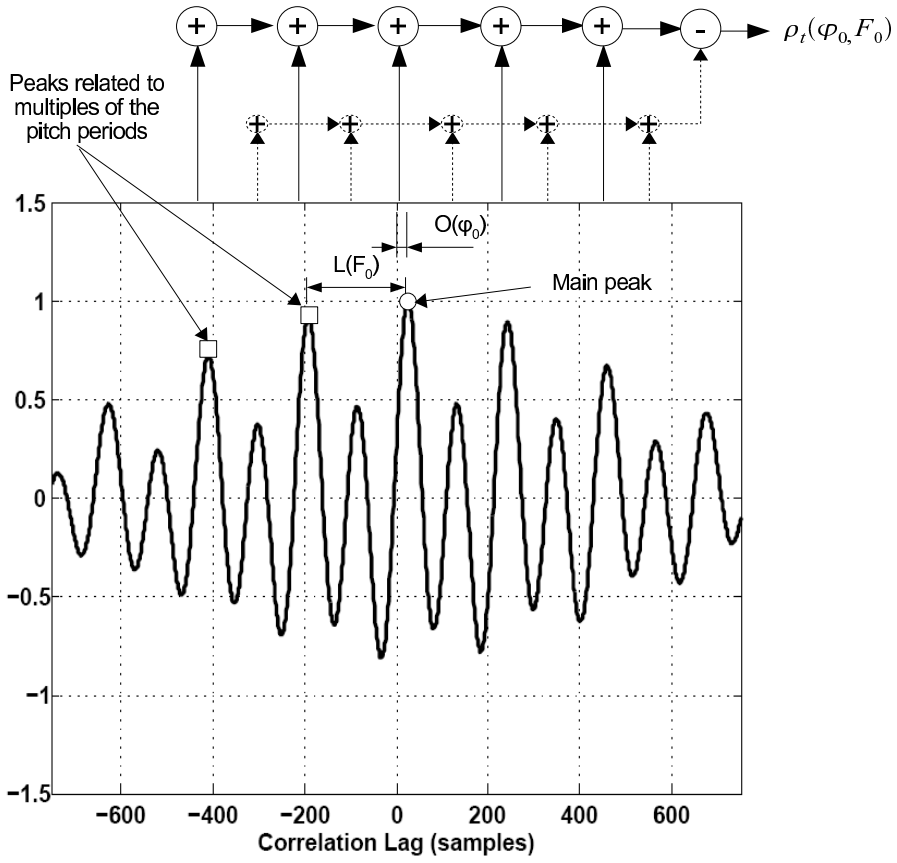


Figure 4.1: PoPi decomposition for a single female speaker at 115°, $F_0 = 219\,\mathrm{Hz}$ using a pair of microphones. $O(\varphi_0)$ denotes the correlation lag corresponding to the DoA of interest, and $L(F_0)$ denotes the correlation lag corresponding to the fundamental frequency. The $\rho_t(\varphi_0, F_0)$ is the averaged summation of the cross-correlation values for a DoA and pitch pair (φ_0, F_0). It is defined for any $\rho_t(\varphi_0, F_0)$, even if these parameters are not matched to the peaks.

4.1.1 The PoPi Plane

The PoPi plane $\rho_t(\varphi_0, F_0)$ is computed by scanning for the pitch F_0 and DoA φ_0 over a predefined range, and exhibits large peaks at indices that correspond to a source present in the acoustic scene. This 2D representation, the so called PoPi plane presents the DoA along the horizontal-axis and the fundamental frequency along the vertical-axis, respectively. The PoPi plane is evaluated only for predefined values of $L(F_0)$ and $O(\varphi_0)$, which are pre-calculated for the frequencies $F_0 = [80, \ldots, 400]$ Hz with a single lag-step defined in samples (which means pitch period ranging from 120 samples to 600 samples for the sampling frequency of $F_s = 48$ kHz), and DoA candidates $\varphi_0 = [0°, \ldots, 180°]$ with a stepsize of 1°.

This decomposition acts like a "comb" filter, which is shifted along the predefined values of DoA over the cross-correlation. For every position of the comb filter, the width of the comb is adjusted according to the predefined pitch values. Thus it defines the cross-correlation values summed up for every combination of pitch and DoA. These values are stored in the $\rho_t(\varphi_0, F_0)$ matrix, a peak in this matrix indicates the likely DoA of the speech source with its respective pitch value. The decomposition process is shown in Fig. 4.1, where the PoPi plane shows a maximum when the DoA and pitch values are chosen such that the comb filter sums up the main peak and the attenuated multiples of the pitch period. The cross-correlation is computed from the signals received at microphone 1 and microphone 13 of the 24-channel UCA. The signals waveforms along with their corresponding amplitude spectrums are shown in Figs. 4.2(a)-(d). There is a strong second harmonic present along with the original fundamental frequency $F_0 = 219$ Hz. The higher magnitude of second harmonic can be due to the presence of first formant frequency of the speech segment.

Fig. 4.3 shows a PoPi plane using a single pair of microphones. The peaks at multiple of the pitch period indicate the well-known problem of pitch multiplicity in correlation based pitch estimation algorithms. The DoA in this case also suffers from the cone of ambiguity, which is defined for a linear aperture: "signals propagating from above, below, or to the side of the linear aperture cannot be distinguished" (cone of ambiguity; axis parallel to array) [22, p.66] . Therefore, a pair of microphones fails to provide the front/back information about the source DoA. A uniform circular array with multiple pair of microphones is capable of a complete 360° resolution. The multi-microphone extension of the PoPi algorithm is presented

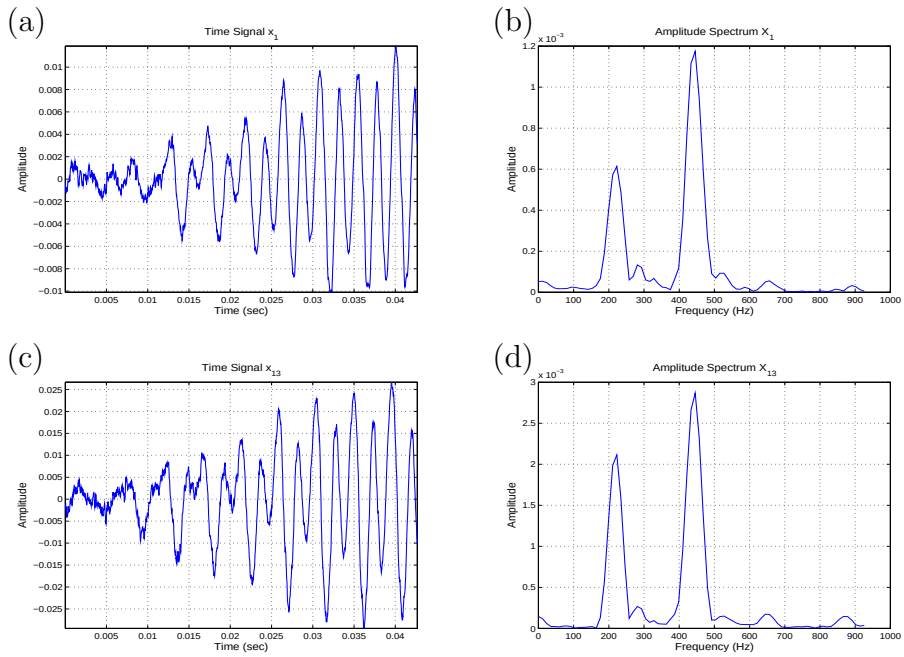


Figure 4.2: Time-waveform and amplitude spectrum of signals received at microphones 1 and 13. These two microphones form the first pair of the 24-channel UCA. The time-waveform is shown for a segment length of 42.67 msec. The amplitude of the signals is plotted along the vertical-axis.

in the next section, where the same example is used to illustrate the 360° resolution capability of the circular array. The similarities that the PoPi algorithm exhibits with respect to the SRP-PHAT algorithm are presented in Appendix A.

4.1.2 Multi-Microphone Position Pitch Algorithm

The PoPi algorithm can be easily extended to a multi-microphone system, where the PoPi plane can be calculated over different pairs of microphones and added together. The PoPi algorithm for multiple pairs is extended as follows

$$\rho_t(\varphi_0, F_0) = \sum_{m_p=1}^{M_p} \left[\frac{1}{2K+1} \cdot \sum_{k=-K}^{K} R_{t,m_p}(\lfloor k \cdot L(F_0) + O(\varphi_{0,m_p}) \rfloor) \right], \qquad (4.3)$$

where M_p is the total number of microphone pairs, m_p is the pair index, and $O(\varphi_{0,m_p})$ are pair dependent steering delays which are predefined for each pair of microphones. A PoPi plane resulting from the data recorded with the SPSC UCA of 24 microphones forming 12 pairs of oppositely placed microphones is shown in Fig 4.4(a).

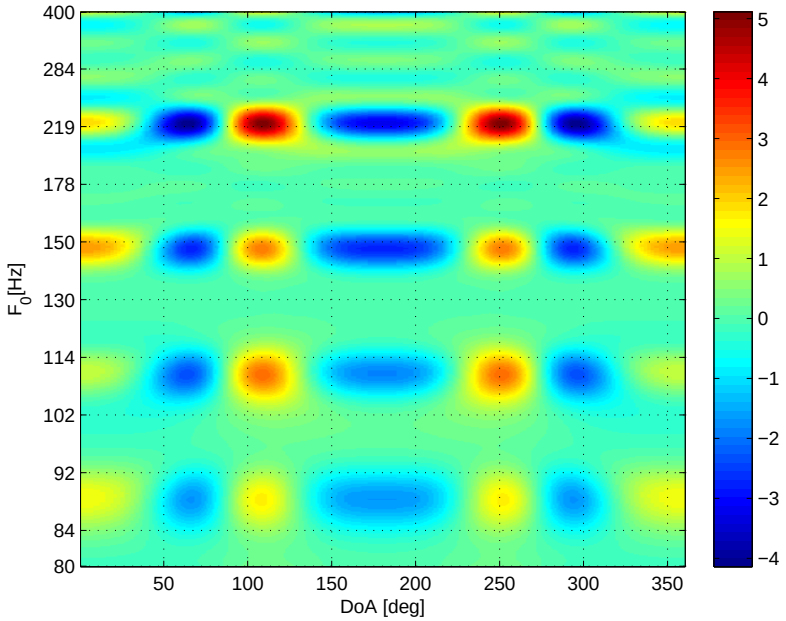


Figure 4.3: PoPi plane for a single female speaker at 115°, $F_0 = 219\,\text{Hz}$ using a pair of microphones. The DoA is shown along the horizontal-axis and the fundamental frequency F_0 along the vertical-axis. The peaks in this 2D representation indicate the likely DoA and the pitch of the active source. The colorbar represents the amplitude of the PoPi plane. The speaker was placed at a distance of 2 m from the array, the estimated SNR for the recordings was 33 dB.

The popi decomposition is made on time lag values corresponding to $0° - 360°$, the shift in the time delays for every pair are defined in $O(\varphi_{0,m_p})$.

By averaging over 12 pairs, the PoPi plane gets sharper, but it does not take care of the pitch multiplicity problem seen in the single pair PoPi plane. This happens when the multiple equidistant peaks in the cross-correlation are misinterpreted as spurious pitch values. To mitigate this effect, the PoPi decomposition is modified in such a way that the "comb" filter does not add up the cross-correlation values located in the middle of "the comb teeth".

The modified PoPi decomposition for multiple pairs is given as

$$\begin{aligned}\rho_t(\varphi_0, F_0) \quad = \quad & \sum_{m_p=1}^{M_p} \Bigg[\frac{1}{2K+1} \cdot \sum_{k=-K}^{K} \Bigg\{ R_{t,m_p}(\lfloor k \cdot L(F_0) + O(\varphi_{0,m_p}) \rfloor) \\ & - \beta \cdot R_{t,m_p}(\lfloor \frac{2k-1}{2} \cdot L(F_0) + O(\varphi_{0,m_p}) \rfloor) \Bigg\} \Bigg]. \end{aligned} \tag{4.4}$$

The second term in the sum defines the position of values, which is subtracted to suppress the unwanted pitch estimate. These values are attenuated by a factor β (a value 0.5 was selected after several trials conducted over different speakers). This modification is illustrated in the second summation line in Fig. 4.1, Fig 4.4(a) and Fig. 4.4(b) show the original and modified PoPi planes respectively, each using 12 pairs of microphones. The new decomposition decreases the presence of pitch multiples.

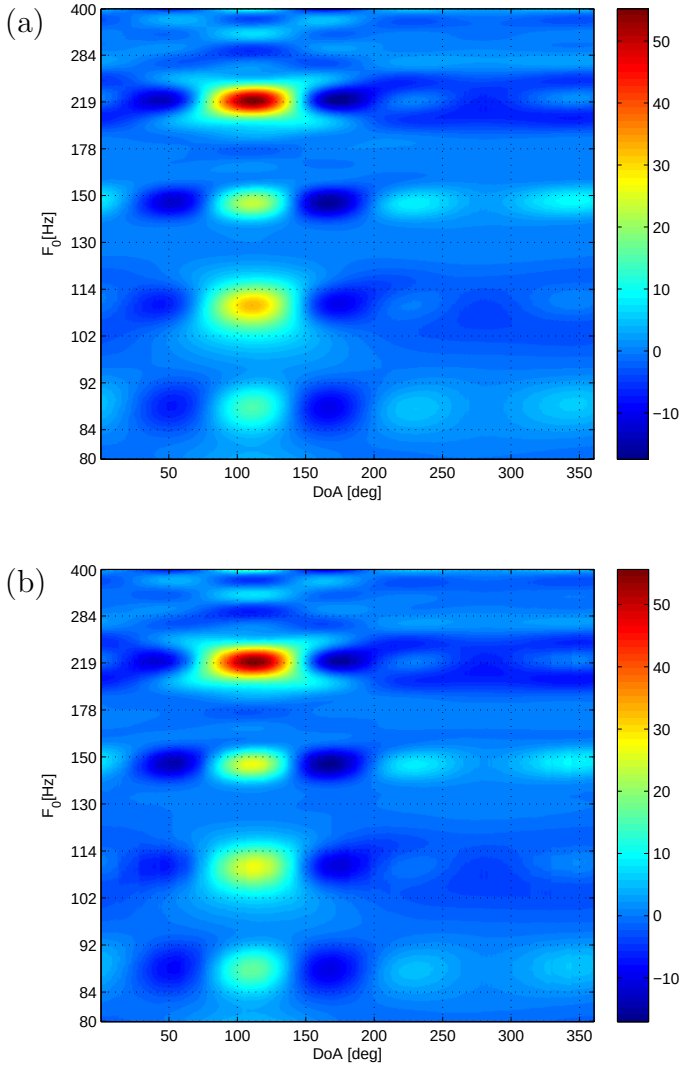


Figure 4.4: PoPi-plane for a single female speaker at 115°, $F_0 = 219\,\text{Hz}$ (a) Original, (b) Modified, each using 12 pairs of microphones from a 24-channel circular microphone array. The colorbar represents the amplitude of the PoPi plane.

4.1.3 Cepstrum Based Weighting Function†

The cross-correlation functions of periodic signals are also periodic. Therefore, when there is a peak in the cross-correlation function at a lag τ, there will be multiple peaks at lags equal to $n \cdot \tau$ (where $n = 1, 2, 3, \cdots$). The PoPi decomposition as defined in (4.1) is based on sampling of cross-correlation function to extract position-pitch relations. Hence, this sampling translates in frequency domain in multiples (including fractions) of the fundamental frequency. In Figs.4.4(a)-(b), these peaks are prominent at $F_0 = 219\,\text{Hz}$, $\frac{2}{3}F_0 = 146\,\text{Hz}$, $\frac{1}{2}F_0 = 109.5\,\text{Hz}$, and $\frac{2}{5}F_0 = 87.6\,\text{Hz}$. The modification proposed in (4.4) tries to resolves this problem but takes into account only half pitch multiple at 109.5 Hz, which get attenuated to a certain degree but the other cross-terms remain unchanged. Therefore, a weighting function is designed to suppress the cross-terms in the PoPi plane, which are the result of the decomposition according to (4.1). The weighting function is derived from the cepstral representation of the cross-correlation function given as

$$C(F_0) = |\text{IFFT}(\log_{10}(\text{FFT}(R^*(\tau)) + \epsilon))|, \tag{4.5}$$

where $R^*(\tau)$ is the half-wave rectified cross-correlation, ϵ is an offset fixed at 1×10^{-6}, and $C(F_0)$ is the cross-cepstrum sample corresponding to range of 80 to 400 Hz. Furthermore, the smoothing operator is applied to (4.5) with a sliding maximum in order to make the peaks of $C(F_0)$ more consistent and less fluctuating, which results in

$$W(F_0) = \max(C(F_0 - 2), \ldots, C(F_0 + 2)). \tag{4.6}$$

Fig. 4.5 shows the cross-cepstrum $C(F_0)$ derived from a voiced segment with $F_0 = 211\,\text{Hz}$ and the resulting smoothed weighting function $W(F_0)$. To demonstrate the power of the weighting function, first a PoPi plane for every microphone pair is obtained through (4.1), and then each plane is weighted by its corresponding weighting function according to (4.6). The resulting weighted PoPi plane is calculated as follows

$$\hat{\rho}(\varphi_n, F_0) = W(F_0) \cdot \rho(\varphi_n, F_0), \tag{4.7}$$

for all DoA values and the pitch frequency is F_0. Figs. 4.6(a)-(d) show the PoPi decomposition using (4.4) of a single source placed at 169° with $F_0 = 211\,\text{Hz}$. The single speaker recordings were made in the SPSC cocktail party room where the

†This technique is an extended version of previously published paper in [14].

speaker was located 2 m from the 24 channel UCA with an SNR of 29 dB and DRR of 1.5 dB. The PoPi decomposition of an un-weighted single pair is shown in Fig. 4.6(a), the weighted version of the same PoPi plane of Fig. 4.6(a) is shown in Fig. 4.6(b). The cross-terms arising due to the periodicity of the cross-correlation along the pitch axis have been successfully removed but the spread along the DoA axis needs to be removed as well. To improve the performance along the DoA axis, the responses of 12 pairs of microphones have been weighted and then summed up, resulting in Fig. 4.6(d). For comparison, the result of summation of all unweighted PoPi planes according to (4.4) is shown in Fig. 4.6(c). In case of the speaker localization, only the positive energy of PoPi planes is of interest thus negative values will be discarded for the later illustrative examples.

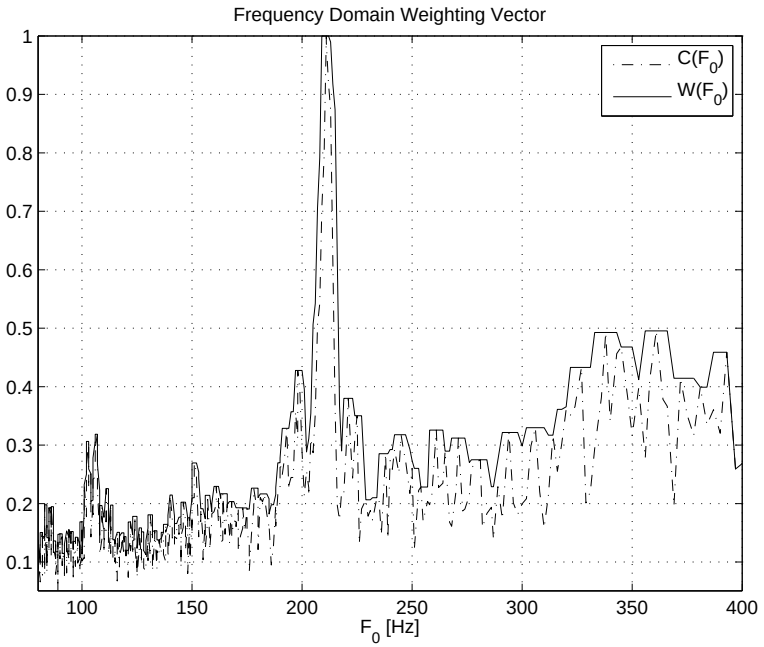


Figure 4.5: The cepstrum based frequency domain weighting function for a source with $F_0 = 211$ Hz. The frequency is plotted along the horizontal-axis and the vertical-axis presents the magnitude of the weighting function. The speaker was placed at 169°, at distance of 2 m from the array with SNR of 29 dB.

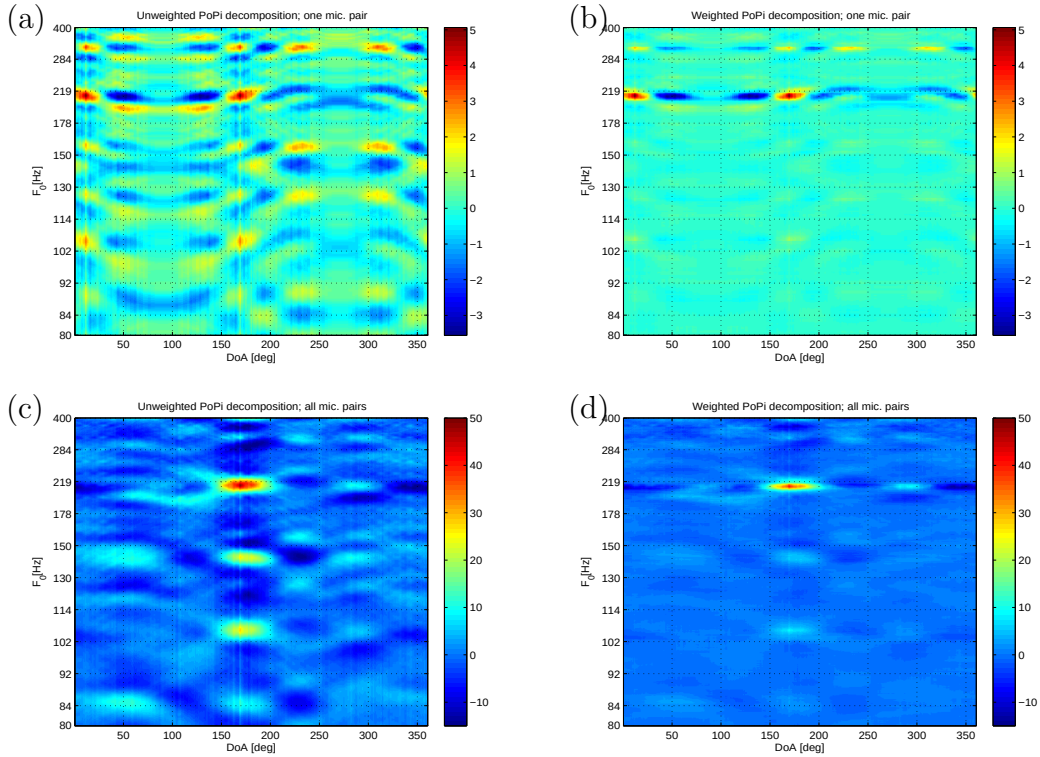


Figure 4.6: The PoPi plane decomposition for a single female speaker placed at 169° using a 24-channel UCA. The speaker was placed at a distance of 2 m from the array with an SNR of 29 dB and DRR of 1.5 dB. (a) PoPi decomposition of an un-weighted single pair, (b) PoPi decomposition of the cepstrum weighted single pair, (c) Unweighted PoPi decomposition of all pairs, (d) Cepstrum weighted PoPi decomposition of all pairs.

4.2 The Multiband Position-Pitch Algorithm†

In case of multiple concurrent speakers, the cepstrum weighting is however unable to extract the pitch information for all active speakers and tends to show erroneous results. Figs. 4.7(a)-(b) shows the PoPi decomposition of a voiced audio segment from a recording of two concurrent speakers, one is male and the other one is female. This recording was made in the SPSC meeting room where both speakers were placed at a distance of 2 m from the array. This example demonstrates the major shortcoming of the full-band PoPi estimation for the two concurrent speakers. The full-band decomposition is unable to detect the female speaker at 142° at all. The male speaker position is correctly detected, but with an incorrect pitch estimate of 211 Hz as shown in Fig. 4.7(a). Even though the cepstrum-based weight-

†This technique is an extended version of previously published papers in [15, 16].

ing introduced in Section 4.1.3 is able to remove cross-terms in the PoPi plane as shown in Fig. 4.7(b), it is unable to detect the pitch estimates of the two concurrent speakers and give erroneous pitch estimate of the male speaker. Therefore, the cepstrum weighting has been omitted from the algorithm and it is not used further in illustrations and experimental results except if stated otherwise.

Therefore, a multiband extension to the PoPi algorithm is investigated for multiple speaker scenarios. Fig. 4.8 shows a complete scheme outlining the Multiband-Position Pitch (MPoPi) algorithm. It is an extended version of the PoPi algorithm. A set of preprocessing steps are taken into consideration before computing the PoPi decomposition to overcome the inability of the PoPi algorithm in detecting more than one concurrent speaker. The PoPi algorithm is used to extract the common periodicities that are present in multi-channel audio in addition to the cross-channel delays. This leads to the parameterized sampling of the cross-correlation. The resulting position-pitch relations can be represented in a plane, the so-called PoPi plane that reveals the peaks at locations that correspond to joint position-pitch estimates of the active sources in an acoustic scene. The description of each processing step in the MPoPi algorithm is outlined below.

Gammatone Filterbank

The audio signals are first processed by a gammatone filterbank, which models part of the auditory process of the inner ear also known as the cochlea. The gammatone filterbank is widely used in Computational Auditory Scene Analysis (CASA) techniques [13]. It impulse responses are the product of a gamma function and a tone (hence "gammatone"), whose impulse response $g_{F_c}(t)$ is given as [13]:

$$g_{F_c}(t) = t^{N-1} \exp[-2\pi t b(F_c)] \cos(2\pi F_c t + \phi)\, u(t), \tag{4.8}$$

where N is the filter order, F_c is the filter center frequency (Hz), ϕ is the phase, and $u(t)$ is the unit step function. The function $b(F_c)$ determines the bandwidth for a given center frequency. The bandwidth of the gammatone filter is set according to the ERB scale. The Equivalent Rectangular Bandwidth (ERB) of a filter is defined as "the bandwidth of an ideal rectangular filter that has the same peak gain, and which passes the same total power for a white noise input" [13, p.16]. For auditory

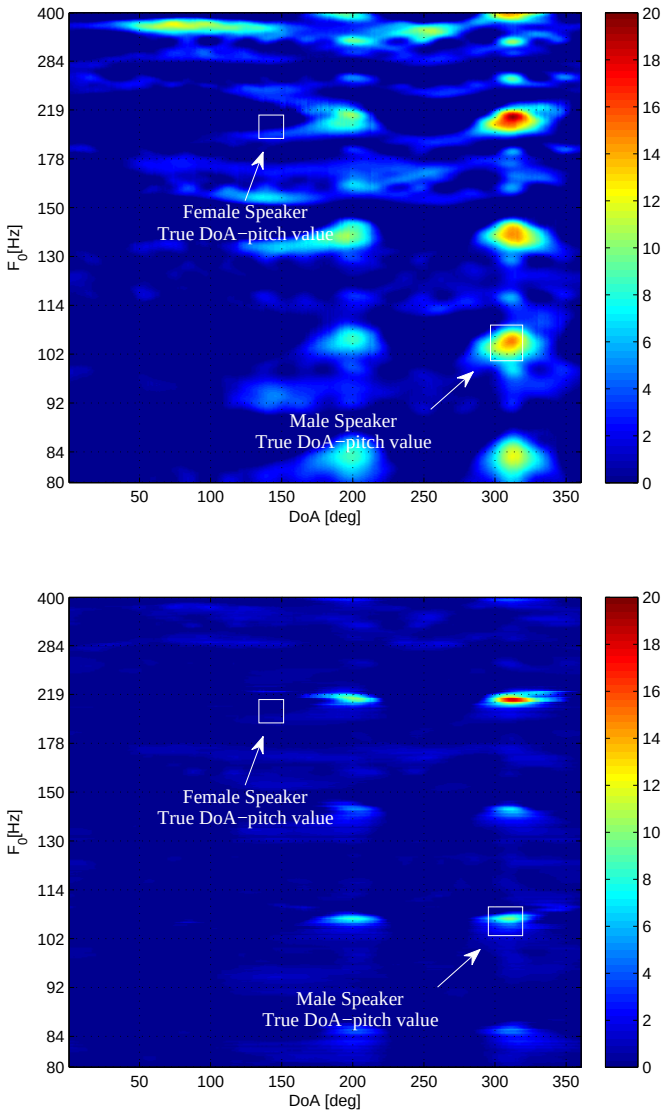


Figure 4.7: Full-band PoPi decomposition of a speech segment with two concurrent speakers (Spkr. 1 (Female): $\varphi_0 = 142°$, $F_0 = 215\,\text{Hz}$; Spkr. 2 (Male): $\varphi_0 = 310°$, $F_0 = 109\,\text{Hz}$) using 24-channel circular microphone array. Both speakers were placed at 2 m from the array, (a) Original PoPi decomposition, (b) Cepstrum weighted PoPi decompositions. Both unweighted and cepstrum-weighted full-band PoPi algorithms are only able to detect one speaker with an incorrect fundamental frequency. The colorbar represents the amplitude of PoPi plane.

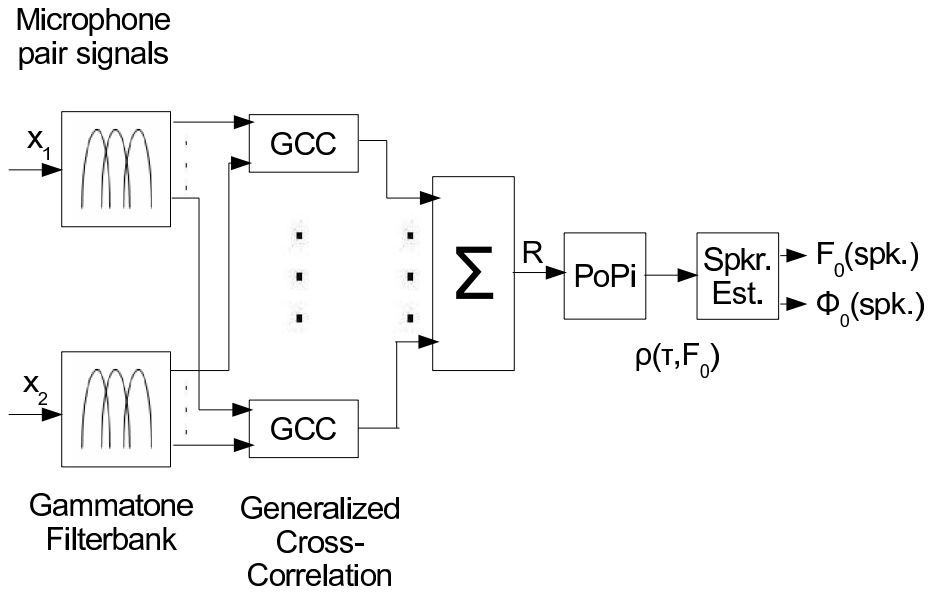


Figure 4.8: Block diagram of the MPoPi algorithm for a single pair of microphones with each signal passing through a gammatone filterbank and then the GCC is calculated for every output channel of the filterbank for one of the microphone signals with the corresponding filterbank output of the other microphone signal. The GCC functions are then summed to generate a "summary" cross-correlation. The PoPi decomposition is applied on the "summary" cross-correlation. For a multi-microphone-pair system, all the PoPi planes are added together to create the final position and pitch estimates for the active sources in an acoustic scene.

filters, it is obtained from

$$\mathrm{ERB}(f) = 24.7 + 0.108f, \tag{4.9}$$

where $\mathrm{ERB}(f)$ and f are in Hz. The relationship between the filter center frequencies and the corresponding bandwidth is defined as

$$b(f) = 1.019\,\mathrm{ERB(f)}. \tag{4.10}$$

A good approximation to the frequency response of the gammatone filter in (4.8) is given as

$$G_{F_c}(f) \approx [1 + \frac{j(f - F_c)}{b(F_c)}]^{-N} \quad (0 < f < \infty) \tag{4.11}$$

It was found in [88] that $N = 4$ proves to fit the estimates of the human auditory filter shapes. For the given application, an ERB scale between 50 Hz and 8000 Hz

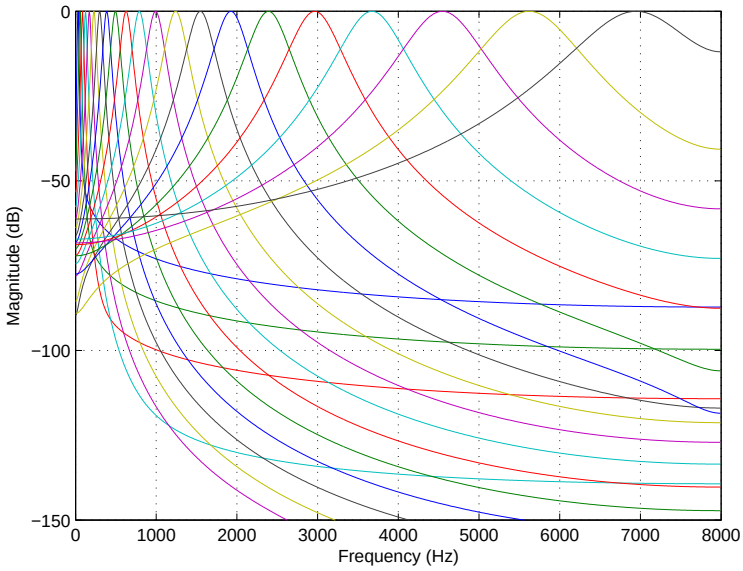


Figure 4.9: Magnitude response of 21 out of 64 gammatone filters.

and 64 overlapping bandpass gammatone filters are used. This number of filters gives a good resolution for the source localization problem. For the experiments presented here, these filters were implemented using the Auditory Toolbox [89]. The magnitude response of every third filter of the filterbank is shown in Fig. 4.9. Multi-pitch tracking is one of the successful applications of these filters [90]. A detailed discussion regarding the effects of the preprocessing on joint estimation of pitch and DoA of an active source is presented at the end of this section illustrated by an example.

Generalized Cross-Correlation

The cross-correlations are computed between all the corresponding bandpass filtered signals of a microphone pair as follows

$$R_t(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(t,\omega) X_1(t,\omega) X_2^*(t,\omega) \exp(j\omega\tau) d\omega, \tag{4.12}$$

where $X_1(t,\omega)$ is the Fourier transform of the windowed signal $x_1(t)$ and $X_2^*(t,\omega)$ is the complex conjugate of the Fourier transform of the windowed signal $x_2(t)$, which is weighted by a weighting function $W(t,\omega)$ and τ is the discrete time-lag. Different

weighting functions can be used with GCC depending on the acoustic conditions. These will be presented in the following section.

Weighting Functions

The well-known weighting functions for the cross-correlation based localization methods have been presented in Section 2.3. The PHAT weighting is particularly advantageous for high SNR and reverberant scenarios. Whereas the maximum likelihood (ML) weighting can be used in cases where the noise statistics can be easily measured or is known *a priori*. In case of PHAT, the magnitude information of the cross-spectrum is removed by whitening the microphone signals. The cross-correlation loses its periodicity, which holds information for the pitch estimation. This makes it an unsuitable candidate for the PoPi algorithm, but PHAT can be used by replacing the central part of the cross-correlation carrying the DoA information with the central part of the GCC-PHAT, where the correlation lag τ corresponds to $0° - 180°$.

$$W(t,\omega,\tau) = \begin{cases} \frac{1}{|X_1(\omega)X_2^*(\omega)|}, & \text{if } \tau \epsilon \langle \tau_{0°}, \tau_{180°} \rangle \\ 1, & \text{otherwise.} \end{cases} \tag{4.13a}$$

$$R(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega,\tau) X_1(\omega) X_2^*(\omega) \exp(j\omega\tau) d\omega. \tag{4.13b}$$

Fig. 4.10 shows the proposed modification, where the modified cross-correlation includes the advantages of GCC-PHAT, while maintaining the periodicity. The top plot shows the full-band cross-correlation, and the second plot illustrates the GCC-PHAT. Both are combined according to (4.13a) in the third plot, which is dubbed as Weighted-GCC. For the multi-band cross-correlations, this process is carried out for every filterbank channel and summed to create a "summary" cross-correlation. The multi-band cross-correlations are denoted as $CC_f(\tau)$ which are computed using (4.13b) for every gammatone filter. The bottom plot of Fig. 4.10 illustrates the resulting multi-band "summary" cross-correlation with the proposed weighting function.

A detailed analysis of PHAT weighting for the ASL task is carried out in [33]. It is one of the first studies done to investigate why the PHAT weighting works well in practical conditions. The study set out to explain two important characteristics of

PHAT weighting: one regarding PHAT being optimal in ML sense when the level of noise is low and second that PHAT is robust to reverberation, because its optimality is independent of environmental reverberation.

In the MPoPi algorithm, the next step is to use a summarization step, which normalizes the cross-correlation before the PoPi decomposition step. With the normalization, the relative information content of the multiple correlation functions is adequately represented and allows combinations of the information from these functions in a better way. With this step, the relative delays associated with the position-pitch information of all sources will be more enhanced in the case of multiple speakers reducing the impact of sources with higher SNR values. These multiple normalized cross-correlations $CC_f(\tau)$ are then summed up to form the so called "summary" cross-correlation $R_{\text{summary}} = \sum_{f=1}^{64} CC_f(\tau)$. The "summary" cross-correlation is used for the PoPi decomposition.

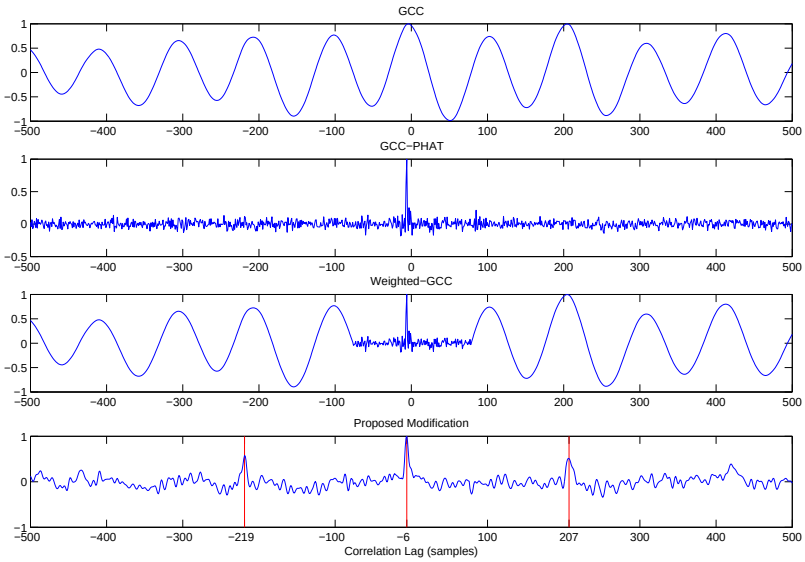


Figure 4.10: The full-band GCC with no weighting is shown in the top plot, the second plot illustrates the GCC with PHAT weighting. The third plot presents the modified PHAT-weighted GCC (Weighted-GCC). The bottom plot shows the proposed modification applied on every cross-correlations of the gammatone filtered signals, which are added to form the "summary" cross-correlation. A pitch period of 213 samples can be seen in the bottom plot where the left peak is shifted to -219 samples as the main peak is not at the zero lag but at -6 samples.

Multiband Position-Pitch Decomposition

The joint position and pitch estimates are determined for the "summary" cross-correlation function using the relation described in detail in Section 4.1 given as

$$\rho_t(\varphi_0, F_0) = \frac{1}{2K+1} \cdot \sum_{k=-K}^{K} R_{\text{summary}}(\lfloor k \cdot L(F_0) + O(\varphi_0) \rfloor), \tag{4.14}$$

where K is set to 3, $L(F_0)$ is pre-calculated for the frequencies $F_0 = [80, \ldots, 400]$ Hz, and $O(\varphi_0)$ are DoA candidates $\varphi_0 = [0°, \ldots, 360°]$ with a stepsize of 1° as discussed in Section 4.1.2. This decomposition is carried out for every pair of microphones and added together to create the final PoPi plane.

Illustrative Example

The effectiveness of the gammatone filterbank preprocessing in the MPoPi method is shown using the same speech segment of Fig. 4.7. The cross-correlations of two signals of a microphone pair filtered with four different gammatone filters with center frequencies (F_c) at 442 Hz, 640 Hz, 1974 Hz, and 4555 Hz are presented in Fig. 4.11. Splitting the signal into subbands emphasizes the harmonic structure of each speaker, furthermore the bandpass filtering helps the weaker source with low energy to get a strong presence in the final summed cross-correlation. Here each subband is making a different contribution to the "summary" cross-correlation, which is used for the PoPi decomposition. The low frequency channels exhibit the pitch information, and the high frequency channels provide better DoA resolution.

The pitch estimation improves by using the preprocessing as the band-pass-filters with different F_c isolate either the fundamental frequencies or multiples it for both speakers. In the top subplot of Fig. 4.11, the band-pass filter number 18 ($F_\text{c} = 442$ Hz) leads to a cross-correlation from which the PoPi decomposition can extract the true male pitch of 109 Hz, because the fourth harmonic ($4F_0$) lies within the pass-band of Filter 18. In a similar way, Filter 23 ($F_\text{c} = 640$ Hz) includes the true female pitch $F_0 = 215$ Hz because of the presence of the corresponding third harmonic ($3F_0$). The last two plots show the cross-correlation functions of high frequency channels, which exhibit peaks corresponding to the position of the male and female speakers, respectively. Moreover, not every frequency channel is

useful in gathering the correct information regarding location and pitch cues. Therefore, the GCC of two more gammatone filters with center frequencies of 555 Hz and 2774 Hz are shown in Fig. 4.12. Both filters show no useful information for time-delay and pitch values. The straight-forward summation of all the channels to generate the PoPi decomposition can lead to erroneous location estimates. Therefore, a frequency-selective criterion is presented in Section 4.3, which pre-groups the frequency channels based on the periodicity information. This pre-grouping improves the accuracy of the MPoPi method. The results are presented in the experimental section of the chapter.

Fig. 4.13 illustrates the PoPi decompositions of the cross-correlations of the band-pass filtered signals presented in Fig. 4.11. In the MPoPi method, the cross-correlations of all filter outputs are normalized and summed up to a so-called "summary" cross-correlation. Fig. 4.14 shows the "summary" cross-correlation of the speech segment. Here, the merits of all filters are combined leading to a function,

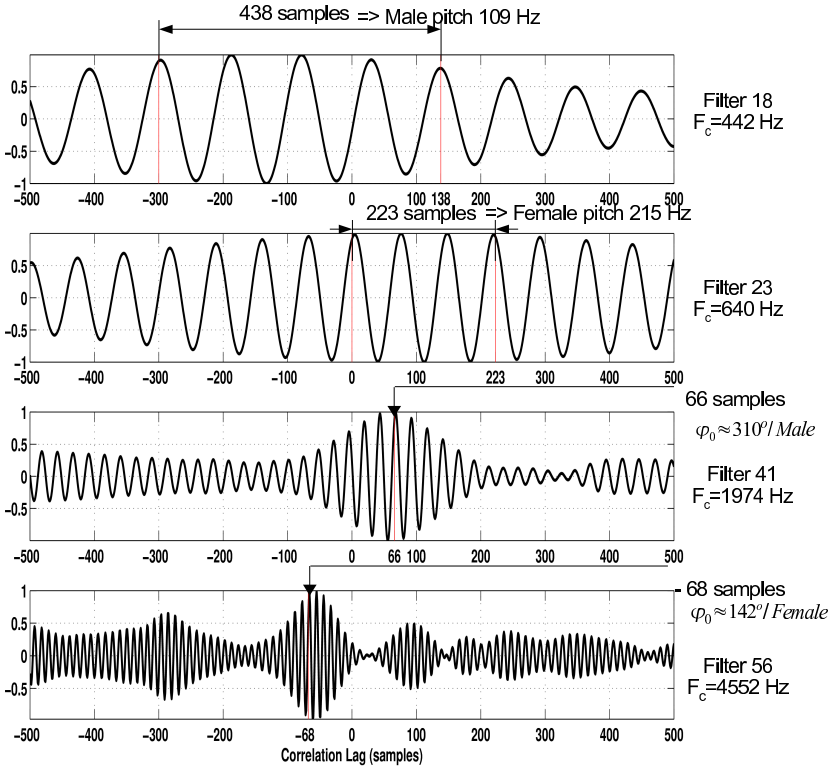


Figure 4.11: The cross-correlations of two signals of a microphone pair ($m_p = 12$) filtered with 4 different gammatone filters with center frequencies (F_c) at 442 Hz, 640 Hz, 1974 Hz, and 4555 Hz.

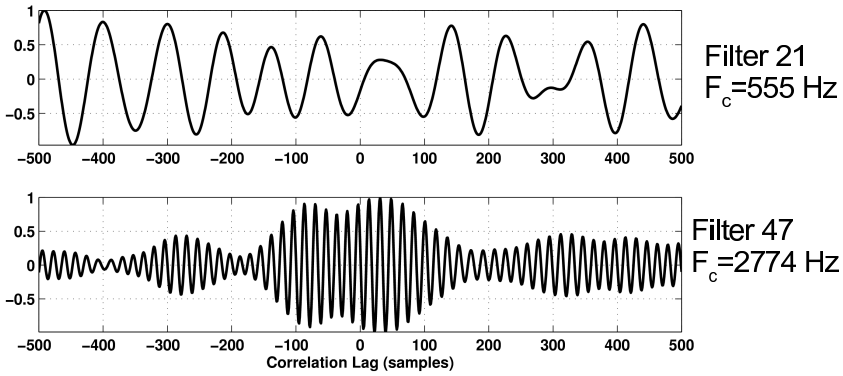


Figure 4.12: GCC of non-informative gammatone filter outputs.

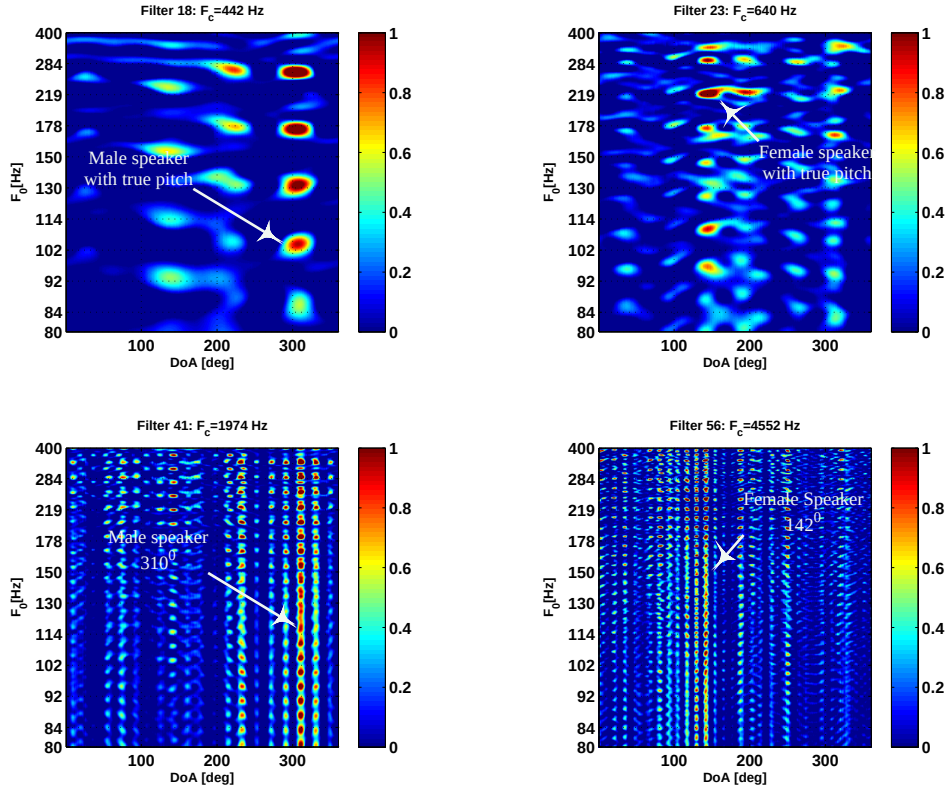


Figure 4.13: PoPi planes of different filter outputs. Both speakers were located 2 m away from the array in the SPSC meeting room with an estimated reverberation time of 500 msec. The estimated SNR for this recording was 32 dB and DRR was 1.51 dB.

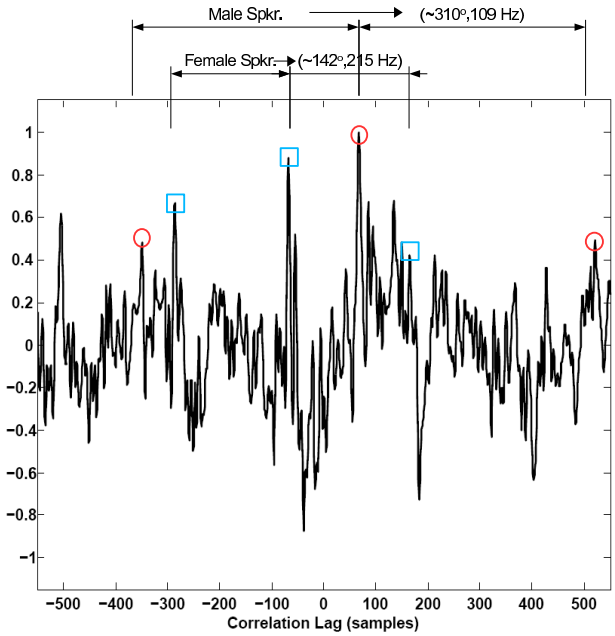


Figure 4.14: Summary cross-correlation of the gammatone filterbank for a mixture of speech, where "○" marks present the DoA and pitch information for the male speaker and "□" marks present the DoA and pitch information for the female speaker.

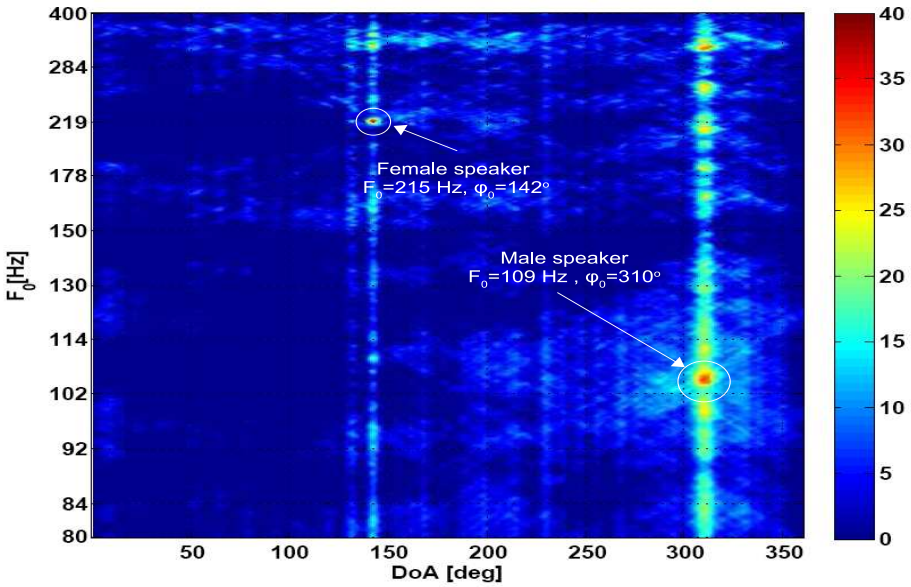


Figure 4.15: MPoPi decomposition of the same speech segment with two concurrent speakers. Contrary to PoPi method, the MPoPi algorithm correctly estimates both DoA and F_0 of each speaker.

including the position and the pitch information of both speakers. The center peak belonging to the male speaker (marked with a '$\circ$') is located at the position of +66 samples ($\approx 310°$), while two correct multiples appear at the distance of 438 samples (109 Hz). The peak defining the position of the female speaker (marked by a '$\square$') is located at -68 samples ($\approx 142°$) with a distance to the multiples of 223 samples (215 Hz). Using the "summary" cross-correlation in Fig. 4.15, the PoPi-plane shows two sharp peaks at positions which indicate the true pitch and DoA values of the two sources.

4.3 Frequency Selection Based MPoPi Method

The MPoPi algorithm works well when both speakers show voiced speech segments in a low reverberant and high SNR scenario. Given real-world speaker scenarios, especially involving speaker interactions and in challenging acoustic conditions, the performance of the MPoPi algorithm deteriorates. This section describes further investigations which were made in order to explore whether a frequency-selective procedure, which is based on statistical speech models and which is inspired from ASA techniques can successfully be combined within the MPoPi algorithm. In case of multiple speakers with varying voiced and unvoiced combinations, the joint position-pitch decomposition over the summation across all frequency channels produces erroneous results. This leads to the definition of a mechanism to pre-group the channels and apply the PoPi decomposition on different sets. The frequency-selective criterion is based on the structure of the auditory filterbank (gammatone filterbank), which have overlapping bandwidths; therefore, in general the neighboring channels contain the same harmonic(s) or formant. In order to quantify their similarity, a cross-channel correlation coefficient C is used as follows

$$C(f,t) = \frac{1}{L}\sum_{\tau=0}^{L-1} \hat{A}(f,\tau,t)\hat{A}(f+1,\tau,t), \tag{4.15}$$

where f is the filter index, t is the time instant, τ is the time-lag, and L denotes the maximum lag of the normalized auto-correlation function $\hat{A}(f,\tau,t)$ of the filtered input signal. The auto-correlation functions are computed using a 42.67 msec Hann window on the output of each auditory filter (64 gammatone filters). The auto-correlations are also used to generate the Auto-Correlogram (ACG), which is

a three-dimensional function representing sound periodicity, it maps the signal periodic energy in a frequency channel of the auditory model at different auto-correlation delays at a given time. The normalization in (4.15) makes the cross-channel correlation insensitive to the signal energy. The channels whose cross-channel correlation is higher than the threshold of 0.95, are grouped together. This accounts for a preliminary grouping of frequency channels, where the reduced ACG is obtained by summation of these channels across frequency. Each set of grouped channels is called "subband" [91].

In [91], the final spectral grouping on these selected channels is carried out to extract the periodicity information of the speech sources from ACG. The reason to select ACG was based on the well-known property that it exhibits a dendritic (or tree-like) structure for periodic sounds or voiced speech signal. The dendritic structures are formed because the frequency channels excited by the periodic signal show a similarity at delays corresponding to the multiples of the fundamental period of the signal. Fig. 4.16 shows the ACG of a female speech utterance. The periodicity of speech is illustrated through the presence of the dendritic structure in the top subplot. The corresponding "summary" ACG which is formed by summing all the frequency channels is shown in the bottom subplot. A peak at 3.42 msec in the "summary" ACG shows the presence of a harmonic source with fundamental frequency $F_0 = 292\,\text{Hz}$ which can also be seen at the stem of the dendrite in the top subplot.

As each source has its own dendritic structure present in the ACG , a 2D cosine function was proposed in [91], which approximates the local shape of the dendritic structure around each gammatone filter. The 2D cosine function consists of five Gabor functions applied to adjacent reduced subbands, in which the middle Gabor function is aligned with the subband. The Gabor functions for a cosine is given as

$$\text{gabor}_c(x; T, \sigma) = \exp(-\frac{x^2}{2\sigma^2}) \cos \frac{2\pi x}{T}, \tag{4.16}$$

where T is the period of the cosine and σ is the standard deviation of the Gaussian. For the application of Gabor functions on real speech signals, the authors in [91] proposed to estimate the frequency of sinusoid in Gabor function. It was emphasized in [91] that due to quasi-periodic nature of speech signals, the repeating frequency or period p_i in each ACG can be off its F_c. Therefore, p_i was empirically estimated for each filter. The standard deviation of the Gaussian was then fixed to $p_i/2$

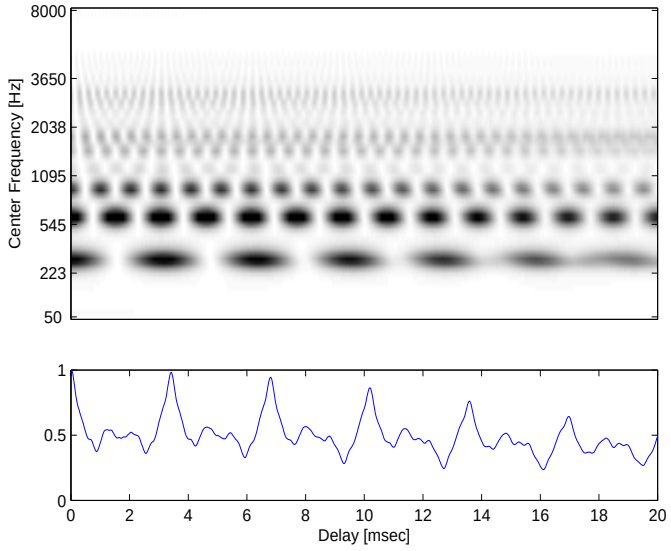


Figure 4.16: The Auto-Correlogram (ACG) of a female speaker. The auto-correlation delay (in msec) is shown along the horizontal-axis and the center frequencies (in Hz) of gammatone filterbank which are linearly distributed on ERB scale are shown along the vertical-axis. The corresponding "summary" ACG which is formed by summing all the frequency channels is shown in the bottom subplot. A peak at 3.42 msec in the "summary" ACG shows the presence of a harmonic source with fundamental frequency $F_0 = 292$ Hz which can also be seen at the stem of the dendrite in the top subplot.

as proposed by the authors in [91]. The enhanced auto-correlation is formed by convolving the auto-correlation of each subband i with support of its four adjacent subbands (two above and two below) with its corresponding 2D Gabor function after zero-padding is given as

$$A_c(i,\tau,t) = \sum_{m_i=-2}^{2} \sum_{n=1}^{L} A(i+m_i, \tau+n, t)\, \text{gabor}_c(n; p_{i+m_i}, p_{i+m_i}/2), \tag{4.17}$$

where L is the maximum auto-correlation delay. The authors [91] further proposed to replace the cosine by a sine function. This reduces the ripples produced as the function aligns not only with the stem of the dendrite but with other peaks as well. Both the sine and cosine functions are convolved with the auto-correlation functions,

and then squared and added to create the enhanced ACG $A(f, \tau, t)$ given as

$$A(f, \tau, t) = A_c(f, \tau, t)^2 + A_s(f, \tau, t)^2. \tag{4.18}$$

It was suggested in [91] that in the enhanced ACG, the stems of the dendritic structures are more emphasized and the peaks in each frequency channel align better with the stems. In an offline process, for each frame of the input signal, dendritic structures ranging from 0 to $\mathrm{N_{max}}$ are found, where $\mathrm{N_{max}}$ is the maximum number of harmonic sources. This process is carried out by picking the largest peak in each subband and generating a histogram, where the highest counting bins indicate the location of possible dendrites associated with harmonic sources. The authors used an empirical threshold of 5 (kept same in this study as well), thus ignoring all bins with counts less than the threshold. A small threshold makes overestimation errors and may assign more frequency channels than what actually contain the harmonicity information. However, it avoids the problem of discarding a channel when it carry harmonic information. In case of two concurrent sources, each frame is labeled as having 0, 1 or 2 dendritic structures. The frequency-selection preprocessing is combined with the MPoPi method to generate a joint position-pitch decomposition based on spectral grouping. The cross-correlations $CC_f(\tau)$ of every gammatone filter as defined in Section 4.2 are summed according to segmentation information retrieved from the frequency-selective criterion. The resulting "summary" cross-correlations is denoted as $R(\tau)_{FS_P}$, where the subscript is added to differentiate from the original MPoPi-based "summary" cross-correlation R_{summary}. The new "summary" cross-correlations $R(\tau)_{FS_P}$ is as follows:

$$R(\tau)_{FS_P} = \frac{1}{N_P} \cdot \sum_{f=1}^{N_P} CC_f(\tau), \tag{4.19}$$

where, for instance in case of a two speaker scenarios, $P \in [\emptyset, 1, 2]$; N_1 and N_2 are the number of frequency channels belonging to speaker 1 and speaker 2, respectively. For every group of channels, a separate PoPi decomposition according to (4.1) is calculated from the function in (4.19):

$$\rho_{FS_P} = \frac{1}{2K+1} \sum_{k=-K}^{K} R_{FS_P}(\lfloor k \cdot L(F_0) + O(\varphi_0) \rfloor). \tag{4.20}$$

The resulting algorithm will from now on be referred to "MPoPi-FS", for the case

when there is no frequency selection, the system reverts to the MPoPi algorithm where the PoPi decomposition is carried out on the "summary" cross-correlation R_{summary} by summing over all channels. So far, only the spectral integration process which is derived on a framewise segmentation of concurrent speech has been discussed. It was further emphasized in [91] that the sequential integration have additional benefits especially when considering the pitch of simultaneous speakers which may overlap in time and tend to be smooth and continuous within a short period of few hundred milliseconds. Thus the combination of spectral and sequential grouping ensures an even stronger grouping of sources, and the addition of evidence from these regions to the MPoPi algorithm produces more robust location estimates. Hence the next section presents how a technique based on this idea was further extended for the developed multi-channel system.

4.4 Spectro-Temporal Fragment Based MPoPi Method

In the previous section, the frequency selection technique proposed in [91] was discussed. The authors designed a system to extract spectro-temporal regions or "fragments" dominated by energy of a single speaker in the presence of multiple concurrent speakers. This scheme was proposed to carry out Speech Fragment Decoding (SFD) which is used for an Automatic Speech Recognition (ASR). In [92], the authors employed a decoding process on the fragment representation to simultaneously identify speech evidence and recognize speech (for details, see [92]). The process of spectral-temporal grouping of the frequency channels is shown in Fig. 4.17. The fragment generation system uses the pre-grouped channels to compute pitch estimates. After which a rule-based tracker was used to create multiple pitch tracks over the estimates and the best matched channels are recruited for each track. This way different spectro-temporal regions of fragments are formed by matching each pitch track with the correlogram peaks presented in every frequency channel. Thus, each spectro-temporal fragment is dominated by a single source.

An example of the spectral and sequential grouping is shown in Figs. 4.18(a)-(d), where Fig. 4.18(a) shows the mixture of speech "place blue in Z zero soon" (female) and "place blue by H 2 please" (male) and Fig. 4.18(b) presents the "cochleagram". According to [13], the cochleagram is defined as the time-frequency representation of

the neural activity pattern, where, in each frequency channel, the response can be interpreted as the instantaneous firing rate within an auditory nerve fiber. The cochleagram is similar to the spectrogram but computed by passing the signal through a cochlear model, which includes the gammatone filterbank and Meddis hair cell model. The Meddis model is known to replicates many characteristics of auditory stimuli such as half-wave rectification, compression, spontaneous firing and saturation effects [13]. The cochleagram shown in Fig. 4.18(b) was produced using a 64-channel gammatone filterbank, where the output of the Meddis hair cell model is smoothed with an 42.7 msec window and updated at 20 msec intervals. The spectral grouping introduced in Section 4.3 is illustrated in Fig. 4.18(c), labeling each time frame as having $\emptyset$, 1, or 2 dendritic structures using different shades of gray showing the presence of no harmonic source, one harmonic source or two harmonic sources, respectively. At this level, it gives a preliminary measure of harmonicity present in different frequency channels. This spectral grouping is finally combined with se-

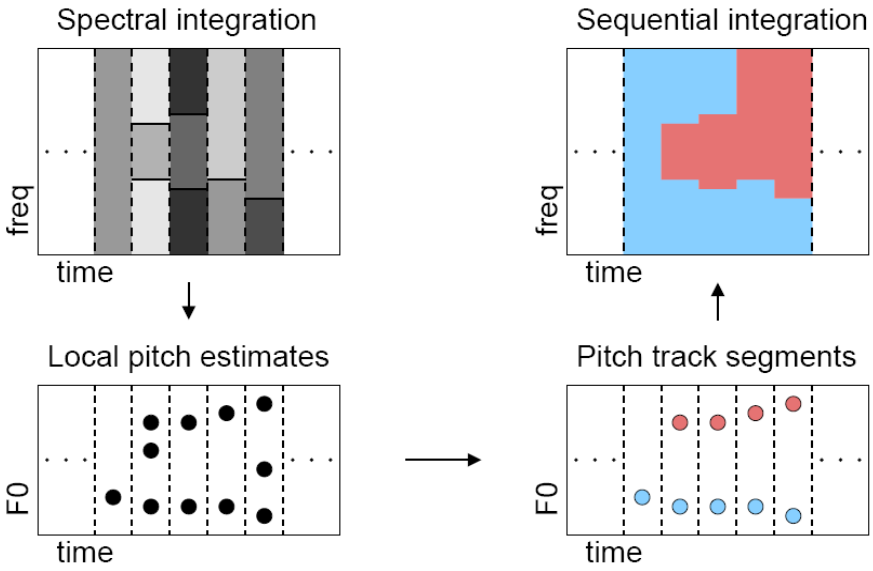


Figure 4.17: Fragment generation process shown in anticlockwise sequence: the upper left plot shows the spectral grouping in different shades of gray in each frame which is based on the frequency-selection criterion. Local pitch estimates in each spectral group are extracted as shown in lower left plot. From these local estimates, two pitch tracks are produced by linking the local pitch estimates as shown in the bottom right plot. Finally, two regions are formed based on two pitch track segments as shown in top right plot. These regions are referred to as spectro-temporal fragments ([92, p.117]).

quential integration to create the harmonic fragments presented in this section as shown in Fig. 4.18(d).

Recent studies on data recorded with a binaural mannequin have demonstrated that the grouping of location cues over the spectro-temporal regions of speech signal yields robust location estimates [66]. A further extension of the fragment based system was proposed in [67] by weighting the elements in a fragment to improve location estimates. The results presented in [67] showed that the proposed scheme helps as not all elements in the fragment are equally affected by reverberation.

In this thesis, a study to investigate the significance of integrating the spectro-temporal regions into the MPoPi algorithm is carried out for concurrent speaker localization. Unlike the previously mentioned study [66, 67], a microphone array consisting of 24 omni-directional microphones is used to record the concurrent speech. The current setup lacks any directionality or human like characteristics and hence is more vulnerable to acoustic conditions of the surrounding. The array is treated as a coherent set of sensors, for which the spectro-temporal regions are generated using one reference microphone pair consisting of microphones 1 and 13. The reason for using this pair was that it is also used as a reference axis to localize speakers with respect to the array.

Fig. 4.19 presents the proposed scheme which combines the frequency-selection and spectro-temporal fragments within the MPoPi method. In this case, the fragments provide the additional information as to how the cross-correlations should be added for the PoPi decomposition presented in (4.21). The resulting method is named MPoPi based spectro-temporal-fragment algorithm or "MPoPi-STF", whereas also only spectral grouping information can be used to combine the cross-correlations, which is then termed "MPoPi-FS". These terms are used in the remainder of the thesis to distinguish between the MPoPi algorithm and its proposed modifications.

The speech signals from microphone 1 and 13 are normalized and added to form the speech mixture, which is processed through the fragment generation system. The system then outputs the fragments formed over both frequency and time as shown in Fig. 4.19. The cross-correlations are computed for every filter output. These cross-correlations as defined in Section 4.2 are added across the spectro-temporal

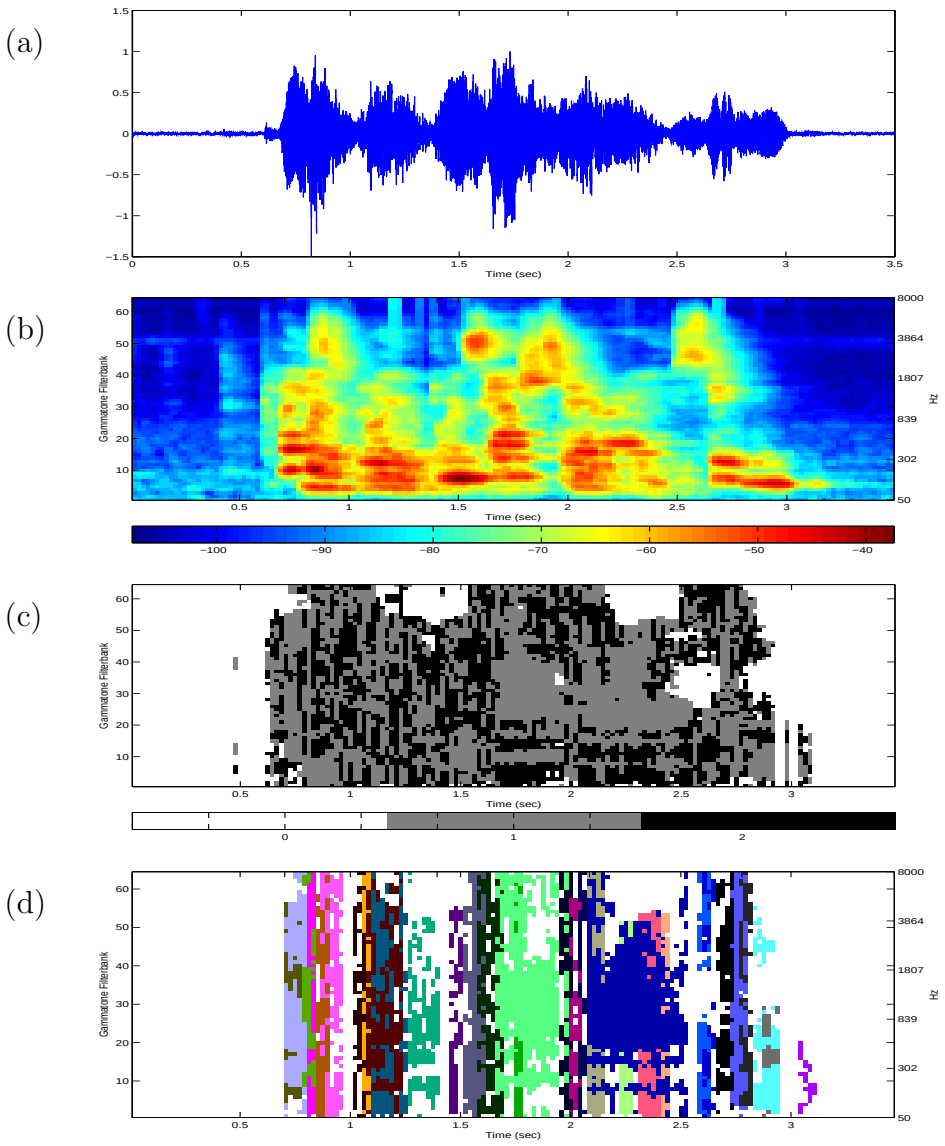


Figure 4.18: (a) A mixture of speech "place blue in Z zero soon" (female) and "place blue by H 2 please" (male) (b) Cochleagram of the mixture, where time (sec) is plotted on the horizontal-axis and vertical-axis presents the gammatone filter index or corresponding center frequency (Hz), and the colorbar represents the energy of the signal at a given time and frequency in dB (c) Spectral grouping of frequency channels labeling each time frame as having 0, 1, or 2 dendritic structures using different shade of gray showing the presence of no harmonic source, one harmonic source or two harmonic sources (d) Harmonic fragments after sequential grouping, the spectro-temporal regions are color-coded where a single color represents the number of frequency channels and time frames belonging to a particular fragment.

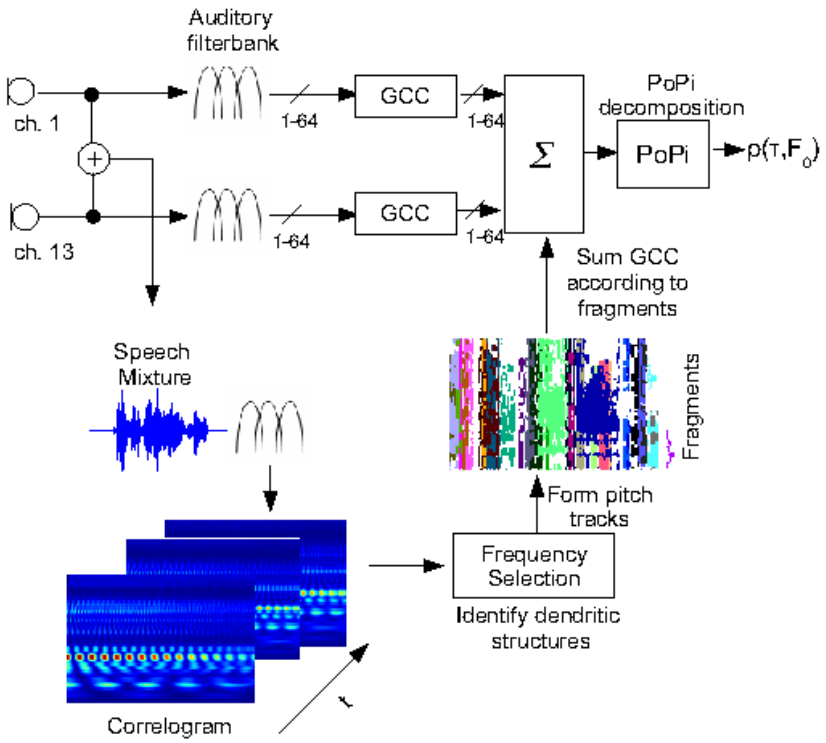


Figure 4.19: Proposed scheme for multi-channel extension of a fragment system for a 24-channel UCA, where ch.1 and ch.13 are normalized and added to form the speech mixture used to generate the spectro-temporal fragments. The resulting fragments are finally combined within the MPoPi algorithm. The PoPi planes of all 12 oppositely placed microphones are generated through this process and then added according to (4.4).

fragments following the systems proposed in [67] such as:

$$R(\tau)_{unweighted} = \frac{1}{L} \cdot \sum_{(f,t)\ \epsilon\ F_p} CC_{(f,t)}(\tau) \tag{4.21a}$$

$$R(\tau)_{weighted} = \frac{1}{L} \cdot \sum_{(f,t)\ \epsilon\ F_p} \psi_{(f,t)} \cdot CC_{(f,t)}(\tau), \tag{4.21b}$$

where F_p is a speech fragment containing frequency channels at various time instants, L is the total number of frequency channels and time frames in a fragment, and $CC_{(f,t)}(\tau)$ is the cross-correlation for a frequency channel f and frame t over

a range of time-lags, $\tau \in \{\tau_{min}, \ldots, \tau_{max}\}$. The spectro-temporal regions are color-coded as shown in Fig. 4.18(d), where each color represents the frequency channels and time frames belonging to a particular fragment. In (4.21a), an averaged cross-correlation is computed for a speech fragment, where each frequency channel in the fragment is assigned an equal weight. Whereas in (4.21b), the summation is carried out by assigning different weights $\psi_{(f,t)}$ based on the interaural coherence (IC) weighting criterion [93] given as:

$$\psi_{(f,t)} = \max_{\tau} \; CC_{(f,t)}(\tau). \tag{4.22}$$

The results reported in [17] suggest that the weighted "summary" cross-correlation from (4.21b) is more suitable for the fragment generation system. Therefore, in this thesis only the weighted spectro-temporal fragment scheme is used. The process is repeated for both algorithms for every microphone pair. The resulting planes are later summed for all microphone pairs to generate the final location estimates.

4.5 Experimental Evaluations

This section presents a list of experiments which evaluate the performance of the proposed ASL algorithms for different acoustic conditions such as: background noise, single to multiple, and static to mobile speakers to name the few. The details of the recordings and speaker setups have already been discussed in detail in Chapter 3. For analysis of the proposed algorithms, a frame length of 42.6 msec is used with a frame shift of 20 msec. These parameters were selected keeping the problem at hand in mind as for pitch estimation in order to determine the pitch of 80 Hz and its three harmonics at least 37.5 msec frame length is required. The high update rate is essential for any localization algorithm to track fast changing events in the environment. The array used for all recordings had 24 microphones placed in a UCA with 0.55 m diameter. The recordings were made in the cocktail party room with reverberation time of 500 msec. The estimates were produced by using 12 pairs of oppositely placed microphones. Furthermore, the proposed algorithms are compared with the well-known ASL method called as SRP-PHAT, which was presented in Chapter 2. In [8], the author presented the relationship between the SRP-PHAT and GCC-PHAT method. The author showed that the SRP-PHAT for an M microphone system is equal to the sum of GCC-PHAT of all possible pairs of microphones. As

the cross-correlation of a microphone pair $(1, 2)$ will be same as $(2, 1)$, and the auto-correlations only add a constant to the final value, the traditional SRP-PHAT requires only the upper triangular matrix of GCC-PHAT. For the 24-channel UCA, the SRP-PHAT is computed with only the 12 diametrically placed pairs. This process varies from the traditional formulation. This summation, however, gives a better functional value in locating the source(s). The reason is that microphone pairs with two microphones which are far apart from each other tend to minimize the cross-correlation values of the noise, while maintaining the cross-correlation values of the true signals coming from the source. On the one hand, close pairs tend to boost up the cross-correlation values but those values could be because of the similarity of the noise profiles at the two microphones of the pair. The computation of SRP-PHAT using of a subset of microphone pairs can be referred to as the "modified" SRP-PHAT algorithm[†].

4.5.1 Controlled Experiments

The controlled setups are used for algorithm evaluations by playing back high-quality speech through loudspeakers. This setup allows to repeat the experiments many times and makes a detailed evaluation of the algorithms much easier. To determine the accuracy scores, the ground truth values are determined from reference label files. These reference files are prepared using the noise-free and reverberation-free high quality speech from the Grid corpus [81] as illustrated in Section 3.5.

Single Static Speaker

The evaluations of the algorithms begin with a single static speaker placed at 2 m from the array. Fig. 4.20(a) and Fig. 4.20(b) show the accuracy in percent (the accuracy counts are calculated using the metric defined in Chapter 3) plotted as a Cumulative Distribution Function (CDF) versus the error threshold for a single speaker located at 169° and 310°, respectively. The threshold is varied from 1° to 90° as shown along the horizontal-axis of the plots. The results are averaged over six speakers, three males and three females, which were randomly selected from

[†]I would like to thank Hoang Do from the Brown University, U.S.A for this useful discussion. For the rest of thesis, the "modified" SRP-PHAT is referred to as the SRP-PHAT defined for the given geometry.

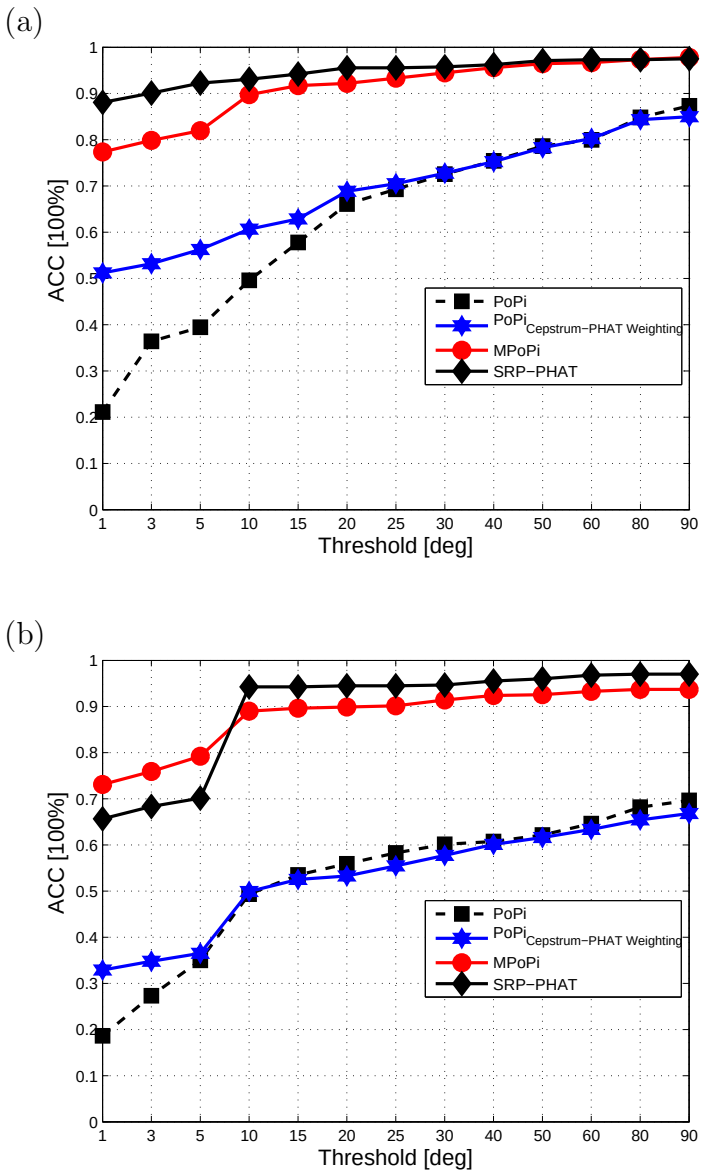


Figure 4.20: Accuracy counts plotted as a CDF versus the error threshold for a single speaker placed at (a) 169° and (b) 310°. The "- ■ -" represents the results of PoPi algorithm, "–✶–" represents the results of PoPi algorithm with cepstrum and PHAT weighting, "–●–" represent the results of MPoPi, and "–◆–" represents the results of SRP-PHAT algorithm. The speaker was placed at a distance of 2 m from the array.

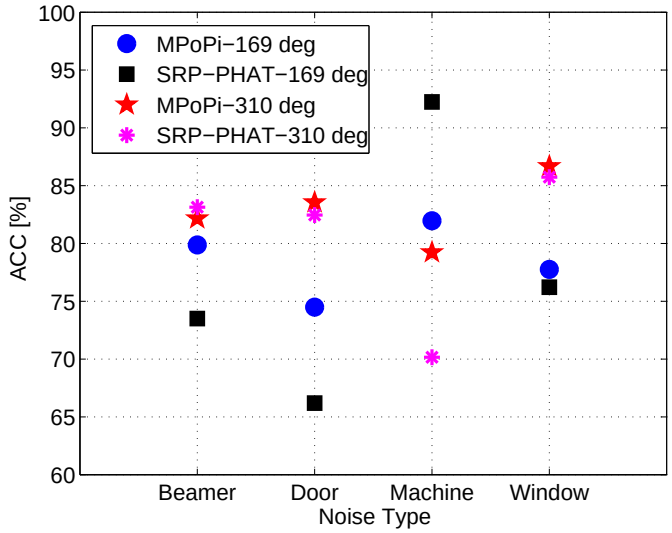


Figure 4.21: Accuracy counts of MPoPi and SRP-PHAT algorithm for a single speaker versus different kinds of background noise. The "●" represents the results of MPoPi for 169° case, "■" represents the results of SRP-PHAT for 169° case, "★" represent the results of MPoPi for 310° case, and "✳" represents the results of SRP-PHAT for 310° case. The speaker was placed at a distance of 2 m from the array.

the Grid corpus for the playbacks. Each speaker utterance is three seconds long. For comparison, the full-band PoPi algorithm and its proposed modification with cepstrum and PHAT weighting methods (Section 4.1.3), the multiband extension termed MPoPi algorithm (Section 4.2), and SRP-PHAT is used. Out of all the algorithms, the full-band PoPi algorithm has the most outliers. The proposed weighting for the PoPi method shows improved performance at smaller error thresholds. In case of the speaker at 310°, however, both PoPi methods with and without the weighting function fail to converge even at 90° threshold. Because of the poor performance of the PoPi algorithms, they will not be used further in the evaluations. The MPoPi and SRP-PHAT algorithms give ninety percent correct estimates in both cases within a 10° error threshold. Hence, for the single speaker scenarios with high SNR conditions, MPoPi shows a similar performance as the SRP-PHAT.

Background Noise

The evaluations are extended to testing MPoPi and SRP-PHAT under various acoustic conditions with different kinds of background noise commonly present in a meet-

ing room or an office space. The kinds of noise against which all algorithms are tested are as follows:

- Beamer Noise (fixed on the roof in the center of the meeting room).
- Environmental noise coming from an open window (located on one of the wall of the meeting room as shown in Fig. 3.5).
- Opening and closing of the door (see Fig. 3.5).
- Machine Noise (a computer placed on the floor below the window of the meeting room).

Fig. 4.21 presents the accuracy counts for a single speaker at 169° and 310°. The accuracy counts are displayed versus the different kinds of background noise along the horizontal-axis. The results are again averaged over all six speakers. The error threshold is fixed at 5°. This value is chosen as it corresponds to the minimal inter-speaker distance of 35 cm in case of 2 m distance from the array. For all the conditions, MPoPi shows more consistent performance in comparison to SRP-PHAT, which has an accuracy difference of up to twenty percent for similar noise scenarios at different speaker positions especially between the beamer and machine noise for 310° due to strong reflection arising around 8° shift.

Moreover, a special set of recordings were carried out, where a "noise" (interference) loudspeaker was placed on a floor at the same distance from the array as the "speech" (desired) source (for details of the experimental setup, please see Section 3.7). The speech source was placed at two different positions, one close to the noise source, at 169°, and the other farther away at 310°. Fig. 4.22 shows the accuracy results at 5° error threshold for SRP-PHAT and MPoPi algorithm for the two DoAs. In this case, the performance of both algorithms deteriorates as the SNR decreases below 10 dB. An interesting observation is that if the desired and interference speakers are closer, the SRP-PHAT algorithm works better than MPoPi algorithm. In contrast, for the case where the desired speaker is farther away from the interference speaker, MPoPi is able to localize the speech source better than the SRP-PHAT.

Multiple Concurrent Speakers

The evaluations are further extended to multiple concurrent speakers, which means that there can be more than one active speaker in a single analysis frame. The algorithms are tested for four concurrent speakers, and results are averaged for different

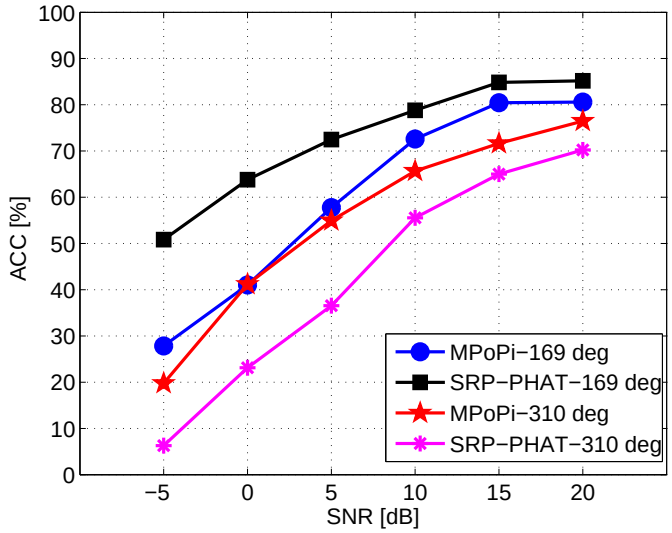


Figure 4.22: Single speaker accuracy counts of MPoPi and SRP-PHAT algorithm versus different SNR levels. The "–●–" represents the results of MPoPi for 169° case, "–■–" represents the results of SRP-PHAT for 169° case, "–★–" represent the results of MPoPi for 310° case, and "–✳–" represents the results of SRP-PHAT for 310° case. The speaker was placed at a distance of 2 m from the array.

speaker combinations. Here, two proposed modifications to the MPoPi algorithm that is MPoPi-FS (Section 4.3) and MPoPi-STF (Section 4.4) are added for evaluations. These methods are designed to improve localization accuracy for concurrent speaker scenarios using different subband processing techniques. Figs. 4.23(a)-(g) presents the averaged accuracy scores of all algorithms for two concurrent speakers plotted as a CDF versus the error threshold ranging from 1° to 40°. The curves show the averaged results over all speaker combinations. There were nine speaker combinations recorded for the two concurrent speakers scenarios consisting of three male-male, three female-female, and three male-female combinations. Fig. 4.23(a) shows the accuracy counts for closely spaced speakers, where speaker separation is gradually increased up to oppositely placed speakers as shown in Fig. 4.23(g).

The MPoPi-FS and MPoPi-STF outperform SRP-PHAT algorithm for smaller thresholds up to 5° in all cases except in Fig 4.23(b), and Fig 4.23(f). The reason for the decrease is that the DoA estimates of the algorithms were one degree higher than the acceptance range for one of the speaker's position. That is why the accuracy increases as the threshold becomes larger. Another problem for the poor performance of algorithms for the scenario in Fig 4.23(f) is because the second

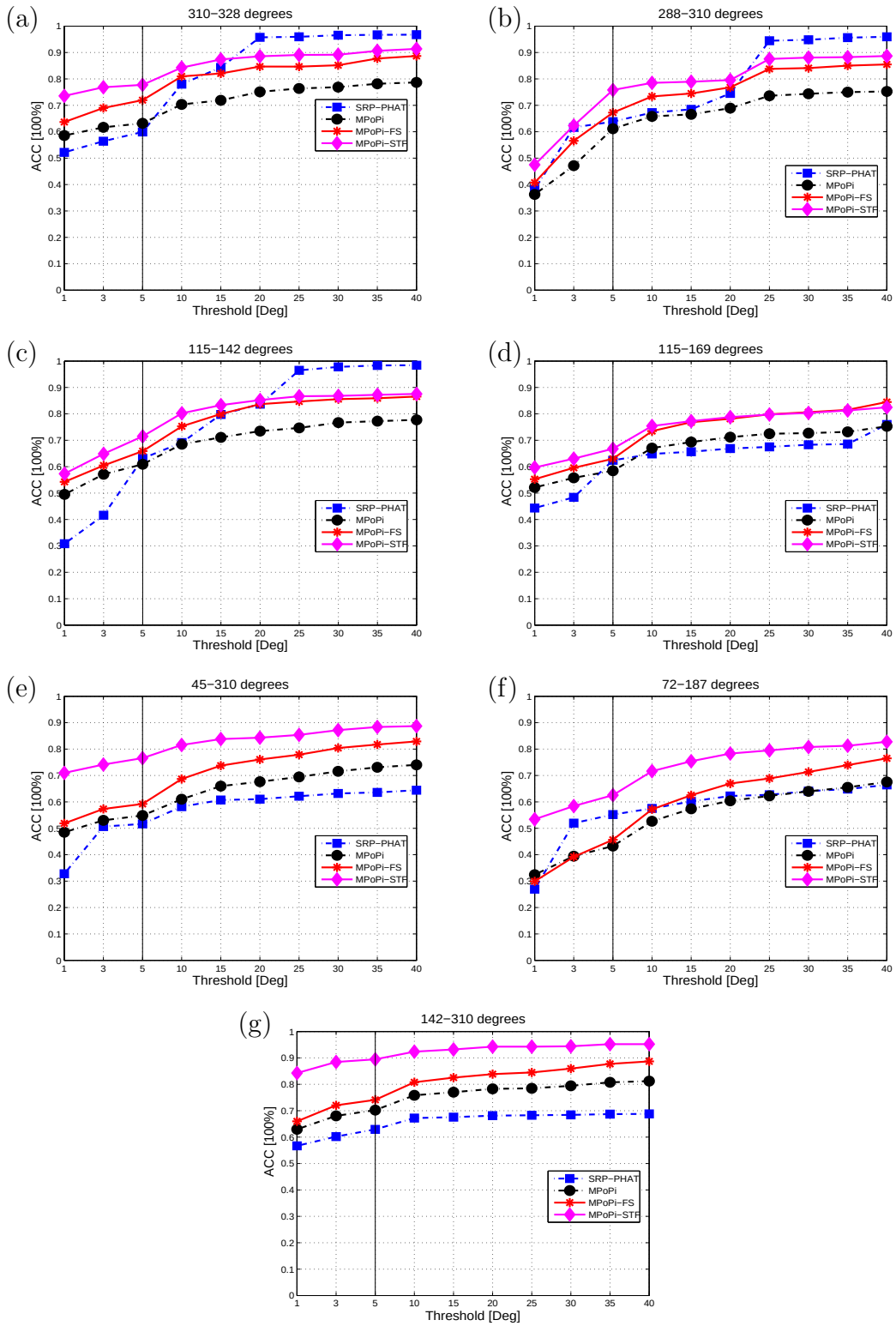


Figure 4.23: Accuracy counts versus the error threshold plotted as a CDF for two concurrent speakers at different speaker positions (a)-(g) starting from closely spaced going up to oppositely placed, where "-■-" represents the SRP-PHAT algorithm, "-●-" represents the MPoPi method, "–✱–" represent the MPoPi-FS and "–◆–" represents the MPoPi-STF method. The speakers were placed at a distance of 2 m from the array with SNR of 32 dB and DRR of 1.51 dB.

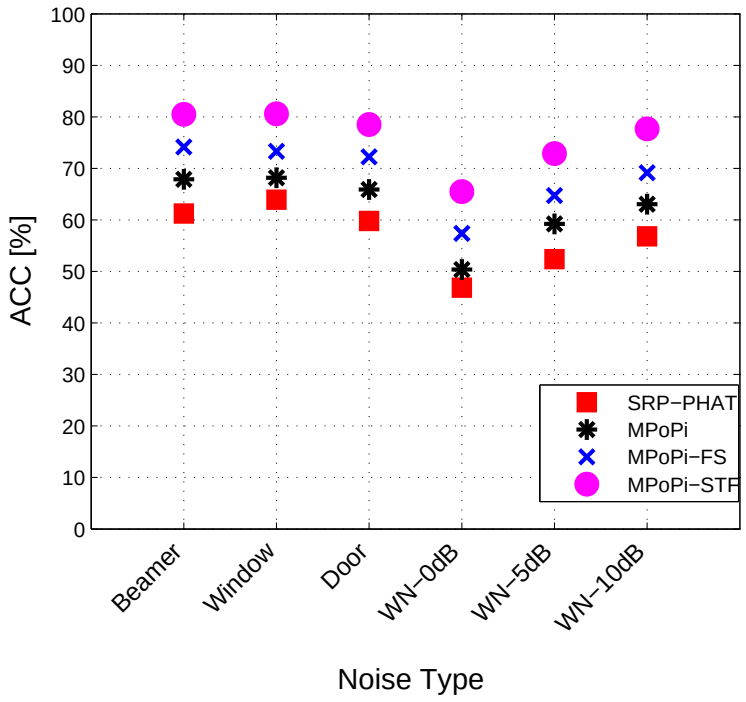


Figure 4.24: Accuracy counts of SRP-PHAT and MPoPi algorithms for different noise types in case of two concurrent speakers placed at 142° and 310°. The "■" represents the SRP-PHAT algorithm, "✱" represents the MPoPi method, "×" represent the MPoPi-FS and "●" represents the MPoPi-STF method. The speakers were placed at a distance of 2 m from the array.

speaker was placed close to the window at 187°. Due to the highly reflective surface of the window, strong reflections were present in form of spurious peaks in the PoPi plane.

Single speaker noise conditions were repeated for the case of two concurrent speakers as well. The SNR recordings with white noise were only repeated up to SNR of 10 dB, which is represented as "WN-10dB" on the plot. The speakers were placed at 142° and 310° while keeping the same distance from the array. Fig. 4.24 shows the accuracy counts for all algorithms for different noisy conditions presented along the horizontal-axis. The error threshold is fixed at 5°. The results are averaged over nine speaker combinations as explained earlier in the section. The accuracy scores show the advantage of MPoPi methods over the phase information only SRP-PHAT method in localizing two concurrent speakers.

The algorithms are further evaluated for three and four concurrent speaker sce-

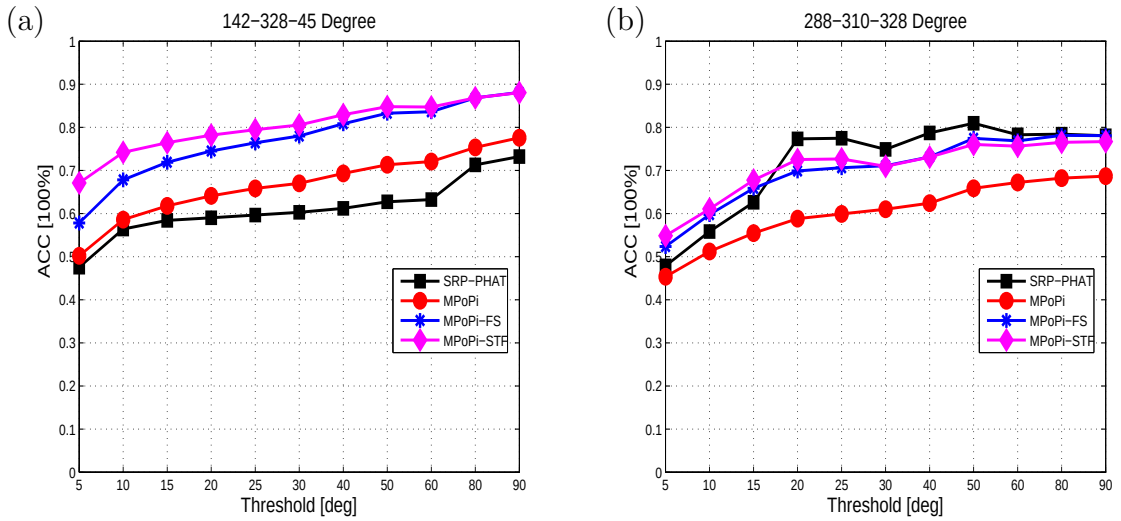


Figure 4.25: Accuracy counts versus error threshold plotted as a CDF for three concurrent speakers placed at (a) 142-328-45 degrees, (b) 270-288-310 degrees, where "–■–" represents the SRP-PHAT algorithm, "–●–" represents the MPoPi method, "–✳–" represent the MPoPi-FS and "–◆–" represents the MPoPi-STF method. The speakers were placed at a distance of 2 m from the array.

narios. In the three speakers case, eight speaker combinations were recorded, which included two sets each of male-male-male, male-female-male, female-female-male, and female-female-female. In the four speaker case, there were five different speaker combination scenarios including one set each of male-male-male-male, male-male-male-female, male-female-male-female, female-female-female-male, and female-female-female-female. Fig. 4.25(a) shows the results of three spatially separated speakers and Fig. 4.25(b) shows the results of three closely spaced speakers. The accuracy counts are plotted as a CDF versus the error threshold, which is varied from 5° to 90°. The results are averaged over all eight speaker combinations in both cases. For the four concurrent speakers, Fig. 4.26(a) shows the accuracy counts of far placed speakers and Fig. 4.26(b) shows the accuracy counts of closely spaced speakers. The threshold range is the same as for four speaker scenarios. The accuracy results are averaged over all five speaker combinations. The relative improvement achieved by the MPoPi-FS and MPoPi-STF over MPoPi for the these scenarios is decreased. The deterioration in performance is due to highly challenging environment where the frequency segmentation and fragment generation is carried out on a single channel where it is difficult to resolve three concurrent harmonic sources

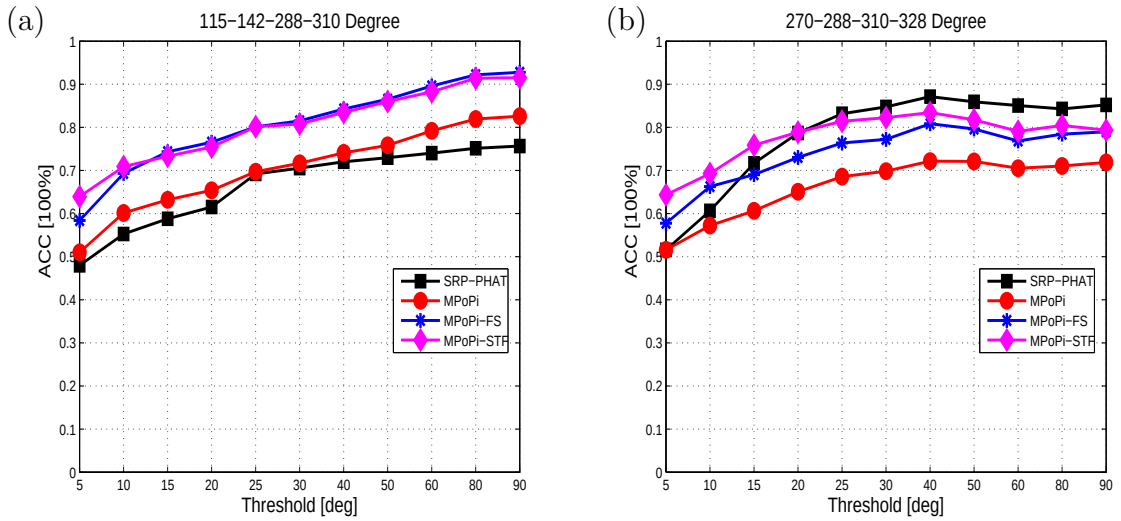


Figure 4.26: Accuracy counts versus error threshold plotted as a CDF for four concurrent speakers placed at (a) 115-142-288-310 degrees, (b) 270-288-310-328 degrees, where "–■–" represents the SRP-PHAT algorithm, "–●–" represents the MPoPi method, "–✳–" represent the MPoPi-FS and "–◆–" represents the MPoPi-STF method. The speakers were placed at a distance of 2 m from the array.

in reverberant environment. A further investigation is required so that the other channels of the microphone array can also facilitate the segmentation and sequential integration process.

Different Array Configurations

Another interesting set of recordings was made by exploiting the variability provided by the UCA. The diameter of the array can be varied from 0.2 m to 0.55 m as shown in Fig. 3.1. Only the MPoPi algorithm is chosen to test its performance variations occurring because of the variable number of microphones and array diameters. Figs. 4.27(a)-(d) present the accuracy counts versus error threshold plotted as a CDF for all four diameters. For this experiment, the speaker was placed at 310° at a constant distance of 2 m from the array. Fig. 4.27(a) shows the accuracy counts by using only three microphone pairs, Fig. 4.27(b) shows the counts using four microphone pairs, Fig. 4.27(c) and Fig. 4.27(d) presents the cumulative counts using six and twelve microphone pairs, respectively. In all cases, the pairs are formed by using only oppositely placed microphones. This simplifies the problem of localizing

an active speaker as all pairs share a common center defined as the origin of the array.

Fig. 4.27(a) shows that if fewer numbers of microphone pairs are used, an array with a smaller diameter performs better. This can be attributed to the problem of spatial aliasing, which occurs when the microphone pairs are placed at a large distance from each other and the distance between a microphone pair is more than the smallest wavelength of interest. Thus grating lobes are created along with the main beam in the beam pattern of the microphone array. The presence of grating lobes leads to incorrect position estimates. The spatial disambiguation improves by adding one more pair as shown in Fig. 4.27(b). Therefore, at least four microphone pairs are needed to accurately localize active speakers. Overall, the array with the largest diameter outperforms the other for the higher number of microphone pairs because the DoA resolution improves if the distance between the microphones increases. This is also the reason why the smaller diameter arrays require large thresholds to achieve high localization accuracy.

The effect of array diameter is further tested for the two concurrent speaker scenario. Fig. 4.28 shows the accuracy results in terms of CDF versus the error threshold of two speakers placed at 142° and 310° for different array diameters. The threshold is plotted along the horizontal-axis, which is varied from 1° to 40°. Here all 12 pairs are used for the DoA estimation. The arrays with the smaller diameters, for example, $d = 30\,\text{cm}$ and $d = 20\,\text{cm}$ require larger thresholds to achieve the accuracy of larger diameter arrays. Therefore, if the application and computational resources permit, using a large number of microphones and diameter spacing improves localization accuracy.

4.5.2 Real Speakers Experiments

This section presents the results for different speaker interaction scenarios recorded with human speakers. The recordings were carried out with four speakers, two males (Martin and Wolfgang) and two females (Anna and Tania) using two scenarios:

- Presentation scenario
- Meeting scenario

In the presentation scenario, one speaker was standing close to the white board or projector screen, and the others formed the audience sitting in fixed positions. In

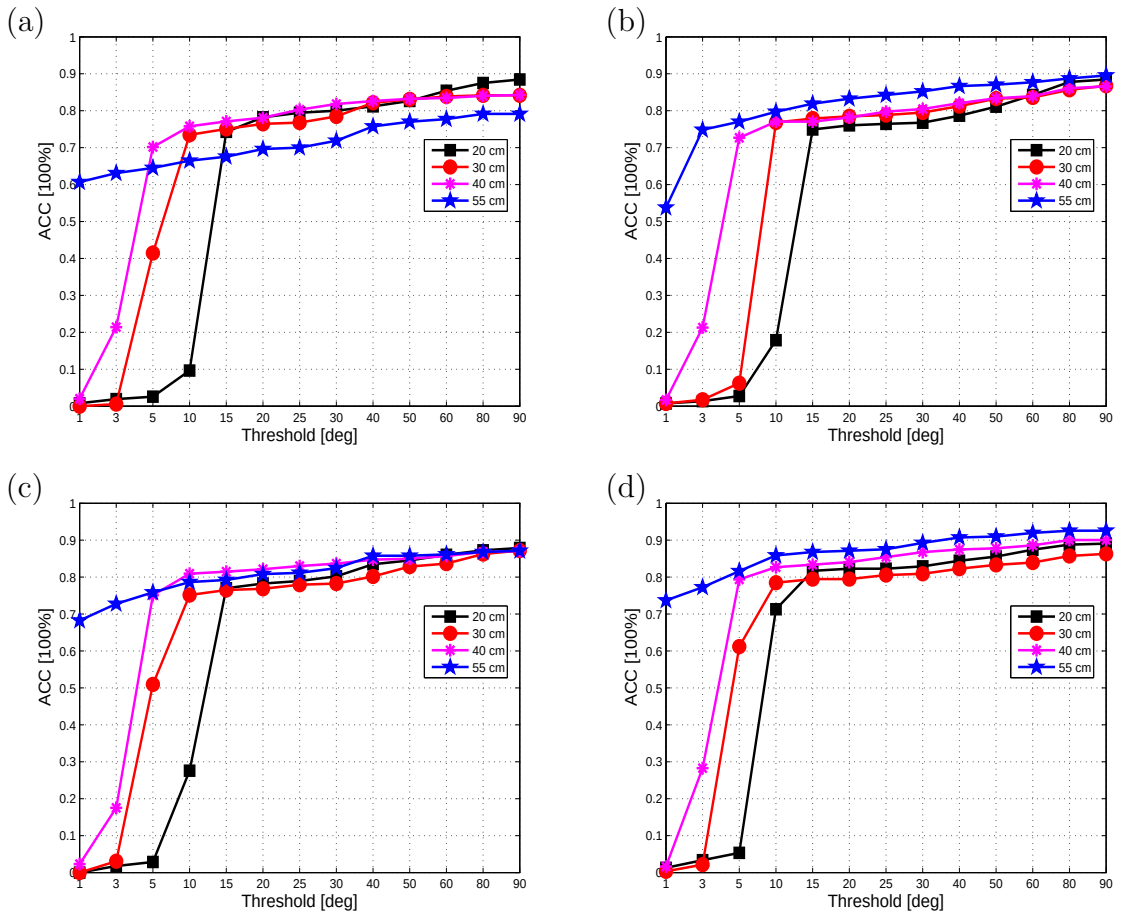


Figure 4.27: Accuracy in percent versus the error threshold plotted as a CDF for a single speaker placed at 310° using (a) 3 pairs (inter-pair angle: 60°), (b) 4 pairs (inter-pair angle: 45°), (c) 6 pairs (inter-pair angle: 30°), (d) 12 pairs (inter-pair angle: 15°) of microphones for different array diameters. The "–■–" represents MPoPi results for the 20 cm, "–●–" represents the results for 30 cm , "–✳–" represent results using 40 cm, and "–★–" represents results with 55 cm array.

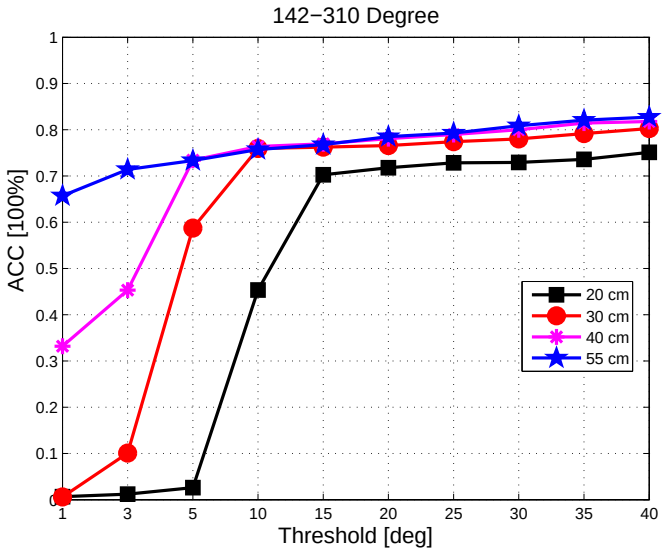
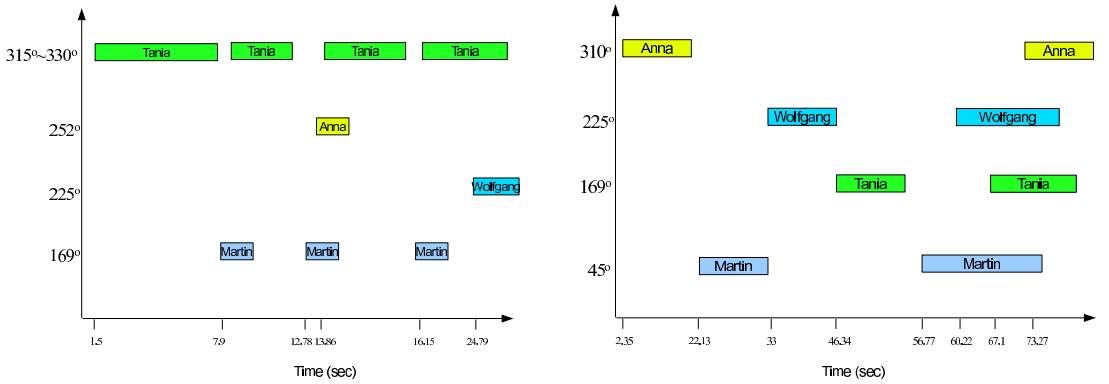


Figure 4.28: Accuracy in percent of MPoPi algorithm versus the error threshold plotted as a CDF for two concurrent speakers placed at 142° and 310° using different array diameters. The "–■–" represents the 20 cm results, "–●–" represents the 30 cm results, "–✳–" represent the 40 cm results, and "–★–" represents the 55 cm results.

the meeting scenario, a group of participants were sitting closely around the array. In both cases, the array was placed in the center of the room, Fig. 4.29(a) and Fig. 4.29(b) shows the speaker combinations over time in presentation and meeting scenarios, respectively. In the presentation scenario, the presenter is not fixed at one position and there are constant head movements which change the speaker's orientation with respect to the array. Moreover, during the recording, there were small interruptions from the audience with speech overlap. Fig. 4.30(a) presents the accuracy in percent plotted as a CDF versus the error threshold for the presentation scenario. The accuracy drops in comparison to the controlled speaker scenarios for the MPoPi and SRP-PHAT algorithm because of an absence of head orientation and exact DoA information for the presenter at each time frame. The recordings with real speakers pose challenges such as: uncertainty in an actual source position resulting due to involuntary movements of mouth, head, and body by a speaker. There were no close talking or lapel microphones used in the recordings. Therefore, the reference files were generated manually by a listening test using channel 1 of the array. This is somewhat crude but a reasonable solution in absence of a head tracking system. The MPoPi modifications tend to perform better than the SRP-



(a) Presentation (b) Meeting

Figure 4.29: Real-world interaction scenarios including four human speakers: (a) Presentation scenario with one speaker giving a talk or lecture and the rest are the audience creating small interruptions during the talk, (b) Meeting scenario, where speakers first takes turns to talk and later have overlapped speech up to four concurrent speakers during the meeting.

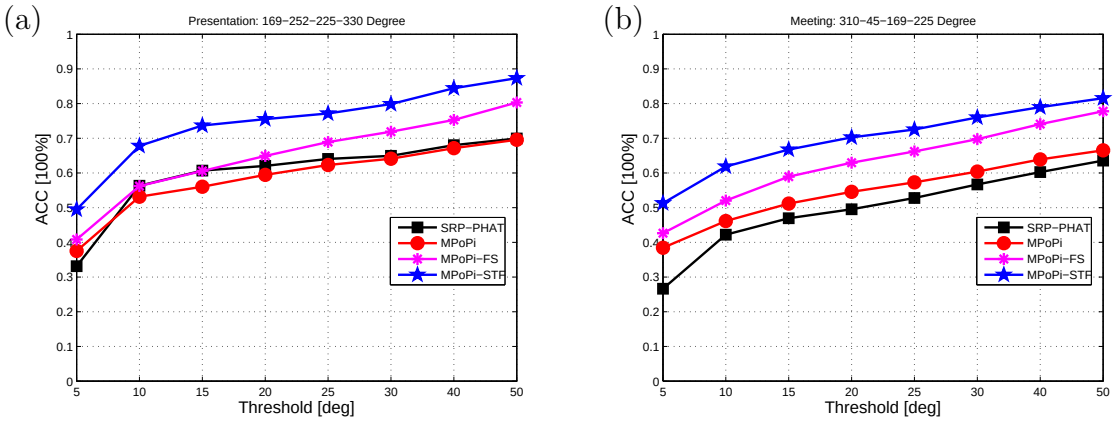


Figure 4.30: Accuracy in percent versus error threshold plotted as a CDF for real-world speaker interaction scenarios: (a) Presentation, and (b) Meeting. The "–■–" represents the SRP-PHAT algorithm, "–●–" represents the MPoPi method, "–✳–" represent the MPoPi-FS and "–★–" represents the MPoPi-STF method.

PHAT in the presentation scenario.

For the meeting scenario, all four speakers actively participated as shown in Fig. 4.29(b), where in the beginning each participant took turns to speak followed by mutual interruptions creating different concurrent speaker combinations. The results are presented in Fig. 4.30(b), the ground truths were estimated the same way as in the presentation scenario. The MPoPi-STF method performs the best and SRP-PHAT generates the lowest score out of all the algorithms.

The modifications to the MPoPi algorithm show consistent performance improvement in both controlled and real speaker experiments. This shows that the pre-grouping and sequential integration of frequency channels gives robust location estimates and improves detection of multiple concurrent speakers.

Mobile Speakers

So far, the speakers are considered restricted to fixed positions, which emulates speaker interaction scenarios in meetings, and office space avoiding speaker mobility. Some recordings for single and multiple speakers were carried out, where all speakers were moving usually facing the array. Fig. 4.31(a) shows the detection results of SRP-PHAT and MPoPi algorithms for a mobile speaker, where a male speaker (Martin) is moving in front of the array keeping constant distance from the array starting from 180° and moving towards 0° in a step-wise manner. Both detection algorithms generate erroneous estimates due to the reverberant environment and the variation in head orientation of a speaker during motion. Fig. 4.31(b) shows the accuracy counts for both algorithms, where MPoPi performs better than SRP-PHAT at smaller error thresholds. The ground truth was generated with the help of visual information: a point and shoot camera was used to capture the movements of the speaker during recordings. The audio and video tracks were synchronized and DoA information is labeled during the active portion of speech. This is a simpler workaround because of the absence of any motion tracking system in the meeting room. Under such scenarios, a likelihood function taking into account the history of observations and a dynamical model for speaker movement yield more robust estimates. Keeping this idea in mind, three likelihood functions are proposed in the particle filtering framework to mitigate this problem. The details of these new methods will be presented in Chapter 5.

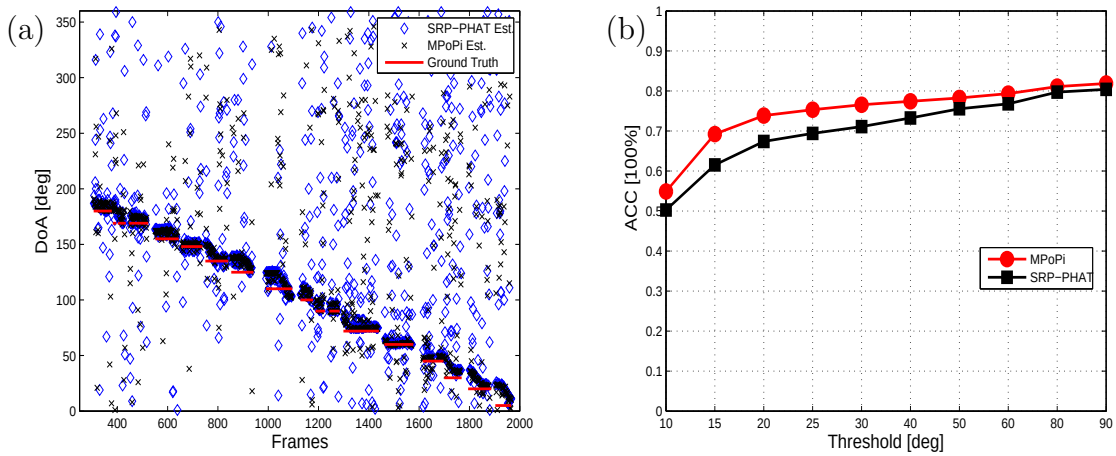
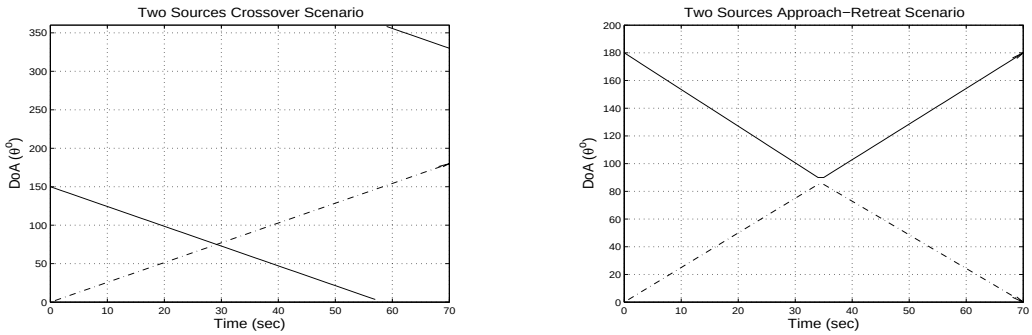


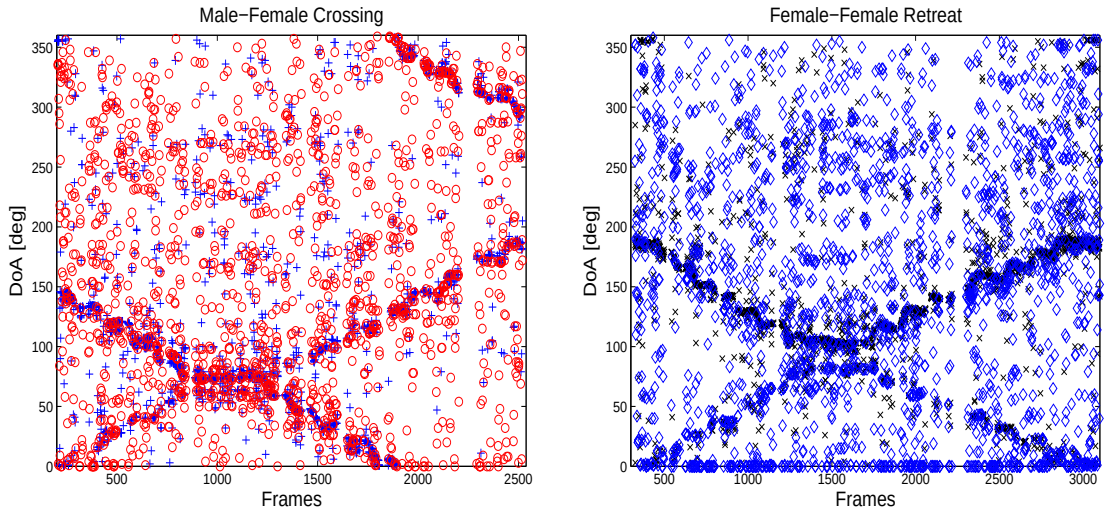
Figure 4.31: A single speaker moving in front of the array (a) Detection results of SRP-PHAT ("◇") and MPoPi("×"), (b) Accuracy counts versus error threshold plotted as a CDF, "–■–" represents SRP-PHAT and "–●–" represents the MPoPi method.

Moreover, a multi-speaker scenario was recorded using two concurrent speakers, where cross-over and approach-and-retreat scenarios were created as shown in Fig. 4.32(a) and Fig. 4.32(b), respectively. In both cases, the speakers start from a given position and move towards each other to create a crossing or approach-retreat scenario. The detection results of MPoPi algorithm for two such cases are presented in Figs. 4.33(a)-(b). Figs. 4.33(a) show the localization result of a male-female (Martin-Anna) crossing scenario, where the male speaker started at 150° and the female speaker started at 0°. The male speaker keeps on moving in a circular motion around the array after approaching the 0° mark. The detection results are shown by markers which are not associated with any particular speaker. The localization result of a female 1-female 2 (Anna-Tania) interaction scenario where both speakers starting at well-separated DoAs, female 1 started at 180° and female 2 started at 0°, both come closer to each other and then move away towards their respective starting points is shown in Fig. 4.33(b). The second female is unstructured because of the strong power ratio mismatch between the two speakers. The first female has much louder voice than the second female. As MPoPi method is an energy based method, it is difficult to localize both speakers concurrently. Even though the MPoPi algorithm makes several errors during the detection, the pattern still can be seen from the plots and the detection accuracy can be improved by using a tracking algorithm.



(a) (b)

Figure 4.32: Mobile speaker scenarios for two speakers over time starting at well-separated DoAs and moving towards each other creating (a) Cross-over and, (b) Approach-and-retreat cases.



(a) (b)

Figure 4.33: Detection Results of MPoPi algorithm for two mobile speakers over time starting at well-separated DoAs and moving towards each other creating (a) Cross-over scenario where male speaker starts at 150° and female speaker starts at 0°. The male speaker keeps on moving in a circular motion around the array after approaching the 0° mark, (b) Approach-and-retreat scenario where the first female speaker starts at 180° and the second female starts at 0°. Both speakers come closer to each other and then move away towards their respective starting points. The position markers only represents the estimated DoAs and are not assigned to any particular speaker in these plots.

4.6 Discussion

The goal of this thesis is to test the proposed ASL algorithms in real environments and compare their performance with the state-of-the-art SRP-PHAT algorithm. There is extensive work carried out in the microphone array community with various kinds of algorithms proposed, which perform well in simulated environments; hence leaving an ambiguity about their performance with real data. Therefore, it is essential to test the methods in actual environments to assess their usefulness for real-world applications. The test set was recorded in a meeting room, where no efforts were made to reduce the noise and multi-path propagation as these two pose the most challenges to the ASL methods (for experimental details, see Chapter 3).

The results for the single speaker scenario are presented in Figs. 4.20(a)-(b). The full-band PoPi algorithm performs the worst out of all algorithms with just 20% accuracy at 1° threshold. There is not much accuracy gained by using the "Cepstrum-PHAT" weighting function for the full-band PoPi algorithm. The MPoPi algorithm gives consistent performance in both cases. The SRP-PHAT gives around 90% accuracy for the speaker placed at 169° but deteriorates to 70% for the 310° case. The problem with the 310° case is that there is a strong reflection at 318°. Therefore, when the error threshold increases to 10° there is an abrupt increase in accuracy score. The MPoPi algorithm does not suffer from this problem, it has an accuracy count of 83% for the 169° case and 80% for the 310° case. The evaluations are further extended to the various kinds of background noise. The results in Fig. 4.21 show comparable performance of SRP-PHAT and MPoPi methods, but these noises did not make much effect on the SNR conditions of the room as the loudness of the speech signal was quite high in comparison to these environmental noises. The accuracy scores for MPoPi are around 80% for all the noisy conditions in both cases. But SRP-PHAT has an inconsistent performance: the accuracy drops to 66% for the 169° "Door" case and then jumps to 83% for the 310° "Door" case. This shift in accuracy scores based on speaker position is observed for all noise cases. The evaluations of algorithms for different SNR recording as discussed in Section 3.7 are presented in Fig. 4.22. Both algorithms deteriorate for SNR $\leq 10\,\mathrm{dB}$, where MPoPi at $-5\,\mathrm{dB}$ gives only 25% of correct estimates within 5° range of the true DoA. The position of the speech source relative to the noise source has little to no effect on MPoPi performance but SRP-PHAT has around 40% shift in accuracy for both cases. The results show that the MPoPi algorithm makes better use of the

arrays' geometry, and it is more consistent in spatial disambiguation in comparison to the SRP-PHAT algorithm.

The results for two concurrent speakers as depicted in Fig. 4.23 includes two new algorithms which have been introduced in Sections 4.3, and 4.4, and which are referred to as MPoPi-FS and MPoPi-STF. The results show that the selective criterion in the form of "spectral grouping" (MPoPi-FS) and "spectral and sequential grouping" (MPoPi-STF) improve the location accuracy over MPoPi and SRP-PHAT methods for two concurrent speakers. MPoPi-STF gives on average 75% correct estimates within an error threshold of 5°, whereas MPoPi-FS gives on average 65% correct estimates. The MPoPi and SRP-PHAT gives on average 59% and 57% correct estimates within 5° of the true source positions. The percentage of correct estimates increases more sharply for MPoPi-STF and MPoPi-FS algorithms as the error threshold is increased to 10°. The plots show that the proposed algorithms are able to localize the sources irrespectively of the speaker setup.

The single speaker noise conditions were also repeated for two concurrent speakers. Fig. 4.24 presents the accuracy scores of all algorithms using the noise recordings of the two speakers case. The MPoPi-STF performs the best out of all algorithms by providing 77% correct estimates within 5° of the true speakers positions. The SRP-PHAT performs the worst and gives an average of 57% accuracy for all noisy conditions.

Further tests carried out in case of three and four concurrent speakers showed similar performance improvement of subband processing based MPoPi methods over original MPoPi and SRP-PHAT algorithms. The results presented in Figs.4.25 and 4.26 showed two different cases of well-separated and closely spaced speakers. The performance pattern of SRP-PHAT is similar; it performs poorly when speakers are well-separated and better for closely placed speakers. This behavior can be explained further when looking at the analytic solution of SRP-PHAT, which is assumed to be a summation of many cosines. When the array is steered towards the target speakers, the cosines interfere constructively; for other positions, the cosines interfere destructively. This behavior is based on the principle of linear superposition in wave-theory. When the speakers are closely spaced, there is coherent superposition near the speakers, but when they are placed at farther locations, the cosines may add up incoherently leading to near-zero values. A similar principle is true for MPoPi decomposition, but the mapping of fundamental frequency with the respective DoA improves the algorithm's spatial disambiguation ability resulting in a consistent

performance independent of source positions as shown by MPoPi-STF and MPoPi-FS methods. On average, the MPoPi-STF has 67% correct estimates within 10° of the true speakers positions for three concurrent speakers scenarios and 68% for the four concurrent speakers scenarios. The SRP-PHAT has on average 56% correct estimates for three speakers and 57% correct estimates for four speakers within 10° of the true speaker positions.

The accuracy results of the MPoPi algorithm for a variable number of microphone pairs and array diameters in case of a single speaker and a two concurrent speaker scenario are shown in Figs. 4.27 and 4.28, respectively. The aperture of the array defines its spatial resolution. The accuracy counts increase with the diameter of the array but fewer numbers of microphone pairs have a detrimental effect on localization accuracy. This problem arises because of spatial aliasing, where there are strong grating lobes present with the main lobe and the array pointing at the desired direction accumulate inputs from other directions as well (this becomes more challenging in multi-path environments). The spatial aliasing can be avoided if sampling is carried out at half of the wavelength, which corresponds to the minimum wavelength of interest. In practice, however, this means that for an array with sensor spacing of 0.55 m, aliasing occurs above 311 Hz. The authors in [94] have addressed this problem and suggested that the classical narrowband aliasing criterion should not be used for broadband signals such as speech. Furthermore, to characterize the time-domain response of a microphone array, the signal's bifrequency spectrum should be considered. Here for larger spacing, more microphones should be used as fewer will not have their cosines add up coherently or incoherently for target and other directions, respectively. To gain good spatial disambiguation with a large aperture array, the minimum number of microphone pairs should be four. For multiple active speakers, the large aperture arrays (0.4-0.55 m) score better than the smaller array with 0.2 m diameter. Therefore, if the application permits, increasing the number of microphones and the diameter improves ASL methods. For the 0.55 m array, the accuracy drops from 80% to 65% for a single speaker when using only three microphone pairs instead of twelve. For two concurrent speakers, the smaller array with 0.2 m diameter requires at least a 15° threshold to bring the accuracy close to the large aperture arrays, whereas the 0.55 m diameter array achieves a similar accuracy at only 5° threshold.

The results for the two real-world scenarios such as presentation and meeting scenarios are depicted in Fig. 4.30. The recordings with real speakers pose more

challenges than the controlled speakers, for instance, the correct labeling of the speaker's position and activity is difficult (without a close-talking microphones and motion tracking system) due to involuntary movements of mouth, head and body. The distant-speech based voice activity detection of concurrent speakers is an open research area with limited performance for smaller time frames. A decrease in performance of all algorithms in comparison to controlled cases can be attributed to the above mentioned problems. Nevertheless, the proposed methods still show an improvement over SRP-PHAT. In the presentation scenario, MPoPi-STF has 70% accurate estimates within 10° of the true speaker positions. In comparison, SRP-PHAT has only 58% accuracy for the same threshold. The MPoPi-STF has 61% accuracy for the meeting scenario within 10° threshold. In comparison, the SRP-PHAT has only 41% accurate estimates.

In the last part of the experiments, the setup was used to evaluate the MPoPi algorithm for moving speakers. The detection and accuracy results for these experiments for a single speaker are presented in Fig. 4.31. This preliminary case study was carried out to see if a moving speaker can be tracked with a circular microphone array as the focus of this thesis has been to improve the location accuracy for the case of static multiple sources in meeting rooms or office space environments. Both MPoPi and SRP-PHAT generates erroneous location estimates. The MPoPi algorithm has 70% accurate estimates within 15° of the true speaker position, whereas SRP-PHAT gives 61% for the same threshold. The results in Fig. 4.33 show the estimates of MPoPi algorithm for two moving speaker scenarios, where the speakers cross-over and approach-and-retreat. The estimation accuracy can be improved by using a tracker taking into account a dynamical model and history of observations with a likelihood function instead of with a direct estimation. The application of sequential Monte Carlo methods to speech signals have become an active area of research in recent years, the recursive Bayesian filter or Particle Filters (PF) are two of such methods. In this thesis, the proposed ASL algorithms are combined with the PF tracking framework to propose new methods, which will be discussed in detail in Chapter 5.

4.7 Conclusions

This chapter presented different ways of combining the periodicity information of the speaker in the location estimation task. The full-band PoPi algorithm failed to pro-

vide robust location estimates in realistic environment. The cepstrum and PHAT weightings marginally improved the PoPi performance for a single speaker. The MPoPi algorithm with an auditory inspired preprocessing gave the largest performance improvement and comparable performance to SRP-PHAT for single speaker scenarios. The addition of a frequency-selective criterion in MPoPi algorithm referred as MPoPi-FS was proposed, which grouped the frequency channels belonging to the same speaker. In this way, a speaker with relatively low energy was emphasized more. The joint position and pitch decomposition on these grouped channels gave robust location estimates for concurrent speaker scenarios. A pitch tracker was integrated at a low algorithm level to create spectro-temporal regions and fragments for PoPi decomposition called as MPoPi-STF. These auditory inspired techniques gave on average 20% more accurate results than SRP-PHAT. Although the performance improvement comes at a cost of high computational complexity (see Appendix B), with the availability of increasing computing resources these days, an implementation of a MPoPi and its variants is possible close to real-time.

Chapter 5

Acoustic Source Tracking

This chapter presents a brief overview of traditional Acoustic Source Localization and Tracking (ASLT) algorithms as well as the basics of the well-known Sequential Monte-Carlo (SMC) methods, also known as particle filters. Thereafter the formulation of the source tracking problem for the given setup is presented. The algorithms proposed in Chapter 4 are combined with the particle filtering framework. Subsequently, a number of solutions based on the Markov Chain Monte-Carlo (MCMC) sampling techniques are presented, which solve the well-known problems faced by particle filters. A joint position and pitch tracking algorithm is also presented in this chapter. In the end, the proposed ASLT algorithms are evaluated for the different acoustic scenarios discussed in Chapter 3.

5.1 Background

In this chapter, the problem of speech source localization is defined in terms of the optimal Bayesian solution. There are a number of techniques available in the literature to recursively estimate this solution. The well-known tracker is the Kalman filter [95], which assumes that the underlying process is linear and Gaussian (including the process and measurement noises in the system). Therefore, it can be parameterized by the mean and covariance. Under such conditions, an analytic solution can be derived by a set of parametric equations to recursively estimate the posterior density providing the current state of the system. These assumptions are not strictly valid in real-world scenarios where the speakers often change their location and there is a significant speech overlap along with short speech utterances

(typically less than a second). In case of source localization using a microphone array, the TDoAs are estimated for various pairs and combined through a certain criterion to find the position of the source. The TDoA of a speech source at position $\boldsymbol{\gamma}_s$ is a non-linear function expressed as

$$\boldsymbol{\mathcal{T}}(\{\boldsymbol{\gamma}_{m1}, \boldsymbol{\gamma}_{m2}\}, \boldsymbol{\gamma}_s) = \frac{||\boldsymbol{\gamma}_s - \boldsymbol{\gamma}_{m1}|| - ||\boldsymbol{\gamma}_s - \boldsymbol{\gamma}_{m2}||}{c}, \tag{5.1}$$

where $\boldsymbol{\gamma}_{m1}$ and $\boldsymbol{\gamma}_{m2}$ are the positions of a pair of microphones in Cartesian coordinates, and c is the speed of sound. The position of sources can be estimated by minimizing the error function $e(\boldsymbol{\gamma}_s)$ given as

$$e(\boldsymbol{\gamma}_s) = \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} [\hat{\tau}_i - \boldsymbol{\mathcal{T}}(\cdot)]^2, \tag{5.2}$$

where $\hat{\tau}_i$ is the observed TDoA for a pair of microphones and σ_i^2 is the error covariance associated with this observation [6]. Because $\boldsymbol{\mathcal{T}}(\cdot)$ is non-linear, different linearization techniques such as a Taylor's series expansion [96] and applying the Extended Kalman Filters (EKF) [97] have been used in literature. However, it is difficult to tune the parameters of these algorithms. In order to avoid the linearization step and to include non-Gaussian measurement noise, the Unscented Kalman Filter (UKF) has been proposed in literature [98, 99]. A recent application of the UKF to the speaker localization problem is presented in [100]. The spontaneous speech produced in realistic environments can be highly dynamic in terms of space such as fast speaker changes and time such as short speech utterances, which makes it a challenging task to apply these techniques in such scenarios.

As an alternative to the conventional tracking approaches, the SMC methods known as Particle Filtering (PF) provide a probabilistic framework to track acoustic sources in a realistic environment. In principle, the PF methods approximate the optimal Bayesian filter by representing probability distributions through a finite set of particles [101, 102]. For a state-space model with a given dynamical model, an observation model, and sampling techniques, the particle filter recursively approximates the filtering distribution of the states. The candidate sources' locations are predicted and measured by carrying out a random search in the defined space and evaluating their respective likelihoods. Applications of particle filters to the single acoustic source localization and tracking can be found in [103, 104]. Although the PF methods have been successfully applied in realistic scenarios, there are still

some open issues to resolve. The conventional particle filtering framework is difficult to extend to spontaneous multi-party speech scenarios. Under such conditions, the PF methods either require multisource models (or multi-modal distributions) or the adaptation of a single-source model to switch between different speakers' positions. In order to resolve the data association problem of PF methods, an accurate estimation of the number of active speakers is required, which is non-trivial in multisource reverberant scenarios. Moreover, complex birth/death rules should be created for rapidly varying numbers of active sources. In this chapter, the particle filtering framework is modified for joint position-pitch tracking of single and multiple concurrent sources. Moreover, solutions to some of the above mentioned problems related to the application of particle filters in such scenarios are presented.

5.2 Particle Filter Based Source Tracking

Particle filters are widely used in practical applications of tracking single and multiple speakers due to their ability in dealing with multimodality, non-linear functions, and non-Gaussian noise. The idea behind the state-space approaches such as particle filters is that there is temporal continuity in peaks arising due to real sources in the observations, whereas the outliers have no temporal consistency. According to [102], the tracking problem can be defined as follows:

Let the evolution of state $\boldsymbol{\Theta}_k$ of a target at time-step k be given by

$$\boldsymbol{\Theta}_k = \boldsymbol{\mathcal{F}}_k(\boldsymbol{\Theta}_{k-1}, \boldsymbol{n}_{k-1}), \tag{5.3}$$

where $\boldsymbol{\mathcal{F}}_k$ can be a non-linear function of the state $\boldsymbol{\Theta}_{k-1}$, and $\boldsymbol{n}_{k-1}$ is an independent identically distributed (i.i.d) process noise sequence. The above equation defines the dynamics of the source and how the states are evolving. There is a measurement process using any source localization algorithm, where the observation $\mathcal{Y}_k$ is available in the form of noisy measurement of the hidden state $\boldsymbol{\Theta}_k$ given as

$$\mathcal{Y}_k = \boldsymbol{\mathcal{T}}_k(\boldsymbol{\Theta}_k, \boldsymbol{v}_k), \tag{5.4}$$

where $\boldsymbol{\mathcal{T}}_k$ is a non-linear function defined in (5.1), and $\boldsymbol{v}_k$ is a possibly non-Gaussian i.i.d measurement noise sequence. The task at hand is to track speech sources with the source state defined as $\boldsymbol{\Theta}_k = [\hat{\varphi}_1, \hat{\varphi}_2, \cdots, \hat{\varphi}_S, N_s]$, where $\hat{\varphi}_s$ is the DoA for

source s and N_s is the total number of sources active at the current time-step k. Let $\mathcal{Y}_{1:k} = [\mathcal{Y}_1 . \cdots , \mathcal{Y}_k]$ denote the concatenation of all measurements up to the time frame k. The aim is then to recursively estimate the posterior filtering distribution $p(\mathbf{\Theta}_k|\mathcal{Y}_{1:k})$ using Bayes' Theorem as follows:

$$p(\mathbf{\Theta}_k|\mathcal{Y}_{1:k-1}) = \int p(\mathbf{\Theta}_k|\mathbf{\Theta}_{k-1}) p(\mathbf{\Theta}_{k-1}|\mathcal{Y}_{1:k-1}) d\mathbf{\Theta}_{k-1}$$
$$p(\mathbf{\Theta}_k|\mathcal{Y}_{1:k}) \propto p(\mathcal{Y}_k|\mathbf{\Theta}_k) p(\mathbf{\Theta}_k|\mathcal{Y}_{1:k-1}). \tag{5.5}$$

The first step is the *prediction step*, which will use the combined dynamical model $p(\mathbf{\Theta}_k|\mathbf{\Theta}_{k-1})$ to propagate the previous posterior $p(\mathbf{\Theta}_{k-1}|\mathcal{Y}_{1:k-1})$ to give the estimate of the predictive distribution $p(\mathbf{\Theta}_k|\mathcal{Y}_{1:k-1})$. The second step is the *update step*, where the likelihood $p(\mathcal{Y}_k|\mathbf{\Theta}_k)$ is combined with the predictive distribution at time-step k.

Particle filters essentially implement the recursions in (5.5) by using a large set of discrete samples, so called particles, with associated discrete probability masses commonly known as weights w_k. The symbolic representation of the particle filter is shown in Fig. 5.1, where the set of particles with the corresponding weights approximates the true posterior distribution by following the steps outlined above. In the current application, the measurements $\mathcal{Y}_k$ are the marginalized MPoPi decompositions over the F_0 dimension given as

$$\mathcal{Y}_k = \sum_{F_0} \rho_k(\varphi_0, F_0). \tag{5.6}$$

Fig. 5.2 shows the PoPi plane of two concurrent speakers (a male at 310° and a female at 142°). The resulting measurement function $\mathcal{Y}_k$ is presented in the bottom plot.

A general particle filter algorithm is presented in Algorithm 5.1. The process is started with a set of N_p state samples $\{\mathbf{\Theta}_0^{(i)}\}_{i=1}^{N_p}$ with corresponding likelihood weights uniformly distributed in the state-space. At time-step $k-1$, the combination of state samples and likelihood weights approximates the distribution $p(\Theta_{k-1}|\mathcal{Y}_{1:k-1})$. The new samples are generated by propagating the state samples from the previous iteration conditioned on a transition equation. With the new set of particles and the availability of the observation $\mathcal{Y}_k$, the likelihood of each weight is updated using the likelihood function $p(\mathcal{Y}_k|\mathbf{\Theta}_k)$. After the normalization, an additional step is carried out to deal with one inherent problem of *degeneracy* (where the variance of particles

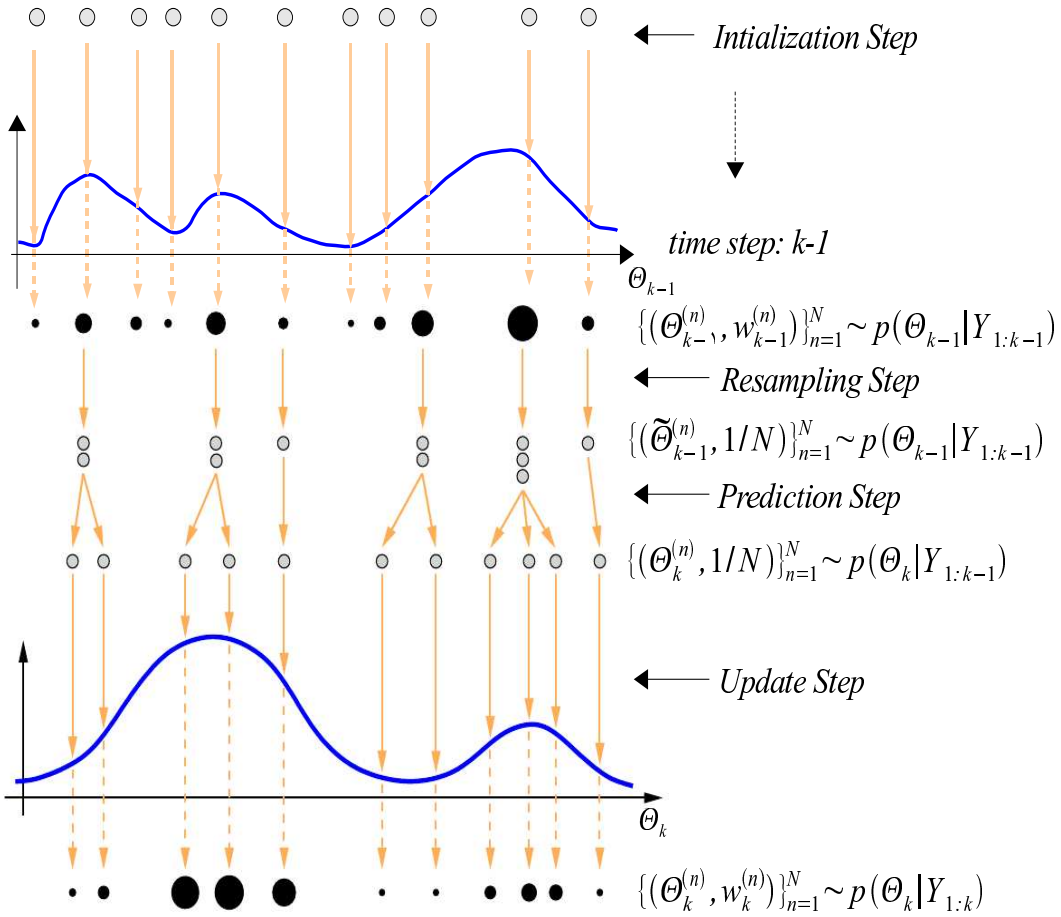


Figure 5.1: Symbolic representation of particle filtering starting from an initialization step where the circles represent particles with uniform weights. One iteration from time-step $k-1$ to k is shown here, the size of particles denotes the corresponding likelihood weight. The set of particles and weights approximates a specific PDF. The steps are outlined on the right-hand side (as defined by [10]).

increases over time and after some iterations all but one particles have negligible weight). One possible solution presented in literature is to apply a resampling step [102]. The basic idea of the resampling process is to eliminate particles that have smaller weights and to increase the number of particles with larger weights. There are many different resampling techniques proposed in literature [105]. The Systematic resampling is used to decrease the particle variance. As the resampling is done on a discrete representation of $p(\mathbf{\Theta}_k|\mathcal{Y}_{1:k})$ given by

$$p(\mathbf{\Theta}_k|\mathcal{Y}_{1:k}) \approx \sum_{i=1}^{N_p} w_k^{(i)} \delta(\mathbf{\Theta}_k - \mathbf{\Theta}_k^{(i)}), \tag{5.7}$$

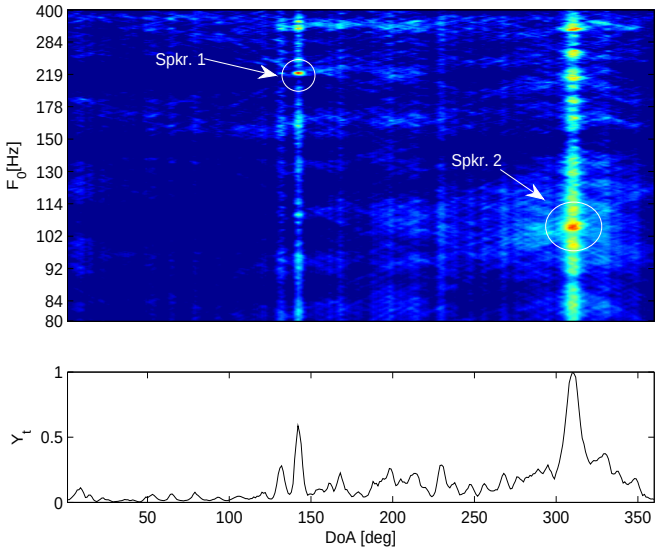


Figure 5.2: Position-Pitch plane of two concurrent speakers, a male at 310°, and a female at 142°. The bottom plot displays the measurement function $\mathcal{Y}_k$, that is the MPoPi decomposition marginalized over the F_0 dimension.

such that $\mathrm{p}(\boldsymbol{\Theta}_k^{i*} = \boldsymbol{\Theta}_k^j) = w_k^j$, thus resulting in an i.i.d sample from the density. Therefore, all the weights are reset to $w^{(i)} = \frac{1}{N_p}$. One side effect of the resampling process is the problem of sample impoverishment, where particles with large weights are selected many times. According to [102], when the process noise is small, there will be a repetition of particles at the same sample points and all the particles will converge to a single point after a few iterations. The authors in [102] suggested to use MCMC sampling and Sequential Importance Resampling (SIR) techniques to avoid this problem. Moreover, the choice of importance density is crucial for the appropriate use of particle filters for any given problem. In the following sections, the different parts of the particle filtering framework are discussed and new algorithms are proposed to track multiple speakers.

Algorithm 5.1 A generic particle filter algorithm for source tracking.

Initialization of Particle Filters:

The particle filters are randomly distributed in the state-space, $\{\mathbf{\Theta}_0^{(i)}\}_{i=1}^{N_p}$ with associated uniform weights $\{{w_0}^{(i)} = 1/N_p\}_{i=1}^{N_p}$, where i is the particle index and N_p is the number of particles.

Iteration:

1. Predict the new set of particles $\{\mathbf{\Theta}_k^{(i)}\}_{i=1}^{N_p}$ by propagating the previous set $\{\mathbf{\Theta}_{k-1}^{(i)}\}_{i=1}^{N_p}$ according to the dynamics model.
2. Transform the signals received at the microphones into localization measurements $\mathcal{Y}_k$ given as:
$$\mathcal{Y}_k = \mathcal{T}(\mathbf{Y}_k)$$
3. On the basis of the observation $\mathcal{Y}_k$, form the likelihood function $p(\mathcal{Y}_k|\mathbf{\Theta}_k)$.
4. The new weights corresponding to the particles are assigned to as:
$${w_k}^{(i)} = p(\mathcal{Y}_k|{\mathbf{\Theta}_k}^{(i)}),$$
and normalized to obtain $\sum_{i=1}^{Np} {w_k}^{(i)} = 1$.
5. Resample the particles by multiplying the particles $\{\mathbf{\Theta}_k^{(i)}\}_{i=1}^{N_p}$ with higher weights and deleting the ones with smaller weights to avoid the *degeneracy* problem using a suitable resampling method. Set the weights to a uniform value.

Location Estimation:

The final estimate for the location of sources can be calculated by clustering the particles' set or a histogram measure using a predefined threshold.

Table 5.1: Selected values for dynamic model parameters.

Model	σ_{RW}	β_Θ [Hz]	$\bar{v}_\Theta$ [degree/s]
RW	1°	-	-
LM	-	0.2	0.1

5.3 Dynamic Model

Various source dynamic models have been presented in literature [101] for the problem of localizing of an active source. We have used two kinds of dynamic models for the source tracking algorithms. One is based on random walk and the other is the Langevin model [103].

1. The Random Walk (RW) with variance σ^2 is given by

$$\Theta_k = \Theta_{k-1} + \sigma_{\mathrm{RW}} \cdot u_k, \tag{5.8}$$

   where $\sigma_{\mathrm{RW}} = 1^\circ$ and $u_k \backsim \mathcal{N}(0,1)$ is a Gaussian variable with zero mean and unit variance.

2. The Langevin Model (LM) is a well-known process to characterize stochastic motion. It assumes that the source motion is independent and identically distributed in each Cartesian co-ordinate. To track the DoAs of active sources, the model is transformed for angular co-ordinates resulting in following equations

$$\dot{\Theta}_k = a_\Theta \dot{\Theta}_{k-1} + b_\Theta u_\Theta, \tag{5.9a}$$
$$\Theta_k = \Theta_{k-1} + T_U \dot{\Theta}_k, \tag{5.9b}$$

   where $u_\Theta \backsim \mathcal{N}(0,1)$, and T_U is the time interval between two consecutive updates of the state vector, and

$$a_\Theta = \exp(-\beta_\Theta T_U),$$
$$b_\Theta = \bar{v}_\Theta \sqrt{1 - a_\Theta^2},$$

   with $\bar{v}_\Theta$ is the steady-state velocity parameter, and β_Θ is the rate constant.

The values of the parameters used for the experiments are shown in Table 5.1.

5.4 Multiband Position-Pitch Estimation Based Likelihood Function

There were some properties defined in [104] regarding the use of any localization function as a likelihood function. The author suggested that the likelihood function should be chosen to reflect that the peaks in the localization function belong to likely source positions. Additionally, it should also reflect that there might be no peak belonging to any source locations such as when no source is active, or when spurious peaks are present due to background noise and sensor calibration errors.

Keeping in view the above mentioned criterion, a pseudo-likelihood function is derived from the M-PoPi algorithm output based on the formulation of [104] given as

$$F(\mathcal{Y}_k, \boldsymbol{\Theta}) = \max\{\mathcal{P}_k(\hat{\varphi}_{\boldsymbol{\Theta}}), \xi_0\}^r, \tag{5.10}$$

where $\hat{\varphi}_{\boldsymbol{\Theta}}$ is the localization parameter corresponding to the state, $\xi_0 \geq 0$, and $r \in \mathbb{R}^+$. The use of r is to sharpen the peaks in the localization function. The presence of ξ_0 ensures that the function is non-negative and includes the case where no peak in the localization function belongs to the true source [104]. The likelihood function used to assign new weights to the particle filters is then calculated as, $p(\mathcal{Y}_k|{\boldsymbol{\Theta}_k}^{(i)}) = F(\mathcal{Y}_k, \boldsymbol{\Theta}_k^{(i)})$. In the experiments, $\xi_0 = 0$ and $r = 2$ are used.

5.5 Voice Activity Detection

In order to switch the likelihood model, a simple speech/non-speech classification. The particle filters should stop updating during the silence periods and spread the particles randomly in the state-space. This allows the particle filters to follow the speaker and to avoid the track loss problem when the speaker reappears at different/same position. To classify a frame of data as speech or silence, an energy based voice activity detection method is used. In case of a microphone array with multiple channels, the first channel was used for the classification. A quantile-based method is used to estimate the adaptive thresholds related to the noise level with a 15 msec /200 msec conversation rule [106]. The energy-based VAD algorithm needs at least 50 frames or 1 sec of speech signal to make the speaker activity decision. This is a limitation for the framewise ASL method. Therefore a framewise VAD method is

presented, which makes the decision by defining a threshold on the maximum value of the position-pitch matrix. The threshold is learned from the recorded data. Even though the previous VAD method is more robust than the latter, in most of the experiments, the framewise VAD method is used except if stated otherwise.

5.6 Particle Filter with Integrated Voice Activity Detection

The tracking algorithm continuously updates the source locations during the silence periods occurring in the middle of speech signals, as if the source was still active. This makes it necessary to include voice activity detection in the tracking framework. The idea introduced in [107] is used to integrate the voice activity detector in the likelihood function such as

$$p(\mathcal{Y}|\boldsymbol{\Theta}) = q_0 \cdot \mathcal{U}_{\mathcal{D}}(\hat{\boldsymbol{\varphi}}_{\boldsymbol{\Theta}}) + \mathcal{Z} \cdot (1 - q_0) \cdot \mathcal{P}(\hat{\boldsymbol{\varphi}}_{\boldsymbol{\Theta}}), \tag{5.11}$$

where the subscript k has been omitted for the sake of simplicity. The value q_0 represents the hypothesis that the measurement originates from clutter, and $1 - q_0$ indicates that the measurement originates from a true source. $\hat{\boldsymbol{\varphi}}_{\boldsymbol{\Theta}}$ corresponds to the state vector $\boldsymbol{\Theta}$ and $\mathcal{U}_{\mathcal{D}}$ is the uniform Probability Density Function (PDF) over the considered state-space $\mathcal{D}$. The second term in (5.11) is the pseudo-likelihood function $\mathcal{P}(\cdot)$ derived from the MPoPi algorithm as explained in Section 5.2 and $\mathcal{Z}$ is the normalization constant.

During the silence periods, this integration allows the tracking algorithm to put more emphasis on the considered dynamics' model in spreading the particles, while at the same time reducing the importance of MPoPi observations because no useful information is present when the speaker is inactive. This allows the particles to spread in the state-space, and when the speaker reappears at a different/same location, the algorithm can resume to track successfully.

5.7 Proposed Modification

So far the discussion is based on the bootstrap method, where the particles are drawn without considering the current observations. The performance of the bootstrap method degrades when the number of acoustic sources and disturbance levels

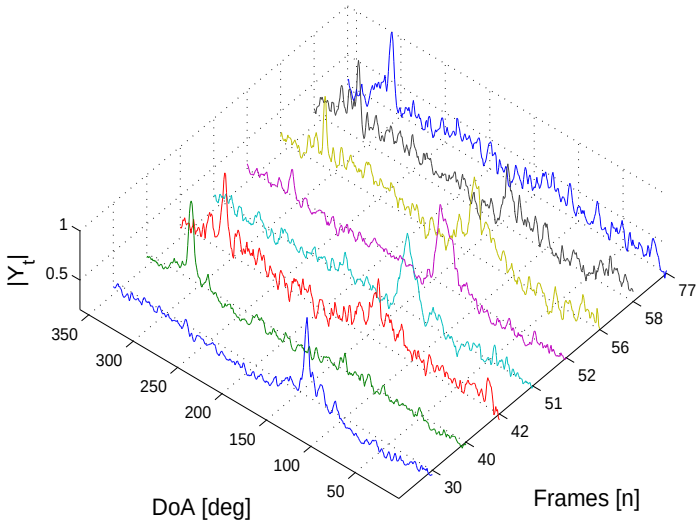


Figure 5.3: The evolution of the measurement function $\mathcal{Y}_k$ over time in case of two concurrent sources. A male speaker is located at 310° and a female speaker is located at 142°. The measurement function shows speakers presence at varying instances.

increases. Fig. 5.3 shows the time evolution of the measurement functions in case of two concurrent sources (a male speaker at 310° and a female speaker at 142°). Since both speakers are present at varying instances, the tracker is required to reinitialize and sample from the majority of the state-space to track a varying number of sources appearing at different locations over time. Therefore, in this thesis, a scheme is proposed, which combines the bootstrap and importance sampling techniques.

The bootstrap method was proposed in [108]. It gained popularity for being conceptually simple and it leads to straightforward practical implementations. The main advantage of this method is that it locks on to speaker position and does not get affected by surrounding factors. The bootstrap method has one inherent problem: it does not account for the current observations when propagating the particles in the current time step. Therefore, it only generates particles defined by the previous time step. This property has a major drawback if the speaker becomes silent, or when he takes a longer pause and/or changes location. This results in track loss and there is no mechanism defined in the method to recover from this problem. The use of VAD information in the likelihood function resolves this problem but no new samples are drawn from the state-space to track new speakers entering the acoustic scene.

As discussed earlier, the bootstrap or the particle filter is a recursive Bayesian filter by Monte Carlo simulations. It exhibits the posterior density function by a set of random samples with associated weights, which asymptotically converges to the posterior PDF. In real-life Bayesian filtering problems, the posterior density $p(\Theta_k|\mathcal{Y}_{1:k})$ is not available, and it is not possible to directly draw samples from it. As an alternative, the importance sampling approach can be used [102]. It relies on the principle that in the absence of a posterior density function, the samples can be drawn from the importance sampling function $q(\cdot)$. For this distribution, the samples $\Theta_k^{(i)}$ where i is the particle index can be drawn from the conditional PDF $\Theta_k^{(i)} \backsim q(\Theta_k|\mathcal{Y}_{1:k})$. The corresponding weights can be drawn in the following way [102]:

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathcal{Y}_k|\Theta_k^{(i)}) p(\Theta_k^{(i)}|\Theta_{k-1}^{(i)})}{q(\Theta_k^{(i)}|\Theta_{k-1}^{(i)}, \mathcal{Y}_k)}. \tag{5.12}$$

An appropriate choice of the sampling function is essential to reduce the problem of degeneracy, where the variance of particles increases over time and after a few iterations all but one particle will end up having negligible weights. In the literature, the optimal importance sampling function has shown to be [102]:

$$q_{\text{opt}}(\Theta_k|\Theta_{k-1}, \mathcal{Y}_k) = p(\Theta_k|\Theta_{k-1}, \mathcal{Y}_k). \tag{5.13}$$

In this case, the importance density takes the previous state Θ_{k-1} and the current observation $\mathcal{Y}_k$ into account.

Choice of importance sampling function

In theory, any density with some assumptions can be used as an importance sampling function. The main objective of such a function is that during the iterative update of the particles, some are redirected to regions of state-space with high posterior likelihoods. In the literature, the choice of sampling function varies depending on the given problem. Therefore, the new importance sampling functions based on MPoPi functions introduced in Chapter 4 are derived here. The MPoPi function can be used directly as an importance sampling function

$$q_{\text{MPoPi}}(\Theta_k|\mathcal{Y}_{1:k}) \triangleq \mathcal{P}, \tag{5.14}$$

where $\mathcal{P}(\cdot)$ is the MPoPi function marginalized over the F_0 dimension as defined in (5.6). It does not take previous observations into account but still presents an improvement over the bootstrap method, where the prior density $p(\Theta_k|\Theta_{k-1})$ is used for importance sampling without taking any observations into account.

It is important to reiterate that q_{MPoPi} is not a PDF but a pseudo-density function that is not normalized. To draw importance samples from $\mathcal{P}(\cdot)$, a strategy similar to [10] is applied by defining a threshold function $\Phi_{\text{MPoPi}}(\boldsymbol{\varphi})$, which is non-nil only for regions of state-space, where $\mathcal{P}(\boldsymbol{\varphi})$ is above a certain threshold level ξ.

$$\tilde{q}_{\text{MPoPi}}(\Theta_k|\mathcal{Y}_{1:k}) \propto \Phi_{\text{MPoPi}}(\boldsymbol{\varphi}) = \begin{cases} 1, & \text{if } \mathcal{P}(\boldsymbol{\varphi}) \geq \xi \\ 0, & \text{otherwise.} \end{cases} \tag{5.15}$$

The DoA vector $\boldsymbol{\varphi} \triangleq \boldsymbol{\varphi}_k$ corresponds to the current state of the variable Θ_k. To draw importance samples, $\Phi_{\text{MPoPi}}(\boldsymbol{\varphi})$ is normalized as a uniform distribution. The value of ξ is set to be thirty-five percent of the maximum value of $\mathcal{P}(\cdot)$. It is important to note that the function $\tilde{q}_{\text{MPoPi}}(\Theta_k|\mathcal{Y}_{1:k})$ is only used to draw state samples. For the computation of the likelihood function $q_{\text{MPoPi}}(\Theta_k|\mathcal{Y}_{1:k}) = \mathcal{P}$ is used instead.

Algorithm 5.2 presents an algorithm which combines bootstrap and importance sampling approaches via the effective sample size parameter N_{eff}. This parameter either selects bootstrap for quickly locking on to speaker position and robustness in challenging environments [103] or importance sampling for reducing the degeneracy problem. The process starts with those particles which are randomly distributed with uniform weights. At every iteration, the effective sample size is computed. The decision to either choose importance sampling or simple bootstrap is based on a pre-selected threshold (the value is fixed to $N_p/2$, where N_p is the total number of particles). This is a tunable parameter and needs to be chosen beforehand.

Another solution to the above mentioned problems is to combine both approaches, where a certain percentage of particles is drawn from the importance function and the rest of the particles are drawn through bootstrapping. This approach is usually helpful in cases where the threshold does not yield good results. The modification of the approach is shown in Algorithm 5.3. In this case, the bootstrap process is carried out first, then a certain percentage of particles is deleted and re-populated with samples and their corresponding likelihood weights using the importance density

function. The parameter J_p defines re-population factor for particles. This provides a surveillance mechanism to search for new sources while keeping track of active ones. As a new source appears at a different location, the particles will be able to track the new source without the problem of track loss.

In this thesis, the bootstrap method is combined with the importance sampling technique to create algorithms which can recover from track loss and have the property of re-initialization, which is necessary for concurrent speaker scenarios. The proposed algorithms are combined with the three ASL algorithms presented in Chapter 4 such as: MPoPi, MPoPi-FS and MPoPi-STF methods. For each ASL method, the PoPi decomposition matrix is used as an importance sampling function. The best fitting PF algorithm is used for every ASL method. The details of the extensions are presented in the experimental section of this chapter.

The next section presents a novel joint position and pitch tracking algorithm. Moreover, a new particle filter-based dynamical model for pitch tracking is presented. The particle filtering framework is redefined for this problem, where state-vector includes both the pitch and the DoA estimates.

Algorithm 5.2 Proposed SIS based algorithm

Initialization of Particle Filters:

The particle filters are randomly distributed in the state-space, $\{\mathbf{\Theta}_0^{(i)}\}_{i=1}^{N_p}$ with associated uniform weights $\{{w_0}^{(i)} = 1/N_p\}_{i=1}^{N_p}$, where i is the particle index and N_p is the number of particles.

Iteration:

1. Compute $\mathrm{N_{eff}}$ given as:
$$N_{\text{eff}} = \frac{1}{\sum_{i=1}^{N_p} (w_k^{(i)})^2}$$

2. If $N_{\text{eff}} < \dfrac{N_p}{2}$ **Importance sampling:**

   a) Sample the particles $\Theta_k^{(i)} \backsim \tilde{q}(\Theta_k^{(i)}|\mathcal{Y}_{1:k})$
   b) Compute the importance weights:
$$\tilde{w}_k^{(i)} = p(\mathcal{Y}_k|\Theta_k^{(i)}) \cdot \min\left\{\frac{p(\Theta_k^{(i)}|\mathcal{Y}_{1:k-1})}{q(\Theta_k^{(i)}|\mathcal{Y}_{1:k})}, 1\right\}$$

3. Otherwise **Bootstrap:**

   a) Sample the particles $\Theta_k^{(i)} \backsim p(\Theta_k^{(i)}|\mathcal{Y}_{1:k-1})$
   b) Compute the likelihood weights
$${w_k}^{(i)} \propto {w_{k-1}}^{(i)} \cdot p(\mathcal{Y}_k|\mathbf{\Theta}_k^{(i)})$$

4. Finally, normalize to obtain $\sum_{i=1}^{N_p} {w_k}^{(i)} = 1$.

Location Estimation:

The final estimate for the location of sources can be calculated by clustering the particles' set or a histogram measure using a predefined threshold.

Algorithm 5.3 Proposed SIR algorithm

Initialization of Particle Filters:

The particle filters are randomly distributed in the state-space, $\{\boldsymbol{\Theta}_0^{(i)}\}_{i=1}^{N_p}$ with associated uniform weights $\{{w_0}^{(i)} = 1/N_p\}_{i=1}^{N_p}$, where i is the particle index and N_p is the number of particles.

Iteration:

1. **Bootstrap:**
2. Sample the particles $\Theta_k^{(i)} \backsim p(\Theta_k^{(i)}|\mathcal{Y}_{1:k-1})$
3. Compute the likelihood weights

$$ {w_k}^{(i)} \propto {w_{k-1}}^{(i)} \cdot p(\mathcal{Y}_k|\boldsymbol{\Theta}_k^{(i)}) $$

4. Select a subset of particles J_p:

$$ \hat{w} = \text{SORT}(w), \text{ in descending order and store indices} $$
$$ \hat{\Theta}_k = \Theta_k(\text{ind}(\hat{w})), \text{ ind}(\hat{w}) \text{ are weight indices} $$

5. **Importance sampling:**
6. Sample the remaining $(N_p - J_p)$ particles $\tilde{\Theta}_k^{(i)} \backsim \tilde{q}(\tilde{\Theta}_k^{(i)}|\mathcal{Y}_{1:k})$
7. Compute the corresponding importance weights:

$$ \tilde{w}_k^{(i)} = p(\mathcal{Y}_k|\tilde{\Theta}_k^{(i)}) \cdot \min\left\{ \frac{p(\tilde{\Theta}_k^{(i)}|\mathcal{Y}_{1:k-1})}{q(\tilde{\Theta}_k^{(i)}|\mathcal{Y}_{1:k})}, 1 \right\} $$

8. Create the new set of particles and weights:

$$ \check{\Theta}_k = [\hat{\Theta}_k; \tilde{\Theta}_k] $$

$$ \check{w}_k = [\hat{w}_k; \tilde{w}_k] $$

9. Finally, normalize to obtain $\sum_{i=1}^{N_p} \check{w}_k^{(i)} = 1$.

Location Estimation:

The final estimate for the location of sources can be calculated by clustering the particles' set or a histogram measure using a predefined threshold.

Importance Function using MPoPi

Unlike Fig. 5.2, where the position-pitch plane is marginalized and used as a measurement function $\mathcal{Y}_k$ for state estimation. This section presents an extension of the particle filters for joint position and pitch estimation.

A new dynamic model based on the random walk is defined for the pitch tracking, where $\mathbf{F}$ is the state-vector of pitch. At every iteration k, it is determined as

$$F_k = F_{k-1} + u_k, \tag{5.16}$$

where u_k models the pitch changes in consecutive time frames, which is based on Laplacian distribution given as

$$p(\Delta_p) = \frac{1}{2\lambda_p} \exp\left(-\frac{|\Delta_p - l_p|}{\lambda_p}\right), \tag{5.17}$$

where Δ_p represents pitch period changes, and l_p and λ_p are distribution parameters. The parameters were estimated with a trial and error approach and fixed at $\lambda_p = 2.5$ lag steps and $l_p = 0.4$ lag steps. The authors in [90] cited the *declination phenomenon* as the reason to use positive l_p. The *declination phenomenon* is associated with the natural speech, where it suggests that the speech utterance has a tendency for pitch periods to increase (and the respective pitch frequencies to decrease). Moreover, the pitch and DoA dynamics are modeled independently. The source state is redefined as $\boldsymbol{\alpha}_k = [\hat{\varphi}, \hat{F}_0]_k$, where $\hat{\varphi}$ is the DoA, and $\hat{F}_0$ is the pitch estimate of an active source at the current time step k. 2D particles are used here, where each dimension contains state samples representing the position and pitch with corresponding combined likelihood weights. For the joint tracking of position and pitch, the MPoPi function containing position and pitch relations of the active speakers is used as an importance sampling function given as

$$q_{\text{MPoPi}}(\alpha_k|\mathcal{Y}_{1:k}) \triangleq \mathcal{P}. \tag{5.18}$$

In order to draw samples from the importance function, similar definition (5.15) is used

$$\tilde{q}_{\text{MPoPi}}(\alpha_k|\mathcal{Y}_{1:k}) \propto \Phi_{\text{MPoPi}}(\boldsymbol{\vartheta}) = \begin{cases} 1, & \text{if } \mathcal{P}(\boldsymbol{\vartheta}) \geq \xi \\ 0, & \text{otherwise.} \end{cases} \tag{5.19}$$

Here again the vector $\boldsymbol{\vartheta} \triangleq \boldsymbol{\vartheta}_k$ corresponds to the current state of the variable $\boldsymbol{\alpha}_k$ containing the joint DoA and pitch estimates of the active sources at time instant k. To draw importance samples, $\Phi_{\text{MPoPi}}(\boldsymbol{\vartheta})$ is normalized as a 2D uniform distribution. The value of ξ is set to be thirty-five percent of the maximum value of $\mathcal{P}(\cdot)$. It is important to re-iterate that the function $\tilde{q}_{\text{MPoPi}}(\alpha_k|\mathcal{Y}_{1:k})$ is only used to draw state samples. For the computation of the likelihood function $q_{\text{MPoPi}}(\alpha_k|\mathcal{Y}_{1:k}) = \mathcal{P}$ is used instead. Moreover, the likelihood function in (5.11) for the given problem is re-defined as

$$p(\mathcal{Y}_k|\boldsymbol{\alpha}_k) = q_0 \cdot \mathcal{U}_{\mathcal{D}} + \mathcal{Z} \cdot (1 - q_0) \cdot \mathcal{P}_k(\cdot), \tag{5.20}$$

where $\mathcal{P}$ is the 2D pseudo-likelihood function.

Figs. 5.4(a)-(i) illustrate the process of joint position and pitch tracking using the PoPi plane. For this case, a 3 sec long speech utterance of a female speaker placed at 328° is used. In the beginning, the particles are randomly spread in the position-pitch plane. As the speaker becomes active, all the particles migrate to the active position-pitch region. Algorithm 5.3 is selected to estimate the pitch and DoA estimates in this case. The value of J_p is selected to be ninety percent of the total number of particles, which is fixed at $N_p = 100$. During each iteration, a small number of particles is drawn from the current observation and combined with the rest of the particles propagated from previous iteration. This technique helps in recovering from track loss when the speaker takes a short pause.

The estimated pitch and DoA contours for the above example are shown in Fig. 5.5. The top plot shows the estimation results of the pitch and the bottom plot shows the DoA estimates. The pitch contour is more dynamic than the position contour (which is fixed in this case). The proposed method makes pitch estimation errors at the beginning of the utterance. Once the particles are converged to the true pitch value, the joint position and pitch tracking algorithm is able to successfully track both DoA and pitch of the speaker. This shows that the dynamic model selected for the pitch tracking is suitable for this application. The VAD is made on the learned threshold for the PoPi plane. Therefore, it is difficult to distinguish the voiced and unvoiced frames based on this information. Hence, during the unvoiced segments, the method continues to track both pitch and DoA.

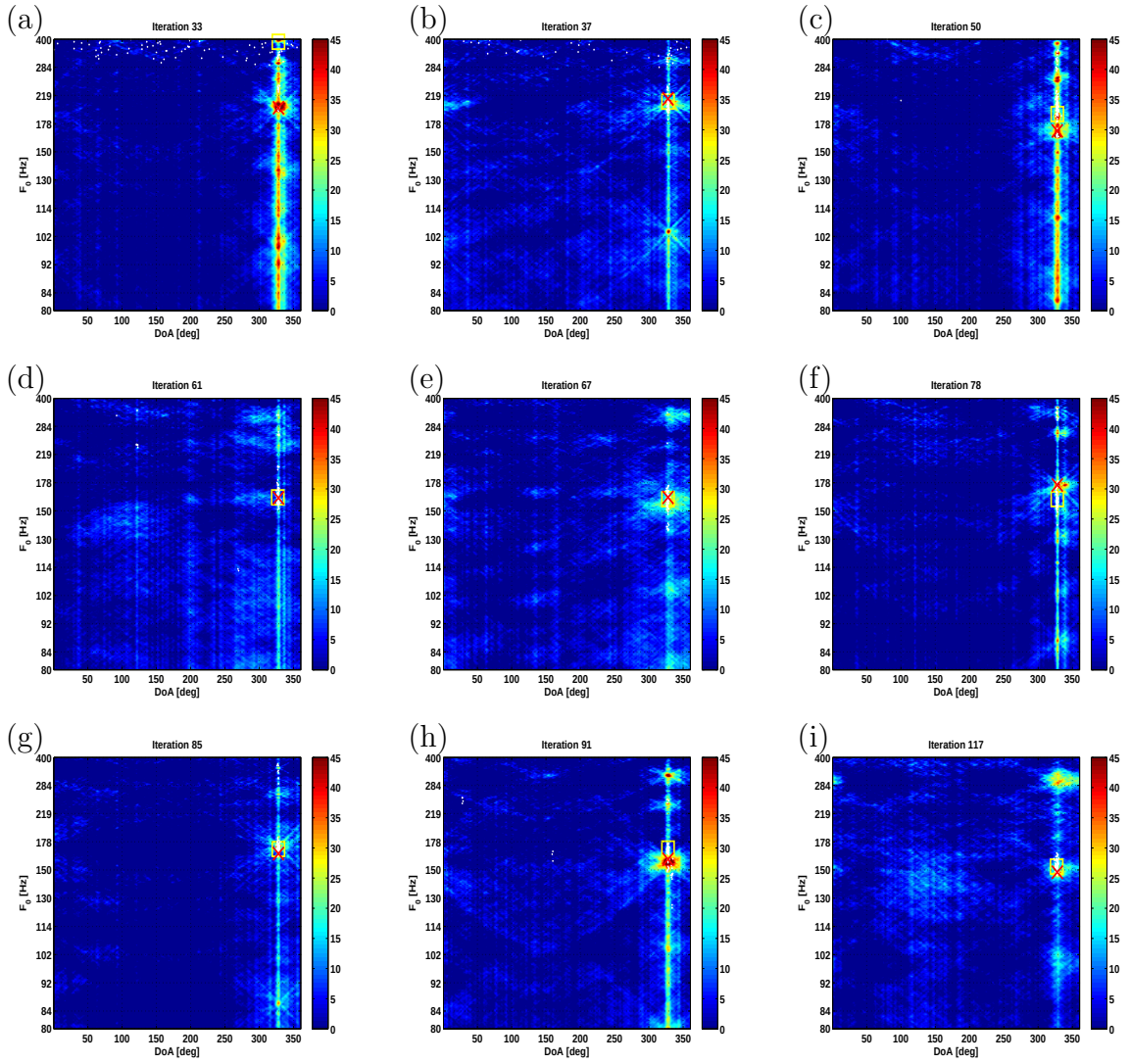


Figure 5.4: Joint position-pitch tracking using particle filters of a female source placed at 328°.(a)-(i) shows the PoPi planes at different instances, "□" shows the particle filter-based estimate, "×" shows the ground truth position and pitch value and the particles are presented by "white dots" on the plane.

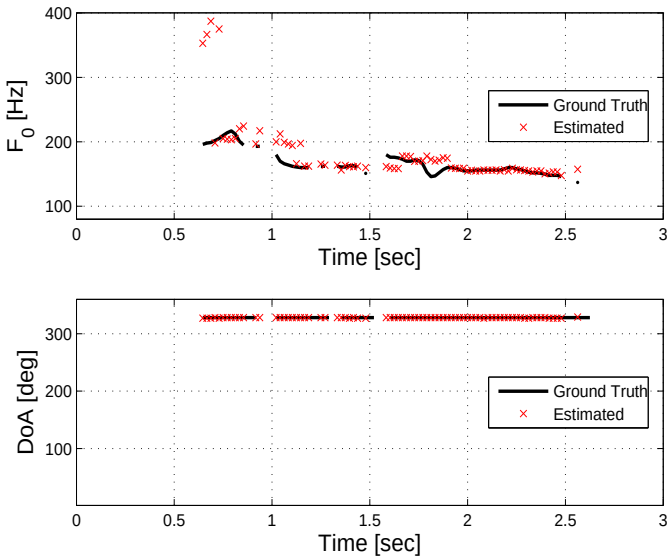


Figure 5.5: The position-pitch tracking of a female speaker placed at 328°. The top plot shows the pitch and the bottom plot shows the DoA estimation over time. The "solid-line" denotes the ground truth value and "×" shows the estimated values for pitch and DoA.

5.8 Experimental Evaluations

In this section, the experiments presented in Chapter 4 are repeated by adding the results of the ASLT algorithms proposed in this chapter. The total number of particles used in all experiments is fixed at $N_p = 100$. Due to the stochastic nature of the algorithms, the results for all the particle filter-based methods are averaged over ten trials. The accuracy counts and the Cumulative Distribution Function (CDF) presented in Section 3.7 are used to evaluate the algorithms performance for different scenarios.

5.8.1 Controlled Experiments

Here the term controlled means that the recordings were made with loudspeakers, the reasons for the setup were discussed in Section 4.5.1.

Single Static Speaker

The single speaker scenarios are first presented, where the MPoPi and SRP-PHAT based likelihood functions are used in the general particle filtering framework pre-

Table 5.2: Localization accuracy in percent for different speaker positions. The bold values represent the best performance achieved out of all 4 algorithms for every case.

DoA	72°	115°	169°	252°	270°	288°	310°
MPoPi	70.30	80.52	81.96	66.68	69.84	77.26	79.23
MPoPi-PF	**92.39**	**98.39**	**98.00**	**96.75**	**94.10**	**95.56**	**96.67**
SRP-PHAT	55.59	87.70	92.24	59.28	77.80	85.60	70.16
SRP-PHAT-PF	24.89	61.64	85.35	42.86	53.75	64.75	58.13

sented in Algorithm 5.1 referred to as "MPoPi-PF" and "SRP-PHAT-PF" methods, respectively. Table 5.2 shows the accuracy counts for all algorithms at different speaker DoAs. The error threshold is fixed at 5° and the results are averaged over six speakers (three males and three females). The MPoPi based particle filtering algorithm outperforms the other algorithms in every case. The main difference between the particle filtering algorithms and the detection based algorithms is the prior knowledge of the number of speakers. In the particle filtering scheme, a histogram is computed at every iteration to detect number of bins exceeding a predefined threshold, which is fixed at 20% of the N_p in all experiments. The consistent performance of the MPoPi algorithm makes it a suitable candidate to formulate the likelihood function, whereas the SRP-PHAT algorithm suffers from spurious peaks and gives poor performance. The choice of source dynamic models for particle filter algorithms in the static speaker case did not make much difference. Therefore, the random walk model defined in (5.8) is used here. The measurements for the MPoPi algorithm are marginalized over the F_0 dimension (example shown in Fig. 5.2). This process is carried out for all the cases in the remainder of the thesis except if stated otherwise. Unlike Section 4.5.1, the CDF is not plotted for this case because the MPoPi-PF method works well for the given error threshold.

Background Noise

Different background noise scenarios discussed in Section 4.5.1 are used to generate accuracy scores for the particle filter-based algorithms. Table 5.3 presents the accuracy counts for a single speaker at 169° under different acoustic conditions, where the results are averaged over six speakers with the error threshold fixed at 5°. The MPoPi-PF algorithm successfully tracks the speaker in all conditions, whereas the

Table 5.3: Localization accuracy in percent versus different background noise of a single speaker placed at 169°, where bold values represent the best performance achieved out of all 4 algorithms for every case.

Noise Type	Beamer	Window	Door	Machine
MPoPi	79.87	77.76	74.49	82.85
MPoPi-PF	**97.52**	**82.05**	**96.67**	**98.00**
SRP-PHAT	73.50	76.22	66.19	92.24
SRP-PHAT-PF	60.36	41.43	88.46	85.35

Table 5.4: Localization accuracy in percent versus different background noise of a single speaker placed at 310°, where bold values represent the best performance achieved out of all 4 algorithms for every case.

Noise Type	Beamer	Window	Door	Machine
MPoPi	82.16	86.68	83.55	81.71
MPoPi-PF	**90.20**	**93.89**	**97.14**	**96.67**
SRP-PHAT	83.14	85.74	82.46	70.16
SRP-PHAT-PF	41.59	39.15	83.95	58.13

SRP-PHAT-PF algorithm gives inconsistent results. The environmental noise due to the open window affects both algorithms more severely than other noisy conditions. Table 5.4 presents the results for the 310° case. The MPoPi-PF method again shows the best results in every case. The estimates of the MPoPi algorithm become more robust by inclusion of a tracker, whereas the SRP-PHAT algorithm benefits more by inclusion of source number information. The SRP-PHAT function has strong peaks at wrong positions, which yields erroneous results for the particle filtering method. Therefore, the SRP-PHAT function does not seem suitable for the pseudo-likelihood function.

The SNR recordings presented in Section 4.5.1 are used to evaluate the tracking algorithms performance. In these recordings, a "noise" loudspeaker emitting white noise signal was placed on the floor at the same distance from the array as the "speech" source. Two sets of recordings with SNR ranging from $-5\,\mathrm{dB}$ to $20\,\mathrm{dB}$ were recorded at two different "speech" source positions. In the first case, the speech source was placed at 169° (closer to the noise source), and in the second case at 310° (further away from the noise source). The results for both cases are

Table 5.5: Localization accuracy in percent versus different SNR conditions for a single speaker placed at 169°, where bold values represent the best performance achieved out of all 4 algorithms for every case.

SNR	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
MPoPi	25.22	38.20	52.98	66.50	75.46	75.89
MPoPi-PF	**60.06**	**73.39**	**77.92**	**85.05**	86.50	**86.03**
SRP-PHAT	50.83	63.79	72.48	78.79	84.85	85.20
SRP-PHAT-PF	35.70	42.28	65.07	81.87	**90.73**	89.06

Table 5.6: Localization accuracy in percent versus different SNR conditions for a single speaker placed at 310°, where bold values represent the best performance achieved out of all 4 algorithms for every case.

SNR	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
MPoPi	16.85	34.31	48.31	58.72	66.31	70.87
MPoPi-PF	**21.46**	**58.92**	**67.69**	**76.40**	**83.79**	**83.25**
SRP-PHAT	6.27	23.13	36.54	55.54	64.99	70.23
SRP-PHAT-PF	7.67	15.50	23.20	43.70	67.31	76.78

presented in Table 5.5 and Table 5.6, respectively. All the algorithms fail to provide robust estimated for ≤ 0 dB scenarios. The MPoPi-PF algorithm gives the most consistent performance out of the four algorithms. This setup is different from artificially adding the noise to the speech sources. The noise field in this case is not diffusive and has spatial presence. The poor performance of SRP-PHAT algorithms validates their known problem in high reverberation and low SNR scenarios. The MPoPi methods suffer as well but yield better results than SRP-PHAT methods. In general, SNR ≥ 25 dB is measured in the meeting room for the other scenarios (i.e., excluding the spatial noise source).

Multiple Speaker Scenarios

So far the algorithms are evaluated for single speaker scenarios under various acoustic conditions. In this section, the methods are tested for multi-speaker scenarios. Here the term multi-speaker means two kinds of cases, one is the turn-taking case where the speakers are talking without any speech overlap. The other case is

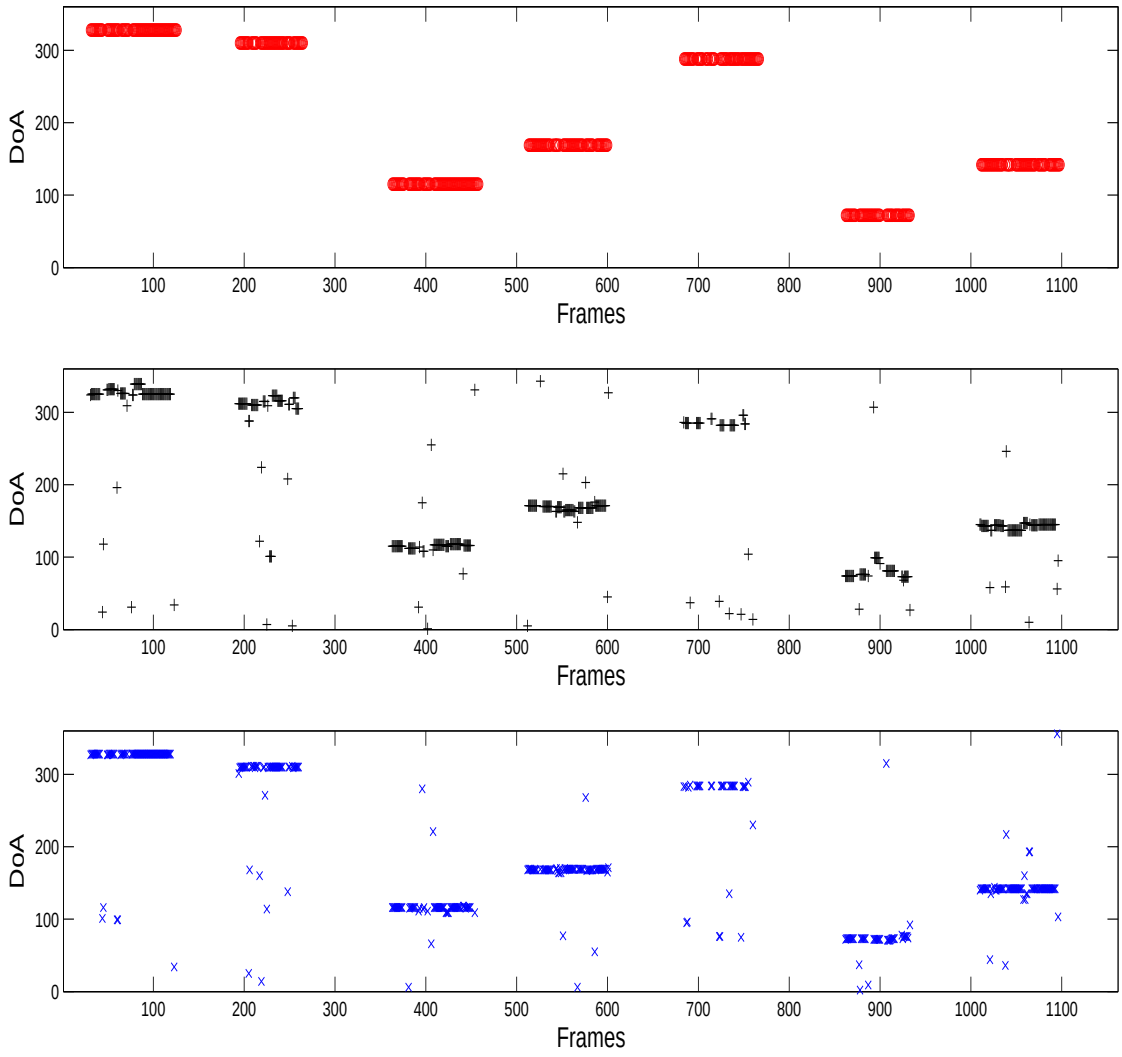


Figure 5.6: The position tracking of a turn taking scenario: The top plot shows the actual DoA positions, the middle plot shows the tracking results of the bootstrap method and the bottom plot shows the results of the combination of bootstrap and importance sampling methods. The inclusion of VAD information helps in re-initialization ability for both methods. The bootstrap method has the difficulty of staying at one DoA position, whereas the importance sampling based method is more consistent in DoA estimation of active speaker. The outliers in both methods are occurring due to voicing errors in the algorithm.

the concurrent speaker scenario, where multiple speakers are talking at the same time. For the turn-taking scenario, the results are only presented for the MPoPi algorithm. Fig. 5.6 shows the tracking results of MPoPi based bootstrap and importance sampling methods. The top plot shows the ground truth values of the speakers' positions. The middle plot shows the results for the bootstrap method (Algorithm 5.1) using (5.11) as the pseudo-likelihood function. The bottom plot shows the result of the combined bootstrap-importance sampling method (Algorithm 5.3) using the same likelihood function. The use of VAD information in both algorithms improves robustness in tracking a new source at different position. This is achieved by spreading the particles randomly in the state-space during the silent segments. When the speaker becomes active again at the new position, some particles will be in the vicinity of that position. This helps in recovering from the track loss problem of the bootstrap method. Without VAD, the tracker remains stuck at the previous position and is unable to track the new speaker. Algorithm 5.3 makes use of the importance sampling step introduced in Section 5.7, where it adds a certain percentage of particles sampled from the importance sampling functions with the previously propagated particles. The proposed method produces more consistent estimates than the bootstrap method and it is able to track the change in speaker position as well. The value of J_p defined in Algorithm 5.3 is set to ninety percent of the total number of particles. Some wrong DoA estimates during the speech utterances are due to voicing errors made by the speaker during short speech pauses.

In Section 5.7, the pitch information was added in the tracking framework. The modification of the likelihood function resulted in joint position and pitch tracking of a single source (see, Fig. 5.4). In this section, the MPoPi based importance sampling method (Algorithm 5.3) is tested for the multiple speaker turn-taking scenario. Fig. 5.7 illustrates the results for the joint pitch and position tracking for the turn-taking scenario with similar parameters as presented above. The proposed algorithm is able to track both pitch and DoA contours. The fundamental frequency contour is more dynamic than the speaker's position (which is fixed in this case). The algorithm makes errors in the beginning of each speaker-turn when the particles are not yet converged to the true position-pitch pair. Due to absence of voiced/unvoiced information, the tracker is unable to accurately track the pitch contour. The choice of dynamic model for pitch tracking seems appropriate but the spectral smearing of the speech signal in a reverberant environment makes the task of pitch estimation

difficult. For these reasons, the task of pitch estimation is not carried out further in this work. Rather than estimating the true pitch value, the inclusion of pitch information in DoA estimation task proves beneficial and yields robust results in realistic environments.

Multiple Concurrent Speakers

This section presents the same results of multiple concurrent scenarios using a similar setup discussed in Section 4.5.1. Four more algorithms are added to the SRP-PHAT and MPoPi methods discussed before. Two of these algorithms are the "MPoPi-FS" and "MPoPi-STF" methods presented in Section 4.3 and Section 4.4, respectively. These algorithms are included in the particle filtering framework presented in Algorithm 5.2. The resulting algorithms are referred as "MPoPi-FS-PF" and "MPoPi-STF-PF".

The MPoPi-FS and MPoPi-STF method are based on auditory pre-processing of the MPoPi algorithm using a gammatone filterbank. In the MPoPi-FS algorithm, the frequency channels are grouped based on the pitch values of the concurrent speakers. The PoPi decomposition is then carried out for all the sub-groups. The particle filters are divided into an equal number of all these sub-groups. In case of two concurrent speakers, the particles are divided into two sub-groups (50 particles for each set). There is no temporal continuity in these segments and the speaker positions are varying in these sub-groups. Therefore, Algorithm 5.2 is a suitable choice in this case as it only allows to resample when the effective size of the particles decreases below a given threshold (fixed at $N_p/2$). The second sub-group does not carry information at each time step. In case of no spectral grouping, the method reverts back to the MPoPi algorithm. Therefore, using Algorithm 5.3 in this case is more suitable. The value of J_p for Algorithm 5.3 is fixed at 40% of the total number of particles. Similar process is used for the MPoPi-STF method. As for the MPoPi-STF case, there is both spectral and temporal grouping but the temporal grouping is based on the pitch information. Therefore, the particle filters provide an additional grouping in terms of source dynamics. The MPoPi and SRP-PHAT based particle filtering methods are using Algorithm 5.3 with similar value of J_p used for MPoPi-FS and MPoPi-STF methods.

Figs. 5.8(a)-(g) present the results of accuracy counts plotted as CDF for two concurrent speakers (averaged over nine speaker combinations discussed in Section 4.5.1) starting from closely spaced speakers to oppositely placed speakers. The

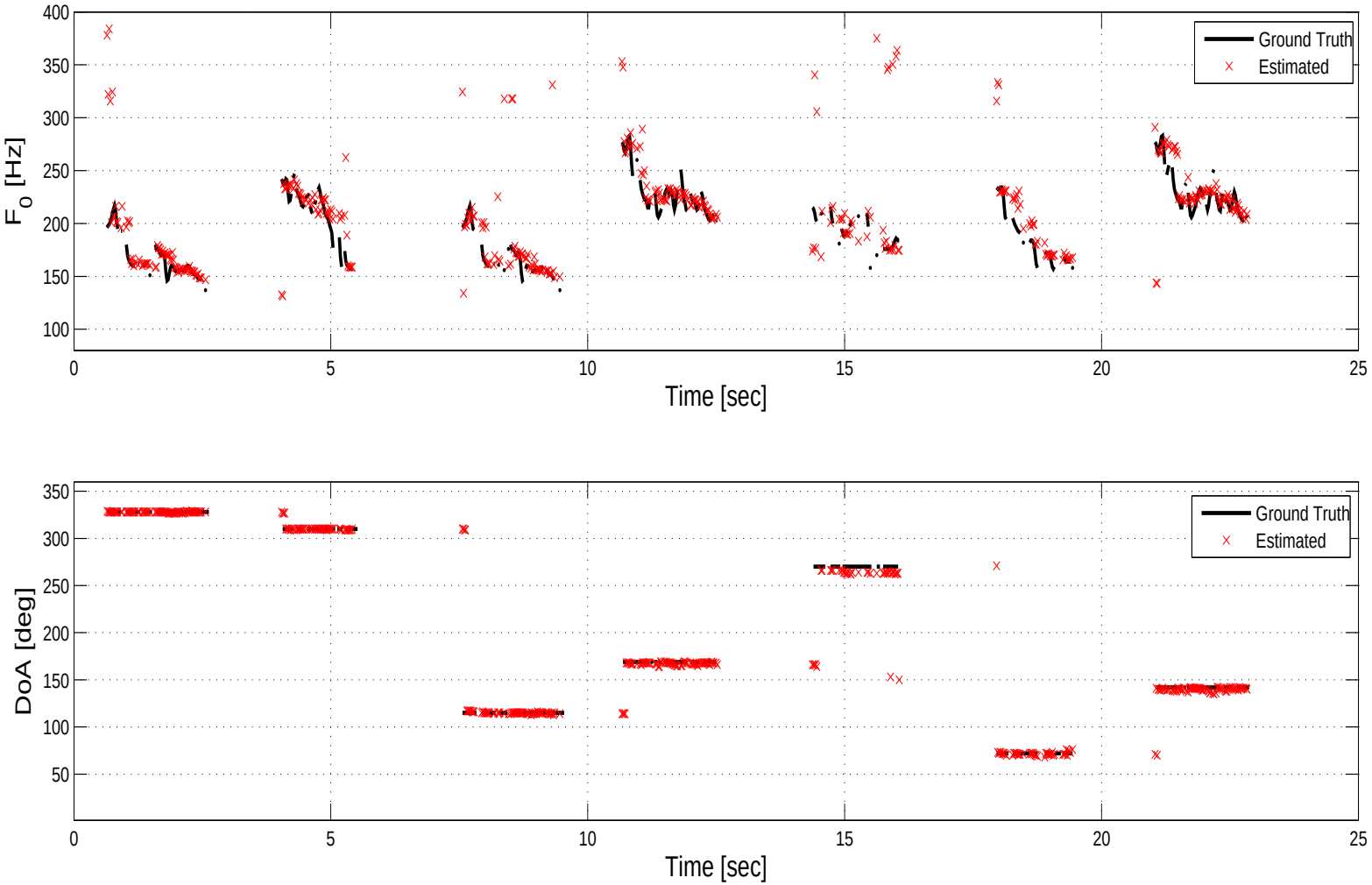


Figure 5.7: The position and pitch tracking of a turn taking scenario. The top plot shows the pitch tracking, the bottom plot shows the results of the combination of bootstrap and importance sampling methods for DoA estimation. A small number of estimation errors occurred during the beginning of each iterations. The algorithm is able to converge to the true pitch and DoA values after few iterations.

Table 5.7: Localization accuracy in percent versus different noisy conditions of two concurrent speakers placed at 142° and 310°, where bold values represent the best performance achieved out of all algorithms for every case.

Noise Type	Beamer	Window	Door	WN 0 dB	WN 5 dB	WN 10 dB
MPoPi	67.88	68.18	65.89	50.37	59.23	63.05
MPoPi-PF	67.98	68.54	65.97	52.59	60.28	63.99
SRP-PHAT	61.23	63.91	59.81	46.86	52.39	56.81
SRP-PHAT-PF	55.44	57.42	53.16	40.71	47.69	49.94
MPoPi-FS	73.44	72.71	71.51	56.39	63.92	68.43
MPoPi-FS-PF	78.32	78.89	**79.47**	**70.79**	**75.72**	77.19
MPoPi-STF	**80.20**	**80.21**	78.38	64.69	72.08	77.15
MPoPi-STF-PF	78.58	78.77	77.81	69.63	73.57	**78.63**

error threshold is varied from 1° to 40°. The MPoPi-FS-PF method works well in all speaker interaction scenarios, whereas the MPoPi-STF does not benefit much from the particle filtering algorithms and gives consistent results without the use of tracking. The results show that the addition of particle filters to the spectro-temporal fragment based MPoPi method is not complimentary, because a low-level tracking is already present in the MPoPi-STF method, which is based on the pitch evidence rather than the position information. The SRP-PHAT algorithms performed the worst out of all the algorithms. The superior performance of MPoPi-FS-PF shows that the temporal integration of spectral regions based on position evidence is more consistent than the temporal integrations based on the pitch evidence as in case of MPoPi-STF.

The noise conditions presented for the two concurrent speakers in Section 4.5.1 are repeated for the evaluation of the tracking algorithms. Table 5.7 shows the accuracy counts for all the above mentioned algorithms. The case where the speakers were placed at 142° and 310° is used here. The error threshold is fixed to 5° and the counts are averaged over all nine speaker combinations. The MPoPi-FS based particle filtering method gives the best results with marginal difference to the MPoPi-STF method in all cases.

The proposed ASLT methods are also tested for three concurrent speaker scenarios. Figs. 5.9(a)-(b) present the results of accuracy counts plotted as a CDF for all algorithms in far placed and closely spaced cases. These results are averaged over

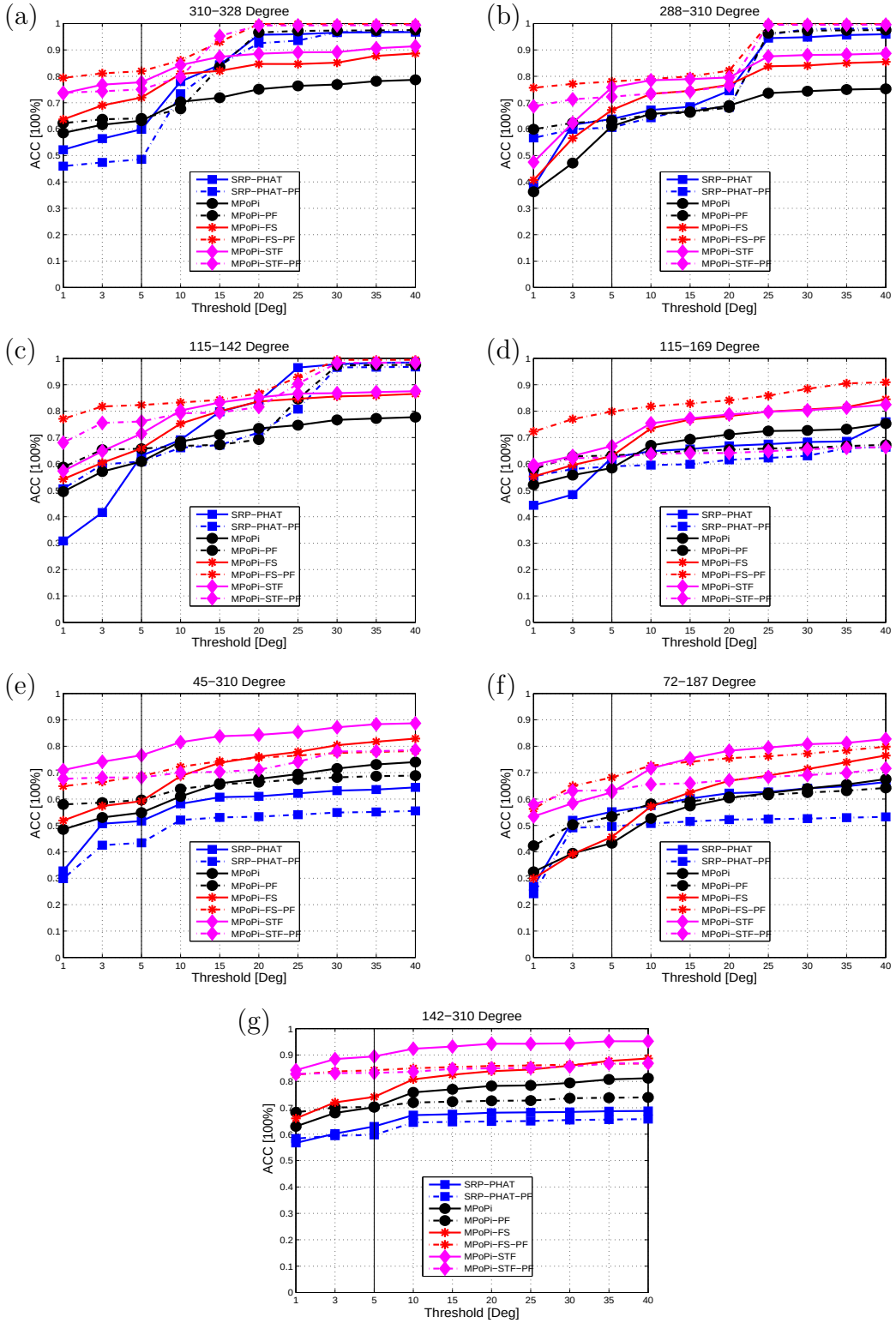


Figure 5.8: Accuracy counts versus the error threshold plotted as a CDF for two concurrent speakers at different speaker positions (a)-(g) starting from closely spaced going up to oppositely placed, where "–■–" represents the SRP-PHAT algorithm, "–●–" represents the MPoPi method, "–✳–" represent the MPoPi-FS and "–◆–" represents the MPoPi-STF method. The dashed lines with similar legend denotes the corresponding particle filtering algorithms for each method.

eight speaker interactions scenarios as discussed in Section 4.5.1. The error threshold is increased from 5° to 90°. Both auditory inspired techniques outperform the SRP-PHAT algorithms. The evidence of correct number of sources becomes much weaker here and the tracking algorithms are unable to localize speakers accurately. The increase in variance of particles helps in localizing closely spaced speakers as shown in Fig. 5.9(b). The prior knowledge about the total number of sources becomes more important as the number of speakers increases. As for the ASL algorithms, the correct peaks are present but the likelihood functions are unable to estimate the correct number of sources at a framewise level. Therefore, a more sophisticated mechanism is required for the source number estimation task. This problem is beyond the scope of the current work. Similarly, Figs. 5.10(a)-(b) shows the accuracy results versus the error threshold for four concurrent speaker case. The results are averaged over five speaker interaction. The worst performance comes from the SRP-PHAT algorithm, while the MPoPi-FS and MPoPi-STF methods are performing the best.

5.8.2 Real Speakers Experiments

The presentation and meeting scenarios for the real speakers are discussed in Section 4.5.2. Similar setups are used to test the ASLT algorithms presented in this chapter. Fig. 5.11(a) shows the results for the presentation scenario and Fig. 5.11(b) shows the results for the meeting scenario. The accuracy counts are plotted as a CDF which are computed for all algorithms for the error threshold varying from 5° to 50°. The speaker interactions for both cases are presented in Fig. 4.29. The MPoPi-FS and MPoPi-STF based tracking algorithms give the best results for both scenarios.

Mobile Speakers

Fig. 5.12(a) shows the tracking results of the SRP-PHAT and MPoPi algorithms for a single mobile speaker. The details of the recordings were presented in Section 4.5.2. The SRP-PHAT tracking algorithm generates more erroneous estimates than the MPoPi based tracking algorithm. Fig. 5.12(b) shows the results of the accuracy counts versus the error threshold plotted as a CDF by varying from 10° to 90°. The CDFs are computed for MPoPi and SRP-PHAT based speaker detection and

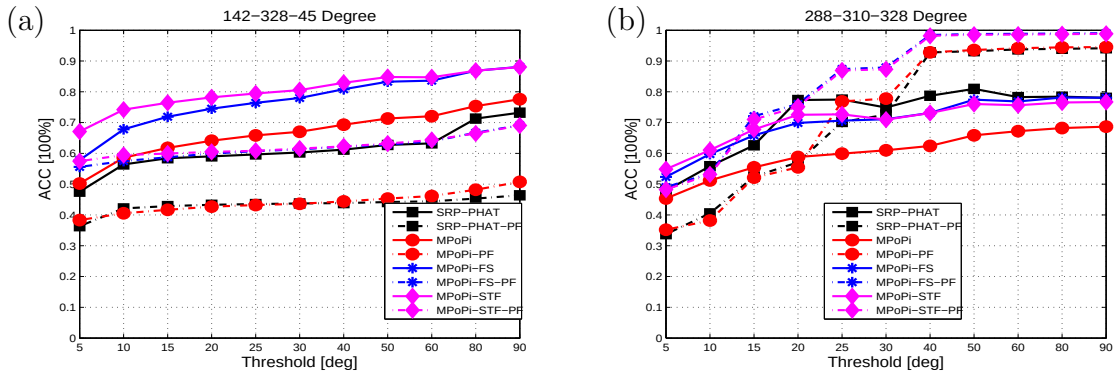


Figure 5.9: Accuracy counts versus error threshold plotted as a CDF for three concurrent speakers at (a) 142-328-45 degrees, (b) 270-288-310 degrees, where "–■–" represents the SRP-PHAT algorithm, "–●–" represents the MPoPi method, "–✳–" represent the MPoPi-FS and "–◆–" represents the MPoPi-STF method. The dashed lines with similar legend denotes the corresponding particle filtering algorithms for each method.

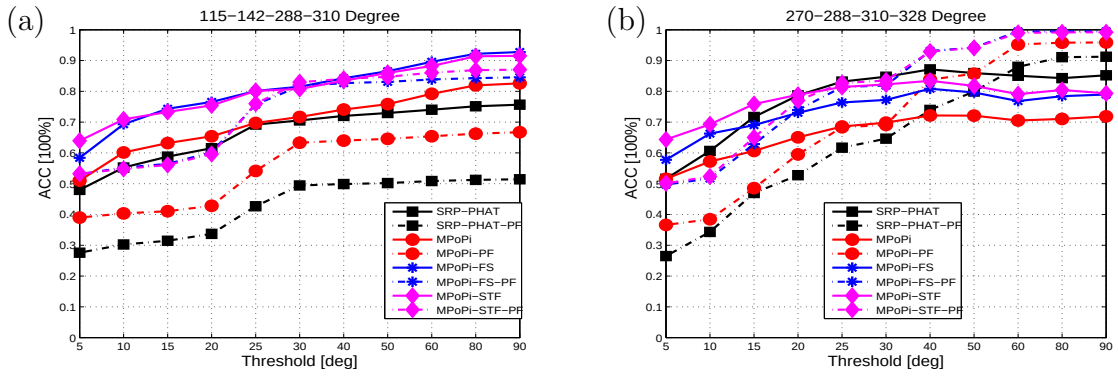


Figure 5.10: Accuracy counts versus error threshold plotted as a CDF for four concurrent speakers at (a) 115-142-288-310 degrees, (b) 270-288-310-328 degrees, where "–■–" represents the SRP-PHAT algorithm, "–●–" represents the MPoPi method, "–✳–" represent the MPoPi-FS and "–◆–" represents the MPoPi-STF method. The dashed lines with similar legend denotes the corresponding particle filtering algorithms for each method.

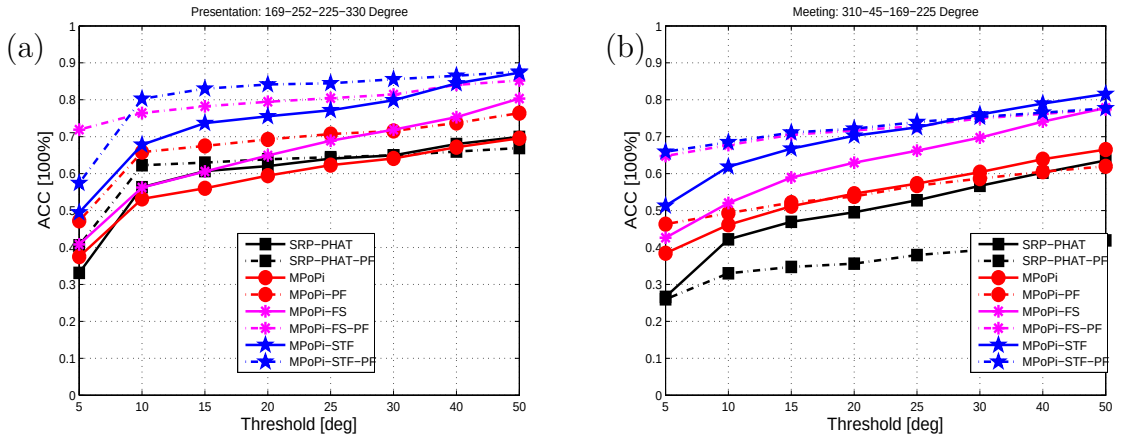


Figure 5.11: Accuracy counts versus error threshold plotted as a CDF for real-world speaker interaction scenarios (a) Presentation style, and (b) Meeting style. The "–■–" represents the SRP-PHAT algorithm, "–●–" represents the MPoPi method, "–✳–" represent the MPoPi-FS and "–★–" represents the MPoPi-STF method. The dashed lines with similar legend denotes the corresponding particle filtering algorithms for each method.

tracking algorithms, where MPoPi-PF performs the best out of all algorithms in this realistic scenario.

5.9 Discussion

The ASLT algorithms proposed in this chapter were tested under a wide range of scenarios using the multi-channel recordings made in a regular meeting room. For the controlled setups, the experiments for the single speaker case highlight the most commonly occurring scenario in an office space or a meeting room. The results of the MPoPi based particle filtering algorithm gives on average 96% accurate estimates within 5° error threshold for a whole range of speaker positions. To obtain speaker independent results for each speaker position, the accuracy counts are averaged over multiple speakers (six in the single speaker case). The conventional MPoPi algorithm has just 75% correct estimates within 5° of the true speaker DoA. The pseudo-likelihood function defined in (5.11) for the tracking framework works really well for the MPoPi algorithm. As depicted in Table 5.2, the SRP-PHAT likelihood function suffers from strong reflections causing anomalies in the location estimates. The SRP-PHAT tracker gets stuck at local maxima and is unable to track the true

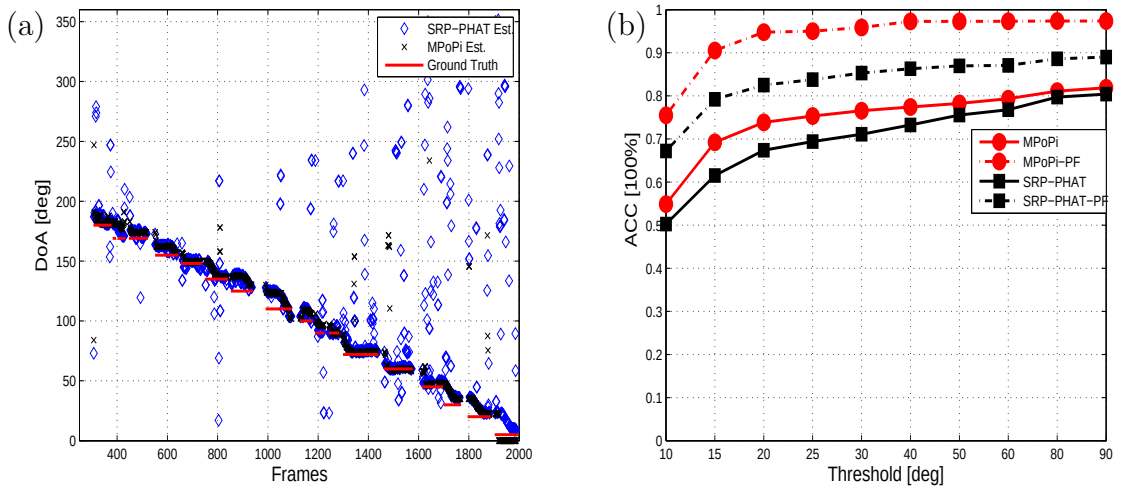


Figure 5.12: A single speaker moving in front of the array (a) Detection results of particle filters using SRP-PHAT ("◇") and MPoPi("×") method, (b) Accuracy counts versus error threshold, "–■–" represents SRP-PHAT and "–●–" represents the MPoPi method. The dashed lines with similar legend denotes the corresponding particle filtering algorithms for each method.

peak in the likelihood function belonging to the actual speaker. On average SRP-PHAT tracker has just 56% accurate estimates for the same threshold. This is a 20% drop in accuracy from the 76% accurate estimates achieved by conventional SRP-PHAT. This shows that the thresholding of SRP-PHAT response to estimate speaker's DoA as it is done in the conventional method gives better results than using the complete response. Another critical information is the *a priori* knowledge of total number of active speakers. Therefore, the MPoPi algorithm's consistent performance makes it suitable for use in practice.

The experiments for the single speaker case for different background noise yield similar results. These results presented for the two cases in Table 5.3 and Table 5.4 show the robustness of the MPoPi-PF method in a wide variety of acoustic conditions by giving on average 95% correct estimates within 5° of true speaker's DoA. The SRP-PHAT-PF method is unable to give consistent performance and gives on average 62% accurate DoA estimates within a 5° threshold. The likelihood functions based on the SRP-PHAT output functions suffer from presence of virtual sources at multiple positions. It is difficult to track the true source position as the particle filtering algorithm gets stuck at the local maxima. A hard-decision method using the information about the total number of sources works better in such scenarios

as evident from the results of the conventional SRP-PHAT algorithm. The SNR experiments repeated for the two background noise cases as shown in Table 5.5 and Table 5.6 follow a similar trend. The accuracy of both methods decreases significantly for SNR $\leq$ 0 dB. On average MPoPi-PF has 72% accurate estimates and SRP-PHAT-PF has 53% accurate estimates for a 5° threshold.

The evaluations were further extended to the multiple speaker scenarios. The interactions of speakers were defined in two separate cases. One was termed turn-taking scenario, where multiple speakers are participating but there is no speaker overlap. The second case was defined as the multiple concurrent scenario, where up to four speakers are active at the same time. For the turn-taking scenario, Fig. 5.6 shows that the conventional bootstrap algorithm with the proposed likelihood function is able to track the speakers at different positions giving up to 90% correct DoA estimates within 10° threshold. A critical step in this case is the inclusion of the VAD information in the likelihood function. During the silent regions, the VAD-based tracker spread the particles in the state-space. Therefore, when the speaker reappears at the same or different position, there are some particles present in the vicinity of the speaker position to resume speaker tracking. The voicing errors made by the VAD algorithm cause wrong position estimates. The proposed importance sampling method is able to track the position estimates more consistently than the bootstrap method, where 96% estimates are accurate within 5° of true speaker's DoA. Therefore, the new resampling method is more suitable in such scenarios than the conventional resampling methods such as the systematic resampling method used in the bootstrap method. The tracking is further extended to pitch dimension, where the importance sampling approach proposed in Section 5.7 is used for joint position and pitch tracking of an active speaker in the turn-taking scenario as presented in Fig. 5.7. Besides some small error made in the beginning and at the end of each utterance, the tracker can successfully follow the pitch and DoA contours. The inclusion of VAD in the likelihood functions is beneficial as only a small number of wrong pitch-position estimates are generated during the silence periods. Pitch tracking is also suffering from the absence of accurate voiced/unvoiced information as the VAD is based on learning the noise floor of PoPi matrix in the silent regions. The unvoiced frames have higher intensity than the noise floor resulting in erroneous pitch measurements. The likelihood for pitch estimates also suffers from pitch halving and doubling errors. Therefore, the tracker requires more robust mechanism to gather pitch evidence.

The two MPoPi-based algorithms defined in Section 4.3 and Section 4.4 were combined with the particle filtering framework. The resulting particle filtering algorithms were presented in Section 5.7. The results of accuracy counts plotted as a CDF for the two concurrent speakers scenarios were depicted in Fig. 5.8. Since the results without the tracker have already been discussed in Section 4.6, here, only the particle filtering results are compared. The MPoPi-STF method used the pitch information to track speakers over time. The inclusion of particle filtering does not improve the method's performance. The MPoPi-FS method on the other hand does not group the pitch cues over time but segments the frequency channels belonging to different speakers at every frame. The inclusion of particle filtering for MPoPi-FS is more beneficial than the MPoPi-STF method. The MPoPi-FS-PF gives on average 78% accurate estimates and MPoPi-STF-PF gives 71% accurate estimates for 5° error threshold. The MPoPi-PF gives an average of 62% accuracy for 5° threshold. Whereas SRP-PHAT-PF algorithm performs worst with average 55% correct estimates for similar threshold. The SIS based particle filtering method involves a compromise between filter's freedom and probability of incorrect initialization. The parameter J_p as defined in Algorithm 5.3 selects this ratio of increasing the filter's freedom to search a wider area of state-space, and increasing the probability of incorrect re-initialization. Hence, the value of J_p needs to be selected keeping in mind the tracking application.

The results of all algorithms for the two concurrent speakers noisy scenarios are shown in Table 5.7. On average, MPoPi-FS-PF gives the best performance out of all algorithms with 77% accuracy within 5° of true speaker's position. The results for MPoPi-STF are on average 75%, which are comparable to MPoPi-FS-PF. The similarity shows that the pitch information can be incorporated at different parts of the localization process as it is done in case of MPoPi-FS and MPoPi-STF. Therefore, the source tracking can be carried out either on the pitch contours or source dynamics as both yield comparable results.

Figs. 5.9 and 5.10 show the accuracy plotted as CDFs for three and four concurrent speakers, respectively. The estimation of the number of sources becomes difficult for these scenarios. The particle filtering algorithms give poor performance in comparison to the conventional methods, where the number of speakers is known *a priori*. The accurate knowledge about the total number of speakers is difficult to gather by using a simple histogram technique. The design of a robust source number estimator is still an open-research topic and beyond the scope of the current work.

The tracking algorithms for real speaker interaction scenarios such as meeting and presentation scenarios are evaluated. The results for all algorithms are shown in Fig. 5.11. The particle filtering algorithms using the MPoPi-FS and MPoPi-STF based importance functions show superior performance in both cases. In the presentation scenario, MPoPi-FS-PF and MPoPi-STF-PF give 80% correct estimates for 10° threshold. The meeting scenario is more challenging with speech overlap of four speakers at one time, but still both MPoPi-FS-PF and MPoPi-STF-PF manage to give 70% correct estimates within 10° of actual source positions. These results show that for practical scenarios involving real-speakers, the particle filter-based estimation is more robust than the conventional approaches. Moreover, the results for a mobile speaker presented in Fig. 5.12 show MPoPi-PF robustness in this scenario, where only few estimation errors are made, whereas SRP-PHAT-PF makes more estimation errors than MPoPi-PF. The accuracy results shown in Fig. 5.12(b) validate the improved performance of the MPoPi based particle filtering over the SRP-PHAT method. The MPoPi-PF gives more than 90% accurate DoA estimates within 15° threshold, whereas SRP-PHAT-PF has around 80% correct estimates for the same threshold.

5.10 Conclusions

This chapter combined the ASL methods proposed in Chapter 4 in the particle filtering framework. The problem of source localization and tracking was developed using the new importance sampling functions. These functions were based on the MPoPi algorithm presented in Chapter 4. Two new particle filtering algorithms were proposed, which combine the bootstrapping and importance sampling approaches for localization and tracking of multiple speakers. The proposed ASLT algorithms were evaluated using the corpus presented in Chapter 3. These new algorithms show improved localization accuracy over the SRP-PHAT method in a wide range of scenarios involving both controlled and real speaker setups.

Chapter 6

Conclusions and Future Work

This work treats the problem of acoustic source localization and tracking using a microphone array. The ASL problem was formulated in Chapter 2 along with the discussion about the signal model used for location estimation. The details about the new multi-channel database were presented in Chapter 3. Different acoustic measurements and evaluation metrics to determine accuracy of location estimates were outlined. Furthermore, a detailed discussion about the segmentation process of the reference speech files was presented. This labeling process allowed to accurately evaluate the algorithms' performance for concurrent speaker scenarios. In Chapter 4, a speech-related feature known as the fundamental frequency was emphasized in the localization process by using a novel joint position and pitch (PoPi) decomposition method. To improve the performance of the PoPi method in realistic scenarios, different weighting functions were proposed. These weighting functions were based on cepstrum analysis and the well-known PHAT weighting of the GCC methods. An auditory inspired pre-processing was proposed for the PoPi method and the resulting algorithm was referred to as the MPoPi method. The usefulness of that pre-processing step was shown through an illustrative example using a speech segment with two concurrent sources. Furthermore, the grouping of location cues was carried out in two different ways. The location cues were estimated by grouping different subbands of the MPoPi method based on the pitch information. That resulted in a low-level tracking of multiple concurrent speakers. A frequency-selective criterion using a single speech segment was used to group different frequency channels belonging to different speakers in the case of concurrent speakers. The location cues were estimated from the individual MPoPi decompositions of those groups. All

those CASA techniques were then combined in the particle filtering framework. The proposed ASLT methods were presented in Chapter 5. The MPoPi decomposition proved to be a suitable importance sampling function in experimental evaluations under various acoustic conditions. The modifications to the MPoPi method using the frequency-selective criterion (MPoPi-FS) was successfully combined with the particle filtering algorithm. The combination of the low-level pitch tracking and location-based particle filtering did not significantly improve location accuracy. Therefore, either approach can be used for speaker localization, but the frequency-selective criterion with particle filters gave consistent results in most of the cases. All the proposed ASLT methods were evaluated with both controlled and real speaker experiments under various acoustic conditions. In conclusion, this thesis showed that the combination of CASA and SMC techniques yields better results in comparison to the conventional methods such as SRP-PHAT.

Result Comparison with the Selected Theses

The experimental results of this thesis are compared with the results reported in the doctoral work listed in Section 1.1.1. The comparison is focused on the experimental setup used to evaluate the algorithms and the results reported for different speaker scenarios. The work done by Michael Brandstein [7] presented the theoretical framework of source location estimates. The Linear Intersection (LI) method presented in the thesis was tested for practical speech source localization. The successful localization of individual talkers in multi-speaker scenarios is achieved using a 10-element bilinear array and three small independent arrays setup in a conference room. The passive source localization method presented by Brandstein work is a two-step process. In the first step, the TDEs are calculated for the microphone pairs and then they are combined based on the array geometry to compute the location estimates. This method leads to erroneous results in challenging environments with high reverberation time and ambient noise. Another issue is the disambiguation between the “real” and “virtual” sources created due to intersection of bearing lines at multiple locations formed by the multiple microphone pairs. In the current work, the two-step process is avoided and the multiple pairs are combined to estimate the source location. For the single speaker scenario, the proposed joint position-pitch estimation methods give around 80% accuracy for 5° error threshold. With the inclusion of the tracker almost all location estimates are correct within this range.

The well-known traditional SRP-PHAT method was proposed by Hector Dibiase [8]. This is a direct method avoiding the two-step process. In his thesis, he used a 15-element microphone array placed on the wall in a mildly reverberant conference room. The results for single speaker scenarios showed nearly 90% correct estimates for speech sources using 4° error threshold. Later the 3D source localization of a single source using the Huge Microphone Array (HMA) with 128 microphones is presented. The SRP-PHAT gives 70% correct location estimates for this task. The performance of SRP-PHAT for concurrent speakers was briefly discussed without any quantitative evaluations. In this thesis, the proposed methods are compared with the SRP-PHAT method. The auditory pre-processing shows improved performance over the SRP-PHAT algorithm for the single speaker as well as the multiple concurrent speaker scenarios. The tracking algorithms based on the MPoPi likelihood functions show robust results in comparison to the SRP-PHAT methods presented in Chapter 5. In single speaker scenarios, nearly all location estimates are correct within 5° and around thirty percent more accurate estimates are achieved than SRP-PHAT for concurrent speaker scenarios. The main issue is with the computational complexity of the MPoPi algorithm. A comparison with the SRP-PHAT method is shown in Appendix B, which shows the complexity of the MPoPi algorithm is linearly increasing with the number of gammatone filters.

The other notable work in this area was carried out by Guillaume Lathoud [9]. His work focused on the post-processing stage of the sound source localizer. The author proposed a short-term clustering method for speaker identification using two table-top 8-channel UCAs in a conference room. His technique used location cues along with spectral features such as MFCCs for the speaker identification task. In the context of the speaker detection and localization task, the author proposed a sector-based Phase Domain Metric (PDM) following the steered beamforming principle. The results presented for speaker detection and localization were processed through an adaptive speech/non-speech classification and short-term clustering. The goal of this thesis is to provide robust instantaneous location estimates, which will improve the performance of the post-processing algorithms. Hence, the comparison of results with this work was not carried out in this thesis.

The successful application of particle filters to track a single speaker in realistic environments was presented by Eric Lehmann [10]. He proposed different likelihood models based on GCC and SBF principles for tracking a single moving speaker. The proposed methods were tested both on synthetic and real audio. The RMSE

for different audio examples was reported to be less than 50 cm in an office environment. The author used a distributed microphone setup to track the moving speaker, which is somewhat different from the current work. In my work, single source tracking results with an RMSE of around 35 cm (corresponds to 5° error threshold) are achieved. In addition to that, results of tracking multiple concurrent speakers are reported in this thesis which were only speculated in Lehmann's work. Another notable difference is that in Lehmann's work, the tracker is assigned the true starting position of the source in the beginning. This is somewhat restrictive in practice as the speaker true starting position is not known and must be determined by the ASL methods. The tracking algorithms proposed in this thesis are randomly initialized, which is a necessary condition for application of any ASLT algorithm in practice.

A recent attempt in use of particle filters for multiple speakers tracking was reported by Maurice Fallon [11]. The results reported in his work were based on recordings made with the distributed microphone setup. The joint location and orientation estimates with an RMSE of 25 cm were reported for a single moving speaker scenario. The track-before-detect framework reported in Fallon's thesis showed better location estimates with an RMSE of less than 10 cm at different levels of SNR. This framework also assumes that the source positions are known at the start of the tracking algorithm and that the sources remain active throughout. Keeping this shortcoming in mind, Fallon proposed a probabilistic variable-dimension particle filtering algorithm. For the variable-dimension particle filtering algorithm, only illustrative examples for speaker tracking were presented. Therefore, it is difficult to compare the algorithms. Moreover, there were certain heuristics applied for the particle addition and removal mechanism. The success of such algorithms in a wide variety of scenarios is difficult to foresee.

Future Work

The problem addressed in this thesis leads to some interesting new directions, which are important and significant to pursue for a practical Sound Source Localization (SSL) system. The use of VAD makes the location estimates robust, where the location estimates can be discarded based on the likelihood value assigned for the current frame by the VAD algorithm. Therefore, a robust distant-speech VAD method for concurrent speaker scenarios is desired, which can be used in conjunction with the

SSL system. Another important aspect is to incorporate the information of speaker orientation in the localization process. For hands-free communication, the use of cross-correlations or the SBF to extract the orientation information can be of greater value in practice. The *a priori* knowledge about the number of active speakers is required for traditional ASL algorithms. This task is not trivial and needs to be combined with the ASL algorithms to perform the localization of active speakers in a stand-alone manner. The improved performance of the MPoPi methods comes at a price of higher computational complexity than the modified SRP-PHAT method. As shown in Appendix B, the computational complexity is directly proportional to the number of gammatone filters. Therefore, a computationally efficient form of MPoPi algorithm that yields same performance should be investigated in the future.

Appendix A

Relationship Between SRP-PHAT and MPoPi Approaches

According to the author in [8], the traditional SRP-PHAT for a M channel microphone array can be computed by summing the GCC-PHAT of all possible microphone pairs shifted by the steering delays (excluding the auto-correlations and symmetric pairs). Therefore, the SRP-PHAT response according to (2.20)

$$\mathbf{P}(\Delta_1 \dots \Delta_M) = 2\pi \sum_{k=1}^{M} \sum_{q=k+1}^{M} R_{kq}(\Delta_q - \Delta_l). \tag{A.1}$$

where R_{kq} is the GCC-PHAT between signals received at microphone k and microphone q, which according to (4.2) is given as

$$R_{lk}(\Delta_{lk}) = \int_{-\infty}^{+\infty} \frac{1}{|X_k(\omega)X_q^*(\omega)|} X_k(\omega)X_q^*(\omega)\, \exp(j\omega\Delta_{kq})\, d\omega \tag{A.2}$$

The first term after the integration is the PHAT weighting. If the PHAT weighting is disabled and set to 1 and by taking only the diametrically placed pairs m_p for the UCA, the relation in (2.20) can be rewritten as

$$P(\Delta) = \sum_{m_p=1}^{M_p} R_{m_p}(\Delta_{m_p}) \tag{A.3}$$

where Δ_{m_p} is defined as the set of steering delays for every direction. Similarly, if the full-band PoPi decomposition defined in (4.3) considers only the end-fire length

(in samples) of the cross-correlation instead of $2K-1$ peaks (no pitch estimation), the definition reduces to:

$$\rho(\varphi_0) = \sum_{m_p=1}^{M_p} R_{m_p}(O(\varphi_{0,m_p})) \tag{A.4}$$

According to (A.3) and (A.4), there is a strong similarity between SRP-PHAT and the PoPi decomposition as both methods use the steering delays to scan over the DoA range. The PoPi algorithm performs the parameterized sampling of the cross-correlation, whereas the SRP-PHAT carry out the spatial averaging across different microphone pairs. Both algorithms will show a maximum if the set of steering delays matches the true DoA of a sound source.

Appendix B

Computational Complexity

The improved performance of MPoPi algorithm over SRP-PHAT comes at a price in form of increased computational complexity. A comparative complexity analysis of the two algorithms can be performed using the following definitions for each operation as defined in [8]:

$N_l \equiv$ # of evaluations of objective function

$N_k \equiv$ # of DFT components in computation

$N_m \equiv$ # of microphones in the array

$N_p \equiv$ # of pairs used

$N_g \equiv$ # of gammatone bandpass filters for MPoPi algorithm

$N_\tau \equiv$ # of points computed for each GCC function in the time-lag domain

Using the "big-O" notation [109], the number of operations required for evaluations of PoPi decomposition and SRP-PHAT algorithms. In this thesis, the SRP-PHAT is computed by summing the GCC-PHATs of a subset of microphone pairs. Therefore, the SRP-PHAT requires $O(N_l\, N_p) + O(N_k\, N_\tau\, N_p)$ evaluations (the complexity for traditional SRP-PHAT is $O(N_l\, N_k\, N_m)$). The PoPi decomposition computes the evaluations for a set of microphone pairs, where the signals of each pair is passed through gammatone filterbank and GCC is computed for every filter resulting in $O(N_l\, N_p) + O(N_g\, N_k\, N_\tau\, N_p)$.

$$\text{Compute Ratio} = \frac{\text{MPoPi operations}}{\text{SRP-PHAT operations}} = \frac{O(N_g\, N_k\, N_p\, N_\tau) + O(N_l\, N_p)}{O(N_k\, N_p\, N_\tau) + O(N_l\, N_p)} \quad \text{(B.1)}$$

B.1 Example

In this thesis, a 24-element UCA was used to search over a grid of DOAs ranging from 1 to 360 degrees with a 1-degree step size. In this case, $N_l = 360$. The range of maximum lag for GCC is not defined for the SRP-PHAT and MPoPi algorithms. Therefore, the GCC is computed for the maximum-lag equivalent to the frame length of 2048 samples. For the MPoPi method, all the microphones N_m are filtered through the gammatone filterbank $N_g = 64$, and GCC functions are computed for all the frequency channels for every pair. This number of evaluations for the PoPi decomposition results in $N_l = 481 \times 360 = 173160$ evaluations. The value of N_p is chosen to be 12 as only diametrically placed microphones pairs are used for computation. Using the derived relationship, the MPoPi method requires 64 times more computation than modified SRP-PHAT and 1.36 times when compared to traditional SRP-PHAT. The full-band PoPi method requires 1.55 times more computation than modified SRP-PHAT and 3.04 times more faster than the traditional SRP-PHAT. Hence, the major computation load comes from the evaluations of GCC for every frequency band rather than the joint position-pitch relations.

Appendix C

Other Work

During the period from Nov., 2008 to Mar., 2009, I was a visiting researcher at the Chair of Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Germany. During my visit, I worked on the problem of tracking speaker identities using Blind Source Separation (BSS) systems.

The problem of speaker switching in BSS occurs due to the permutation ambiguity inherent to any BSS system, e.g, in multi-speaker scenarios where trajectories of the moving sources approach and cross each other. To address this problem, a new scheme is proposed to be used as a post-processing stage to BSS systems. This scheme exploits the fundamental frequency F_0, which is a strong speaker discriminant feature especially in cross-gender cases. Various state-of-the-art pitch estimation methods have been evaluated for this task including some new promising techniques to compute the fundamental frequency. The proposed scheme has been evaluated for different scenarios involving moving sources for various gender combinations. The proposed scheme performed well in most of the cases. The proposed techniques along with the results were published in a technical report.

Related Publication:

- Tania Habib, Anthony Lombard, and Walter Kellermann. Tracking Speaker Identities in Blind Source Separation Systems. Technical report, Graz University of Technology, 2010.

List of Figures

1.1 Speech localization problem in a reverberant environment 2

2.1 Illustration of DoA estimation based on the free space signal model. . 15
2.2 Sound source localization using adaptive filters 16
2.3 Spatial diagram 17
2.4 TDoA mapping into spatial coordinates 19

3.1 SPSC UCA with variable diameter 40
3.2 Delay-and-sum beam pattern of the SPSC UCA 44
3.3 Beam pattern in polar coordinates at different frequencies 44
3.4 Example of a reference file for the two concurrent speaker scenario. . 48
3.5 SPSC cocktail party room layout. 49
3.6 Measured room impulse response. 50
3.7 Energy decay curve for RIR 51

4.1 PoPi decomposition 57
4.2 Time-waveform and amplitude spectrum of two microphone signals . 59
4.3 PoPi plane 60
4.4 Multichannel PoPi plane 61
4.5 The cepstrum based frequency domain weighting function 63
4.6 The PoPi plane decomposition for a single source 64
4.7 Full-band PoPi decomposition of two concurrent speakers 66
4.8 Block diagram of the MPoPi algorithm 67
4.9 Magnitude response of 21 out of 64 gammatone filters. 68
4.10 Generalized cross-correlation 70
4.11 GCC of four different gammatone filter outputs 72

4.12 GCC of non-informative gammatone filter outputs. 73
4.13 PoPi planes of different filter outputs. 73
4.14 Summary cross-correlation 74
4.15 MPoPi decomposition of two concurrent speakers 74
4.16 The auto-correlogram of a female speaker 77
4.17 Fragment generation process 80
4.18 The spectral grouping and fragment generation process. 82
4.19 Proposed scheme for multi-channel extension of a fragment system . . 83
4.20 Accuracy counts of a single speaker. 86
4.21 Accuracy counts versus different kinds of background noise. 87
4.22 Single speaker accuracy counts versus SNR 89
4.23 Accuracy counts versus threshold for two concurrent speakers. 90
4.24 Results of two concurrent speakers versus different noise types 91
4.25 Accuracy counts plotted as a CDF for three concurrent speakers . . . 92
4.26 Accuracy counts plotted as a CDF for four concurrent speakers . . . 93
4.27 Accuracy in percent for different array diameters for a single speaker 95
4.28 Accuracy counts for 2 speakers vs. different array diameters 96
4.29 Real speakers interaction scenarios. 97
4.30 Accuracy results for real speakers interaction scenarios. 97
4.31 Mobile speaker scenario. 99
4.32 Mobile speaker scenarios for speaker switch detection. 100
4.33 Detection results for two mobile speakers. 100

5.1 Symbolic representation of particle filtering 111
5.2 Position-Pitch plane of two concurrent speakers 112
5.3 The evolution of the measurement function $\mathcal{Y}_k$ over time 117
5.4 Joint position-pitch tracking of a single speaker. 125
5.5 The position-pitch tracking of a single speaker 126
5.6 The position tracking of a turn taking scenario 130
5.7 The position and pitch tracking of a turn taking scenario 133
5.8 Accuracy counts versus threshold for two concurrent speakers. 135
5.9 Accuracy counts versus error threshold for three concurrent speakers . 137
5.10 Accuracy counts versus error threshold for four concurrent speakers . 137
5.11 Accuracy in percent results for real speakers interaction scenarios. . . 138
5.12 Mobile speaker scenario. 139

List of Tables

3.1 Quantization error for SPSC UCA 43
3.2 $f_{\max}$ values for different sensors' spacing. 45

5.1 Selected values for dynamic model parameters. 114
5.2 Results of the single speaker case. 127
5.3 Localization accuracy versus different background noise (Case 1). . . 128
5.4 Localization accuracy versus different background noise (Case 2). . . 128
5.5 Localization accuracy versus different SNR conditions (Case 1). . . . 129
5.6 Localization accuracy versus different SNR conditions (Case 2). . . . 129
5.7 Results of two concurrent speakers versus different noisy conditions. . 134

List of Algorithms

5.1 A generic particle filter algorithm for source tracking. 113
5.2 Proposed SIS based algorithm 121
5.3 Proposed SIR algorithm 122

Bibliography

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Processing Mag.*, vol. 13, pp. 67–94, 1996.

[2] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag, Heidelberg, Germany, 2001.

[3] J. Benesty, J. Chen, and Y. Huang, *Springer Topics in Signal Processing: Microphone Array Signal Processing*, J. Benesty and W. Kellermann, Eds. Springer Verlag, Heidelberg, Germany, 2008.

[4] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proc. of ICASSP*, Istanbul, Turkey, June 2000, pp. 909–912.

[5] A. Johansson, N. Grbic, and S. Nordholm, "Speaker localisation using the far-field SRP-PHAT in conference telephony," in *Proc. of Int. Symp. on Intelligent Signal Processing and Communication Systems*, Kaohsiung, Taiwan, 2002.

[6] M. Wölfel and J. McDonough, *Distant Speech Recognition*. John Wiley, West Sussex, United Kingdom, 2009.

[7] M. S. Brandstein, "A framework for speech source localization using sensor arrays," Ph.D. dissertation, Brown University, USA, 1995.

[8] J. H. Dibiase, "A high-accuracy low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, USA, 2000.

[9] G. Lathoud, "Spatio-temporal analysis of spontaneous speech with microphone arrays," Ph.D. dissertation, EPFL, Lausanne, Switzerland, 2006.

[10] E. A. Lehmann, "Particle filtering methods for acoustic source localization and tracking," Ph.D. dissertation, Australian National University, 2004.

[11] M. Fallon, "Acoustic source tracking using sequential monte carlo," Ph.D. dissertation, University of Cambridge, UK, 2008.

[12] M. Képesi, F. Pernkopf, and M. Wohlmayr, "Joint position pitch tracking for 2-channel audio," in *International Workshop on Content based Multimedia Indexing*, Bourdeaux, France, June 2007, pp. 303–306.

[13] D. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, Algorithms, and Applications*. Wiley/IEEE Press, New Jersey, USA, 2006.

[14] T. Habib, M. Képesi, and L. Ottowitz, "Expermental evaluation of the joint position-pitch estimation (PoPi) algorithm in nosiy environments," in *IEEE Workshop SAM*, Darmstadt, Germany, 2008, pp. 369–372.

[15] M. Képesi, L. Ottowitz, and T. Habib, "Joint positon-pitch estimation for multiple speaker scenarios," in *IEEE Workshop HSCMA*, Trento, Italy, 2008, pp. 85–88.

[16] T. Habib, L. Ottowitz, and M. Képesi, "Experimental evaluation of multi-band position-pitch estimation (M-PoPi) algorithm for multi-speaker localization," in *Proc. of INTERSPEECH*, Brisbane, Australia, 2008, pp. 1317–1320.

[17] T. Habib and H. Romsdorfer, "Concurrent speaker localization using multi-band position-pitch (M-PoPi) algorithm with spectro-temporal preprocessing," in *Proc. of INTERSPEECH*, Makuhari, Japan, 2010, pp. 2774–2777.

[18] ——, "Comparison of SRP-PHAT and multiband-popi algorithms for speaker localization using particle filters," in *Proc. of 13th International Conference on Digital Audio Effects,DAFx-10*, Graz, Austria, September 6-10, 2010, pp. 1–6.

[19] ——, "Combining multiband joint position-pitch algorithm and particle filters for speaker localization," in *Proc. of IEEE Workshop SAM*, Israel, Oct. 2010, pp. 149 – 152.

[20] ——, "Improving multiband position-pitch algorithm for localization and tracking of multiple concurrent speakers by using a frequency selective criterion," in *INTERSPEECH*, Florence, Italy, 2011.

[21] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Springer Verlag, 2007.

[22] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Prentice Hall, New Jersey, 1993.

[23] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source location," *J. Acoust. Soc. Am.*, vol. 107, pp. 384–391, 2000.

[24] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, New York, 2001.

[25] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech and Audio Processing*, vol. 13, pp. 120–134, 2005.

[26] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," *Signal Processing*, vol. 86, pp. 1260–1277, 2006.

[27] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound source using averaged directivity patterns of blind source separation systems," in *Proc. of ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 233–236.

[28] Y. Huang, J. Benesty, and J. Chen, Eds., *Acoustic MIMO Signal Processing*. Springer, Heidelberg, Germany, 2006.

[29] T. Habib and M. Képesi, "State-of-the-art report on speaker localization methods," Graz University of Technology, Tech. Rep., December 2007.

[30] C. F. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-24, pp. 320–327, 1976.

[31] G. C. Carter, "Coherence and time delay estimation: An applied tutorial for research, development, test and evaluation engineers," *IEEE Press*, 1993.

[32] A. H. Quazi, "An overview on the time delay estimation in active and passive systems for target localization," *IEEE. Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 527–533, 1981.

[33] C. Zhang, D. Florêncio, and Z. Zhang, "Why does PHAT work well in low

noise, reverberant environments?" in *Proc. of ICASSP*, Las Vegas, Nevada, 2008, pp. 2565–2568.

[34] M. Omologo and P. Svaizer, "Talker localization and speech enahancement in a noisy environment using a microphone array based acquisition system," in *Proc. of Eurospeech*, Berlin, September 1993, pp. 605–609.

[35] ——, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. of ICASSP*, Adelaide, Souch Australia, 1994, pp. 273–276.

[36] ——, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *Proc. of ICASSP*, Atlanta, Georgia, 1996, pp. 921–924.

[37] S. Bédard, B. Champagne, and A. Stéphenne, "Effects of room reverberation on time-delay estimation peformance," in *Proc. of ICASSP*, Adelaide, South Australia, April 1994, pp. 261–264.

[38] J. Scheuing and B. Yang, *Reverberant EnvironmentsSpeech and Audio Processing in Adverse Environments*. Springer Verlag, Heidelberg, 2008, ch. Correlation-Based TDOA-Estimation for Multiple Sources in Reverberant Environment, pp. 382–416.

[39] H. Do and H. F. Silverman, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. of ICASSP*, Honolulu, Hawaii, 2007, pp. 121–124.

[40] ——, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proc. of ICASSP*, Dallas, Texas, 2010, pp. 125–128.

[41] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. on Speech and Audio Processing*, vol. 12, pp. 499–508, 2004.

[42] J. Dmochowski, J. Benesty, and S. Affes, "Fast steered response power source localization using inverse mapping of relative delays," in *Proc. of ICASSP*, Las Vegas, Nevada, Apr. 2008, pp. 289–292.

[43] S. Trevo and T. Lokki, "Interpolation methods for the SRP-PHAT algorithm," in *Proc. of IWAENC 2008*, Seattle, USA, 2008, pp. 1–4.

[44] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, pp. 276–280, 1986.

[45] R. Roy and K. Kailath, "ESPRIT- Estimation of signal paramters via rotational invariance techniques," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, pp. 984–995, 1989.

[46] B. D. Rao and K. V. S. Hari, "Peformance analysis of Root Music," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 37, no. 12, pp. 1939 – 1949, December 1989.

[47] S. Haykin and K. J. R. Liu, Eds., *Handbook on Array Processing and Sensor Networks*. Wiley, New Jersey, USA, 2009.

[48] H. Teutsch and W. Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular arrays," *J. Acoust. Soc. of Am.*, vol. 120, no. 5, pp. 2724–2736, Nov. 2006.

[49] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using eb-esprit with spherical microphone arrays," in *Proc. of ICASSP*, Prague, Czech Republic, May 2011.

[50] K. W. Wilson, "Estimating uncertaininty models for speech source localization in real-world environments," Ph.D. dissertation, Massachusetts Institute of Technology, USA, 2006.

[51] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann, "Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering," in *Proc. of ICASSP*, Pennsylvania, U.S.A, Apr. 2005, pp. III–97 – III–100.

[52] A. Lombard, H. Buchner, and W. Kellermann, "A real-time demonstrator for the 2D localization of two sound sources using blind adaptive MIMO system identification," in *Proc. of HSCMA*, Trento, Italy, 2008, pp. 41–44.

[53] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Exploiting the self-steering capability of blind source separation to localize two or more sound sources in adverse environments," in *Proc. of ITG-Fachtagung Sprachkommunikation*, Aachen, Germany, Oct. 2008.

[54] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, Upper Saddle River, New Jersey, 2002.

[55] M. Wohlmayr and M. Képesi, "Joint position pitch extraction from multichannel audio," in *Proc. of INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 1629–1632.

[56] X. Qian and R. Kumaresan, "Joint estimation of time delay and pitch of voiced speech signals," in *IEEE 29th Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, California, November 1995, pp. 735–739.

[57] G. Liao, H. C. So, and P. C. Ching, "Joint time delay and frequency estimation of multiple sinusoids," in *Proc. of ICASSP*, Salt Lake City, Utah, May 2001, pp. 3121–3124.

[58] R. J. McAULAY and T. F. Quatieri, "Speech analysis/Synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 34, no. 4, pp. 744 – 754, 1986.

[59] L. Y. Ngan, Y. Wu, H. C. So, P. C. Ching, and S. W. Lee, "Joint time delay and pitch estimation for speaker localization," in *Proc. of ISCAS*, Bangkok, Thailand, 2003, pp. III–722 – III–725 vol.3.

[60] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. of ICASSP*, 1997, pp. 375–378.

[61] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[62] B. Yegnanarayana, S. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverbant speech for time-delay estimation," *IEEE Trans. of Speech and Audio Processing*, vol. 13, no. 6, pp. 1110–1118, Nov. 2005.

[63] B. Yegnanarayana and S. Prasanna, "Analysis of instantaneous F_0 contours from two speakers mixed signal using zero frequency filtering," in *Proc. of ICASSP*, Texas, U.S.A, Mar. 2010, pp. 5074–5077.

[64] S. N. Wrigley and G. J. Brown, "Recurrent timing neural networks for joint F0-localisation based speech seperation," in *Proc. of ICASSP*, Honolulu, Hawaii, April 2007, pp. 157–160.

[65] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, pp. 3907–3908, 1995.

[66] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "Integrating pitch and localisation cues at a speech fragment level," in *Proc. of INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 2769–2772.

[67] ——, "A speech fragment approach to localising multiple speakers in reverberant environments," in *Proc. of ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 4593–4596.

[68] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus," in 5^{th} *International Conference on Methods and Techniques in Behavorial Research*, Wageningen, The Netherlands, Sept. 2005.

[69] "AMI Project Webpage," Last Checked: 21-01-2011. [Online]. Available: http://corpus.amiproject.org/

[70] "CHIL Project Webpage," Last Checked: 1-05-2011. [Online]. Available: http://chil.server.de/

[71] "NIST MARK III Array Webpage," Last Checked: 1-05-2011. [Online]. Available: http://www.nist.gov/smartspace/index.html

[72] A. Brutti, L. Critoforetti, W. Kellermann, L. Marquardt, and M. Omologo, "WOZ acoustic data collection for interactive TV," *Springer Language Resources and Evaluation Journal*, vol. Special Issue LREC2008, pp. 205–219, 2010.

[73] "Behringer ECM8000 data sheet," Last Checked: 04-05-2009. [Online]. Available: http://www.behringer.com/EN/Products/ECM8000.aspx

[74] "RME Fireface 800 audio interface," Last Checked: 04-08-2009. [Online]. Available: http://www.rme-audio.de/en_products_fireface_800.php

[75] "Behringer ADA8000 data sheet," Last Checked: 04-08-2009. [Online]. Available: http://www.behringer.com/EN/Products/ADA8000.aspx

[76] W. Jäger, "Analysis and implementation of the position-pitch source localization algorithm on a hybrid reconfigurable CPU," Master's thesis, SPSC Lab, Graz University of Technology, Austria, March 2011.

[77] H. Liu and E. Milios, "Acoustic positioning using multiple microphone arrays," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 2772–2782, 2005.

[78] H. L. Van Trees, *Optimum Array Processing*. Wiley-Interscience, New York, 2002, ch. Planar Arrays and Apertures (Sec.4.2), pp. 280–284.

[79] I. McCowan, *Microphone Arrays: A Tutorial*, Document retrieved in Nov, 2007:. [Online]. Available: http://www.idiap.ch/~mccowan/arrays/tutorial.pdf

[80] E. Zwyssig, "Digital microphone array: Design, implementation and speech recognition experiments," Master's thesis, The University of Edinburgh, Aug. 2009.

[81] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, pp. 2421–2424, 2006.

[82] J. M. Zmöling, "Pure Data Community Site," Last Checked: 21-07-2009. [Online]. Available: http://puredata.info/

[83] HTK Speech Recognition Toolkit, Cambridge, United Kingdom. [Online]. Available: http://htk.eng.cam.ac.uk/

[84] H. Romsdorfer, "A tool for accurate segmentation of speech signals into voiced, unvoiced, and silent segments," Report, SYNVO GmbH, Zurich, Switzerland,, Tech. Rep., 2010.

[85] M. Schroeder, "New method for measuring reverberation time," *Journal of Acoustical Society of America*, vol. 37, pp. 409–412, 1965.

[86] H. Christensen and J. Barker, "Using location cues to track speaker changes from mobile, binaural microphones," in *Proc. of INTERSPEECH*, Brighton, U.K, Sept. 2009, pp. 140–143.

[87] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets, "Signal-based performance evaluation of dereverberation algorithms," *Journal of Electrical and Computer Engineering*, vol. 2010, pp. 1–5, 2010.

[88] R. D. Patterson, I. N. Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the Gammatone function," MRC Applied Pschology Unit, Cambridge, Tech. Rep., 1987.

[89] M. Slaney, "Auditory toolbox: A Matlab toolbox for auditory modeling work," Apple Computer, Inc. Advanced Technology Group, Tech. Rep. 45, 1994.

[90] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[91] N. Ma, P. Green, J. Barker, and A. Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources," *Speech Comm.*, vol. 49, no. 12, pp. 874–891, 2007.

[92] N. Ma, "Informing multisource decoding in robust automatic speech recognition," Ph.D. dissertation, Dept. of Computer Science, The University of Sheffield, 2008.

[93] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, 2004.

[94] J. Dmochowski, J. Benesty, and S. Affès, "On spatial aliasing in microphone arrays," *IEEE Trans. on Signal Processing*, vol. 57, no. 4, pp. 1383–1395, 2009.

[95] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. of ASME, Journal of Basic Engineering*, vol. 82, no. D, pp. 35–45, 1960.

[96] U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Journal on Applied in Signal Processing*, vol. 2006, pp. 1–15, 2006.

[97] H. Sorenson, Ed., *Kalman Filtering: Theory and Applications*. IEEE Press, 1985.

[98] S. J. Julier and J. K. Uhlmann, "A new extension of Kalman filter to nonlinear systems," in *Proc. Int. Sym. on Aerospace/Defense Sensing, Simulation and Controls (AeroSense)*, Orlando, Florida, USA, 1997, pp. 182–193.

[99] J. LaViola, "A comparison of unscented and extended Kalman filtering for estimating quarternion motion," in *Proc. of American Control Conf.*, Denver, Colorado, USA, 2003, pp. 2435–2440.

[100] T. V. Dorvkind and S. Gannot, "Speaker localization using unscented Kalman filter," in *Proc. HSCMA Workshop*, Piscataway, New Jersey, USA, Mar. 2005.

[101] A. Doucet, N. de. Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, 2001.

[102] M. S. Arulamalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[103] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. of ICASSP*, Salt Lake City, Utah, 2001, pp. 3021–3024.

[104] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[105] R. Douc, O. Cappé, and E. Moulines, "Comparison of resampling schemes for particle filtering," in *Proc. of 4th International Symposium Image and Signal Processing and Analysis*, Zagreb, Croatia, Sept. 2005, pp. 64–69.

[106] M. Kepési, T. V. Pham, G. Kubin, L. Weruaga, A. Juffinger, and M. Grabner, "Noise cancellation frontends for automatic meeting transcription," in *Proc. of Euronoise*, Tampere, Finland, 2006, pp. 1–6.

[107] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–11, 2007.

[108] N. J. Gordon, D. J. Salmond, and A. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proc. of Radar and Signal Processing*, vol. 140, no. 2, pp. 107–113, 1993.

[109] T. H. Cormen and C. E. Leiserson, *Introduction to Algorithms*, T. H. Cormen and C. E. Leiserson, Eds. MIT Press, Cambridge, Massachusetts, 1990.