Saarland University
Faculty of Natural Sciences and Technology I
Department of Computer Science

Dissertation

for obtaining the title of Doctor of Engineering of the Faculties
of Natural Sciences and Technology of Saarland University

Adaptive Time-Frequency Analysis for Cognitive Source Separation

submitted by
Sylvia Kümmel

Supervisor
Prof. Dr.-Ing. Thorsten Herfet

Saarbrücken, December 2009

Date of Colloquium: 21.04.2010

Dean of Faculty: Prof. Dr. Holger Hermanns

Members of examination board:

Prof. Dr.-Ing. Thorsten Herfet, Saarland University
Prof. Dr.-Ing. Udo Zölzer, Helmut Schmidt University Hamburg
Prof. Dr. Antonio Krüger, Saarland University
Dr. Mark Hillebrand, Saarland University

Statutory declaration

Hereby I affirm in lieu of an oath, that I made the present thesis autonomously and without other than the indicated auxiliary means. The data used indirectly or from other sources and concepts are characterized with lists of sources. The thesis has not been submitted for academic degree consideration either nationally or internationally in identical or similar from to date.

Saarbrücken, April 27, 2010

Sylvia Kümmel

Declaration of Consent

Herewith I agree that my thesis will be made available through the library of the Computer Science Department.

Saarbrücken, April 27, 2010

Sylvia Kümmel

Abstract

This thesis introduces a framework for separating two speech sources in non-ideal, reverberant environments. The source separation architecture tries to mimic the extraordinary abilities of the human auditory system when performing source separation. A movable human dummy head residing in a normal office room is used to model the conditions humans experience when listening to complex auditory scenes.

This thesis first investigates how the orthogonality of speech sources in the time-frequency domain drops with different reverberation times of the environment and shows that separation schemes based on ideal binary time-frequency-masks are suitable to perform source separation also under humanoid reverberant conditions.

Prior to separating the sources, the movable human dummy head analyzes the auditory scene and estimates the positions of the sources and the fundamental frequency tracks. The source localization is implemented using an iterative approach based on the interaural time differences between the two ears and achieves a localization blur of less than three degrees in the azimuth plane.

The source separation architecture implemented in this thesis extracts the orthogonal time-frequency points of the speech mixtures. It combines the positive features of the STFT with the positive features of the cochleagram representation. The overall goal of the source separation is to find the ideal STFT-mask. The core source separation process however is based on the analysis of the corresponding region in an additionally computed cochleagram, which shows more reliable Interaural Time Difference (ITD) estimations that are used for separation.

Several algorithms based on the ITD and the fundamental frequency of the target source are evaluated for their source separation capabilities. To enhance the separation capabilities of the single algorithms, the results of the different algorithms are combined to compute a final estimate. In this way SIR gains of approximately 30 dB for two source scenarios are achieved. For three source scenarios SIR gains of up to 16 dB are attained. Compared to the standard binaural signal processing approaches like DUET and Fixed Beamforming the presented approach achieves up to 29 dB SIR gain.

Zusammenfassung

Diese Dissertation beschreibt ein Framework zur Separation zweier Quellen in nicht-idealen, echobehafteten Umgebungen. Die Architektur zur Quellenseparation orientiert sich dabei an den außergewöhnlichen Separationsfähigkeiten des menschlichen Gehörs. Um die Bedingungen eines Menschen in einer komplexen auditiven Szene zu imitieren, wird ein beweglicher, menschlicher Kunstkopf genutzt, der sich in einem üblichen Büroraum befindet.

In einem ersten Schritt analysiert diese Dissertation, inwiefern die Orthogonalität von Sprachsignalen im Zeit-Frequenz-Bereich mit unterschiedlichen Nachhallzeiten abnimmt. Trotz der Orthogonalitätsabnahme sind Separationsansätze basierend auf idealen binären Masken geeignet um eine Trennung von Sprachsignalen auch unter menschlichen, echobehafteten Bedingungen zu realisieren.

Bevor die Quellen getrennt werden, analysiert der bewegliche Kunstkopf die auditive Szene und schätzt die Positionen der einzelnen Quellen und den Verlauf der Grundfrequenz der Sprecher ab. Die Quellenlokalisation wird durch einen iterativen Ansatz basierend auf den Zeitunterschieden zwischen beiden Ohren verwirklicht und erreicht eine Lokalisierungsgenauigkeit von weniger als drei Grad in der Azimuth-Ebene.

Die Quellenseparationsarchitektur die in dieser Arbeit implementiert wird, extrahiert die orthogonalen Zeit-Frequenz-Punkte der Sprachmixturen. Dazu werden die positiven Eigenschaften der STFT mit den positiven Eigenschaften des Cochleagrams kombiniert. Ziel ist es, die ideale STFT-Maske zu finden. Die eigentliche Quellentrennung basiert jedoch auf der Analyse der entsprechenden Region eines zusätzlich berechneten Cochleagrams. Auf diese Weise wird eine weitaus verlässlichere Auswertung der Zeitunterschiede zwischen den beiden Ohren verwirklicht.

Mehrere Algorithmen basierend auf den interauralen Zeitunterschieden und der Grundfrequenz der Zielquelle werden bezüglich ihrer Separationsfähigkeiten evaluiert. Um die Trennungsmöglichkeiten der einzelnen Algorithmen zu erhöhen, werden die einzelnen Ergebnisse miteinander verknüpft um eine finale Abschätzung zu gewinnen. Auf diese Weise können SIR Gewinne von ungefähr 30 dB für Szenarien mit zwei Quellen erzielt werden. Für Szenarien mit drei Quellen werden Gewinne von bis zu 16 dB erzielt. Verglichen mit binauralen Standardverfahren zur Quellentrennung wie DUET oder Fixed Beamforming, gewinnt der vorgestellte Ansatz bis zu 29 dB SIR.

Detaillierte Zusammenfassung

Diese Dissertation beschreibt ein Framework zur Separation zweier Quellen in nicht-idealen, echobehafteten Umgebungen. Die Architektur zur Quellenseparation orientiert sich dabei an den außergewöhnlichen Separationsfähigkeiten des menschlichen Gehörs. Um die Bedingungen eines Menschen in einer komplexen auditiven Szene zu imitieren, wird ein beweglicher, menschlicher Kunstkopf genutzt, der sich in einem üblichen Büroraum befindet. Auditive Szenen werden mithilfe eines normalen 7.1 Lautsprecher-Systems erzeugt.

Orthogonalität von Sprachsignalen in echobehafteten, humanoiden Szenarien

Ein oft genanntes Ziel von Quellenseparationsarchitekturen ist das Finden der idealen binären Zeit-Frequenz-Maske: Jeder Eintrag der Zeit-Frequenz-Maske wird genau dann auf eins gesetzt, wenn die Energie der Zielquelle in diesem Bin größer als die interferierenden Energien ist.

Das Konzept der binären Maske basiert auf der annähernden Orthogonalität von Sprachsignalen in der Zeit-Frequenz-Ebene, welche für echofreie Sprachsignale nachgewiesen ist. Um das Konzept der binären Masken auch in realen Szenarien wie etwa dem humanoiden Aufbau in diesem Projekt zu nutzen, untersucht diese Dissertation wie sich die Orthogonalität von Sprachsignalen unter verschiedenen echobehafteten Bedingungen verändert und evaluiert, ob sich solche Separationsalgorithmen auch dazu eignen, eine Trennung unter echobehafteten, humanoiden Bedingungen zu erzielen.

Echos und die Filtereigenschaften des menschlichen Kopfes beeinflussen die Orthogonalität von Sprachsignalen in der Zeit-Frequenz Domäne. Das Signal-Interferenz-Verhältnis (SIR) nimmt für echobehaftete, humanoide Szenarien mit zwei Quellen um ca. 5 dB ab. Nichtsdestotrotz erreicht das Konzept der idealen binären Maske eine ausreichende Qualität der separierten Sprachsignale um auch in echobehafteten, humanoiden Szenarien anwendbar zu bleiben.

Auditive Szenenanalyse

Wenn Menschen eine auditive Szene betreten, analysieren sie automatisch die Umgebung um ihnen und schätzen Parameter wie die Anzahl und die Positionen, sowie den Verlauf der Grundfrequenz der klangerzeugenden Quellen ab. Die Quellenseparationsarchitektur dieser Dissertation versucht diese kognitiven Fähigkeiten des menschlichen Gehirns zu imitieren. Bevor die Quellen getrennt werden, analysiert der menschliche Kunstkopf die auditive Szene und ermittelt die Anzahl und Positionen der Quellen und die Grundfrequenzverläufe der Sprachquellen. Diese Parameter werden dann genutzt um die folgende Quellentrennung zu verbessern.

Ein neuer Lokalisierungsansatz nimmt an, dass die Klangquellen auf einem Kreis um den Hörer angeordnet sind und zeigt bessere Ergebnisse als die Standardverfahren zur humanoiden Quellenlokalisation wie die Woodworth Formel und der Freifeldansatz. Zusätzlich wird ein Lokalisierungsansatz basierend auf einer approximierten HRTF vorgestellt und ausgewertet.

Iterative Varianten verbessern die Lokalisierungsgenauigkeit und lösen Mehrdeutigkeiten auf. Mithilfe der beschriebenen Methoden wird eine Lokalisierungsgenauigkeit von ungefähr drei Grad erreicht, welche vergleichbar mit der menschlichen Lokalisierungsgenauigkeit ist. Eine Vorne-Hinten-Bestimmung erlaubt eine zuverlässige Lokalisation der Klangquellen in der kompletten Azimuth-Ebene in bis zu 98.43 % der Fälle.

Zur Bestimmung des Grundfrequenzverlaufs wird eine Variante des YIN-Algorithmus [22] implementiert. Die Eingangssignale werden in Zeitfenster von 50 ms Länge unterteilt, so dass zwei Perioden eines 40 Hz Signals gerade noch erfasst werden. Für jedes dieser Fenster wird eine Grundfrequenz abgeschätzt. Eine Nachbearbeitungsstufe glättet die Grundfrequenzkurve und entfernt Ausreißer basierend auf den Charakteristiken der menschlichen Stimme.

Quellentrennung

Die Quellenseparationsarchitektur dieser Dissertation extrahiert die orthogonalen Zeit-Frequenz-Punkte der aufgenommenen Sprachmixturen. Dazu kombiniert der vorgestellte Ansatz die positiven Eigenschaften der STFT mit den positiven Eigenschaften des Cochleagrams. Das Ziel ist, die ideale STFT-Maske zu finden. Die eigentliche Quellentrennung basiert jedoch auf der Analyse der entsprechenden Region eines zusätzlich berechneten Cochleagrams. Auf diese Weise wird eine weitaus verlässlichere Auswertung der Zeitunterschiede zwischen den beiden Ohren verwirklicht.

Mehrere Algorithmen basierend auf den interauralen Zeitunterschieden und der Grundfrequenz der Zielquelle werden bezüglich ihrer Separationsfähigkeiten evaluiert. Um die Trennungsmöglichkeiten der einzelnen Algorithmen zu erhöhen, werden die einzelnen Ergebnisse miteinander verknüpft um eine finale Abschätzung zu gewinnen. Auf diese Weise können SIR Gewinne von ungefähr 30 dB für Szenarien mit zwei Quellen erzielt werden. Für Szenarien mit drei Quellen werden Gewinne von bis zu 16 dB erzielt. Verglichen mit binauralen Standardverfahren zur Quellentrennung wie DUET oder Fixed Beamforming, gewinnt der vorgestellte Ansatz bis zu 29 dB SIR.

Acknowledgments

Many people have supported me during the work on this thesis.

First of all, I would like to thank my supervisor Prof. Dr.-Ing. Thorsten Herfet. This work would not have been possible without his persistent support and scientific guidance at all times. Many thanks for the inspiring discussions and the pleasant working environment at the telecommunications lab.

I would like to thank Prof. Dr.-Ing Udo Zölzer for kindly accepting the role of second advisor.

Furthermore, I would like to thank my colleagues and friends at the telecommunications lab, who always supported me and with whom I had a wonderful time during the last years. I cannot mention them all, but I am especially grateful to Eric Haschke, Jochen Krämer, Igor Fischer, Zakaria Keshta, Diane Chlupka, Manuel Gorius, Jochen Miroll, Tan Guoping, Zhao Li and Aleksej Spenst.

Last but not least my thanks go to my parents, my husband and my children. Without their endless love and support this thesis would not have been possible.

Contents

Nomenclature

T_{60} 45
Time that the energy of an idealized Dirac impulse needs to decrease by 60 dB.

ACF 23
Autocorrelation Function.

AM 26
Amplitude Modulation.

ASA 6
Auditory Scene Analysis.

ASR 2
Automatic Speech Recognizer.

CASA 1
Computational Auditory Scene Analysis.

CMU 65
Carnegie Mellon University.

DDR 45
Direct-Sound-To-Reverberation-Ratio.

DUET 32
Degenerate Unmixing Estimation Technique.

ERB 19
Equivalent Rectangular Bandwidth.

ESPRIT 40
Estimation of Signal Parameters via Rotational Invariance Techniques.

F0 22
Fundamental frequency.

1 Introduction

When humans enter a complex auditory scene such as a cocktail party they have no problem to attend to a specific sound source while suppressing all other sources and background noise. If a human wants to talk to a specific person on a cocktail party, the speech of the other talking guests and additional sounds like a running TV, clinking glasses and a barking dog are masked to enable the human listener to conduct the chosen conversation. This effect – commonly known as Cocktail Party Effect – has turned out to be an extraordinary ability of the human brain.

For a long time, engineers have tried to imitate this excellent source separation performance of the human brain. Till now no machine is able to perform such a good auditory scene analysis as humans are able to do. The scientific area which aims at developing computational models for the understanding and interpretation of auditory scenes analog to human scene analysis is commonly called Computational Auditory Scene Analysis (CASA). Most CASA architectures aim to implement known human strategies for auditory scene analysis in computational algorithms to examine if machines are able to perform a scene analysis comparable to humans and in the hope of finding new mechanisms of how the human auditory system works. The goal of CASA systems is to analyze the auditory scene with at most two microphones (according to the two human ears). Human abilities like source localization, source separation and estimating several characteristics of the sources are tried to be imitated.

Application areas of CASA systems are manifold:

Robots are getting more and more humanoid every day. In the future robots will act as full partners of humans. Speech signals and an analysis of the auditory scene are critical to perform a reliable interaction between humans and robots. Especially in cases, when relevant objects are not in the visible range, robot audition can be employed. By implementing a robotic audition that performs similar to human audition, an intelligent behavior of the robot can be modeled that is conclusive and understandable for the interacting human.

Robots that are able to perform a reliable auditory scene analysis can for example be used to work in areas where persons are buried alive (i.e. earthquake or avalanche areas) or which are difficult to access (i.e. caves or dense woods) and are not visible to the robotic or human helper. Hearing sensors that are more sensitive than the human ears can filter out the cries for help of the victims. Additional several robots and humans can work together

in rescue teams, which are communicating with each other by speech and the robots filter out the instructions from their supervisors, while ignoring the other voices around.

Other application areas lie in the home sector. Consider for example a housekeeping robot. If the robot is able to localize and separate speech sources, it can serve the orders of several people that are living in the house without having face-to-face interaction.

Automatic Speech Recognizer (ASR) usually are tuned to work on clean speech sources without interfering sources. To use the existing ASR systems as interpreter for complex auditory scenes, a source separation process which segregates the target source from the interfering sources can be prepended to allow a recognition of the target source. For example in scenes like a driving car where people on the back seats are talking to each other and the driver wants to control the car via speech commands such an approach is useful.

Hearing Prosthesis aim to imitate parts of the auditory system. Examined algorithms and strategies from CASA systems are used to enhance the techniques of current hearing prosthesis. If detailed knowledge about the working of the human auditory system is available and engineering approaches that can be used to imitate these mechanisms are known, hearing aids can be substantially enhanced. Especially the problem of listening to a specific target source has turned out to be a problem, as most hearing aids amplify the whole scene around. An automatic source separation approach improves the capabilities of hearing aids listening to a specific target source.

When using CASA systems in the context of hearing prosthesis, specific attention has to be turned to the time delay of the algorithms. Only algorithms which are able to analyze the auditory scene in real-time can be considered to be used in hearing prosthesis, but the performance of current architectures for CASA is far from being comparable to that of human perception.

Biological Correspondence is the goal of some of the CASA systems. If algorithms are constructed that perform comparable to the human auditory system and exhibit the same behavior in limiting cases, conclusions can be drawn about how the human auditory system works. Further psychological studies can then be initiated to confirm or disprove the hypothesis.

Automatic transcription of musical audio recordings is possible if the heard sound is understood and the single instruments are identified and separated. Then each instrument or voice can be altered in a specific way and remixed to get a new and better sounding recording. Some mistakes of a musical player or a singer can be removed in this way without having to re-record the whole title.

1.1 Cognitive Source Separation

Cherry [20] was the first who investigated the Cocktail Party Effect of the human auditory system by presenting subjects mixtures of several sound sources over a single loudspeaker. Separating different speakers while hearing the intermixed voices only from one loudspeaker turned out to be a very difficult task. This study revealed that the human ability to separate sound sources strongly relies on spatial cues like interaural time and level differences and on source characteristic attributes such as fundamental frequency, amplitude and frequency modulation and harmonicity.

Broadbent [107] conducted dichotic listening experiments and asked subjects to recall digits in the heard sequence for each ear. The subjects were only able to attend to one source and showed very low recall accuracy for the non-attended channels. Broadbent therefore postulated the Filter Theory for selective attention: The human mind separates the incoming sound into channels based on source characteristics and allows only certain channels to pass for a detailed semantic analysis.

Treisman [107] extended the Filter Theory to explain also the recognition of single words in unattended channels like i.e. the own name[1]. Treisman claims that all channels are treated alike. A selective filter reduces the signal-to-noise ratio (SNR) of the unattended channels by splitting the input signal based on gross physical characteristics like i.e. interaural time and level differences or fundamental frequency. The next stages then evaluate the meaning of the channels by activating only those lexical units that exhibit certain thresholds. For important words like i.e. the own name, these thresholds are very low and get activated even in the case of the unattended low-SNR-channel.

Many technical approaches for source separation (i.e. Independent Component Analysis (ICA) [1] or Beamforming [47]) usually rely only on the input signal and do not presume or utilize further information regarding the source mixture. The human auditory system however is linked to many previous hearing experiences and has access to much information regarding the current auditory scene by analyzing the scene by itself and performing mode fusion with other sensual systems like the eyes. This information regarding the auditory scene can then serve as starting point for the separation of the auditory streams.

Consider a person arriving at a cocktail party. When entering the auditory scene, the person automatically analyzes the environment around. The person for example first looks around and estimates coarsely how many other people there are, where these persons are positioned, who they are and if they are known. Additional artificial sound sources like perhaps a running TV or background music are recognized. When source separation is required – i.e. when starting a conversation with another person – a lot of information regarding the auditory scene is already known to the human mind and can be used to enhance the separation process. Sound fragments

[1] Everyone knows the phenomenon talking to someone on a party and suddenly hearing his own name out of the crowd.

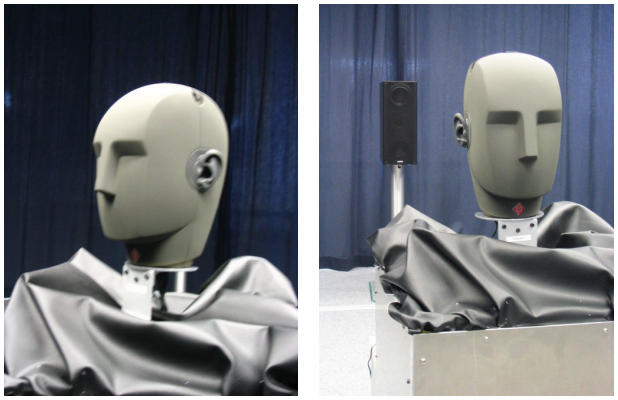

Figure 1.1: *Bob – The robotic head.*

that do not belong to the conversational partner, but are instead identified to emanate from another direction are filtered out based on the already available information of the auditory scene. In this way the attention of the human auditory system is concentrated on the target signal.

The source separation architecture presented in this thesis tries to mimic these cognitive abilities of the human brain and claims that cognitive signal processing approaches such as algorithms relying on a previous scene analysis can enhance the separation process. To imitate the human listening situation closely, a robotic human dummy head, called Bob, is used (see figure 1.1). Bob resides in a normal echoic office room to achieve realistic non-ideal conditions of the auditory scene. The head is able to move in three degrees of freedom and can explore the auditory scene around in human-like manner. Prior to separating the sources, Bob analyzes the auditory scene and estimates several parameters like the positions of the sources and the fundamental frequency. The following source separation approach then uses the gathered information to perform the separation of the target source from the mixture based on cognitive strategies.

1.2 Main Objectives

The main objectives of this thesis can be summarized as follows:

- To enhance the separation results of standard binaural signal processing approaches for source separation by cognitive signal processing mechanisms.
- To find an appropriate way to test these cognitive separation approaches in a realistic non-ideal experiment setup.
- To confirm if the goal of finding the ideal binary mask is sufficient to perform source separation also in humanoid reverberant experiment setups.

- To find out what prior information about the auditory scene is useful to enhance the cognitive computational approach to auditory scene analysis and how this information can be integrated in the whole architecture.
- To implement a cognitive binaural source separation approach that analyzes the auditory scene and integrates the detected information to perform source separation under realistic conditions.

1.3 Thesis Outline

Chapter 2 gives an overview of the field of human auditory scene analysis and computational auditory scene analysis and describes all fundamentals that are required to follow the approaches described in the next chapters. Furthermore existing technical auditory scene analysis and source separation approaches are evaluated.

Chapter 3 investigates the Window-Disjoint-Orthogonality (WDO) of speech signals in the time-frequency domain and evaluates if ideal binary masks are appropriate as final goal for source separation in humanoid reverberant scenarios.

Chapter 4 describes the realistic non-ideal experiment setup used to evaluate the following algorithms.

Chapter 5 shows the analysis of the auditory scene, which is performed prior to the source separation. The human dummy head localizes the sources in the auditory scene and estimates a fundamental frequency track of each speech source in the auditory scene.

Chapter 6 describes the implemented cognitive signal processing approaches for source separation and compares the results to standard binaural source separation approaches.

Chapter 7 sums up the results of the last chapters and gives a short outlook for future directions of the project.

2 Computational Auditory Scene Analysis

2.1 Human Auditory Scene Analysis

The field of Human Auditory Scene Analysis (ASA) deals with the mechanisms and phenomena humans apply to built a mental representation of the auditory scene around them. Based on this mental representation humans are able to percept the outer world in terms of auditory features. Humans can localize different sound sources, segregate them against each other and estimate different characteristics of each source.

Much research in the field of psychology and psychophysics has been done to find out how the human auditory system works and enables such a detailed description of the auditory scene (see for example [13] and [10] for an extensive overview). Many results on the ASA topic arise from medical studies on deaf and hearing impaired people and aim to identify possible enhancements to the hearing loss. Psychological studies contribute to the perceptual organization of sound in the human auditory system.

The human auditory system is coarsely organized as a two-part system [13]: The first part takes the input of both ears and forms internal representations of the auditory scene that specify the characteristics of the single constituents. The second part interprets and processes these estimated representations to perform specific actions or to gain more information. If a person wants to concentrate on one specific signal, the human brain performs source separation by grouping together only those representations that most probably belong to the preferred source.

Auditory Scene Analysis distinguishes between primitive ASA and schema-based ASA [13]. Primitive ASA is assumed to be an innate bottom-up process which analyzes the auditory scene based only on low-level properties, such as the interaural time and level differences between the two ears and builds mental descriptions of the constituents of the scene. Schema-based ASA on the other hand extensively uses former experiences and expectations of heard sounds to estimate the mental descriptions and is learned throughout the whole life.

Wang et. al [119] tested the hypothesis if the understanding of a language is needed to perform speech source separation. Auditory scenes consisting of different speech mixtures in different languages (French, German, Hindi, Japanese, Mandarin and Spanish) were created and presented to subjects that were asked to follow a single speech source. Informal tests showed

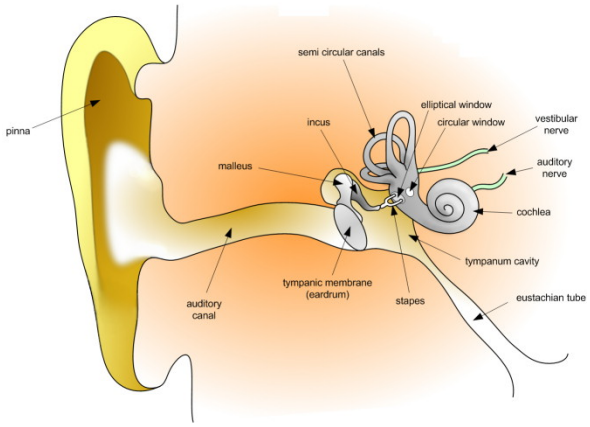

Figure 2.1: *Anatomy of the human ear*[1].

that listeners were able to follow a specific speech source although they didn't understand any of the spoken words. Independent of the language it was much easier to separate two sources with different characteristics in the produced speech such as a male and female voice opposed to similar source characteristics. Wang et. al conclude that language familiarity doesn't seem to be essential for human speech source separation.

Primitive ASA opposed to schema-based ASA is common to all humans, regardless of their culture, mother tongue, intelligence or social background [13]. This thesis therefore focuses on primitive ASA to gain some insight in the human processing of sounds and tries to apply these human mechanisms to an engineering solution for source separation.

2.1.1 Auditory Periphery

The human ear enables the transduction of physical sound waves to human auditory sensation. Figure 2.1 shows the anatomy of the human ear, which can mainly be divided in three parts[2]: The outer ear, the middle ear and inner ear.

The outer ear consists of the pinnae, the ear canal and the ear drum. The pinnae imposes a direction dependent filtering on the incoming sound and the ear canal enhances specific frequency dependent resonances. The ear drum vibrates according to the incoming sound wave and initiates movements of three bones in the air-filled middle ear: The hammer, anvil and stirrup translate the motions of the ear drum to different pressure levels at the circular window to stimulate a watery liquid in the cochlea, which performs a spectral filtering of the incoming signal comparable to a bank of filter channels that are positioned logarithmically on the frequency scale. The fluid inside the cochlea moves in response to this stimulation and flows against receptor hair cells

[1]Image taken from Wikipedia: The human ear, http://en.wikipedia.org/wiki/Ear
[2]Wikipedia: The human ear, http://en.wikipedia.org/wiki/Ear, cited: 21.08.09

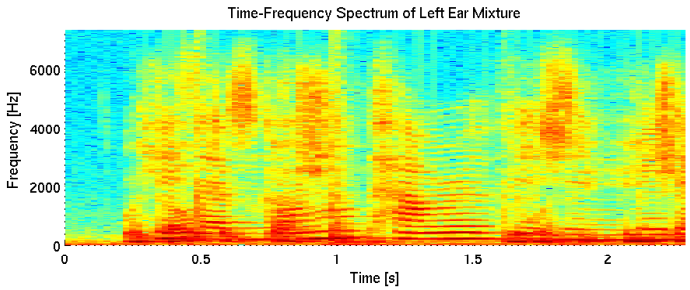

Figure 2.2: *Time-frequency spectrum of a single speech signal.*

which transform the motion to electrical signals. The created electrical impulses are in turn transmitted to thousands of nerve cells which send the information about the incoming sound to the brainstem for further processing.

For detailed information about the anatomy and functioning of the human ear see i.e. [119] or [13].

2.1.2 Perceptual Grouping Principles

Psychological experiments with simple auditory stimuli (i.e. single tones or harmonically related tones) yielded some insight in the process of human auditory scene analysis [119]. As stated above the human auditory system first segments the input signal in different junks by grouping them according to low-level properties. In a second step these segments are integrated to whole streams representing one part of the auditory scene [13].

Analog to the perceptual grouping principles that the Gestalt psychologists propose for the human processing of visual stimuli (see for example [79] for an overview), a number of such grouping principles can be identified for auditory stimuli. To visualize these perceptual grouping strategies, auditory stimuli are commonly represented in a time-frequency spectrum, which describes the frequency content of the signal over time. Figure 2.2 shows such a time-frequency spectrum of a speech signal. The temporal and spectral perceptual grouping principles can be classified as follows [13]:

Proximity in time and frequency. Different components that are closely related in frequency tend to be grouped together as the frequency resolution of the human ear is based on a logarithmical scale and decreases at higher frequencies [13].

The less the frequency difference between two different components, the higher the probability that they are assigned to one auditory stream. The same holds true for components that are closely related in time.

Harmonicity. Harmonic instruments and the human speech (especially voiced parts of human speech) exhibit a clear harmonic structure [13]: The different partials are approximately

integer multiples of the fundamental frequency. The relative location of components in a time-frequency spectrum is used to group the single components.

Onset and Offset. The start and end times of belonging together components are similar, depending on the physics of the source. The start and end times – denoted by onset and offset – are used to group together components over time.

Amplitude Modulation. The temporal amplitude envelope of a source varies time dependent. Depending on the physics of the source, the Amplitude Modulation (AM) of different components across frequency is similar. Auditory streams are formed by grouping together those components with a correlated amplitude modulation.

Frequency Modulation. For a wide range of signals, the single components exhibit similar changes in frequency across time (i.e. pitch changes in human speech). Similar to the grouping principle based on amplitude modulation, the components are grouped together by correlating frequency modulation.

Spatial Location. Signals that emanate from the same direction in the auditory scene are grouped by the human brain. To compute the spatial location of a signal, a minimum of two sensors is needed. Signals from different directions yield specific time and level shifts of the signals received at the left and the right ear. As humans can also separate signals from monaural recordings, the known spatial location of a signal only enhances the source separation, but is not crucial to perform the segregation [119].

2.1.3 Human Source Localization

Humans can estimate the position of a sound source in the auditory scene quite accurate and are sensitive to direction-differences of 2-3 degrees in the fore side azimuth plane [10]. To describe the perceived location of a source in the environment, a polar coordinate system is used. Figure 2.3 shows the three planes that construct the auditory scene: The horizontal plane, the median plane and the frontal plane. A dedicated location is defined by a triple (ϕ, Θ, r), where ϕ specifies the azimuth direction in degree, Θ indicates the elevation direction in degree and r denotes the distance of the incident sound.

For source localization the human brain mainly uses Interaural Time Differences (ITD) and Interaural Level Differences (ILD) [10] of the signals received at the left and the right ear, which arise due to the distance between the ears. This spatial separation enables a sampling of the received signals in the auditory space. The solid head between the two ears introduces diffraction and scattering of the sound waves and accounts for significant head shadows at the ear that is turned away from the sound source.

Figure 2.4 shows the scenario when the head is assumed to be a perfect sphere with radius r_h and the sound waves are supposed to be able to travel through the head without diffraction and

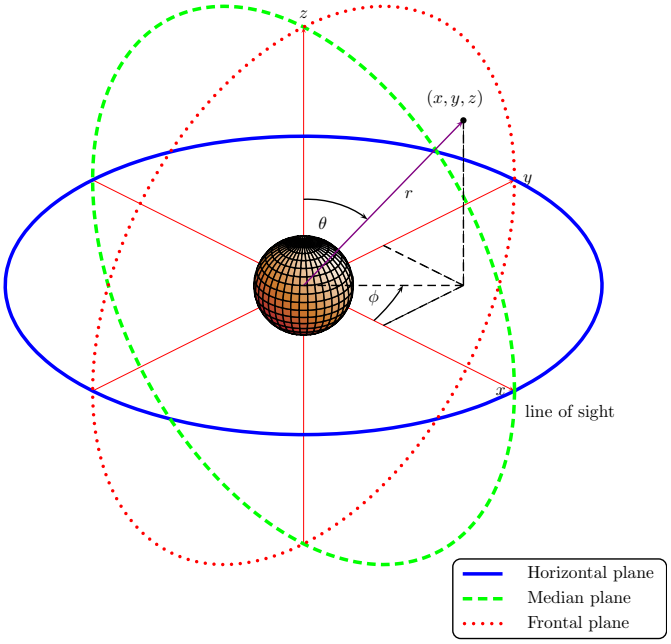

Figure 2.3: *Interaural polar coordinates.*

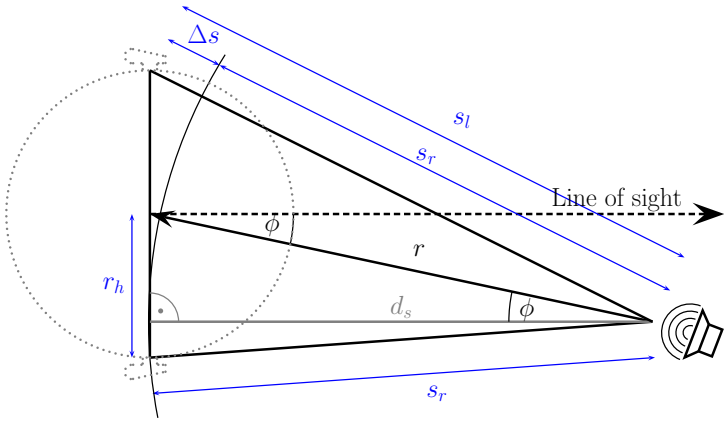

Figure 2.4: *Free-field propagation of sound waves through a transparent head.*

reflection. The distance of the sound emanating source is assumed to be large compared to the head radius. The traveled distances of the left and right signal differ by an amount

$$\Delta s = s_l - s_r, \tag{2.1}$$

which is dependent on the incidence direction of the wave front. Assuming a propagation speed of sound in air of $c = 343\ m/s$, the arrival time difference Δt is proportional to Δs:

$$\Delta t = \frac{\Delta s}{c} \tag{2.2}$$

By applying simple geometric transformations, the interaural time difference between the two ears can be approximated dependent on the incidence angle ϕ and the head radius r_h by the sine law [113]:

$$\Delta t \approx \frac{2r_h \sin\phi}{c} \tag{2.3}$$

In the case of a head modeled as a perfect solid sphere, the sound waves diffract and reflect at the turned-away side. Accounting for the diffraction characteristics, the length of the traveled path of the incident sound wave is longer than in the free-field case. Motivated by this, Woodworth and Schlosberg [122] applied diffraction theory to a completely spherical head, yielding the following formula to approximate the ITD:

$$\Delta t = \frac{r_h(\phi + \sin\phi)}{c} \tag{2.4}$$

The spatial separation of the ears and the head shadow not only affect the arrival times of the signals, but also account for interaural level differences of the signals. The signal at the turned away ear has traveled further and so has lost more energy on its way, which leads to slight level differences dependent on the incidence direction. The head shadow contributes additional level differences, which can be up to 25 dB at high frequencies [113]. Analog to the ITD, the direction dependent ILD can be used to estimate the location of a sound source. But opposed to the ITD values, the ILD values are not well predictable by diffraction theory and depend heavily on the arrival angle, the frequency and the distance of the source [119].

Lord Rayleigh [86] first identified the underlying physics of human binaural hearing. His theory is commonly known as Duplex Theory and describes the human source localization based on the combined evaluation of the physical cues ITD and ILD. In the low frequencies the ITD is used to estimate the direction of the incident sound. In higher frequencies the ITD suffers from phase ambiguities, when the wavelength becomes comparable to the distance of the two ears. For the human head, these phase ambiguities start at approximately 1.5 kHz [10] and the human binaural system begins to loose its localization capabilities based only on ITD in this frequency

range [83]. If high frequency signals ($>$ 1.5 kHz) are modulated with lower frequencies, humans are able to localize these sound events by extracting the ITD of the envelopes of the signals [83].

ILD cues physically exist only for frequencies greater than approximately 500 Hz as signals of lower frequency are not diffracted by the head because of their long wavelength [119]. For frequencies greater than 3 kHz, the ILD cues become a reliable measure of the incidence direction as the signals of short wavelength are not refracted by the human head – they either are reflected completely or pass with little refraction. In intermediate frequencies between 1.5 kHz and 3 kHz neither the ITD cues, nor the ILD cues work reliable, which results in a sensitivity loss of the auditory system in this range [44].

There is psychoacoustic evidence that supports the Duplex theory. ITD cues are primarily used to identify the position of the sound source and ILD cues are used to resolve phase ambiguities in the high frequencies and to avoid possible front-back confusions [119]. Further indications consistent with the Duplex theory are found in psychological experiments. Domnitz and Colburn [28] tested the human just noticeable difference for ITD and ILD cues: The human sensitivity for ITD is in the order of $10\mu s$, for ILD the sensitivity lies in the range of $1dB$. Humans are insensitive to ITD changes above 1.5 kHz, but sensitive to ILD changes in the complete frequency range.

The ITD estimation heavily relies on the coherency of the left and right ear signals. In reverberant environments this coherency cannot be ensured as reflections and other sources in the auditory scene smear and disturb the signals. Almost every reflecting surface has the acoustical characteristic of increasing absorption with increasing frequency [44]. This results in fewer and weaker reflections of high frequency signals that arrive at the human ear. The human auditory system can use the ILD cues across the complete audible frequency range and so it is advantageous to use the highest available and audible frequency to estimate the location of a source based on ILD. These frequencies are considered to contain least power from reflections as high frequencies loose more power during reflection [44]. Experiments found out that humans in fact use ILD cues above 8000 Hz for source localization in reverberant environments [44].

In engineering applications the ITD is often computed by cross-correlating the left and right ear signal and extracting the highest peak as ITD estimate. Current neurophysiological research [72] found evidence that also the human brain is able to perform correlation based methods to estimate the time differences. There are special cells in the inferior colliculus in the brain stem which are maximally sensitive to a specific ITD, independent of the frequency of the incoming signal. There also exist special cells that are tuned to respond to specific ILDs [119].

The described mechanisms for human source localization can only localize sources in the fore side azimuth plane. Signals coming from the back are localized erroneously at the mirrored frontal position. Humans can easily estimate, if the incident waves are coming from behind or ahead. The pinnae of the human ear resolves this front-back-confusion by applying a frequency and directional dependent filtering on the incoming waves and filters signals coming from the

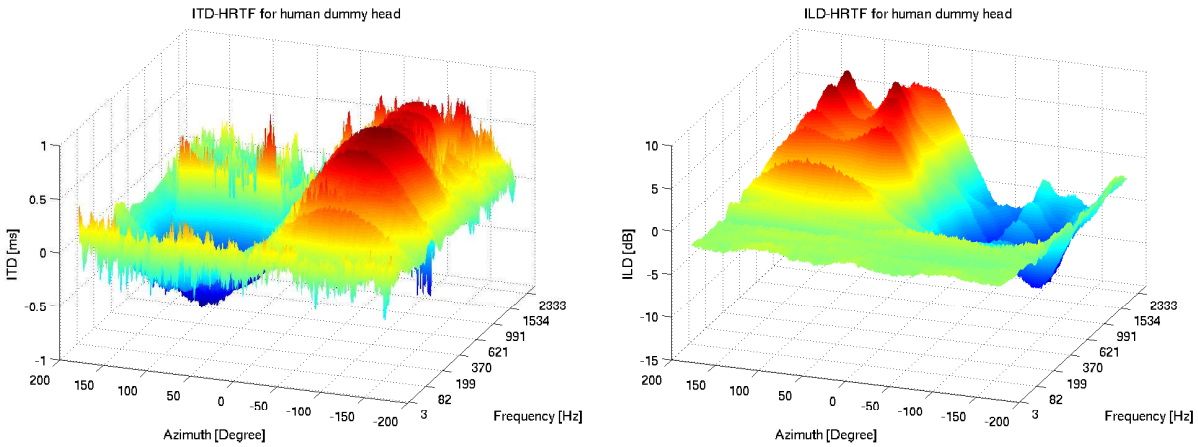

Figure 2.5: *HRTF for ITD and ILD for dummy head Bob residing in a normal office room.*

front in a different manner than signals coming from the back. The human auditory system tends to boost waves coming from behind in the 1000 Hz region and waves coming from the forward direction in the frequency region near 3000 Hz [10]. Additional to the frequency and direction dependent filtering of the outer ear, humans use head motions to resolve ambiguities [10]. A slight movement of the head yields a specific change in the ITDs and ILDs and is used to estimate, if the source is coming from the front or the back.

The Head Related Impulse Response (HRIR) of the ear is measured to specify the complete refraction and resonance characteristics of the outer ear. White noise is played back over loudspeakers from different directions and gets filtered by the human ear and the room, the human head is residing in. The frequency-domain analogon – the Head Related Transfer Function (HRTF) – describes the ITDs and ILDs dependent on frequency and incidence angle. Figure 2.5 shows the HRTF of a human dummy head in a normal office room. The left figure shows the interaural time differences that arise due to the different spatial locations of the source. The ITD values of each frequency channel can be roughly approximated by sinusoids of different frequencies and amplitudes as shown in detail in chapter 5.1.2. The right figure specifies the ILD values, also in dependency of the direction of the sound source and the frequency. Analog to the ITD, the ILD values can roughly be approximated by sinusoids.

If the HRTF of a human head is known, the localization of a source in the auditory scene can be approached by using the HRTF as a table look-up. For each ear, the ITD and ILD of the current sound is measured and compared to the HRTF to find the dedicated position (see i.e. [113]). Problems with this scheme arise due to the severe varying nature of the HRTF. The HRTFs of several persons differ drastically and so one specific HRTF cannot be used to describe the localization abilities of several people [114]. Mean HRTFs yield bad results, although humans are able to adapt to a different HRTF by learning over several experiments [119]. Besides the

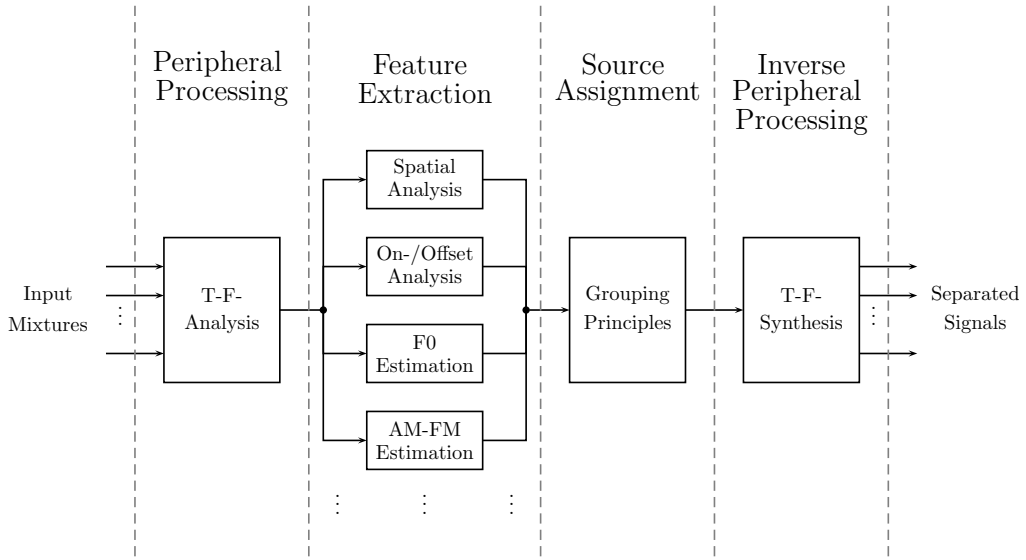

Figure 2.6: *Schematic overview of a common CASA pipeline.*

characteristic head, shoulder and outer ear forms that contribute to the HRTF, the current environment is responsible for many characteristics of the HRTF and the HRTF varies for different environments and even for different head positions in the same environment.

The elevation estimation of a sound source is far more difficult than the azimuth detection. The human auditory system performs only moderately in this task and is able to localize the elevation of a source with a localization blur of approximately 9° [10]. The localization capabilities arise – as in the case of front-back-confusion – due to the specific directional filtering of the outer ear, that imposes a spectral coloration of the signal. Especially in the 4000 Hz range, the outer ears and the pinnae become significant scatterers [10]. Above 6000 Hz the filtering becomes very individually, but shows prominent peaks for different elevations [10]. Analog to the HRTF in the horizontal plane, a HRTF for the vertical plane can be constructed to discover systematic variations as a function of elevation and to estimate the position of a sound source in the vertical plane.

2.2 Basics of CASA Systems

Analog to human ASA many CASA systems (see for example [119], [75], [91]) follow a straight forward approach to analyze an auditory scene and to perform source separation. Figure 2.6 shows a common CASA system that consists of the following succeeding stages that process the input mixtures:

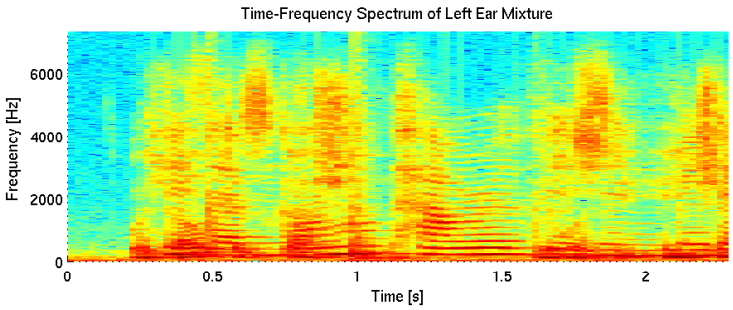

Figure 2.7: *STFT spectrogram of a speech signal.*

1. In a first stage, the input mixtures are converted to a Time-Frequency (TF) representation, that specifies the frequency content of the signals over time. The most frequently used representations are the Short-Time-Fourier-Transform (STFT) and the cochleagram which are described in detail in the next sections. The STFT analyzes the input signal in linearly spaced filter channels, while the cochleagram approximates the peripheral processing of the human ear by a bank of logarithmically spaced gammatone filters.

2. The second stage extracts the features and estimates the characteristics of the single parts of the signal in the time-frequency domain. This stage could for example estimate the spatial location of a time-frequency bin, extract the fundamental frequency of the signal at a specific time or compute the onsets and offsets and the amplitude and frequency modulation of the signal over time.

3. The third stage constructs auditory streams that consist of time-frequency regions belonging to the same source. The features estimated by stage 2 are used to group together those regions that have the same characteristics and do most probable belong to the same source.

4. The last stage reconstructs the estimated auditory stream by converting the time-frequency regions belonging to the source of interest back to the time-domain. Dependent on the intended use of the CASA architecture, this stage is sometimes omitted.

2.2.1 Time-Frequency Representations

Time-Frequency representations are used to visualize the received signals. The frequency content of the signal is plotted against the time, which reveals several characteristics of the signal in the time-frequency plane. The human cochlea also performs a time-frequency analysis to analyze the incoming signals [13].

Short-Time-Fourier-Transform

A commonly used TF-representation for speech analysis is the lossless and computationally efficient Short-Time-Fourier-Transform (STFT) . The STFT is computed by taking the Fourier transform of short segments of the time domain signal at fixed time intervals. The segments are obtained by multiplying the time-domain signal with a finite window function. The STFT analyzes a continuous time-domain signal $x(t)$ in linearly spaced frequency channels up to the Shannon frequency:

$$X(t,f) = \int_{-\infty}^{\infty} w_a(t)x(t+\tau)e^{-i2\pi ft}dt, \tag{2.5}$$

where $w_a(t)$ denotes an arbitrary analysis window function and τ specifies the running index, which shifts the window to the fixed time intervals. The corresponding spectrogram shown in figure 2.7 is obtained by taking the logarithm of the magnitude of each value.

For a general discrete signal $x(n)$ and an arbitrary discrete analysis window function $w_a(n)$, the STFT degrades $\forall q \in \{0,1,...,N-1\}$ to

$$X(k,q) = \frac{1}{\sqrt{N}} \cdot \sum_{n=0}^{N-1} w_a(n)x(n+k)e^{-i2\pi \frac{qn}{N}} \tag{2.6}$$

and computes the Fourier transform at uniform time and frequency intervals. The STFT can be regarded as a bank of filters, that filters out the signal parts with corresponding frequencies. The frequency response of a single filter channel is obtained by modulating the channel center frequency with the window function:

$$\begin{aligned} X(k,q) &= \frac{1}{\sqrt{N}} \cdot \sum_{n=0}^{N-1} w_a(n)x(n+k)e^{-i2\pi \frac{qn}{N}} && (2.7)\\ &= \frac{1}{\sqrt{N}} \cdot \sum_{n=0}^{N-1} x(n+k)\left(w_a(n)e^{-i2\pi \frac{qn}{N}}\right) && (2.8)\\ &= \frac{1}{\sqrt{N}} \cdot \left(x(n) * (w_a(n)e^{-i2\pi \frac{qn}{N}})\right) && (2.9)\\ &= \frac{1}{\sqrt{N}} \cdot (x(n) * h_{w_a}(k,q)) && (2.10) \end{aligned}$$

Figure 2.8 shows the impulse response $h_{w_a}(k,q)$ and the positive frequency response $H_{w_a}(k,q)$ of the STFT for a Hamming window of length 32 using a sampling frequency of 3.2 kHz. The shape of the linearly spaced filter channels and the overlap between two consecutive channels is specified by the shape of the analysis window function. Figure 2.9 shows the frequency responses for several other window functions, which differ in the overlap and sidelobes of the channels.

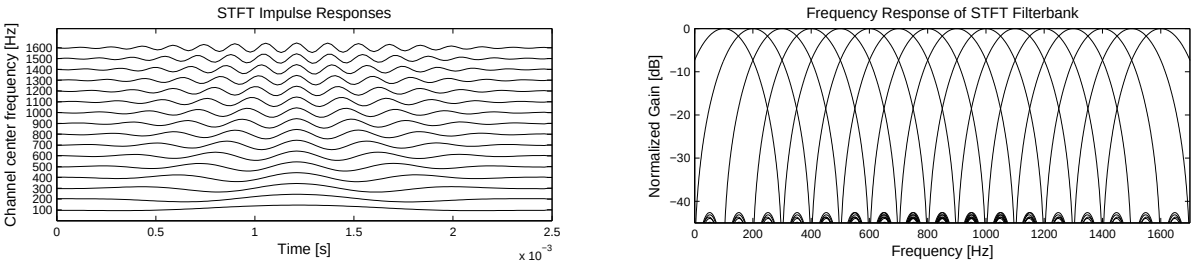

Figure 2.8: *Impulse and Frequency Response of STFT using a Hamming window.*

The frequency resolution of the discrete STFT is determined by the length of the analysis window. Choosing longer windows improves the frequency resolution, but then the amplitudes and frequencies are averaged over a longer period and the time-resolution of the STFT decreases. According to the uncertainty principle, the STFT can either have a good time resolution or a good frequency resolution [81]. The optimal trade-off between time and frequency resolution under the given constraints is fulfilled by the Gaussian window, which produces minimum uncertainty in the time-frequency representation [23].

A discrete STFT spectrogram can be transformed back to the time domain by applying the inverse discrete STFT (ISTFT), that uses a possibly different discrete synthesis window function w_s:

$$x(n) = \sum_k w_s(n-k) \sum_{q=0}^{N-1} X(k,q) e^{i2\pi \frac{qn}{N}} \tag{2.11}$$

The ISTFT is able to reconstruct the signal perfectly up to a constant scaling factor, when the analysis and synthesis window functions w_a and w_s satisfy the following condition [113]:

$$\sum_k w_a(n-k) w_s(n-k) = C, \quad \forall n. \tag{2.12}$$

Almost all energy of speech signals is distributed in frequencies up to 8 kHz. For analyzing speech signals, a finer frequency resolution in the low frequency range is favorable, whereas in higher frequencies a coarse resolution is sufficient. Because the STFT analyzes linearly up to the Shannon frequency, the frequency resolution in the low frequencies cannot be enhanced by increasing the sampling rate.

Cochleagram

Many source separation architectures try to imitate the frequency analysis of the human auditory system. A commonly used computer model that approximates the frequency selectivity of the

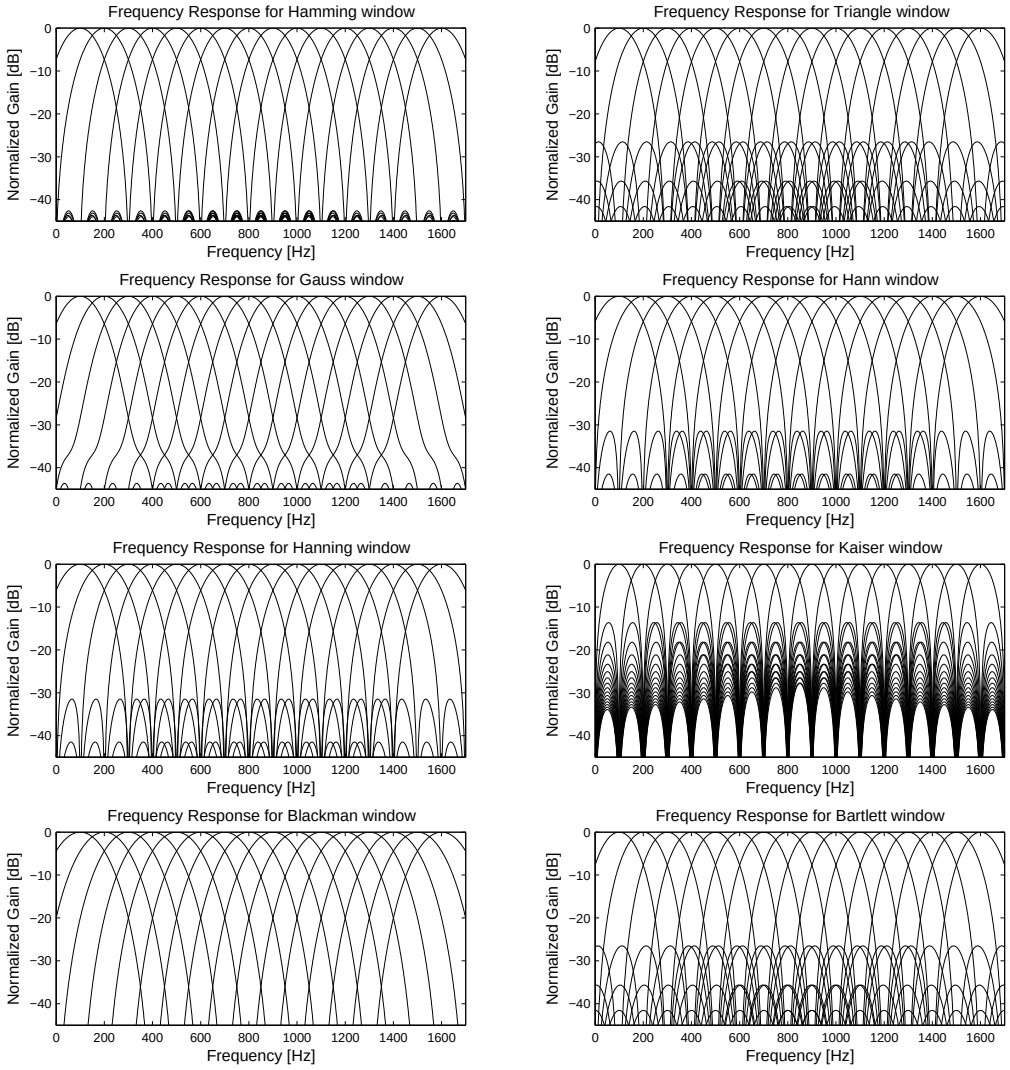

Figure 2.9: *Frequency Response of STFT for different windows.*

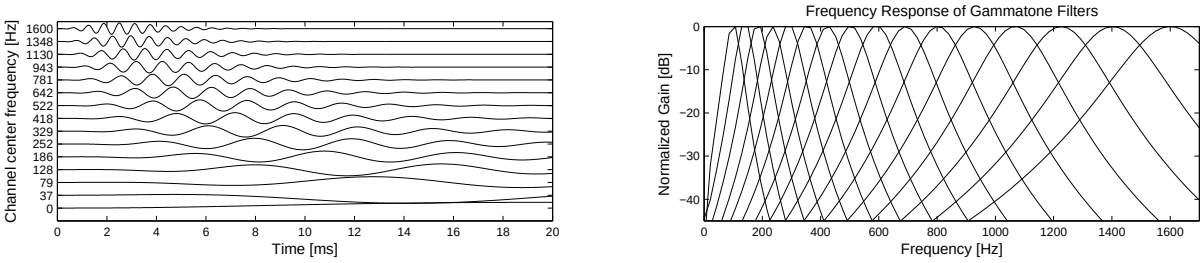

Figure 2.10: *Impulse and Frequency Response of Gammatone Filterbank.*

human peripheral auditory system is the gammatone filter that has been popularized by Johannesma [53] and since then has been used by many researchers (i.e. [75], [118], [96], [14], [106]). The frequency analysis of the human cochlea is approximated using a bank of different gammatone filters. The impulse response of a gammatone filter is defined as the product of a gamma function and a tone [119]:

$$g_{f_c}(t) = t^{N-1} e^{-2\pi b(f_c)} \cdot \cos(2\pi f_c t + \phi) \quad \forall t \geq 0 \tag{2.13}$$

where N denotes the order of the filter and f_c denotes the center frequency of the filter. The value $b(f)$ determines the bandwidth of the filter channels and is set to the Equivalent Rectangular Bandwidth (ERB) of human auditory filters. The ERB defines the bandwidth of an ideal rectangular filter that passes the same total power and has the same peak gain for white noise. The bandwidth function $b(f)$ is defined in terms of the ERB as [119]

$$b(f) = 1.019 \cdot ERB(f), \tag{2.14}$$

where

$$ERB(f) = 24.7 + 0.108f. \tag{2.15}$$

A bank of such gammatone filters gives a good fit to experimentally derived estimates of the frequency analysis of the human cochlea [53] and for such the spectrogram resulting from the logarithm of the magnitude of each time-frequency bin is commonly called cochleagram.

The frequency response of a single filter channel of the cochleagram can be approximated by [119]

$$G(f) \approx \left(1 + \frac{j(f - f_c)}{b(f_c)}\right)^{-N} \qquad \forall f \in [0, \infty] \tag{2.16}$$

where $f_c/b(f_c)$ is assumed to be sufficiently large.

Figure 2.10 illustrates the impulse and frequency response of a bank of 16 gammatone filters in the frequency range from 100 - 1600 Hz. The filtering process of a gammatone filterbank is

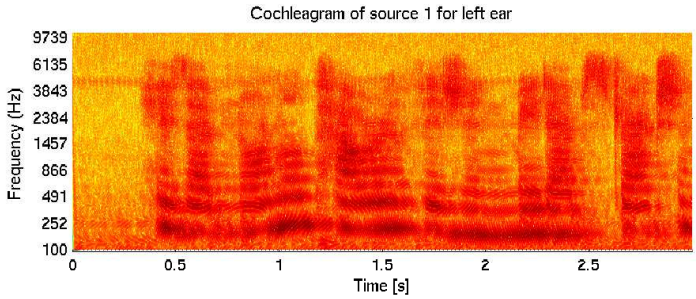

Figure 2.11: *Cochleagram of a speech signal.*

similar to a wavelet transformation [81]. The basis functions are scaled and compressed versions of the kernel function of the first channel and consecutive filters are spaced logarithmically on the frequency scale. Filter channels in the low frequencies have fine frequency resolution, but coarse time resolution. Contrariwise the high frequency channels have coarse frequency resolution, but fine time resolution. The coarse time-resolution in the low frequencies is acceptable as signals consisting of low frequencies change slowly, whereas high-frequency signals need finer time-resolution to illustrate the rapid changes.

The invertion of a given cochleagram to a time-domain signal is non-trivial and lossy. There exist some approaches that yield quite good invertion results (i.e. [120], [14]), but these are complex to compute and only approximately orthogonal, which result in non-perfect reconstruction. Weintraub [120] for example first compensates the across channel phase shifts by reversing the response of each filter channel. Then the reversed response is passed back through the filter and time reversed again, to yield a phase corrected output from each filter channel [119]. This output is then windowed according to time units and summed over all frequencies to reconstruct the original signal.

2.2.2 Correlogram

The interaural time difference between two signals emanating from one source is used to estimate the location of the source and can be measured by computing the crosscorrelation of the signals:

$$R(l) = \sum_{t=t_s}^{t_e} x_L(t+l) \cdot x_R(t) \tag{2.17}$$

where x_L and x_R specify the left and right ear signal and t_s and t_e denote the start and end time of the signals. The estimated ITD is computed as the time lag of the highest peak of $R_{x_L x_R}$.

If the two ear signals include energy from several spatially separated sources, the ITD estimation based on the correlation function is not necessarily correct. In the ideal case for i.e. two sources, there will be two peaks in the correlation function corresponding to the positions of the

two sources. In most cases – especially in reverberant environments – these two peaks merge to one peak at an intermediate position, so neither the position of the first, nor the position of the second source can be estimated reliably.

Wang et al. [75] [80] [119] describe an approach to enhance the localization capabilities, by computing the crosscorrelation separately for each frequency channel of the time-frequency representation. Additionally only short time windows of the signal are regarded instead of the whole signal. Computing the described function for each channel and each time window results in a four dimensional function

$$R(l, c, \tau) = \sum_{t=t_s}^{t_e} x_L(t + \tau + l, c) x_R(t + \tau, c) w(t) \tag{2.18}$$

where l denotes the time lag between the windowed left and right signal, c represents the frequency channel and τ is the current position of the window in the whole signal. A plot of a specific time window of the function $R(l, c, \tau)$ over all frequency channels is called correlogram. Figure 2.12 shows the correlogram of a speech mixture where the sources are positioned at azimuth position 0 degree and 45 degree. For clarity each correlation function in figure 2.12 is normalized to a maximum value of one as described in [119]. The correlation peaks at position 0 and 45 can clearly be seen across all frequencies, but in the higher frequencies other candidates occur due to phase ambiguities. To find an estimate of the positions across all frequencies, the correlation functions of each channel are summed, which yields the position estimates seen in the lower plot of figure 2.12.

When localizing nearby sources using the correlogram, the described technique sometimes fails as the peaks in the correlogram are broad and the summation results in a single broad peak instead of two separate peaks. Wang et al. [75] [80] resolve these ambiguities by introducing an improved correlogram where the highest peaks of the correlation functions are replaced by impulses of the same height and are then convolved with Gaussians. The width of the Gaussians is chosen to be inversely proportional to the channel center frequency. Compared to the conventional correlogram, the peaks in the improved correlogram are narrower and the summation of all correlation functions achieves a finer resolution.

Figure 2.13 shows the conventional and the improved correlogram for a mixture of two speech sources located at positions 0 and 20 degree. The conventional correlogram identifies only one peak at position 0, but fails to detect the second peak. The improved correlogram on the other hand correctly detects both locations.

2.2.3 Fundamental Frequency Estimation

Humans tend to use frequencies that are an integer multiple of their fundamental frequency (F0) [13] . The estimation of the fundamental frequency is essential for many speech and music

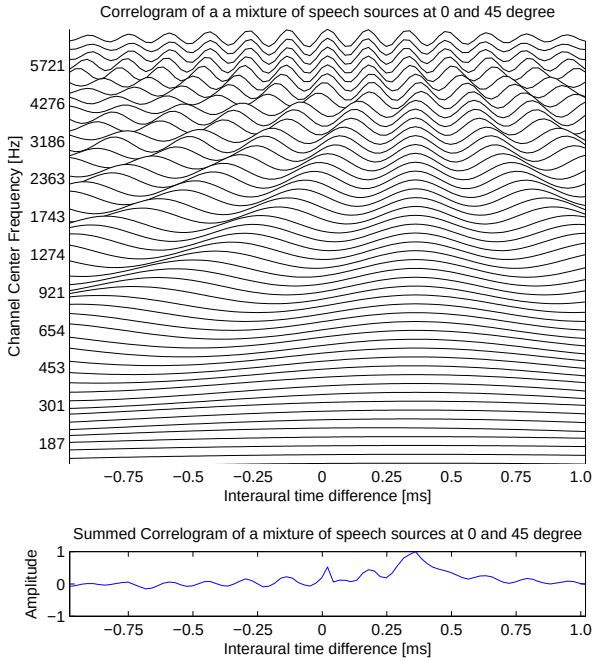

Figure 2.12: *Correlogram of a speech mixture with sources positioned at azimuth positions 0 and 45 degree.*

processing architectures like source separation approaches. The F0 is variously used to extract information patterns for speech or to transcribe musical scores [119]. In CASA systems the F0 estimation often is an essential part of the whole architecture (i.e. [49], [16], [95], [124]).

Over the years there has been a vast amount of research about extracting the F0 of an isolated voice (see for example [89] for an overview). The harder task of estimating the fundamental frequencies of several sources in a mixture has gained less interest, but the demand on this topic grows, as in the last years new applications arose that depend on the knowledge of multiple F0s. Musical indexing for example often relies on the estimation of the F0s of the single instruments and voices.

Many approaches for multiple F0 estimation are connected upstream to a source separation scheme [119], that splits the mixture in single sources, on which the single source F0-estimation algorithms are applied. On the other hand, many source separation architectures rely on the knowledge of the underlying F0s. So this situation leads to the famous "chicken or egg" problem and the task is to find an optimal entry point to the recursive computation.

A signal is periodic if for all t, there exists a T such that

$$\exists T \neq 0 : \forall t : x(t) = x(t+T). \tag{2.19}$$

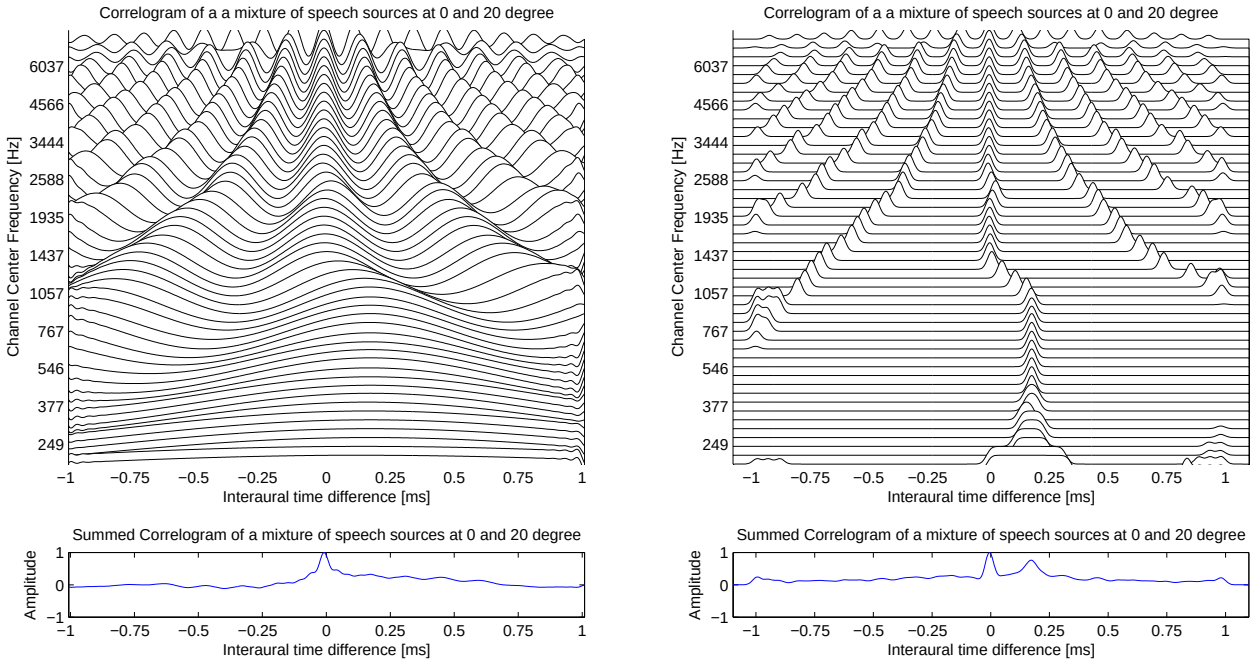

Figure 2.13: *Correlogram and improved correlogram of a speech mixture with sources positioned at azimuth positions 0 and 20 degree. The conventional correlogram fails in detecting the locations of both sources, while the improved correlogram correctly detects the positions.*

The inverse of the period T is the fundamental frequency $F0$:

$$F0 = \frac{1}{T}. \tag{2.20}$$

For real world signals, this ideal condition is usually not satisfied because of noise and reverberation in the signal and so the period and the fundamental frequency of a recorded speech source can only be estimated approximately. The fundamental frequency used by humans is not constant over time and varies up to 30 Hz during a whole sentence [13]. To account for the change in the F0, most F0-algorithms (i.e. [108], [22]) estimate the fundamental frequency in small time frames and construct a complete track of the fundamental frequency over time. The mean F0 of this track corresponds to the mean F0 of the speaker, but in general this value is not useful because of the large variations of the fundamental frequency over time.

F0-estimation schemes can be divided in temporal and spectral approaches. The temporal approaches mainly use the Autocorrelation Function (ACF) or the Squared Difference Function (SDF) – a variant of the ACF – to estimate the fundamental frequency based on the periodicity of the time domain signal. Spectral approaches assess the F0 by using combined evaluations of the position of the spectral peaks and relative spacings between them or utilize pattern matching algorithms to extract the F0.

Licklider [64] and Rabiner [84] first used the autocorrelation function for fundamental frequency estimation of speech sources. The autocorrelation function is defined by computing the correlation of a specific signal with time-shifted versions of itself:

$$R_A(l) = \sum_t x(t+l) \cdot x(t) \tag{2.21}$$

The resulting function exhibits peaks at all multiples of the period of the signal. The fundamental frequency F0 can be estimated by localizing the first global peak greater than zero. The middle plot of figure 2.14 shows the autocorrelation function for the signal $x(t)$ depicted in the left plot. It can be seen clearly that for periodic signals – such as $x(t)$ – the autocorrelation function shows peaks at multiples of the period T.

Opposed to the autocorrelation function that measures the similarity of a signal, the squared difference function (SDF) measures the difference of a signal with time-shifted versions of itself:

$$D(l) = \sum_t (x(t) - x(t+l))^2 \tag{2.22}$$

The squared difference function exhibits dips at multiples of the period T. The F0 is therefore estimated by extracting the first global dip greater than zero in the squared difference function. The right plot of figure 2.14 shows the SDF for the signal $x(t)$. The SDF function is more robust against amplitude changes of the input signal opposed to the autocorrelation function

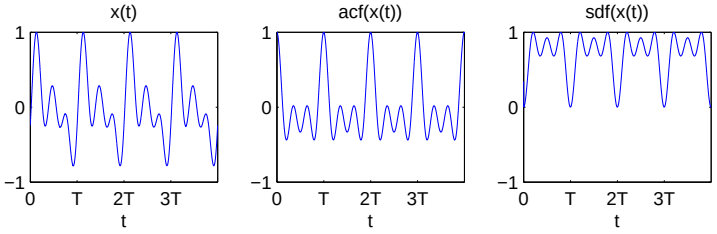

Figure 2.14: *Comparison of the autocorrelation function and the squared difference function. The left plot shows the signal $x(t)$. The middle figure plots the autocorrelation function. The F0 of the signal $x(t)$ can be estimated by finding the first highest peak. The right plot shows the squared difference function, where the F0 can be determined by extracting the first dip in the function.*

and is therefore preferred to estimate the fundamental frequency of real world signals such as speech (for details see [59]). The famous YIN-method [22] for example uses the squared difference function to estimate the F0.

While the estimation of the fundamental frequency of a single source can be achieved quite well with the described methods, the F0-estimation of several sources of a mixture is more difficult. The spectral overlap between the sources distorts the cues used by the single source F0-estimation algorithms [119].

Approaches of multiple voice F0 estimation algorithms can be divided in iterative estimation algorithms and joint estimation algorithms. Iterative algorithms apply a single source F0-estimation algorithm to detect one F0 and then suppress this frequency and the harmonics from the signal, so that there are only signal parts left belonging to other sources. The same procedure is applied to the remaining signal until all F0s are detected. Iterative time domain methods often use comb filters to suppress specific frequencies (i.e. [21]) while iterative frequency domain methods eliminate the corresponding peaks in the spectrum (i.e. [82]).

For a detailed description of temporal and spectral methods for F0-extraction of single and multiple sources, the reader is referred to the extensive background literature like [119] or [89].

2.2.4 Onset and Offset Detection

Depending on the physics of the source, the start and end times of the spectral components of a speech source in a time-frequency representation are more or less the same. The onsets and offsets are used by several researchers (i.e. [50] [14]) to segregate different sources by grouping together those components with same onsets and offsets.

Auditory segmentation corresponds in many parts to image segmentation, where the main task is to detect the edges of the visual objects of the image. The edges of a visual object are defined by a large increase in the color intensity. A commonly used technique for edge detection

is to compute the first-order derivatives of the intensity over the whole image and then finding the edges by identifying the peaks and valleys in the resulting derivatives [119].

The onsets in a time-frequency representation are consistently characterized by a large increase in the intensity of the signal amplitude and the offsets are specified by an abrupt decrease in the signal intensity. Analog to the edge detection in image segregation, the on- and offsets in CASA can be detected based on the first-order derivatives of the intensity.

Hu and Wang [50] use a three step process similar to the Canny edge detector [19] used in image segregation to obtain the onsets and offsets of one frequency channel. In a first step the signal intensity is smoothed by convolving with a Gaussian function to smear out little variations that would lead to false onsets and offsets. In the second step the first-order derivatives of the intensity function are computed and the peaks and valleys and the corresponding points in time are identified. The third step uses absolute onset and offset thresholds to mark those points that are considered as valid on- and offsets.

Brown and Cooke [14] apply an onset detection kernel consisting of a negative side succeeded by a positive side to the envelope of the signal intensity instead of the signal intensity itself to estimate the on- and offsets. Prior to the onset and offset detection the output of the cochleagram is filtered with a hair cell model that simulates the nerve firing rates of the human hair cells to the stimulus.

2.2.5 Amplitude and Frequency Modulation Detection

The amplitude envelope of a speech or music signal varies over time. This Amplitude Modulation (AM) is dependent on the physics of the source and can be used to group components across the frequency range by grouping those components with a correlated amplitude modulation [119].

For a wide range of signals such as human speech, the spectral components exhibit similar changes in frequency across time. In human speech for example the fundamental frequency of the speaker is not constant and varies in the harmonics the same way as in the fundamental frequency [13]. Similar to the grouping principle based on AM the spectral components can also be grouped together by correlating Frequency Modulation (FM) .

The amplitude modulation of a specific filter channel corresponds to the envelope of the signal of this channel. To compute the envelope of a real valued signal $x(t)$, the analytic representation $x_a(t)$ is used to describe the signal:

$$x_a(t) = x(t) + j \cdot \mathcal{H}(x(t)) \tag{2.23}$$

where $\mathcal{H}(x(t))$ is the Hilbert Transform of $x(t)$ [78]

$$\mathcal{H}(x(t)) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau}\, d\tau. \tag{2.24}$$

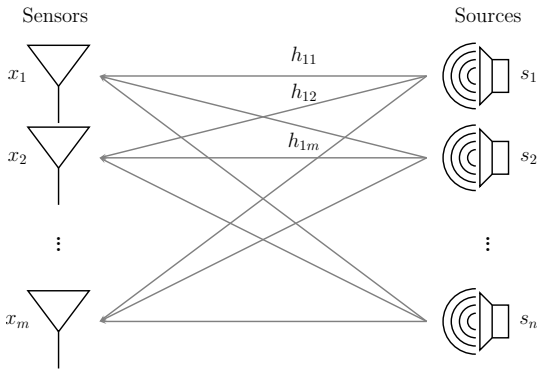

Figure 2.15: *Schematic overview of source separation scenario.*

The analytic signal can be expressed in polar notation by

$$x_a(t) = A(t)e^{j\phi(t)} \tag{2.25}$$

where $A(t)$ specifies the envelope of $x(t)$ and corresponds to the amplitude modulation of the signal. $A(t)$ can be computed as the absolute value of the analytic signal.

$$A(t) = |x_a(t)| \tag{2.26}$$

The instantaneous phase

$$\phi(t) = \measuredangle(x_a(t)) \tag{2.27}$$

corresponds to the phase of the signal at time t. The time derivative of the unwrapped instantaneous phase is called the instantaneous frequency $\omega(t)$ [78]

$$\omega(t) = \frac{d}{dt}\phi(t). \tag{2.28}$$

The envelope and the instantaneous frequency of each filter channel for specific time windows can be used to estimate the amplitude and frequency modulation of the corresponding signal parts and to group together those components.

Some researchers (i.e. [7], [4], [56]) construct a complete AM spectrum to define the amplitude modulation in each frequency channel over time. Voiced speech produces a high response of the AM spectrum near the fundamental frequency and its harmonics [119] and the AM spectrum is used to identify spectral components belonging to a specific source.

2.3 Separation Based On Spectral Segmentation

The goal of all source separation architectures is to estimate the underlying sources only from one or more mixtures of the sources. Figure 2.15 shows a schematic overview of the anechoic source separation scenario. Up to m sensors receive mixtures of the n sources in the auditory scene. On the way from the sound source origin to the sensors the signals get filtered by complex weighting factors h_{ij} which account for the amplitude and phase deviations of the single signals.

Denoting the signals received at the m sensors by $x_1, x_2, ..., x_m$ and the n sources by $s_1, s_2, ..., s_n$, the anechoic mixing model can be defined as

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & & \vdots \\ h_{m1} & h_{m2} & \cdots & h_{mn} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} + \eta \tag{2.29}$$

where η accounts for additive noise. In matrix notation the anechoic model becomes

$$x = H \cdot s + \eta.$$

Many source separation approaches (i.e. Independent Component Analysis) try to estimate the mixing matrix H and demix the single signals from the mixtures by multiplying the mixtures with the inverse H^{-1} of the mixing matrix. For scenarios with a less or equal number of sources than sensors this approach works, but in scenarios with more sources than sensors the matrix inversion technique cannot be applied as the problem gets degenerate and the inverse of the mixing matrix doesn't exist. In reverberant environments the complex weighting factors h_{ij} turn in complete Finite-Impulse-Response (FIR) filters which raise the complexity of the separation process drastically.

To overcome the limitations of the degenerate case, a representation of the sources has to be found in which all sources have disjoint support. Then the single sources can be demixed by choosing only those signal parts that belong to the specific source. Potential representations have to fulfill the disjoint support of several sources and have to be invertible so that the sources can be demixed from the mixtures. For speech signals, time-frequency representations such as the STFT and the cochleagram have turned out to fulfill these requirements in an approximate way as shown in the next sections.

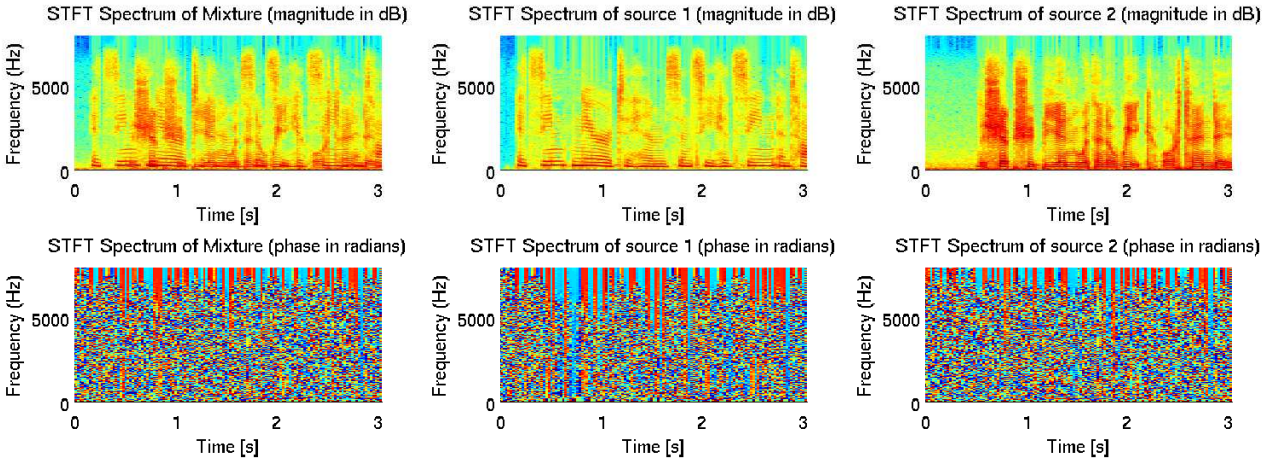

Figure 2.16: *The left plot shows the STFT-spectrum of a mixture of a female and male speech source, while the middle and the right plot show the single sources (female and male).*

2.3.1 Ideal Masks as Goal of CASA

Rickard et al. [127] showed that speech signals are sparsely distributed in high-resolution time-frequency representations. Time-Frequency (TF) representations of different speech signals overlap only in few points and so are approximately orthogonal to each other. This approximate orthogonality in the TF-domain justifies the use of TF-masks that emphasize regions of the TF-spectrum that are dominated by a specific source and attenuate regions dominated by other sources or noise.

Several researchers in computational source separation suggest the ideal binary mask as final goal of computational source separation algorithms (i.e. [117] [75] [127]). Each entry of the TF-mask is set to one if the target energy in this TF-bin is greater than the interfering energy. The binary decision is motivated by masking effects of the human auditory system: Within a critical bandwidth humans don't recognize sounds that are masked by louder sounds [13].

Assume $s_i(t,f)$ denotes the energy of the target signal$_i$ in TF-bin at time t and frequency f and $n_j(t,f)$ denotes the energy of the j-th interfering signal in this TF-bin. The ideal binary mask $\Omega_i(t,f)$ for target source$_i$ and a threshold of x is defined as follows:

$$\Omega_i(t,f) = \begin{cases} 1 & s_i(t,f) - n_j(t,f) > x \quad \forall j \\ 0 & \text{else} \end{cases} \tag{2.30}$$

Figures 2.16–2.18 show the concept of the ideal binary mask for the STFT representation. Figure 2.16 displays the STFT-spectra of a mixture of a female and male speech source and the single sources. Figure 2.17 coarsely illustrates the approximate orthogonality of these two speech recordings in the STFT domain by showing those bins that include only energy from the target

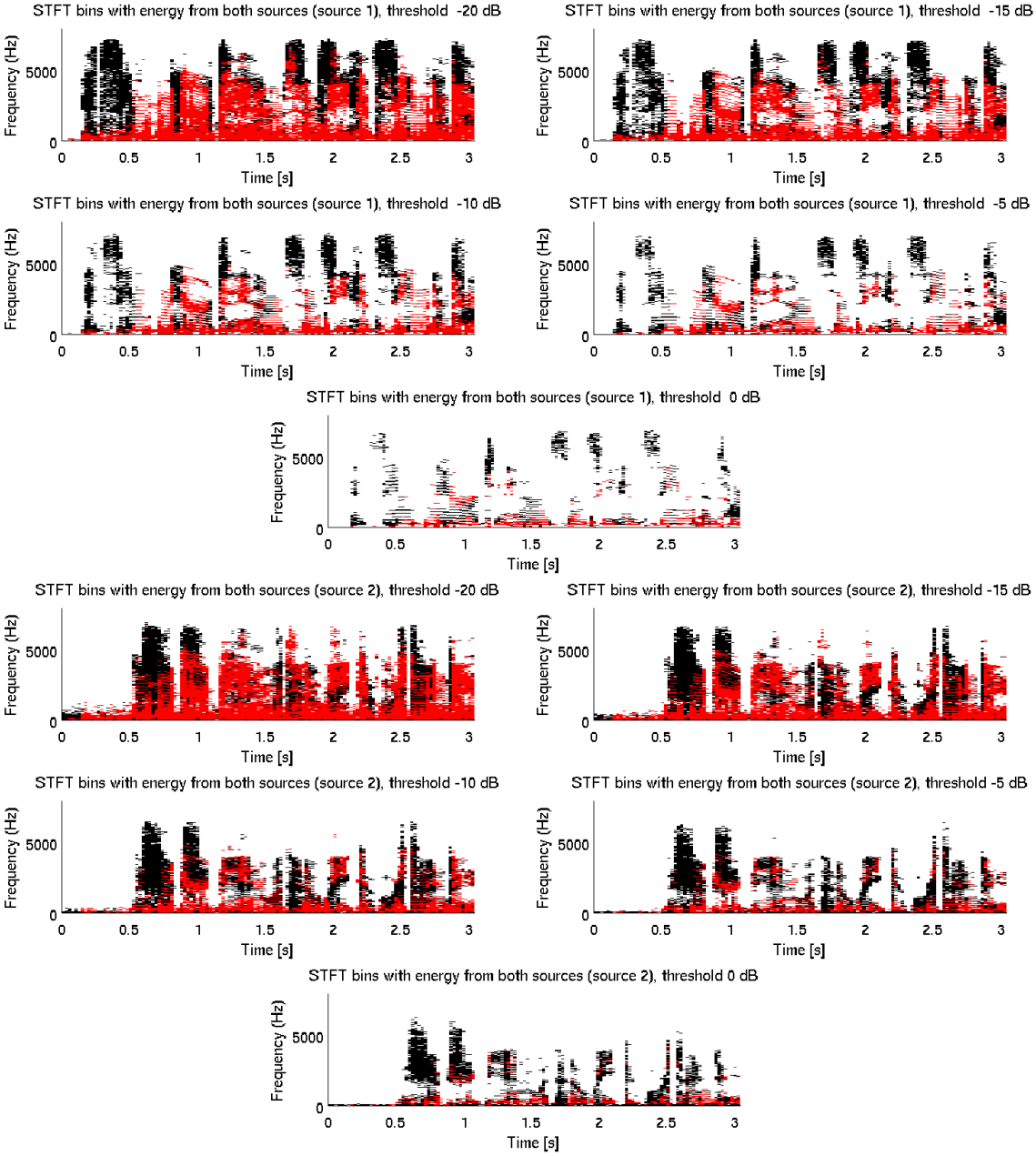

Figure 2.17: *Visualization of STFT bins that include energy only from the target source (black points) and bins that include energy also from the interfering source (red points) for different thresholds.*

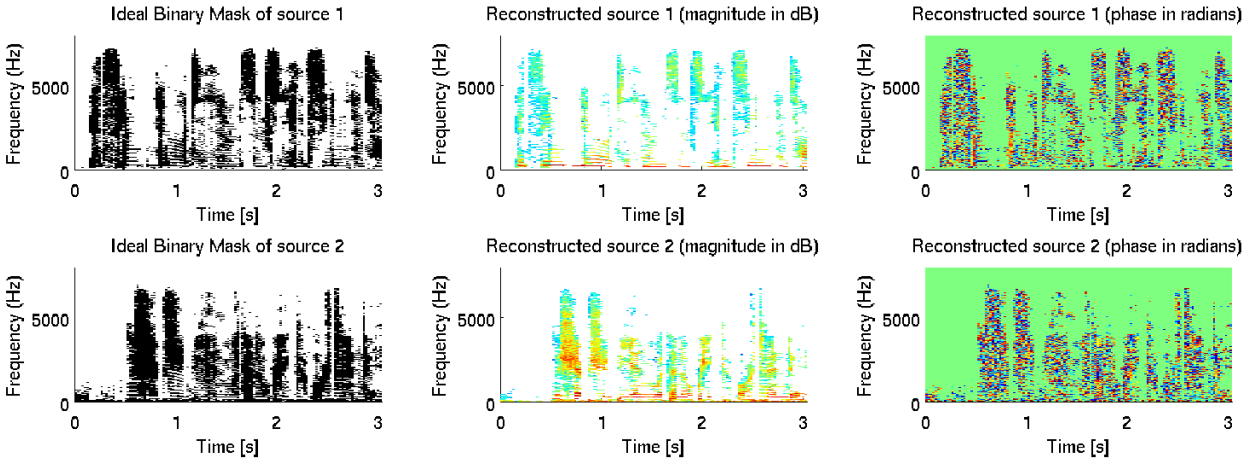

Figure 2.18: *The left plots define the ideal binary mask for the two sources of figure 2.16 at 0 dB mask threshold. The right figures illustrate the result of demixing the target source from the mixture with the help of the ideal binary mask.*

source in black and those bins which also include energy from the interfering source in red. The speech recordings are noisy to some extent, which leads to energy values greater than zero in almost all STFT-bins. Therefore the STFT spectra are filtered by different absolute thresholds to visualize the approximate orthogonality of the sources for these thresholds. Consider for example a threshold of -10 dB: Only those STFT-bins are regarded which have energy larger than -10 dB. Inspecting figure 2.17 shows that the orthogonality of the sources increases with increasing thresholds. For a threshold of 0 dB, where only bins with high energy values are considered, only few STFT-bins belong to both sources – the sources are approximately orthogonal. The lower the threshold, the more bins naturally belong to both sources. But these bins are of low energy and so do not influence the speech signal as severely as the high energy bins. For a detailed analysis of the orthogonality of speech sources in the time-frequency domain see chapter 3.

Knowing the recordings of the separated sources, the ideal binary mask can be computed by equation 2.30. The left plots of figure 2.18 show the ideal binary mask for the female target source at a mask threshold of 0 dB and an absolute threshold of -20 dB. The target source can be obtained from the mixture by multiplying the ideal binary mask with the mixture, which results in an approximation of the target source's STFT-spectrum that can be seen in the right figures.

Brungart et al. [17] support the concept of the ideal binary mask as goal of CASA by noting that the intelligibility of separated sounds increases if more and more energy of the ideal binary mask is reconstructed. Signals demixed with the ideal mask yield an intelligibility score of almost 100% in scenarios with two, three and four speech sources.

The use of ideal masks is conform to the two-stage processing of the human auditory system described by Bregman [13]. The first stage divides the time-frequency spectrum generated by the human cochlea in independent segments – in some sense analog to the grid of the time-frequency analysis used in CASA systems. The second stage groups these segments to an auditory stream which finally consists of a collection of segments of the first stage similar as the ideal mask includes a collection of its primary segments. In this way the ideal mask seems to be natural, as the overall goal of CASA should be consistent to the goal of human ASA – namely to generate auditory streams for specific sources of the auditory scene [13].

The benefit of the ideal mask is strongly dependent on the source types. Speech signals are known to be approximately orthogonal to each other in the time-frequency domain, but other source types such as music are not necessarily orthogonal which leads to a degradation of the source separation capabilities. The human auditory system also suffers when separating sources with large spectral overlap [117]. While humans can easily separate signals that are orthogonal in the time-frequency domain such as tones of different frequencies and speech signals, they fail in separating signals that are merely orthogonal such as white noise and pink noise [117].

The definition of the ideal mask does not require that the sources are orthogonal, so the concept of the ideal mask can in principle also be applied to non-orthogonal sources. Non-binary masks can avoid the decrease in the source separation capabilities due to the overlapping spectral components and can fairly assign the energy in each TF-bin to a specific source.

The human capacity of attention is limited to approximately four sources [25]. While a human listener is able to segregate and attend to four single tones, he is not able to follow the conversations of four parallel speech sources [117]. The human mind does not reconstruct and separate all sources in an auditory scene, it rather seems to concentrate on the most important and preferred source. In analogy to the human auditory system, the demixing process of a CASA architecture should concentrate on one specific target source, so that the whole problem of source separation degrades to a figure-ground segregation similar to Marr's [70] figure-ground segregation known from human vision. A separation of all sources in the auditory scenario can then be realized by applying the algorithm to each preferred source separately.

2.3.2 Source Separation Architectures based on T-F-Masks

Several researchers use the orthogonality of speech sources in the time-frequency spectrum to perform source separation.

The Degenerate Unmixing Estimation Technique (DUET) described by Yilmaz et al. [127] is able to separate several spatially separated speech sources from two anechoic mixtures. The mixtures are recorded with two separated microphones to enable the localization of the sources in the auditory scene. The mixtures are transformed to the STFT domain and the location of each time-frequency point in the resulting spectrum is estimated by computing the amplitude and

phase differences of each TF-point of the two mixture spectra. A three-dimensional histogram of the amplitude and phase differences yields a graph with peaks at the corresponding positions of the speech sources in the auditory scene. Smoothing the histogram and locating the peaks produce valid estimates of the source positions. Binary time-frequency masks are constructed by assigning each time-frequency point to the source with most correlating amplitude and phase shift. Finally the signals are demixed by multiplying the mixtures with the estimated masks and converting back to the time domain. For anechoic mixtures consisting of six sources up to 12 dB increase in the Signal-to-Interference Ratio (SIR) can be approached by this algorithm.

Melia and Rickard [74] extend the DUET algorithm to take into account the information of more than two mixtures and claim the robustness of the algorithm in reverberant environments. Specific evaluations of the performance under reverberant conditions are not given, so that a comparison to other architectures is hard to obtain.

Viste [113] uses a technique analog to the DUET algorithm, but the recording and the demixing of the sources is adapted to a simulated model of the human head and auditory system. The distance of the human ears induce phase ambiguities in the higher frequencies and so the DUET phase estimation suffers. Viste instead tries to estimate the location of each STFT-bin by estimating the azimuth position of the corresponding source in the auditory scene. The algorithms are only tested in anechoic conditions and concrete evaluations regarding the achieved SIR improvement remain.

Kollmeier et al. [57] also use the orthogonality of speech signals in the STFT domain to perform source separation. Their approach attempts to determine for each TF-bin whether this bin is dominated by the target signal or interferences. The decision is made based on time and level differences between two mixtures. From these time and level differences non-binary masks are constructed with weights between 0 and 1 according to the conformance of the estimated time and level differences.

2.3.3 Auditory Segmentation

TF-units are considered as indivisible parts of the auditory spectrum as the dimensions of a time-frequency unit are given by the used time-frequency representation such as the STFT or the cochleagram. By grouping together neighboring consistent TF-units, whole TF-segments can be constructed whose acoustic energy mainly originates from the same source. Adjacent TF-units either originate from the same source and so have the same properties or they belong to an other source and have different characteristics. A collection of those TF-segments forms a stream which represents a single source of the auditory scene. Some researchers exploit auditory segmentation to realize source separation and CASA architectures.

Ellis [31] for example computes auditory segments from a computational model of auditory perception, which are allowed to overlap to each other. The segments describe the periodic signal elements. Special segments called noise clouds represent wideband noise and interferences.

Cooke [24] [119] utilizes auditory segments – called synchrony strands – to segregate sources. Based on a model of the human auditory periphery the algorithm estimates the dominant frequency in each filter channel by extracting the instantaneous frequencies. Then the TF-segments are grouped together based on equal harmonicity and equal amplitude modulation. Segments are constructed by connecting those TF-points that form continuous spectral segments.

Wang and Brown [118] construct auditory segments based on correlating periodicity of neighboring channels. The frequency response of neighboring filter channels in a cochleagram overlaps substantially especially in the high frequencies. If the periodicity of neighboring TF-units correlates highly, these two bins are likely to arise from the same source and are grouped together. Hu and Wang [49] extend this algorithm by using the envelope of the signal in high frequencies instead of the signal itself. In this way the algorithm is more robust against the fast changes of the signals in the high frequencies. By exploiting only the periodicity of the input signal, this method is only suitable for periodic or quasi-periodic signals such as voiced speech opposed to the aperiodic unvoiced speech.

Hu and Wang [50] [48] segregate auditory segments by grouping those TF-units of a cochleagram with same onset and offset, both in the time and the frequency dimension. The onset and offset detection is realized in a three step process. In a first step the cochleagram is smoothed to eliminate minor amplitude deviations. In the second step the onsets and offsets are detected across the time and the frequency dimension. A final stage performs the first and second step for several smoothing granularities and estimates the final borders of the segments. High smoothing factors avoid falsely detected onsets and offsets, but on the other hand also smear other important points that cannot be detected as onset or offset anymore. Fine smoothing factors uncover the opposite problems: Too many found onsets and offsets make it difficult to choose the right ones. The described algorithm dissolves ambiguities by regarding several different smoothing factors and combining the results of each computation.

Brown and Cooke [119] specify a model for CASA consisting of several stages. A psychoacoustical inspired preprocessing estimates a cochleagram representation of the input signal. Then auditory maps for the frequency modulation, the fundamental frequency, the onsets and offsets and a cross correlogram are computed. Segments are constructed based on these maps and grouped by correlating fundamental frequency tracks.

Most models can only reliably compute the harmonics of low frequencies. In higher frequencies the harmonics become ambiguous – so called unresolved harmonics [49]. In most CASA systems these unresolved harmonics cannot be handled satisfactory. The Hu and Wang model [49] uses a specific strategy to handle resolved and unresolved harmonics. The segmentation process is based on the cross-channel correlation of the harmonicity as in the Wang and Brown [118] algorithm

described earlier. The grouping is realized by equal dominant F0 in each time frame and by the similarity of the periodicity pattern. From these segments a target pitch track is estimated. Given this target pitch track each segment is investigated to belong to the target stream or not.

The described approaches heavily rely on the periodicity of the input signals such as voiced speech. Separation schemes for non-periodic signals such as unvoiced speech are not treated often in literature. Besides the lack of periodicity, unvoiced speech is usually characterized by a lower energy level than voiced speech parts [119] and so is more difficult to handle than the separation of voiced speech. The dissertation of Hu [51] addresses the problem of unvoiced speech segregation in detail and the reader is referred to this work for an in-depth description.

2.4 Separation Based On Spatial Filtering

While some of the source separation architectures described in the last section already include the spatial position of the sources (i.e. DUET), the introduced source separation schemes are mainly based on acoustic features such as the fundamental frequency or the onset and offset synchrony. Another approach often used for source separation is spatial filtering of the single sources. Sources are assumed to emanate from specific spatially separated positions in the auditory scene. Separation schemes based on spatial filtering try to achieve the segregation by grouping the signal parts coming from the same direction. The following sections give an overview of the common binaural and engineering approaches used to perform spatial filtering.

2.4.1 Beamforming

Beamforming is used to perform spatial filtering in a multisource or noisy environment. The objective is to enhance the signal coming from a specific direction and attenuate signals from other directions and noise. Beamforming achieves an enhancement of the target source by placing a number of independent sensors at different points in space. By this means the received signal is sampled in space and the beamformer is able to distinguish between the spatial properties of the target signal and the jamming signals [47].

In the ideal case of only one source in an anechoic environment, the output of each sensor is identical but time-shifted and scaled against each other according to the distance between the sensors elements. Summing the time-aligned and rescaled signals preserves the desired signal as signals add up constructively while off-axis noise is smeared and added in an unpredictable way. In a multi source environment the signals from other sources arrive at different time and amplitude lags and so do not add up constructively. For some specified directions the signals finally add up destructively resulting in an erasure of the incoming signal – a so called null in the array response. Figure 2.19 shows the structure of a conventional narrow band beamformer. The incoming signals $x_0(t), x_1(t), \cdots, x_n(t)$ are multiplied by complex weighting

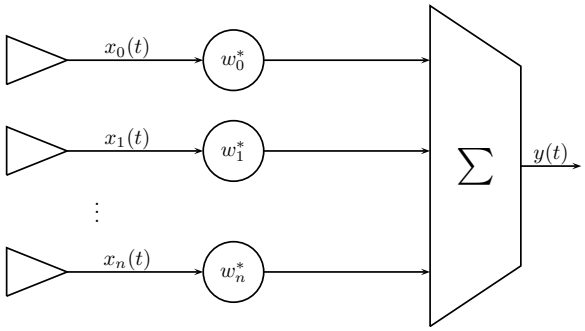

Figure 2.19: *Structure of a conventional narrow band beamformer.*

factors $w_0, w_1, \cdots, w_n$ which compensate the amplitude and phase lags of each array element. The array output $y(t)$ is given by

$$y(t) = \sum_{i=0}^{n} w_i^* x_i(t)$$

where $*$ denotes the complex conjugate. Denoting the weights of the sensor array by $w = [w_0, w_1, \cdots, w_n]^T$ and the received signals as $x(t) = [x_0(t), x_1(t), \cdots, x_n(t)]^T$ the output of the sensor array in matrix notation becomes

$$y(t) = w^H x(t),$$

where H denotes the complex conjugate transpose of vector w.

Beamforming schemes either follow a data independent processing approach or try to achieve a statistical optimum [85]. Data independent approaches – commonly known as fixed beamforming techniques – use fixed array weights to obtain a defined beam pattern and resemble the design of conventional Finite Impulse Response (FIR) filters. The implementation is independent of the environment or the received data, so these approaches cannot adapt to specific situations. Approaches that yield a statistical optimum use adaptive algorithms to estimate the statistical properties of the environment, as the statistics of the environment and the received signals are usually not known and not stationary.

A commonly used data independent beamformer is the primitive delay and sum beamformer shown in figure 2.20. All weights have equal magnitudes and degrade to simple time lags τ, which are adjusted to steer the beamformer in a specific look direction with unity response. Delay and sum beamformers are simple to implement, robust to errors and are even optimal regarding the maximum SNR for environments consisting of only uncorrelated noise and no directional interferences [41]. However one of the major drawbacks of the delay and sum beamformer is that it only enhances the signal coming from the look direction and does not deal with additional interferers from other directions.

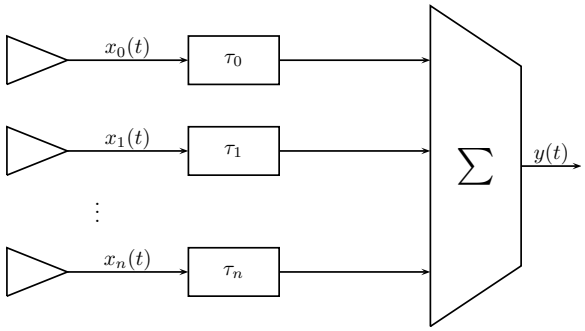

Figure 2.20: *Delay and Sum Beamformer.*

More sophisticated data independent beamformers known as null-steering beamformers are able to steer additional nulls in the directions of jamming signal sources while retaining unity response in the look direction. A beam pattern with unity response in the steering direction and nulls at the desired positions can be formed by estimating the weights of the sensors under specific constraints. Denoting the steering vector of the look direction where unity response is required with s_0 and the steering vectors of the k interferences that should be canceled with $s_1, \cdots, s_k$, the required weight vector that results in the desired beam pattern can be computed by solving the following equations [41]:

$$\begin{aligned} w^H s_0 &= 1 \\ w^H s_i &= 0 \qquad i = 1, 2, ..., k \end{aligned}$$

Assuming A is a matrix with its columns being the steering vectors $s_0, s_1, \cdots, s_k$ and e_1 is a vector with all its elements being zero except the first element which is one, the system of equations to solve is:

$$\begin{aligned} w^H A &= e_1^T \\ \Leftrightarrow \quad w^H &= e_1^T A^{-1} \end{aligned}$$

If $k = n$ and the steering vectors are linearly independent, the matrix A is invertible and w specifies the desired beam pattern. If $k < n$ the matrix A is not square and the weights can be estimated by [41]

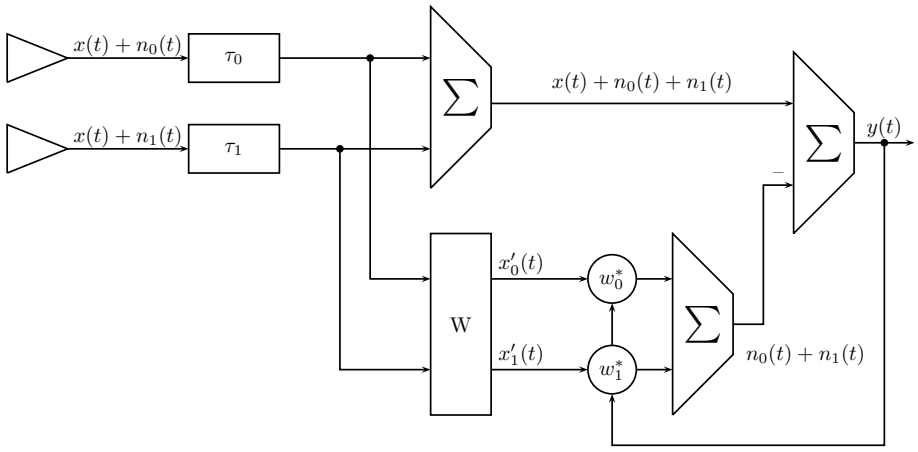

Figure 2.21: *Schematic of a Griffiths and Jim Beamformer.*

$$w^H = e_1^T A^H (AA^H)^{-1}$$

A sensor array consisting of N sensors has N degrees of freedom. The constraint $w^H s_0 = 1$ reduces the number of degrees of freedom to $N - 1$. So a conventional beamformer is able to realize a beam pattern with at most $N - 1$ nulls and unity response in the direction of the preferred source. Especially small sensor arrays are quite limited in their interfering suppression performance. The human ears as a binaural beamformer can suppress only one directional narrowband interference using the null-steered beamforming scheme.

To overcome the limitations of data independent beamformers and to improve the resolution and interference performance, adaptive schemes to estimate the array weights are applied. Opposed to data independent beamformers, the adaptive beamformers contribute to the received signals and the environment by dynamically placing nulls in the directions of interferences. Most adaptive beamformers are based on the minimum variance principle (see i.e. [68]) and are usually referred to as Minimum Variance Distortionless Response (MVDR) or Linearly Constrained Minimum Variance (LCMV) adaptive beamformers. The objective of these minimum variance beamformers is to preserve the signal from the steering direction by linear distortionless combining of the sensor weights. On the other hand the average energy of the summed output is minimized to exclude signal components coming from interfering directions. The MVDR beamformer for example ensures – as in the case of the data independent beamformers – with the constraint $w^H s_0 = 1$ the unity response of the array in the look direction independent of the array weights while the variance of the beamformer output is minimized.

One of the most commonly used adaptive beamformers is the beamforming scheme presented by Griffiths and Jim [42] depicted in figure 2.21 for the binaural case. The upper path realizes a

conventional delay and sum beamformer and is used to steer the array in the desired direction. The lower path implements a sidelobe canceller. A matrix preprocessor W blocks the desired signal $x(t)$. In the case of only two sensors, W realizes a simple subtraction of two input signals $x(t) + n_0(t) - (x(t) + n_1(t))$ yielding an estimate of $n_0 - n_1$. This interference component is processed by an unconstrained adaptive algorithm to remove as much residual interference as possible [119].

One major drawback of the presented beamformers is that they are designed to be applied to narrowband signals. The characteristics of wideband signals such as speech vary dependent on the frequency. Usually this problem is managed by splitting the signal in different frequency channels and using a specific beamformer to process each channel as a narrowband signal [119]. Frost [34] replaces the complex valued weights of the sensor array with frequency dependent filters to achieve efficient wideband processing.

For a detailed description of fixed and adaptive beamforming schemes the reader is referred to [41], [40] and [47].

2.4.2 Independent Component Analysis

The technique of Independent Component Analysis (ICA) tries to estimate the mixing matrix by means of some statistical optimum. Given m mixtures $x_1, x_2, ..., x_m$ of n sources $s_1, s_2, ..., s_n$ the problem of source separation can be described with the following system of equations:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & & \vdots \\ h_{m1} & h_{m2} & \cdots & h_{mn} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} + \eta \tag{2.31}$$

$$x = H \cdot s + \eta \tag{2.32}$$

The goal of ICA is to estimate the mixing matrix H such that the components of s are maximally independent. The independence of the components of s is measured by a specific independence function $F(s_1, \cdots, s_n)$. Considering a linear noiseless mixing model, the mixtures $x_1, ..., x_m$ are generated as the sum of the independent sources $s_1, ..., s_n$ multiplied by the mixing weights h_{ij}.

$$x_i = h_{i1} \cdot s_1 + h_{i2} \cdot s_2 + ... + h_{in} \cdot s_n \quad \forall i \tag{2.33}$$

Given this model, the task is to estimate the sources s and the mixing matrix H. The mixing matrix is estimated adaptively by calculating the mixing vectors and evaluating the current

mixing vectors by the cost function F, which either tries to maximize the nongaussianity of the sources or minimizes the mutual information [1]. The sources are then recovered by computing the inverse of the estimated mixing matrix – the demixing matrix H^{-1}. This inverse matrix only exists for $n \leq m$ and so the standard ICA approach is only useful for scenarios with a less or equal number of sources than sensors. To perfectly demix, the sources have to be statistical independent and non-Gaussian. In this way ICA is able to recover the original sources up to an unknown permutation of the sources and unknown gains.

In real recordings the components of the mixing matrix are not only phase and amplitude shifts, but specify complete room impulse responses which often have the characteristics of FIR filters [61]. So called Multi-Channel Blind Deconvolution (MCBC) methods (i.e. [6], [110]) are used as extension of the standard ICA techniques to enable the separation in reverberant environments, which often exhibit a convolutive mixing of the sources. Similar to the ICA algorithms, the MCBC methods mostly use an adaptive scheme to update the estimation of the mixing matrix according to some statistical optimum. Opposed to ICA, the components of the mixing matrix in MCBC methods are FIR filters that specify the echoic filtering of the signals. For realistic auditory scenes the mixing matrix becomes very complex and the estimation process is computationally very demanding [113]. Even small changes in the auditory scene result in big differences of the mixing matrix and so most MCBC algorithms are not robust to be used in different setups.

Current research in ICA investigates the usability of ICA methods in the underdetermined case where less mixtures than sources are available. Araki et al. [3] combine ICA with spatial prefiltering of the auditory scene. Before applying the ICA, a set of sources of the auditory scene is coarsely extracted to reduce the dimension of the problem and so resolve the underdeterminity of the problem. Then the ICA is applied to each of the source sets to finally extract the sources. Lee et al. [62] on the other hand describe a generalized ICA method to learn overcomplete representations of the sources where more sources than mixtures are present and so enable the ICA to be applied on the underdetermined problem itself.

If the assumptions of statistical independence are fulfilled, the separation of sound sources with ICA gives impressive results. In reverberant environments the mathematical model is not always valid and the demixing capabilities of ICA techniques degrade. Also when the assumptions change or are not fully fulfilled the separation process is deteriorated. Real-world scenarios such as auditory scenes recorded with a dummy head can hardly be described by an idealized mathematical model and specific assumptions of the characteristics of the sources are limited to special cases. The assumptions of most CASA methods tend to be conform with human assumptions before source separation. In this way it remains an open problem if ICA methods are ever applicable to dynamically created real world scenarios [119].

2.4.3 Source Localization

The localization of the sources in the auditory scene is a fundamental part of many CASA architectures. If the positions of the sources are known, spatial filtering of the sources can be applied to enhance the sources from a specific direction. The following paragraphs describe the implementation of several computational approaches for source localization based mainly on interaural time and level differences. There exist several statistical methods for source localization which – analog to ICA – assume at least the same number of sensors than sources. These methods are not useful to perform source localization in a manner analog to the human binaural system and for that reason are not considered in this section. For a description of source localization methods based on statistical methods – like for example MUSIC (Multiple Signal Classification) [99] or ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques) [98] – the reader is referred to the respective literature.

Most of the computational models for source localization are based on the Jeffress model [52] – a classical model for human binaural source localization that enables the extraction of interaural timing information. Jeffress postulates neural structures in the human brain that perform a kind of crosscorrelation of the signals received at the right and the left ear. From the timelag between the signals, the human brain estimates the incidence direction of the sound waves. The Jeffress model only accounts for interaural time differences and leaves out the interaural level differences. Several binaural phenomena like the precedence effect cannot be explained by this simple model.

The Equalization-Cancellation model is another model of human binaural hearing that influences many computational models. It is conceptually simple and is able to describe a number of binaural phenomena. The model was first introduced by Kock [54] and was further developed by Durlach [30]. The model assumes that the noise parts (i.e. interfering sources) are equalized in an early stage of the auditory system, so that the aligned noise signals can be subtracted from the target source and canceled out in this way.

Several computational approaches for source localization exist, but the fundamental principles usually are similar to some extend and are based on the time and level differences of the received signals. A simple form of source localization is realized with a simple delay-and-sum beamformer which is steered in all directions. The locations of the sources in the azimuth plane can be estimated by detecting the peaks in the direction dependent power distribution of the received signals.

Wang et. al [75] [91] use the correlogram described in section 2.2.2 to perform a localization of the sources in the auditory scene. Correlating the separate channels of the cochleagram of the left and right ear yields estimates of the azimuth positions of the sources. By summing the correlogram responses over all channels, the locations of the single sources are estimated.

Braasch [12] enhances the cross-correlation based source localization by using a method which is similar to the method of spectral subtraction often used in speech enhancement. The algorithm

is optimized to localize a target source in the presence of background noise and it is assumed that the target source and the background noise are uncorrelated and that the cross-correlogram of the background noise is known. The cross-correlogram of the background noise can then be subtracted from the cross-correlogram estimated by the received signals – as the correlogram of two uncorrelated sources adds up linearly [119] – yielding a cross-correlogram consisting of only the target source.

Viste [113] [115] investigates the localization capabilities based on ITD and ILD for a human dummy head under anechoic conditions. The HRTF of the human head is measured for each azimuth direction and smoothed to get continuous ITD and ILD values based on the incidence direction of the sound waves. The localization of the sources in the auditory scene is realized by constructing a STFT-time-frequency representation of the incoming signals and computing the ITD and ILD values separately for each channel and time-frame of the STFT. For an average human head, the ITD cues start to become ambiguous at frequencies of about 1.5 kHz [10]. Viste uses the smoothed HRTF as look-up table to choose the correct ITD candidate in high frequencies based on the ILD noted in the respective channel and time frame. The azimuth angle is estimated from the ITD based on Woodworth's formula [122]

$$\Delta t(\Theta) = \alpha_q \frac{r_h(\sin\Theta + \Theta)}{c},$$

where α_q is a frequency dependent scaling factor. The ILD is approximated with the formula

$$\Delta l(\Theta, q) = \beta_q \frac{\sin\Theta}{c},$$

where in turn β_q is a frequency dependent scaling factor. The values of α_q and β_q are found by fitting the curves to the measured HRTF. The model is evaluated in scenarios with a different number of sources which are localized well under anechoic conditions. To make the localization more robust and applicable to other heads, an extension of the model uses an average HRTF to define the scaling factors [114]. The average HRTF is obtained by averaging the HRTFs of 45 humans, which are available in the CIPIC HRTF database [2]. The results are satisfactory but not comparable to the individual HRTF.

Birchfield and Gangishetty [8] estimate the position of a sound source based only on the ILD. The so-called inverse square law states that the energy of a sound source decreases as the inverse of the squared distance between the sensors and the source. Using only two sensors results in ambiguities of the source position. The possible source positions lie on a circle if the level differences are not equal than 0 dB, else they lie on a straight line between the sensors. To resolve these ambiguities, Birchfield and Gangishetty use several sensors which pairwise estimate the position and the final location is inferred by intersecting the circles estimated by each sensor pair.

Roman and Wang [94] investigate the localization of moving sources. Moving sources exhibit uniform changes in the ITD and ILD received at the sensors, from which the velocity and the track of the sound source can be estimated. As the relative distance between the listener and the sound source changes, a Doppler shift can be introduced, but for normal human walking velocity, this Doppler shift is negligible [94]. The received signals are filtered by a gammatone filter bank and processed by a haircell transduction model to imitate the human listening behavior. For each frequency channel the ITDs and ILDs are computed. The reliability of a position estimation is computed by the height of the peak in the correlation used for ITD estimation. An easy model of the source's dynamics supports the estimation of the source track. Roman [90] extensively describes the source tracking approach and compares the results against source tracking with Kalman filters.

2.4.4 Source Separation

The first localization and separation architecture was introduced by Lyon [69]. Before the localization takes place, the received signals are transformed to a cochleagram. For each filter channel and specific time frames, the correlation between the left and the right ear is computed. Summing all channels yields prominent peaks at the positions of the sources. The sources are separated by spatial filtering according to the estimated directions. Ideas from this early approach have been adopted in many other source separation schemes.

Bodden [11] considers the localization and separation of multiple sources based on directional filtering. The model follows the Jeffress model for human source localization [52] and the source separation is based on directional filtering. To simulate the human ear, the received signals are filtered in 24 channels. By using only 24 channels, several signals such as speech exhibit considerable spectral overlap in this time-frequency representation and so the separation and localization capabilities of the architecture decrease. To resolve the spectral overlapping problem for the source localization, the estimated ITD values are converted to azimuth values. Based on the known HRTFs for white noise a frequency dependent mapping from ITD to azimuth degree is estimated by a supervised learning approach. Bodden observes a frequency dependency on the conversion from ITD to azimuth degree and integrates those scaling factors in the architecture. The model gives good results for two digitally combined sources, but problems arise when estimating and separating nearby sources.

Liu et. al [65] [66] constructed a binaural localization and separation system based on an adaptive noise-cancelling scheme. In a first step up to six concurrent sources of the auditory scene are localized by a dual-delay-line approach and a coincidence detection algorithm. For source separation the signal is divided in different frequency bands and time bins. The target source is preserved, while an adaptive null steering algorithm places a null in the frequency response of the current time-frequency bin in the direction of the strongest interferer. In each

time-frequency bin the strongest interference is canceled while the target source is preserved. In this way a SNR enhancement of 8-10 dB for the target source is achieved under anechoic conditions and an auditory scene consisting of four talkers. Anechoic scenarios with six sources yield demixing enhancements of 7-10 dB.

Lockwood et al. [67] developed a binaural system which is capable of separating a target source out of a mixture in real-time. Instead of localizing all sources in the auditory scene – which is computationally expensive – the target source is assumed to be directly in front and so the interaural time and level differences of the target source are assumed to be zero. The core algorithm is a binaural version of a frequency domain MVDR beamforming approach and is referred to as Frequency Domain Minimum Variance Beamforming (FMV). For each frequency bin a 2×2 correlation matrix is computed, which allows the tracking of the signal changes in different frequencies. The algorithm has been evaluated under different simulated reverberant conditions and different antenna setups (two cardioid microphones, two omnidirectional microphones and a KEMAR dummy head [18]). The performance of the FMV was compared to several other beamforming schemes like time-domain distortionless beamformers, the Frost adaptive beamformer [34] and the General Sidelobe Canceller [47]. The FMV algorithm outperformed the other beamforming schemes by 5-6 dB SNR gain.

Roman et. al [75] [96] [91] [92] perform source segregation with a simulated human head by estimating ideal time-frequency masks. The separation process is based on a supervised learning approach that segregates the sources by different interaural time and level differences. A training stage estimates the nonlinear transformations from the interaural time and level differences to the corresponding azimuth position and collects several statistics of the ITD and ILD values. The assignment of the single time-frequency bins to the spatially separated sources is based on a correlogram which is constructed for each time unit and yields an estimate of the ITD of the corresponding TF-point. The standard correlogram is enhanced by the skeleton correlogram described in chapter 2.2.2. The algorithm is tested in several scenarios consisting of speech sources and artificial sources like noise, sirens or a telephone. SNR gains of up to 12 dB are achieved for anechoic auditory scenes with two or three sources.

2.5 High-level Approaches for Source Separation

Human Primitive Auditory Scene Analysis is assumed to be an innate bottom-up process which analyzes the auditory scene based only on low-level properties [13], such as the interaural time and level differences between the two ears. Several engineering approaches for primitive scene analysis and source separation have been described in the last two sections. Opposed to the innate primitive scene analysis, the schema-based scene analysis extensively uses former experiences and expectations of heard sounds and is learned throughout the whole life. Hearing speech in the mother tongue is much easier to understand as specific expectations exist, which assist the

decoding of the received sentences. Musicians also enhance their skills in musical hearing by learning and adapting to the currently heard music based on former experiences.

Analog to the human schema-based scene analysis, high-level engineering approaches try to separate sources based on known prior information of the sources. These approaches usually create some kind of model of the source's characteristics and are therefore commonly referred to as model-based systems. Those high-level approaches for source separation are not in the focus of this thesis and are only mentioned for completeness of existing source separation schemes.

Model-based approaches for source separation expect some prior information regarding the sources. In the easiest case, the single sources are completely known and the task is to determine which source is in the mixture and where the corresponding starting time is. For this problem matched filters are usually used to estimate the most probable occurrences [119]. Knowing the single sources of a mixture is a rather restrictive assumption which cannot be realized in most real world applications.

To weaken the prior assumptions, sources are often modeled by Hidden Markov Models (HMM) (i.e. [111], [37], [38]). The signals are assumed to consist of elements which correspond to the states of the HMM. Transition-probabilities specify the chances of a transition from one specific state to another. The received signal is understood as a multidimensional vector consisting of elements from different signals. The task is to find the most probable elements of the different sources which in summation result in the observed vector. To correctly identify the most probable elements, the estimated elements from the past and the future and their corresponding transition probabilities are included in the computation. The core source separation is done by finding the most probable combination of the modeled sources for all observations.

Simple artificial sources can be modeled quite accurately by an HMM and can be demixed satisfactory with the described method. Complex sources like speech signals, however need a very high number of states to be specified in a realistic way [119]. The resulting state-diagrams of the HMM are very complex and intractable for most computations. In a HMM consisting of only some few states, an exhaustive search over all possible state combinations yields best results. For larger HMM, this exhaustive search is computationally not realizable and approximations have to be used. The Viterbi algorithm [33] for example can be used to estimate a best state sequence. In each stage the Viterbi algorithm keeps track of only one path and so finally estimates only one best state sequence. Another opportunity to decrease the number of states is to exclude some of the possible sources from the computation based on some other measurements, so that the state space is minimized.

For a detailed description of model-based source separation approaches, the reader is referred to i.e. [119] [31] [37] [111].

2.6 Reverberation in CASA

In real life humans have to deal with reverberations of different magnitude at any time and any place. Nature has adapted the human auditory system to work in reverberant environments and to even take advantage of the reflections and echoes [119]. Till this day, human processing of reverberant auditory scenes is only understood in the beginnings and many open secrets remain unsolved. Only few of the vast amount of source separation approaches consider also reverberant, real-life setups to test the separation capabilities. Most systems are restricted to work under anechoic conditions.

The reverberation characteristics of a room are specified by the room impulse response and the reverberation time T_{60}. The room impulse response (RIP) defines the direction-dependent filtering characteristics of the environment which are applied to each signal traveling through the room. The reverberation time T_{60} specifies the time that the energy of an idealized Dirac impulse needs to decrease by 60 dB. Typical T_{60} times for small offices lie in the range of 0.3 s, for concert halls 1.5 s and more are measured [119]. Besides the RIP and the T_{60}, the Direct-Sound-To-Reverberation-Ratio (DDR) is sometimes considered to estimate the influence of reverberations. The DDR specifies the ratio of the direct sound energy to the energy of the echoes expressed in dB.

The effects of echoes on the signals received are manifold. Echoes are reflected and absorbed frequency dependent on walls and other objects which leads to time and level differences of the different wavefronts arriving at the microphones. Time delays of different wavefronts can add up destructively and introduce undesirable nulls in the frequency response of the microphones, which lead to a spectral distortion of the received signal. Reverberation introduces sparse early reflections and dense late reflections. Early reflections are of high energy and are highly correlated with the original signal. The early reflections can be used to enhance the received signal by adding the wavefronts. Late reflections correlate only little with the original signal and smear the signal in an unpredictable and noise-like way.

2.6.1 Effects of Reverberation

Gelfand and Silman [39] [119] investigated the human intelligibility of speech in reverberant environments. The intelligibility of speech decreases as the energy of the signals is smeared to false positions and concurrent signals overlap at positions they would not under anechoic conditions.

The accuracy of human sound localization decreases in reverberant environments. Hartmann [43] detected that the efficiency of the interaural time and level differences reduce under echoic conditions. Broadband noise is less accurately localized in reverberant environments than in anechoic environments. The localization of sounds with a high energy onset is independent of the reverberation time. This phenomena contributes to the human precedence effect [13]:

The first arriving wavefront is weighted higher than the subsequent reflections to estimate the direction of a sound.

The human estimation of the distance of a sound source is enhanced in reverberant environments [119]. Opposed to anechoic environments, the reverberations introduce time and level delays between the first and later wavefronts. The ratio of the direct and the reflected sound is used by the human auditory system to estimate the distance of a sound source. If the source is far away, the ratio between the direct and reflected sound is large, while in the case of nearby sources, this ratio tends to be one. The ability to estimate the distance of the sources allows the human auditory system to provide a better spatial impression of the auditory scene. But on the other hand the accurate localization of the sources decreases as the monaural and binaural cues get distorted.

Humans are able to use the harmonic relationship of a specific sound to segregate it from other sounds. Two concurrent sounds can be segregated better if they have a substantially different fundamental frequency than with the same F0. Culling et. al [26] describe that the extend to which the harmonicity of the sounds can be used in reverberant environments is dependent on the fluctuation rate of the fundamental frequency. If the F0 is constant over the whole signal, reverberation doesn't influence the separation capabilities. If however the F0 fluctuates like for example in human speech, the reflections destroy parts of the harmonicity by smearing the signal over time, which leads to a decrease in the separation capabilities.

Humans tend to group together signal parts with same onset and offset times [13]. Feng and Jones [119] investigated the preservation of onset and offsets under reverberant conditions. Strong onsets endure the reflections without distortion, but weak onsets vanish due to the additional energy of previous reflections. Offsets are influenced more heavily than onsets and most of them completely disappear. Feng and Jones conclude that offsets are unlikely to be used as grouping cue by the human auditory system, as they are not valid in echoic real life situations.

Libbey and Rogers [63] tested if reverberation is equivalent to uncorrelated noise which can be canceled out by the binaural system. They tested the intelligibility of real reverberated speech and speech with reverberation-like added noise and found out that there is a significant difference between those experiments. Real reverberation seems not to be canceled out opposed to the uncorrelated noise. Contrariwise the additionally included localization and separation cues seem to be used to support the speech intelligibility.

Devore and Shinn-Cunningham [27] note that in the case of the human binaural auditory system, one ear usually has a favorable SNR which is dependent on the auditory scene and the positions of the sources. They postulate that the better ear is not necessarily the ear nearer to the target source, as reflections could amplify the signal depending on the environment setup. So human listeners would choose the better ear dynamically dependent on the current state of the scene around.

2.6.2 Acoustic Processing of Reverberant Sources

Few CASA systems have been evaluated in reverberant environments. Nonetheless there exist several signal processing approaches that can be used to perform or enhance source separation under reverberant conditions. The following sections describe the techniques of spatial filtering, inverse filtering and the processing of reverberation robust cues to perform source separation.

Spatial Filtering

Spatial Filtering is an approach that can be applied to enhance a target signal coming from a specific direction, while suppressing interfering signals from other directions. Interfering signals can be the concurrent sources or the reflections from the target source. The technique of spatial filtering has been extensively described in section 2.4 and is only mentioned for completeness here. Most source separation approaches based on spatial filtering can be evaluated in reverberant environments without major changes in the design of the architecture.

Inverse Filtering

The approach of inverse filtering tries to estimate the room impulse response of the environment. Then the inverse of the estimated impulse response is applied to the reverberant signal and in the optimal case the original signal is reconstructed in this way. The received reverberant signal $x(n)$ is a convolved version of the original signal $s(n)$ and the room impulse response $h(n)$:

$$x(n) = h(n) * s(n) \tag{2.34}$$

If the room impulse response can be estimated, then an approximation $y(n)$ of the original signal can be reconstructed by multiplying the received reverberant signal $x(n)$ with the inverse $w(n)$ of the estimated impulse response:

$$y(n) = w(n) * x(n) \tag{2.35}$$

For the existence of a causal and stable inverse filter $w(n)$, the room impulse response has to be minimum phase which is not assured in real-life scenarios [119]. The room impulse response is normally not known and is different in each environment and each incidence direction of the sound wave. Measuring the impulse response for a room for a specific direction is not applicable to other scenarios and so is limited in use for general source separation. In most cases the original source signals additionally are not known. The resulting problem is commonly referred to as blind deconvolution or blind dereverberation. Many approaches have been described in the literature. Most of those methods such as the Principle Component Analysis (PCA) described in section 2.4.2 rely on a greater or equal number of sensors than sources and are restricted in the generality.

One interesting example for blind dereverberation is the harmonic dereverberation algorithm (HERB) described by Nakatani et. al [76]. The harmonic parts of a speech signal are used to estimate the inverse filter $w(n)$. The inverse filter is optimized to enhance the periodicity of the speech signal in short time frames during which the fundamental frequency is assumed to be constant. Reverberation decreases the periodicity of the signal as described in the last section. By estimating an inverse filter that makes the signal periodic in local time frames, reverberation is filtered out to a large extend. The described method turns out to be more effective for female voices than male voices, as the harmonics of female speech are further apart because of the higher F0. In an extension of the HERB algorithm Nakatani et. al [77] weaken the assumption that the F0 has to be constant in the complete analysis window by using dynamic time warping techniques. Substantial improvements compared to the original algorithm are reported in the paper.

Wu and Wang [123] estimate the T_{60} of a reverberant speech recording and enhance the signal by subtracting the echoes according to the estimated T_{60}. As the F0-strength is inversely related to the reverberation time [123], the T_{60} can be estimated based on a correlogram and additionally computed F0-track. For F0-tracking the algorithm from [125] is used. Then a histogram of the time lags between the F0 estimated by the F0-tracking algorithm and the nearest peak in the corresponding time-frequency bin of the correlogram is constructed. In anechoic environments this results in a sharp peak centered at zero. In reverberant mixtures the peak is broadened. The final T_{60} is estimated based on the width of the histogram peak.

Roman and Wang [95] combine a dereverberation scheme with a F0-based sound segregation. In a first step an inverse filter is estimated to equalize the room impulse response subject to the position of the preferred source. In a second step the F0-based segregation scheme is used to demix the sources from the mixture.

Reverberation-Robust Cues

Opposed to the approaches described in the last section that try to filter out the reverberation of the received signals, other source separation architectures use only those characteristics of the sources that are – at least to a specific extend – robust to reverberation.

Faller and Merimaa [32] for example use the properties of the precedence effect for accurate binaural source localization in reverberant environments. After preprocessing the two ear input signals with a model of the human basilar membrane by passing them through a gammatone filterbank, each critical band is processed by a model of neural transduction. Then the interaural time differences between the left and right ear signal are computed for each channel and time frame based on a running normalized cross correlation:

$$R(t,\tau) = \frac{\int_{-\infty}^{t} x_L(\alpha) x_R(\alpha-\tau) w(t-\alpha) d\alpha}{\sqrt{\int_{-\infty}^{t} x_L^2(\alpha) w(t-\alpha) d\alpha} \sqrt{\int_{-\infty}^{t} x_R^2(\alpha) w(t-\alpha) d\alpha}} \tag{2.36}$$

The interaural coherence (IC) between the two signals is defined as the maximum of the correlation function:

$$IC(t) = \max_{\tau} R(t, \tau) \tag{2.37}$$

The IC value lies in the range of $[0, 1]$, where a value of one means perfect coherence between the signals except a time shift, so that it is very probable that only one source is active in this time-frequency unit, while a low value suggests several sources contribute to this time-frequency unit. By choosing only those TF-units for source localization where the IC value is close to one, the localization capabilities can be enhanced. Usually the first wavefronts exhibit high IC values as the reflections will arrive later. The IC-property of this algorithm models the precedence effect used by the human auditory system, where the first wavefronts are weighted higher than the reflections.

Bechler and Kroschel [5] examine the reliability of estimating the interaural time differences of reverberant signals based on the crosscorrelation function. Each ITD estimate is scored by a single value that represents the reliability of the estimate. Mainly three criteria are evaluated: The precedence effect is imitated by enhancing the information contained in the onsets of the signals while attenuating the information of the reflections. So the reliability values of signal parts directly succeeding an onset are weighted higher than later signal parts. The absolute value of the maximum peak in the crosscorrelation function is used as second criterion: The higher the peak, the more reliable is the estimate. The third criterion is based on the ratio between the first and the second highest peak in the crosscorrelation function. The estimate is weighted higher if the ratio is large.

Brown et al. [15] investigate binaural speech separation in reverberant environments based on the spatial location of the sources. The demixed results are used as input to an automatic speech recognizer which is based on the missing data approach. The input signals are filtered by a simulated HRTF of the KEMAR dummy head [18] and a room impulse response with a reverberation time $T_{60} = 0.34s$. After processing the input signals with a gammatone filterbank, interaural time and level differences are obtained by extracting the envelope of each frequency channel, smoothing it with a first order lowpass filter and computing the cross-correlation between the resulting signals. The ITD is estimated based on the highest peak of the correlation function, where the final position is refined by fitting a quadratic curve to the peak and a reliability measure is given by the height of the peak. The ILD is estimated as the power ratio of the left and right ear signal in dB. For each channel a histogram of the ITD and ILD values is constructed and compared to learned histograms. Finally the target signal is extracted by using only those bins with corresponding ITD and ILD values and the target signal is used as input to the automatic speech recognizer.

2.7 Comparison and Evaluation of CASA Architectures

The evaluation of CASA systems differs substantially according to the objectives of the architecture. There is no common way to estimate the performance of a CASA system in the literature. Wang [119] identifies four categories of CASA evaluation schemes for which individual architectures are optimized.

Comparison with Ground-Truth Signal Some researchers (i.e. [118], [88], [112]) use the original signal as ground-truth and compare it with the estimated signal. A conventional Signal-To-Noise Ratio (SNR) is used to estimate the performance of the source separation framework. Separate speech and noise recordings allow to measure the performance as a change in SNR before and after source separation takes place. However the SNR does not necessarily indicate the human intelligibility of the segregated signal. If the separation reconstructs very few of the energy of the source of interest, but totally eliminates the noise (such as other sources and interferers) this yields a very high SNR value, but the intelligibility of the signal of interest is very low.

Automatic Speech Recognizer Source separation frameworks are sometimes used in front of Automatic Speech Recognizing (ASR) Systems to enhance the performance in multisource or reverberant environments (see for example [120]). To this account it is useful to evaluate these frameworks by noting the increase in the ASR scores of segregated sources compared to the non-processed input. A central problem when using a CASA system in front of an ASR is that CASA systems tend to distort the target signals, yielding a mismatch between the signals used for training of the ASR and the signals that should be recognized [117].

Human Listeners Human Listeners are often used to judge the capabilities of source separation frameworks. The mixture and the segregated sources are presented to the subjects which have to rate the results. Yilmaz et al. [127] request the subjects to rate the intelligibility of the separated source on a simple, linear scale to exploit the performance. Ellis [31] on the other hand asks the subjects to rate the similarity of a segregated source and the same source in the mixture. However human ASA will always further process the separated sources and so the rating results cannot be directly transferred to technical application scenarios.

Biological Correspondence In some cases the main objective of the evaluated sound source architecture is the correspondence with biological and psychoacoustical phenomena. These frameworks model the organization of the human ear and brain and try to verify single concepts of human auditory scene analysis. Wang [116] for example uses neural oscillator networks to model several ASA phenomena. The evaluation of these architectures takes into account to which extent the neurophysiological and psychoacoustical characteristics of humans can be approximated and how well these strategies account for these properties.

The comparison of CASA architectures is difficult to accomplish, as each system has its own objectives and is based on special assumptions which cannot be applied to other systems. Most CASA systems inherently differ in the setup and the recording of the auditory scene and are hardly comparable. Additionally there are no standard auditory scenes available that would allow to directly compare the abilities of several architectures.

This thesis focuses on the evaluation of the separated sources based on the comparison with the ground-truth signal. The final goal of the source separation architecture developed in this research is to separate an isolating source from a mixture of several sources without interfering sources and artificial or sensor noise.

2.7.1 Criteria for Estimating the Quality of a Separated Source

A metric for measuring the capabilities of a specific source separation architecture based on a comparison of the separated signals with the known ground truth signals has to fulfill several criteria to be consistent with the human listening experiences. Master [71] identifies the following three main criteria that have to be met by an appropriate metric for evaluating source separation algorithms:

Correspondence to Human Listening Capabilities Separated signals with high human intelligibility ratings should receive higher scores than separation results with low human intelligibility ratings. Minimally different listening impressions should receive minimally different scores.

Separate Treatment of Artificial and Interference Noise Artificial and Interference noise have to be handled separately. Artificial noise tends to arise from the non-linear processing of the mixture during the separation process in underdetermined architectures [71]. Interference noise is induced by signal parts from the interfering sources that are not fully removed by the separation algorithm. The auditory systems deals differently with both noise terms. Artificial noise sounds strange, but does not decrease the separation of the target source and the interfering sources, only the human listening impression is distorted. Postprocessing algorithms can enhance the signals by removing artifacts from the separated source. Interference noise on the other hand arises due to the energy of the interfering sources that has not been removed by the separation algorithm. Interference noise is directly mapped to the interfering sources by human listeners and decreases the separation capabilities.

Independence of Separation Scheme The perfect quality measure is able to be applied to an arbitrary separation scheme. It should be irrelevant if the separation algorithm is based on binary or non-binary time-frequency masks or if the algorithm works in the time-domain.

Additionally there arise some further considerations that have to be regarded.

- If binaural source separation architectures are considered, there can be two different demixing results. Binaural source separation architectures use two received signals to perform the source segregation. After estimating the demixing scheme, this scheme can be applied to the left and the right mixture to demix the separated target source. Dependent on the position of the two sensors, this leads to two different segregated signals with possibly different quality criteria. To estimate a final quality measure, a specific signal has to be chosen. Yilmaz et. al. [127] for example always use the channel, where the input SNR for the target source is favorable – i.e. the channel that is nearer to the target source and so the target signal is expected to be louder compared to the interfering sources. The problem with this approach lies in the automatic selection of the better channel.

- Standard signal processing algorithms estimate the SNR values in the time-domain. The human auditory system however is frequency dependent and works in the time-frequency domain [13]. Quality estimation approaches based on a time-frequency representation of the target signal are used to model the human-specific quality characteristics. The ITU-standard PEAQ (Perceptual Metrics for Audio Quality) [109] for example makes use of the time-frequency domain as front-end for the quality estimation.

- The used metric should be scale invariant. The original and the demixed target signal are not necessarily of the same loudness due to the separation process. If the separation algorithm is applied to an amplified mixture, the quality estimation should be identical to the non-amplified result.

2.7.2 Existing Quality Criteria

Several metrics to evaluate the source separation capabilities of a specific algorithm have been described in the literature, but each of these metrics fails on one ore more of the specified demands.

Time Domain Approaches

Consider the original single target source $s(n)$. This original target source can be obtained by either using the file played back to construct the auditory scene or a recording of the single source that includes the specific characteristics of the auditory environment – such as reflections and reverberation. The target signal reconstructed by the separation algorithm is denoted by $s_{est}(n)$. A conventional signal to noise ratio (SNR) for the demixed target signal is computed as

$$SNR_{processed} = 10 \log_{10} \frac{\sum_n s(n)^2}{\sum_n (s_{est}(n) - s(n))^2} \tag{2.38}$$

The SNR for the unprocessed mixture $x(n)$ is obtained analog as

$s(n)$	original source (recorded single source or played back wav file)
$s_{ideal}(n)$	original source gained from multiplying ideal mask with mixture
$s_{est}(n)$	estimated source gained from multiplying estimated mask with mixture
$x(n)$	mixture
$s_{correctReconstructed}$	signal resulting from applying the estimated mask to the spectrum of the original source $s(n)$
$s_{falseReconstructed}$	signal resulting from applying the estimated mask to the spectra of the interfering sources

Table 2.1: *Definition of the variables used for estimating the quality of a separation algorithm.*

$$SNR_{unprocessed} = 10 \log_{10} \frac{\sum_n s(n)^2}{\sum_n (x(n) - s(n))^2} \tag{2.39}$$

The quality of the source separation algorithm is then estimated as the gain in dB between the processed and the unprocessed SNR

$$SNR_{gain} = SNR_{processed} - SNR_{unprocessed} \tag{2.40}$$

Wang et. al. [75] use the processed SNR value (equation 2.38) to estimate the quality of the separated sources. For anechoic two source mixtures consisting of one speech source and an interfering artificial or speech source, demixing results obtained by multiplying the mixture with the ideal binary mask yield $SNR_{processed}$ values of 10 – 13 dB. Separation results of the implemented segregation algorithm vary in the range of 4 dB to 13 dB.

Vincent et. al. [112] derive a quality metric that addresses several of the criteria specified in the last section. To avoid the scaling problem, specific distortions are allowed to belong to the original source. The algorithms to compute the quality metric are able to tolerate time variant and invariant gains and filters. To estimate the quality of the separated target signal $\hat{s}$, the signal is divided in four signal parts that represent the parts coming from the target signal, the interferences, sensor noise and introduced artificial noise.

$$\hat{s} = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{2.41}$$

The decomposition of the estimated signal is based on orthogonal projections, which are dependent on the allowed distortions of the sources. Assume $\Pi\{y_1, ..., y_k\}$ denotes the orthogonal projector onto the subspace spanned by the vectors $y_1, ..., y_k$. Consider a scenario with n sources, which are recorded by m microphones. The original signals are denoted by $s_j \;\; 1 \leq j \leq n$, while $\hat{s}_j$ refers to the estimated signals and $n_i \;\; 1 \leq i \leq m$ characterizes additive sensor noise. In the easiest case, where only time-invariant gains are allowed, the orthogonal projectors are defined as follows:

$$P_{s_j} = \Pi\{s_j\} \tag{2.42}$$

$$P_{\mathbf{s}} = \Pi\{(s_{j'})_{1\leq j'\leq n}\} \tag{2.43}$$

$$P_{\mathbf{s},\mathbf{n}} = \Pi\{(s_{j'})_{1\leq j'\leq n}, (n_i)_{1\leq i\leq m}\} \tag{2.44}$$

The four signal parts of equation 2.41 are then estimated as:

$$s_{target} = P_{s_j}\hat{s}_j \tag{2.45}$$

$$e_{interf} = P_{\mathbf{s}}\hat{s}_j - P_{s_j}\hat{s}_j \tag{2.46}$$

$$e_{noise} = P_{\mathbf{s},\mathbf{n}}\hat{s}_j - P_{\mathbf{s}}\hat{s}_j \tag{2.47}$$

$$e_{artif} = \hat{s}_j - P_{\mathbf{s},\mathbf{n}}\hat{s}_j \tag{2.48}$$

Then four ratios (Signal-To-Distortion Ratio (SDR), Signal-To-Interference Ratio (SIR), Signal-To-Noise Ratio (SNR) and Signal-To-Artifacts Ratio (SAR)) are estimated that characterize the quality of the separated source:

$$SDR = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \tag{2.49}$$

$$SIR = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2} \tag{2.50}$$

$$SNR = 10\log_{10}\frac{||s_{target} + e_{interf}||^2}{||e_{noise}||^2} \tag{2.51}$$

$$SAR = 10\log_{10}\frac{||s_{target} + e_{interf} + e_{noise}||^2}{||e_{artif}||^2} \tag{2.52}$$

If the allowed distortions of the estimated sources include time-varying gains and filters, the orthogonal projectors of equations 2.42-2.44 are slightly adapted to the corresponding scenario. For a detailed description and analysis see the work of Vincent et. al. [112].

Testing the described quality estimation on an anechoic mixture of three instruments (cello, drums and piano), which is demixed using the DUET technique [127] yields the following quality values [112]:

Source	SDR	SIR	SAR
Cello	5 dB	14 dB	6 dB
Drums	4 dB	7 dB	8 dB
Piano	6 dB	19 dB	6 dB

Problems in using the time-domain SNR values arise due to the operating of the quality metric in a non-perceptual domain. SNR values cannot be mapped directly to human perceived intelligibility of a separated source. The problem of high SNR values, but low recovered total energy as noted above, makes it hard to interpret SNR values as a complete quality metric. By using the SDR and SAR values as described by Vincent et. al. [112] this problem is resolved to some extent.

Time-Frequency Domain Approaches

Wang et. al. [93] describe a quality metric that is appropriate for source separation architectures that aim at estimating the ideal binary mask. The output SNR is estimated as the ratio of the correct reconstructed energy (energy belonging to the target source) and the false reconstructed energy (energy belonging to the interferers). Decisions about correct and false signal parts are judged based on the available ideal binary mask.

$$SNR_{output} = 10 \log_{10} \frac{\sum_n s_{correctReconstructed}(n)^2}{\sum_n s_{falseReconstructed}(n)^2} \tag{2.53}$$

To counter the problem that a high SNR_{output} value does not induce a high human intelligibility as noted above, an additional value specifying the retained speech ratio (RSR) is computed that denotes the percentage of reconstructed energy.

$$RSR = \frac{\sum_n s_{correctReconstructed}(n)^2}{\sum_n s(n)^2} \tag{2.54}$$

For reverberant two source scenarios with reverberation time $T_{60} = 0.4\ s$ the SNR_{output} is computed as 11.49 dB, while $RSR = 75\%$ of the target energy is retained.

Alternatively Wang et. al [93] specify the SNR quality metric as

$$SNR_{output2} = 10 \log_{10} \frac{\sum_n s_{ideal}(n)^2}{\sum_n (s_{ideal}(n) - s_{est}(n))^2} \tag{2.55}$$

For two source scenarios $SNR_{output2}$ values of 3 to 8 dB are achieved.

Yilmaz et al. [127] specify three criteria to estimate the quality of a separated source and to describe the Window-Disjoint Orthogonality (WDO) of sources in the time-frequency domain.

1. The Preserved Signal Ratio (PSR) specifies how well the ideal mask preserves the energy of the target source compared to the clean target signal. The PSR is defined as the ratio of the energy of the ideal mask multiplied with the STFT-spectrum of the clean target signal and the energy of the STFT-spectrum of the clean target signal:

$$PSR = \frac{||\Omega_i(t,f)s_i(t,f)||^2}{||s_i(t,f)||^2}$$

where $||f(x,y)||^2$ is defined as $\int\int |f(x,y)|^2$. In the best case, when the ideal mask includes all time-frequency points of the target source with energy greater than zero, the PSR approaches one.

2. The Signal to Interference Ratio (SIR) defines how well the ideal mask attenuates the interfering sources. In principle the definition of the SIR value is analog to the definition used by Vincent et. al as described in equation 2.50, but the following SIR is adapted to work with ideal binary masks, while equation 2.50 works in the time-domain. The SIR specifies the ratio of the remaining energy of the target source after multiplying with the ideal mask and the energy of all other sources remaining after multiplying with the ideal target source mask.

$$SIR = \frac{||\Omega_i(t,f)s_i(t,f)||^2}{||\Omega_i(t,f)\sum_{j\neq i} s_j(t,f)||^2}$$

High SIR values show that a high percentage of the reconstructed energy belongs to the target source and the interfering sources are suppressed very well. Ideal masks which yield low SIR values include much energy from other sources and so cannot be used to perfectly demix the target source.

3. The orthogonality of different speech sources is estimated by a value called window-disjoint orthogonality (WDO). The WDO is a combined measurement of the PSR and SIR and is specified as the normalized difference between the portion of energy remaining after demixing with the ideal mask and the portion of energy of other sources remaining after demixing:

$$\begin{aligned} WDO &= \frac{||\Omega_i(t,f)s_i(t,f)||^2 - ||\Omega_i(t,f)\sum_{j\neq i} s_j(t,f)||^2}{||s_i(t,f)||^2} \\ &= PSR - PSR/SIR \end{aligned}$$

A value of one defines perfect orthogonality in the STFT-domain, a value of zero specifies almost no orthogonality and so only bad demixing results can be achieved with the ideal mask.

3 Window-Disjoint Orthogonality of Speech Signals

Yilmaz et al. [127] showed that speech signals are sparsely distributed in high-resolution time-frequency representations. Time-Frequency (TF) representations of different speech signals overlap only in few points and so are approximately orthogonal to each other. This approximate orthogonality in the TF-domain can be used to separate a target source out of a mixture of speech sources by defining TF-masks that emphasize regions of the TF-spectrum that are dominated by a specific source and attenuate regions dominated by other sources or noise.

Many speech source separation approaches are based on the assumption of approximate orthogonality of speech sources in the time-frequency domain and utilize TF-masks to separate the single sources from a mixture (i.e. [127] [117] [75] [74] [113] [57]). Several researchers in computational source separation suggest the ideal binary mask as final goal of computational source separation algorithms (i.e. [127] [117] [75]). Each entry of the TF-mask is set to one if the target energy in this TF-bin is greater than the interfering energy. The binary decision is motivated by masking effects of the human auditory system: Within a critical bandwidth humans don't recognize sounds that are masked by louder sounds [13].

The orthogonality of speech sources in the time-frequency domain has been investigated in detail for anechoic speech mixtures (i.e. [127]) and most of the available source separation algorithms are only tested for anechoic and artificially mixed speech mixtures (for example [127] [117] [75] [113] [57]). To be applicable to real world scenarios (such as an robotic human dummy head in the case of this research or the operation of a source separation algorithm as a front-end for an Automatic Speech Recognizer), source separation schemes should be able to operate also in reverberant environments. This section therefore investigates how the orthogonality of speech sources in the time-frequency domain drops with different reverberation times of the environment and evaluates if separation schemes based on ideal binary TF-masks are suitable to perform source separation under reverberant conditions.

To imitate the excellent source separation capabilities of the human auditory system this project uses a humanoid experiment setup, which is described in detail in the next chapter. The speech mixtures are recorded by a human dummy head that performs human-like filtering of the signals by the head and the outer ear structures. The HRTFs of the left and the right ear filter the incoming signals and disturb them. To find out if source separation schemes relying

on the time-frequency orthogonality are also appropriate for such humanoid setups, this section additionally investigates how the orthogonality of speech sources in the time-frequency domain is affected by the HRTF filtering process.

Realistic humanoid experiment setups record the speech mixtures with a human dummy head under normal reverberant conditions (i.e. [102], [103], [104]). In these scenarios the reverberation and the HRTF filtering affects the orthogonality of the speech sources. This section also investigates if ideal binary masks are furthermore sufficient to achieve a satisfactory source separation in reverberant environments with a humanoid recording equipment.

3.1 Evaluating the Orthogonality of Speech Signals

Assume $s_i(t,f)$ denotes the energy of the target signal$_i$ in TF-bin at time t and frequency f and $n_j(t,f)$ denotes the energy of the j-th interfering signal in this TF-bin. The ideal binary mask $\Omega_i(t,f)$ for target source$_i$ and a threshold of x is defined as follows:

$$\Omega_i(t,f) = \begin{cases} 1 & s_i(t,f) - n_j(t,f) > x \quad \forall j \\ 0 & \text{else} \end{cases} \tag{3.1}$$

An ideal binary mask Ω_i with a threshold x of 0 dB includes all time-frequency points where the energy of source$_i$ is larger than the energy of all other sources in this TF-bin. An usual goal of source separation architectures is to maximize the Signal-to-Interference Ratio (SIR) while retaining most of the target source's energy. When using the 0-dB ideal binary mask as final goal, TF-bins that have nearly equal energy from two or more sources are only assigned to one specific source.

The following evaluations of the orthogonality of speech sources in the time-frequency domain use three values described by Yilmaz et al. [127] to define the quality of the ideal binary mask separation. The SIR (Signal-To-Interference ratio), PSR (Preserved-Signal ratio) and WDO (Window-Disjoint-Orthogonality) values operate in the STFT domain (see section 2.2.1) and are described in detail in section 2.7.

3.2 WDO under simulated reverberant conditions

Most source separation architectures that are based on the orthogonality of speech sources in the time-frequency domain are only evaluated on anechoic speech mixtures (i.e. [127] [117] [75] [74] [57]). Many practical applications of source separation architectures like i.e. front-ends of Automatic Speech Recognizing Systems however require the operation of such systems in non-ideal reverberant environments like inside a car or a crowd. To estimate if ideal binary masks as final goal of source separation approaches are also applicable in reverberant scenarios, this

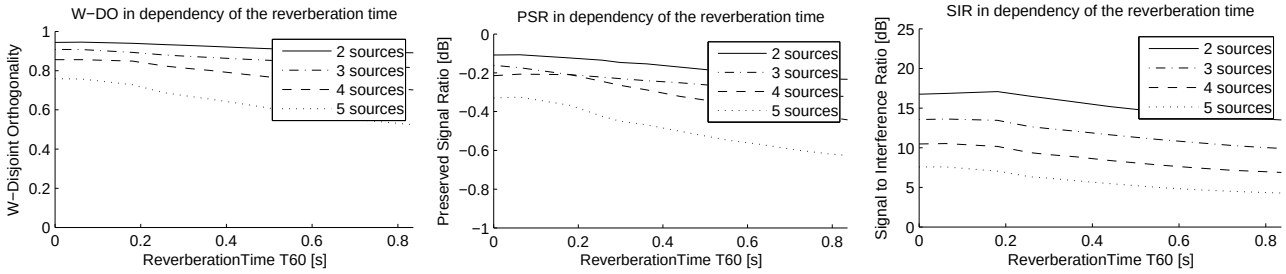

Figure 3.1: *Window-Disjoint Orthogonality in dependency of the reverberation time T_{60} for scenarios consisting of different sources for 0-dB ideal mask.*

section investigates the influence of reverberation on the window-disjoint orthogonality of speech sources in the time-frequency domain and discusses methods for increasing the SIR gains of source separation architectures based on ideal masks also in reverberant environments.

Figure 3.1 shows the window-disjoint orthogonality, the preserved signal ratio and the signal-to-interference ratio of mixtures of two to five speech sources for different reverberation times T_{60}. The values are obtained by computing the average values for 20000 speech mixtures of different speakers and 3-seconds duration taken from the speech database CMU Arctic [58]. The mixtures are constructed by adding together the single speech sources. To simulate the reverberation, the mixture files are filtered with Room Impulse Responses defined by FIR filters according to [73].

The influence of reverberation degrades the orthogonality of speech sources in the time-frequency domain. The room impulse responses smear the energy of specific time-frequency bins in time and in frequency and so disturb the sparseness of the speech sources in this way: the overlap of the time-frequency spectra of the single speech sources increases with increasing reverberation time. The SIR decreases for two, three, four and five source scenarios by approximately 3 dB for reverberation times $T_{60} = 0.6\ s$ compared to the anechoic case. The WDO and the PSR decrease analog to the SIR with increasing reverberation time.

The impact of the window size of the STFT transform on the WDO, PSR and SIR is analyzed in figure 3.2. Windows of length 92 ms and 185 ms perform best, while shorter windows reduce the WDO, PSR and SIR values. For windows of length 11 ms, the SIR is about 6 dB lower than for windows of length 185 ms. In scenarios with more than two sources, the performance of the short windows is worse than in the two-source scenes. Choosing long time windows to compute the STFT results in a better frequency resolution of the spectrum. For speech signals it seems as if a fine frequency resolution is more important than a fine time resolution with respect to the window-disjoint orthogonality of the speech signals. The decrease in WDO of speech sources under different reverberation conditions is not influenced by the choice of the window length in most cases. Only when very long window lengths such as 371 ms are chosen, the WDO for

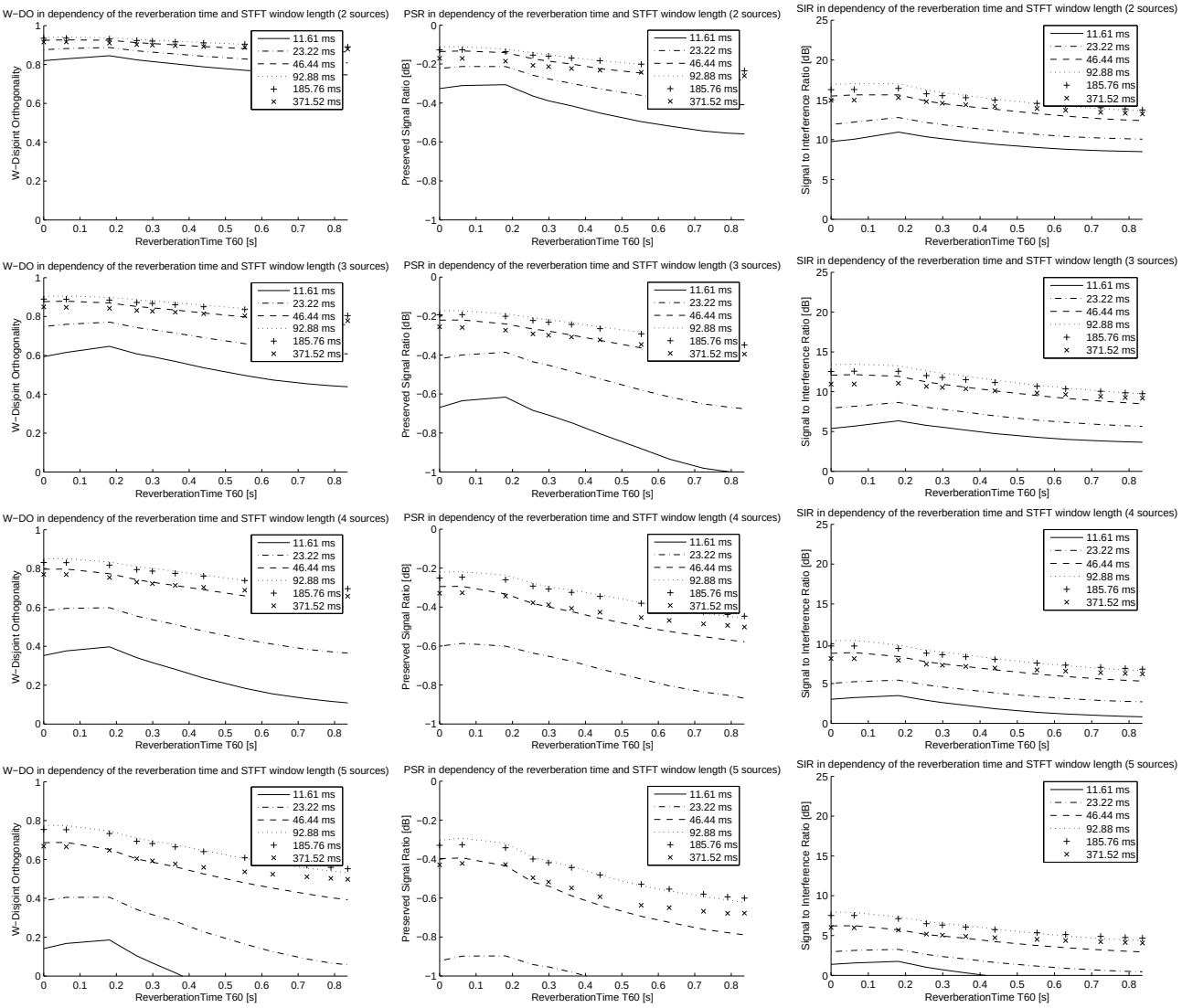

Figure 3.2: *Window-Disjoint Orthogonality in dependency of the window size of the STFT used for different reverberation times* T_{60}.

speech sources in highly reverberant environments decreases more slowly than in the case of shorter windows.

Figure 3.3 shows the dependence of the WDO, PSR and SIR values of the used window function. Most window functions perform very similar except the Kaiser window which is due to large spectral overlap of neighboring channels in the frequency response of the STFT filterbank (see figure 2.9). In all scenarios the Hamming window is a suitable and good choice for a STFT window function. The WDO decrease of speech sources in reverberant environments is not influenced by the choice of the window function.

To locate the time-frequency areas of the target source spectrum that exhibit large orthogonality, the threshold of the ideal mask computation is adjusted to include only those time-frequency bins where the energy of the target source is x dB larger than the energy of the interfering sources.

$$\Omega_i(t,f) = \begin{cases} 1 & s_i(t,f) - n_j(t,f) > x \quad \forall j \\ 0 & \text{else} \end{cases} \tag{3.2}$$

Brungart et. al [17] showed that the intelligibility of sources demixed by such binary masks with a threshold of up to 10 dB is still high and that recognition rates of more than 95 % are achieved even when using the 10-dB mask.

Figures 3.4 and 3.5 show the WDO, PSR and SIR values for ideal mask thresholds of 6 dB and 9 dB. Compared to the 0-dB mask, the SIR for two source scenarios increases by 4 dB for the 6-dB mask and by 6 dB for the 9-dB mask. For mixtures of five speech sources, the SIR increases analog to the two source scenario by 4 dB respectively 6 dB. The PSR of the 6-dB and 9-dB masks decrease compared to the 0-dB mask by 0.2 respectively 0.3 dB for mixtures of two sources and by 0.7 and 0.9 dB for five source scenarios.

This decrease in the PSR inevitably leads to losses in the quality of the reconstructed target source, as many parts of the original time-frequency spectrum are missing. On the other hand, these masks reconstruct signals that include only few energy of the interfering sources, which is the final goal of source separation architectures. The high SIR values indicate the suitability of the 6-dB and 9-dB masks to extract those parts of the target source, that include only very few energy from interfering sources also under reverberant conditions. To compensate for the loss in the SIR and PSR gain in reverberant environments compared to the anechoic scenario, more clever strategies in the computation of the time-frequency spectrum of the target source have to be applied.

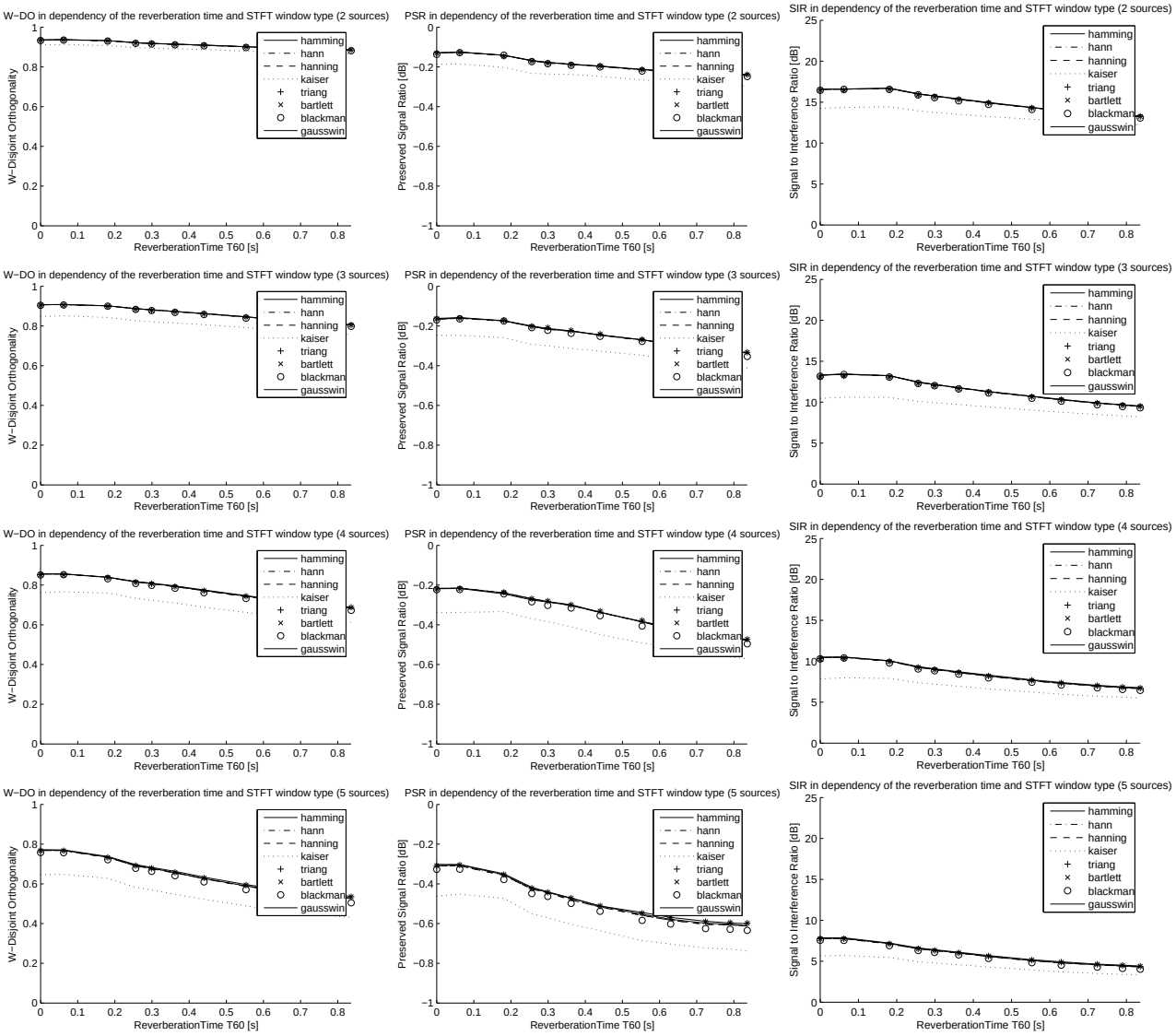

Figure 3.3: *Window-Disjoint Orthogonality in dependency of the used window function for different reverberation times* T_{60}.

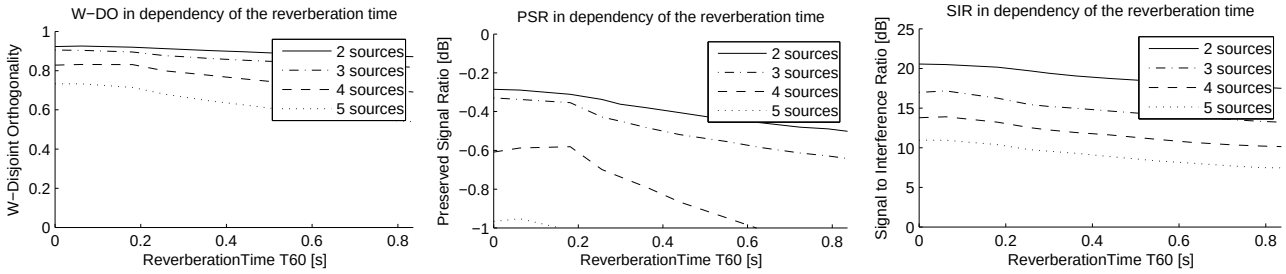

Figure 3.4: *Window-Disjoint Orthogonality in dependency of the reverberation time T_{60} for scenarios consisting of different sources for 6-dB ideal mask.*

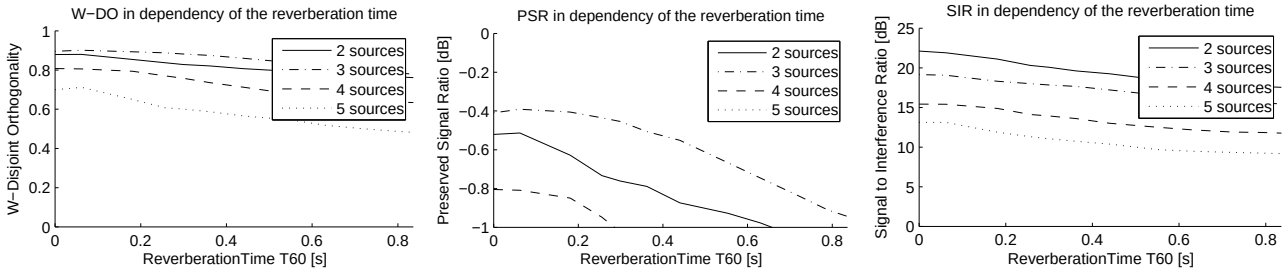

Figure 3.5: *Window-Disjoint Orthogonality in dependency of the reverberation time T_{60} for scenarios consisting of different sources for 9-dB ideal mask.*

3.3 WDO in simulated humanoid conditions

Specific source separation architectures (i.e. [102], [103], [113]) try to imitate the excellent source separation capabilities of the human auditory system by using a humanoid experiment setup. Auditory scenes consisting of several speech sources coming from different directions are recorded by a human dummy head to simulate the conditions humans experience in real life. Realistic outer ears, pinnae and the shape of the head perfectly imitate a real human head. The pinnae and the outer ear structures filter the incoming signals by specific filter functions in dependency of the incidence direction – the Head-Related-Transfer Functions (HRTF). The HRTF of each ear disturbs the incoming signals in time and in frequency and so affects the time-frequency spectra of speech signals. This section investigates if the concept of ideal time-frequency masks for source separation is also suitable for such humanoid setups or if the HRTF filtering process disturbs the signals in such a way that the orthogonality of speech sources in the time-frequency domain drops drastically.

Figures 3.6 – 3.8 show the WDO, PSR and SIR values for 0-dB ideal masks for two spatially separated sources at different positions in a humanoid scenario. The values are obtained by averaging over 20000 speech mixtures of different speakers of 3 seconds duration taken from the CMU arctic database [58]. The HRTFs used to simulate the humanoid setup and the spatial

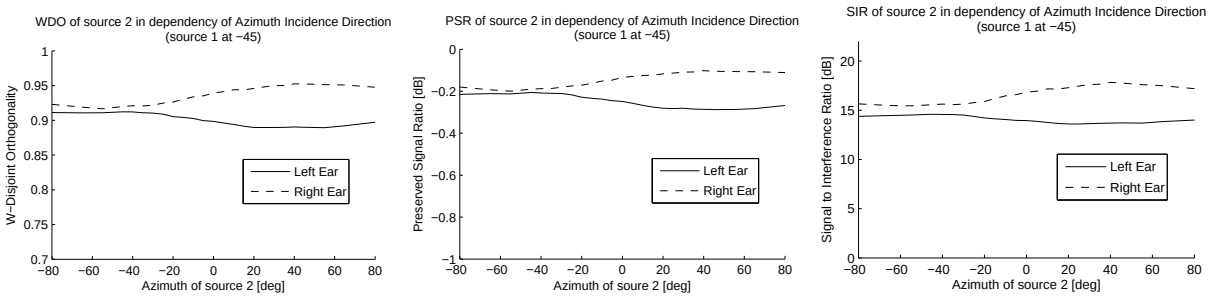

Figure 3.6: *Window-Disjoint Orthogonality of two speech sources in dependency of the incidence direction for left and right simulated human ear for 0-dB ideal masks (source 1 at -45 degree).*

positions of the sources are taken from the CIPIC HRTF database [2] and have been measured in an anechoic chamber with a KEMAR manikin [18] with small pinnae.

Figure 3.6 shows the results for a simulated humanoid two source scenario. Source 1 is considered to be fixed at position $-45°$ (negative degree values are assumed to be on the left side regarding the viewing direction of the head, positive degree values on the right side). The target source (source 2) is moving from $-80°$ to $80°$. It can be seen that the orthogonality of the speech sources is dependent on the relative positions of the two sources in the auditory scene and on the considered ear. For this scenario the best values in WDO, PSR and SIR for the right ear channel are obtained if source 2 is placed far away from source 1 in the right hemisphere. Then the target source is near the right ear, while the interfering source is on the other side of the head and gets attenuated by the natural head shadow. The signal parts of the target source arrive directly at the right ear without attenuation by the head. This relative increase of the loudness of source 2 compared to source 1 leads to a higher degree of orthogonality of source 2, which is seen by high WDO values for the right ear channel in the figure at positions larger than $40°$. For the left ear channel, the orthogonality is best, when the target source is assumed to be on the left side (near the left ear). Because of the interfering source on the left side, the loudness of the two sources is approximately equal and so the WDO, PSR and SIR values are lower than in the right ear channel.

Compared to the anechoic, non-HRTF filtered case of figure 3.1, the WDO decreases by 2 to 10 percent dependent on the source positions. The spatial HRTF filtering leads to SIR gains equal to the anechoic, non-HRTF filtered case for large spatial distances of the two sources (approximately 17 dB), if the better ear is chosen. For spatially nearby sources, SIR gains of only 14 dB can be achieved with ideal binary masks – a decrease of 3 dB compared to the anechoic case. Considering the relatively low maximal SIR gain of 17 dB, a decrease of 3 dB induced by the HRTF surely influences the quality of the separated target source.

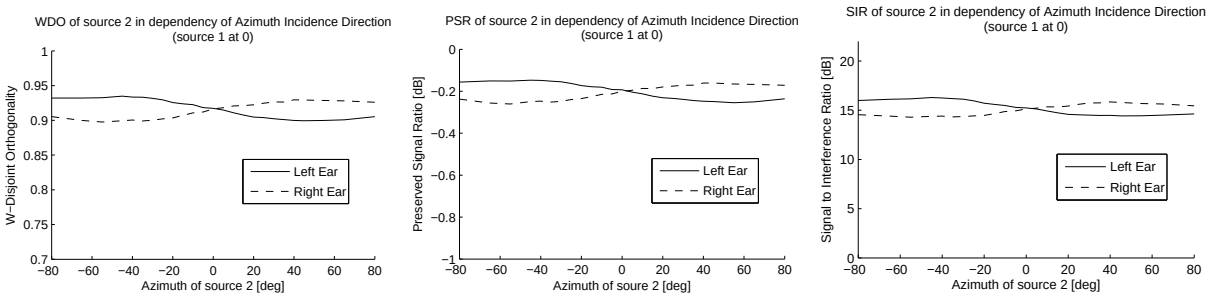

Figure 3.7: *Window-Disjoint Orthogonality of two speech sources in dependency of the incidence direction for left and right simulated human ear for 0-dB ideal masks (source 1 at 0 degree).*

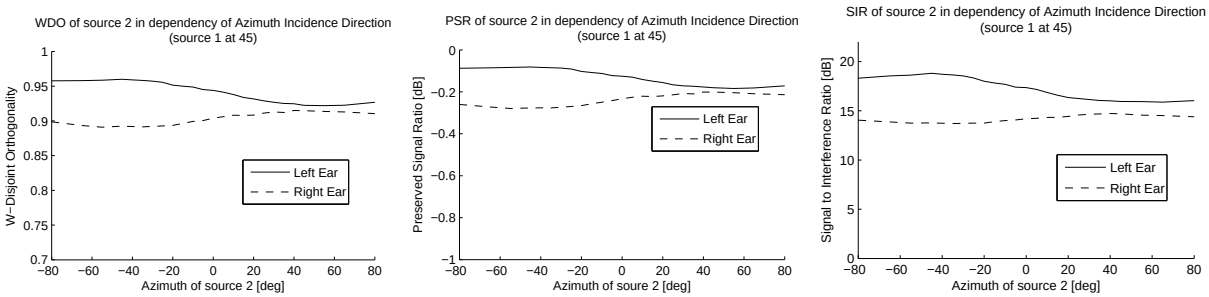

Figure 3.8: *Window-Disjoint Orthogonality of two speech sources in dependency of the incidence direction for left and right simulated human ear for 0-dB ideal masks (source 1 at 45 degree).*

Figure 3.7 evaluates the same scenario for a fixed source at position 0°. The influence of the head shadow can clearly be seen by the run of the WDO, PSR and SIR graphs. The orthogonality of the two speech signals is highest, if the sources are maximally separated in space (in this scenario, when source 2 is assumed to be at positions greater than ±40°). Then the incidence direction of the target source 2 is more direct than the incidence direction of the fixed source 1 at 0°, which leads to the previously described increase in the loudness of source 2, relative to source 1.

Figure 3.8 shows the analog evaluation of the scenario, assuming that the fixed source is at position 45°. The WDO, PSR and SIR graphs confirm the conclusions drawn from figures 3.6 and 3.7: The orthogonality of speech sources is higher for spatially separated sources than for nearby sources, because of the HRTF filtering that is dependent on the spatial position of the sources.

This evaluation leads to the following strategies that can be applied in humanoid source separation architectures to enhance the separation capabilities. If the spatial positions of the sources in the auditory scene are known in advance, the source separation architecture can choose the

ear with the higher expected SIR to perform the separation. If the current scenario of a static auditory scene recorded by a fixed human dummy head is extended to the dynamic case, where the dummy head and potentially also the sources are able to move, the separation capabilities by ideal-mask algorithms can be enhanced by aligning the dummy head to the currently optimal position regarding the orthogonality and SIR of the target speech source. For an evaluation of the ideal head position for several source separation schemes and the previous estimation of the source positions see chapter 6 or i.e. [103].

3.4 WDO under real reverberant humanoid conditions

To examine if the simulations made in the last two sections are also valid in real reverberant humanoid scenarios, this section investigates the concept of the ideal binary mask for binaural recorded speech signals. Five scenarios are accounted to determine the relative decrease of the orthogonality of speech sources in different environments and setups.

Scenario 1 simulates the anechoic case. Speech sources are artificially mixed by adding them together.

Scenario 2 simulates the reverberant case for a normal office room with reverberation time $T_{60} = 0.4\ s$. The room impulse response is generated as described in section 3.2.

Scenario 3 simulates the HRTF filtering of a humanoid setup with a human dummy head. The HRTFs for the specified positions are generated as described in section 3.3.

Scenario 4 simulates the reverberant humanoid case. The room impulse response of scenario 2 and the HRTF filtering of scenario 3 are combined to reproduce a realistic environment, where a human dummy head is situated in a reverberant office room and listens to speech sources coming from specific directions.

Scenario 5 uses real recordings of a human dummy head in a normal office room to investigate and relate the orthogonality of speech sources in real environments compared to the simulated cases. The speech signals are recorded by a human dummy head in a normal office room (see figure 1.1). The human dummy head (Neumann KU-100) is positioned in the center of a rectangular room like depicted in figure 4.2. The recording room is reduced to size $5 \times 8\ m$ by an acoustic curtain and measures a reverberation time $T_{60} = 0.4\ s$. The speech sources are played back by a conventional but high quality 7.1 surround sound system. The recording equipment like microphone amplifiers and the processing computers are placed in a neighboring control room to avoid additional noise.

Figures 3.9 and 3.10 show the mean WDO, PSR and SIR values for 200 speech mixtures of 3 seconds duration taken again from the CMU Arctic speech database [58] for speech mixtures of

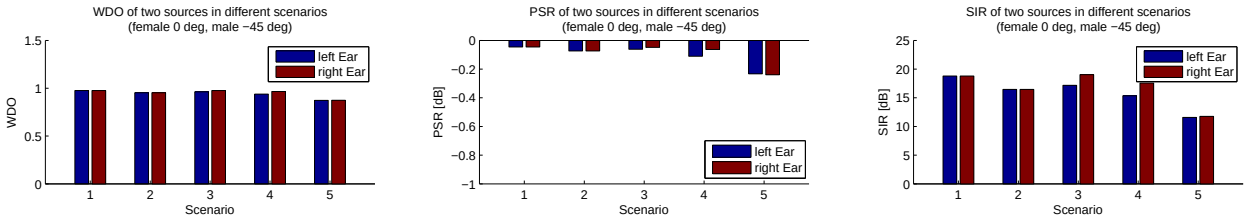

Figure 3.9: *Window-Disjoint Orthogonality of two speech sources (female at* $0°$ *and male at* $-45°$*) for the five different reverberation scenarios.*

two and three sources and for each of the five described scenarios. For each scenario the same corpus of speech mixtures is used to make the values comparable.

Figure 3.9 evaluates the orthogonality of speech sources for two source mixtures. The female target speaker is assumed to be at position $0°$. The second male speaker is considered to be at position $-45°$ (in the left hemisphere of the head's viewing direction). The figure shows the WDO, PSR and SIR values for the female target speaker for the five described scenarios. For the simulated reverberant scenario 2, the orthogonality decreases opposed to the anechoic case as described in section 3.2. Because the second source is considered to be in the left hemisphere, the orthogonality of the right ear channel is slightly higher than in the left ear channel. For scenario 3 and 4 the choice of the correct ear channel increases the SIR gain by approximately 3 dB. For the real binaural recording of scenario 5, the differences between the two ears only amount to approximately 1 dB. Opposed to the anechoic case, the SIR decreases by 5 dB compared to the unprocessed anechoic scenario 1. Nevertheless the choice of the correct ear channel in scenario 5 achieves in the mean of 200 speech mixtures about 1 dB SIR gain. Compared to the simulated reverberant humanoid conditions, the real reverberant humanoid scenario performs about 3 dB worse.

Figure 3.10 shows the WDO, PSR and SIR values for three source speech mixtures. The female target source is again considered to be at position $0°$. The two interfering male sources emanate at positions $\pm 45°$. The right ear channel shows slightly better WDO and SIR values. If the interfering sources would be equal and so have equal energy at all time instances, than the left and right ear channel should perform equal. In this evaluation the two interferers are two different male speech sources. Due to nature of the specific speech signal, the right ear channel has better SIR and WDO values than the left ear channel. Similar to the two source mixture, the real reverberant scenario 5 looses about 6 dB SIR against the anechoic scenario and about 5 dB against the simulated reverberant humanoid scenario 4.

Source separation methods based on ideal binary masks perform worse in real reverberant humanoid scenarios than in simulated reverberant humanoid scenarios. The decrease in SIR gain between simulated humanoid and real reverberant humanoid is about 3 dB for two source scenarios and about 5 dB for three source scenarios. The real reverberant scenario has a more complex

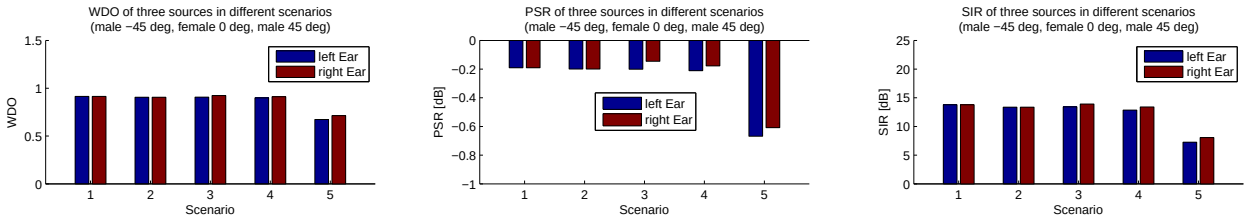

Figure 3.10: *Window-Disjoint Orthogonality of three speech sources (female at* $0°$ *and male at* $45°$ *and* $-45°$*) for the five different reverberation scenarios.*

room impulse response than the simulated room impulse response, which leads to stronger perturbations of the energy in the time-frequency spectrum. The simulated room impulse responses only take into account a limited number of reflections [73] and so also limit the number of disturbed time-frequency bins.

4 Experiment Setup

A movable human dummy head residing in a normal office room is used to closely imitate the conditions humans experience when being situated in a complex auditory scene such as a cocktail party. The robotic dummy head – called Bob – consists of a Neumann KU-100 dummy head which is normally utilized for authentic binaural recordings. Played-back over headphones, such binaural recordings create the illusion that the listener feels like being in the recorded scene.

Figure 4.1 shows the assembly of the robotic dummy head Bob. Bob's torso is realized by a mechanical pan-tilt-roll-unit (PTR-unit). The system consists of three independent and orthogonal axes which correspond to the three axes in a three-dimensional coordinate system. Each axis is driven by a separate servo motor that is controlled by a corresponding motion controller (PMS 5005). The motion controllers communicate via RS232 with a C-driver that is used to move the head in any favorable position. The PTR-unit is able to move the head in almost any human like position. The pan direction can move the head in 360° and so simulates the motion of a human torso and body. The tilt and roll behavior is constrained to work in degree ranges of $-30°$ to $30°$, similar to the human ability to bend and nod the head.

Bob is resident in the Media Lab of the Telecommunications Lab at Saarland University. Figure 4.2 shows the layout of the Media Lab. The Media Lab is a normal office room of size 10×6 m which is adapted to work as the cocktail party location. The walls are covered with big wooden boards, the floor is carpeted and the ceiling consists of plastic covers and neon light lamps. Directly connected to the Media Lab is a control center which accommodates the control computer and the microphone and loudspeaker amplifiers. The control center allows to create and record auditory scenes without disturbing the characteristics of the auditory scene by computer or human-made noise. A window from the control center to the Media Lab admits observing of the scenario.

To construct the auditory scene, a conventional but high-quality 7.1 loudspeaker system (Nubert *nuLine-120-Set 5*) is utilized. Because of the limited frequency response of the subwoofer, only the seven normal loudspeakers are utilized to generate an auditory scene. These seven loudspeakers are arranged in a circle around the head to easily specify sound sources from specific azimuth directions.

An acoustic curtain reduces the size of the media lab to 8×5 m and attenuates severe reflections and reverberation. The curtain encloses the auditory scene from three sides as shown in figure 4.2. The reverberation time T_{60} – the time required for reflections of the direct sound to decay by

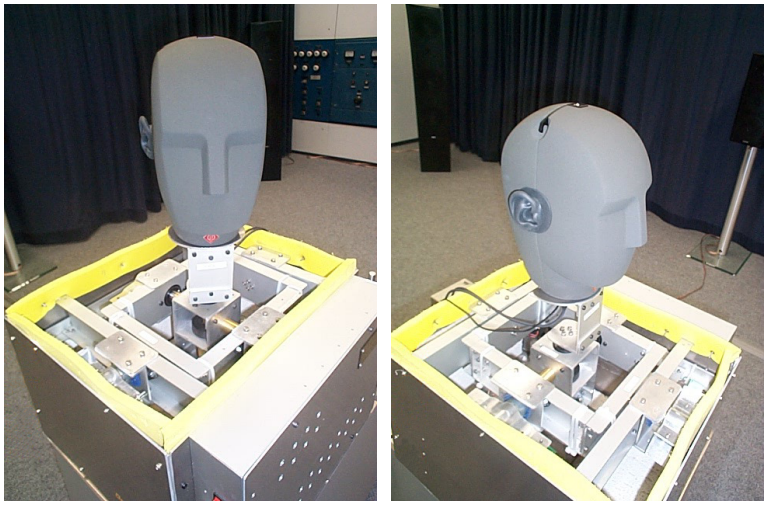

Figure 4.1: *Bob - the robotic head. The head consists of a Neumann KU-100 dummy head. The movable torso is realized by a mechanical pan-tilt-roll unit which is controlled via RS232.*

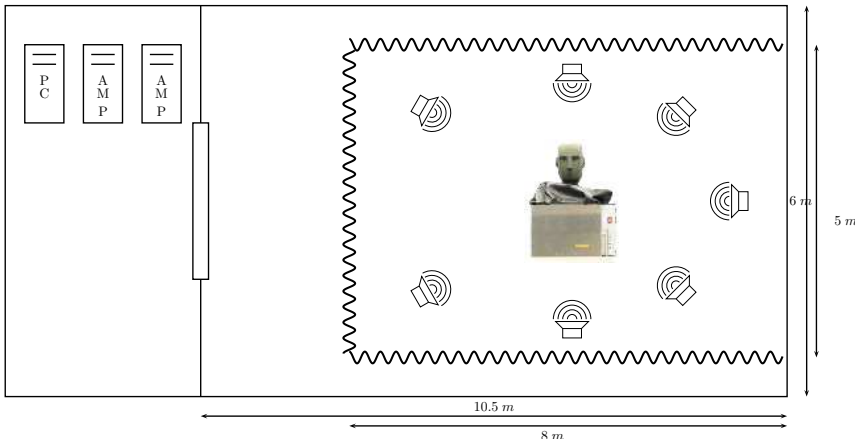

Figure 4.2: *Layout of the Media Lab and the adjacent control center.*

Figure 4.3: *Bob residing in the Media Lab.*

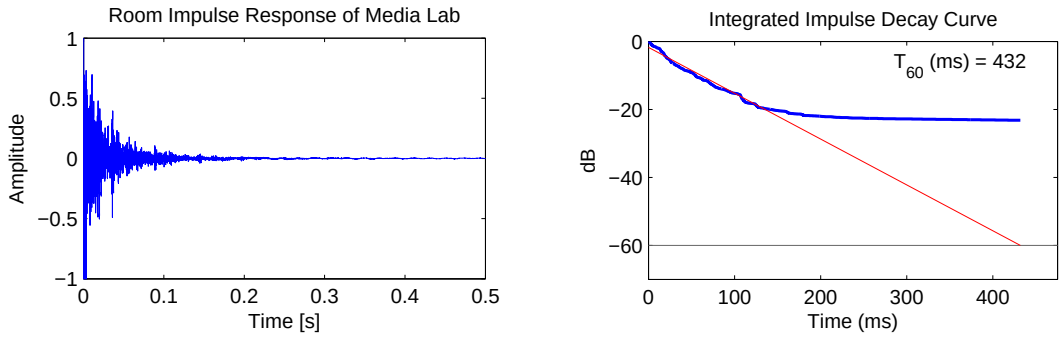

Figure 4.4: *Room Impulse Response and Integrated Impulse Decay Curve of Media Lab used to estimate the reverberation time* T_{60}.

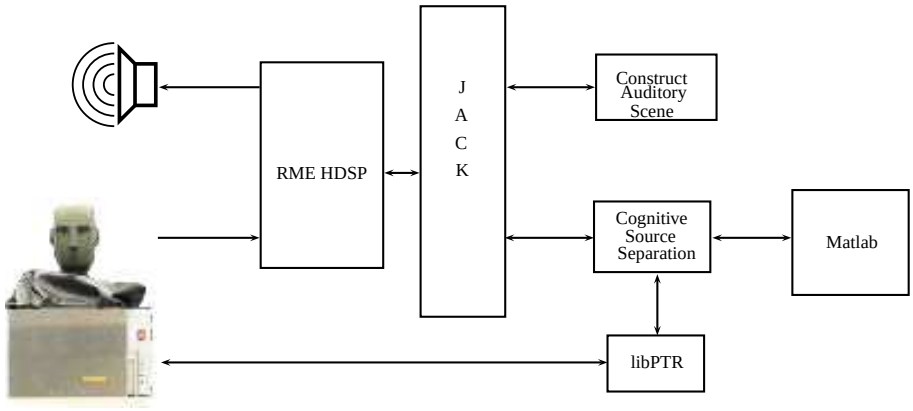

Figure 4.5: *Hardware and software architecture.*

60 dB – of the remaining area is measured as $T_{60} \approx 0.4\ s$. The value is estimated by using Schroeder's integrated impulse decay method [100] [101]. The room impulse response of the recording area (depicted in the left plot of figure 4.4) is transformed to an integrated impulse decay plot by reverse time integration (see right plot of figure 4.4). The reverberation time T_{60} is estimated by evaluating the decay time of the straight part of the slope.

Auditory scenes are created by moving the loudspeakers to the corresponding positions of sources. The parallel playback of the sources and the recording of the sound at Bobs ears is managed by a low-latency audio server called *JACK Audio Connection Kit*[1]. Figure 4.5 shows the overall software design for the source separation framework. *JACK* controls a *RME HDSP 9632* soundcard which is responsible for the output of sound sources to specific loudspeakers and for the input of the two ear microphone signals. One client is accountable to construct the auditory scene, which is specified by several sound files and the corresponding loudspeakers where each soundfile should be played-back. Upon start-up the scene-constructing-client in turn starts a soundplayer for each source in the auditory scene and sends requests to the *JACK* server to connect the sources with the dedicated loudspeakers. The source separation client captures the two ear signals and saves them to Matlab binary files. Additionally the source separation client is able to move the head in any favorable position by using the library *libPTR*, which connects via RS232 to the PTR-unit and controls the servo motors. After recording the desired auditory scene from the desired head perspective the further processing to separate the sources is done in Matlab.

The sound files used to construct the auditory scene are taken from the freely available speech database CMU Arctic [58] recorded by the speech lab of the Carnegie Mellon University (CMU) in Pittsburgh. The CMU Arctic database consists of speech files recorded under studio conditions. Originally it is used for speech synthesis. A big effort has been done to make high quality

[1] http://www.jackaudio.org

recordings without reverberation and background noise. The speech sentences are recorded in the CMUs speech lab in a soundproof anechoic booth. The text is read by expert speakers which sit between 6 and 12 inches from a near field condenser microphone. The speech is recorded by a sampling rate of 44.1 kHz. The text material is taken from freely available ebooks from the author Jack London. Two male and two female native English speakers with a North midland American accent, one male speaker with native Southern Ontario Canadian dialect and one male person speaking South Eastern Scottish accent from Edinburgh served as reading experts. The sound files of the CMU arctic database consist of sentences of three to five seconds duration. Longer audio files of arbitrary length are realized by concatenating different sentences of the same speaker.

5 Auditory Scene Exploration

When humans enter an auditory scene, they automatically analyze the environment around them. Consider a person entering a cocktail party. The person i.e. first looks around and estimates coarsely how many other people are there, where these persons are positioned and who they are. Perhaps additional sound sources like a running TV or a barking dog are recognized. When the person first starts a communication with one of the guests, he has already gathered a lot of information about the auditory scene, that can be used to enhance and simplify the source separation that is needed to understand the current talk. Sound fragments that do not belong to the conversational partner, but are instead identified to belong to a person in the opposite direction are filtered out and so enable the concentration to the talk.

The source separation architecture presented in this thesis tries to mimic these cognitive abilities of the human brain and claims that such prior information regarding the auditory scene is useful to enhance the separation process. Prior to separating the speech sources, the human dummy head Bob analyzes the auditory scene and estimates several parameters that can be used to enhance or enable later separation approaches [60].

The auditory scene analysis is described in the following sections. The dummy head Bob estimates the number and the positions of the sources in the auditory scene and appraises the fundamental frequency track of the speech sources in the auditory scene. This information is used as input in the later presented source separation algorithms.

5.1 Source Localization

The human brain mainly uses Interaural Time Differences (ITD) and Interaural Level Differences (ILD) of the signals received at the two ears to perform source localization in the azimuth plane as described in detail in section 2.1.3. The time differences arise due to the distance between the ears. Level differences are ascribed to the solid head, which introduces diffraction and scattering of the sound waves and accounts for significant head shadows at the ear that is turned away from the sound source.

The following sections describe the localization of the sound sources in the auditory scene in the azimuth plane. The localization of the sources in the median plane is not investigated in this thesis as the median localization does not – for the current source separation architecture

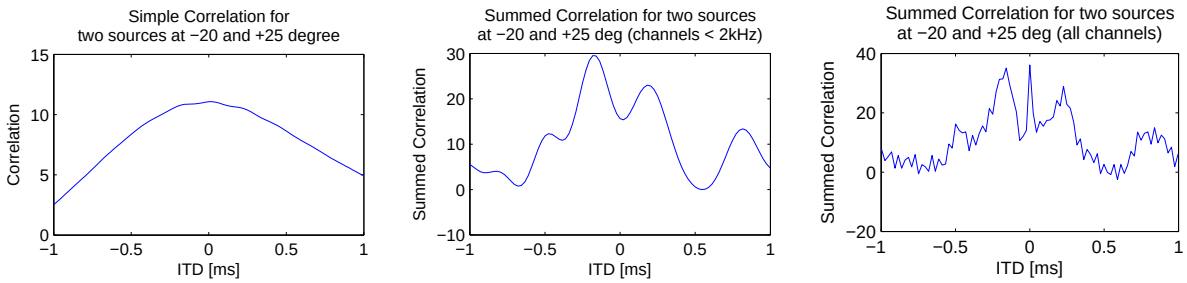

Figure 5.1: *Comparison of the simple correlation function (eq. 5.1) and the summed correlation function of the correlogram (eq. 5.3) for two speech sources at positions* $-20°$ *and* $+25°$ *recorded with Bob.*

– introduce further useful information. For the interested reader a localization approach for the median plane with Bob is described by Haschke [45].

5.1.1 Estimation of ITD and ILD

The ITD between the two ears can be estimated by correlating the two ear signals and finding the peak in the resulting function. Assume x_L and x_R denote the time domain signal of the left and the right ear. The correlation function is defined as

$$R(l) = \sum_t x_L(t+l) \cdot x_R(t). \tag{5.1}$$

The interaural time difference is then estimated as the time value corresponding to the highest peak in the correlation function.

If the two ear signals include energy from several spatially separated sources, the ITD estimation based on the correlation function is often incorrect. In the ideal case for i.e. two sources, there will be two peaks in the correlation function corresponding to the positions of the two sources. In most cases – especially in reverberant environments – these two peaks merge to one peak at an intermediate position, so neither the position of the first, nor the position of the second source can be estimated reliably. The left plot of figure 5.1 shows the simple correlation function of equation 5.1 for a one second speech mixture consisting of a female voice at position $-20°$ and a male voice at position $+25°$ recorded with the human dummy head Bob under reverberant conditions. Instead of two clear peaks at the ITD values corresponding to the positions of the sources (approximately -1.5 ms for $-20°$ and 2.0 ms for $25°$), there are only two little peaks at those locations (which can be seen after detailed inspection of the graph). The highest point of the correlation function is misleadingly positioned at zero ms.

These ambiguities can be resolved by converting the incoming signal to the cochleagram representation and computing the crosscorrelation separately for each frequency channel (see i.e. [75] [80]), which results in a three dimensional function

$$R(l,c) = \sum_t x_L(t+l,c) \cdot x_R(t,c), \tag{5.2}$$

where l denotes the time lag between the left and right signal and c represents the frequency channel. By summing the correlation functions of each channel, a final estimate of the correlation function is computed as

$$R_{\text{summed}}(l) = \sum_c R(l,c,\tau) \tag{5.3}$$

The middle and right plot of figure 5.1 show the summed correlation function for the same signals as the right plot. Opposed to the simple correlation, two clear peaks at the correct ITD positions can be examined. In the middle plot only frequency channels with center frequencies below 2 kHz are used. Channels of higher frequency tend to exhibit peaks at zero ms as the wavelength becomes much smaller than the head diameter. This peaks lead to false peaks at position zero ms in the summed correlation function as can be seen in the right plot.

The interaural level difference between the left and the right ear signal is computed by subtracting the power of the left signal and the power of the right signal.

$$\Delta L = L_{left} - L_{right}. \tag{5.4}$$

$$\Delta L = 10 \cdot \log_{10} \frac{\sum_t x_L^2(t)}{\sum_t x_R^2(t)} \quad dB. \tag{5.5}$$

5.1.2 Incidence Angle Estimation

This section first describes a new circle-formula for estimating the incidence direction of the auditory source based on the available ITD which assumes that the sources are located on a circle around the listener. This formula is compared to two standard formulas to transform the ITD values to the estimated incidence direction: the Freefield formula and the Woodworth formula. Additionally an approximation of the HRTF for the described scenario is evaluated for its source localization capabilities.

Formula for circular loudspeaker constellation

In the current scenario, the dummy head Bob is located in the center of a normal office room, while the loudspeakers are arranged on a circle around as depicted in figure 5.2. For such a scenario, where the sources are assumed to emanate from a circle with a specified radius, the following approach is used to localize the sources.

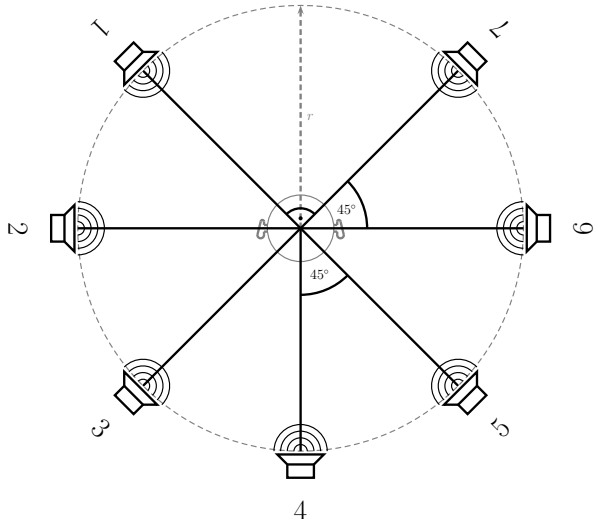

Figure 5.2: *Schematic view of loudspeaker constellation.*

Consider the scenario shown in figure 5.3, where the head is assumed to have radius r_h. The traveled distance of the left and right source to the two ears s_l and s_r differs by

$$\Delta s = s_l - s_r \tag{5.6}$$

The resulting interaural time difference is estimated by

$$\Delta t = \frac{\Delta s}{c} \tag{5.7}$$

where $c = 343\frac{m}{s}$ is the speed of sound in air. Assuming that the sound source is always located on the circle, the distance between the sound source and the center of the head is always r. Substituting equation (5.7) into equation (5.6) yields

$$s_l - s_r = \Delta t \cdot c \tag{5.8}$$

The distances of the sound source to the left and the right ear unfold after simple trigonometric transformations to (compare to figure 5.3)

$$s_l = \sqrt{r^2 + 2 \cdot r_h \cdot r \cdot \sin\phi + {r_h}^2} \tag{5.9}$$

$$s_r = \sqrt{r^2 - 2 \cdot r_h \cdot r \cdot \sin\phi + {r_h}^2} \tag{5.10}$$

Inserting equations (5.9) and (5.10) into equation (5.8) yields after rearranging terms a function of ϕ in dependency of the interaural time difference Δt

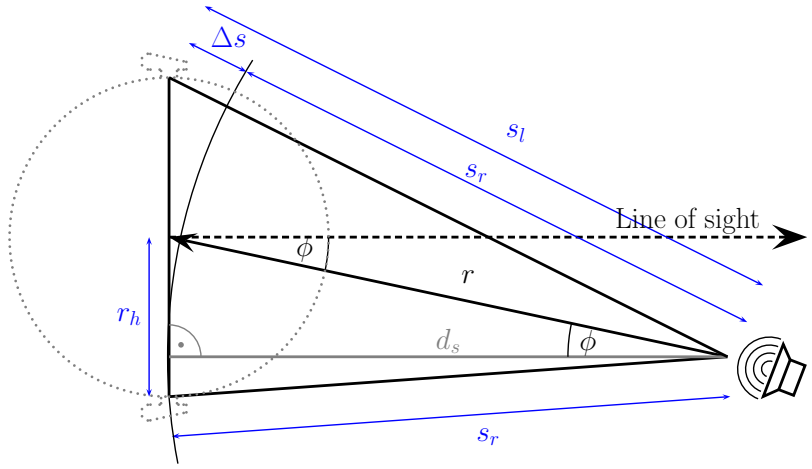

Figure 5.3: *Free-field propagation of sound waves through a transparent head.*

$$\phi(\Delta t) = \frac{180^\circ}{\pi} \cdot \arcsin\left[\frac{\Delta t c}{2 r r_h}\sqrt{\frac{1}{2}\left(r^2 + {r_h}^2 - \frac{1}{2}\Delta t^2 c^2\right)}\right] \tag{5.11}$$

For a detailed derivation of the formula see the work of Haschke [45].

Freefield Formula

Figure 5.3 shows the scenario when the head is assumed to be a perfect sphere with radius r_h and the sound waves are supposed to be able to travel through the head without diffraction and reflection. As derived in detail in section 2.1.3 the incidence direction of a sound wave can be estimated by

$$\Delta t(\phi) = \frac{2 r_h \sin\phi}{c}. \tag{5.12}$$

Woodworth Formula

In the case of a head modeled as a perfect solid sphere, the sound waves diffract and reflect at the turned-away side. Accounting for the diffraction characteristics, the length of the traveled path of the incident sound wave is longer than in the free-field case. Motivated by this, Woodworth and Schlosberg [122] applied diffraction theory to a completely spherical head, yielding the following formula to approximate the ITD:

$$\Delta t(\phi) = \frac{r_h(\phi + \sin\phi)}{c} \tag{5.13}$$

Figure 5.4 shows a comparison of the Circle, the Freefield and the Woodworth formula. In the range of -15° to $+15^\circ$ the three formulas estimate the positions almost equally. For degree values greater than $\pm 40^\circ$ the three formulas differ by up to 20° in the estimations.

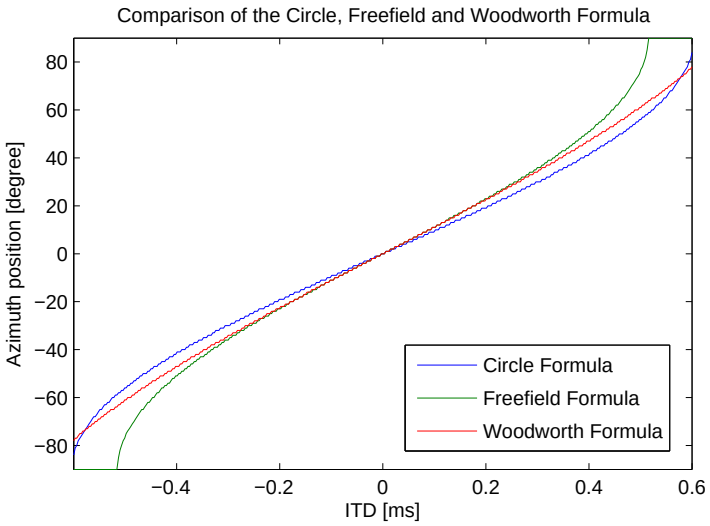

Figure 5.4: *Comparison of the Circle, Freefield and Woodworth formula for different ITD values.*

Head Related Transfer Function

The complete refraction and resonance characteristics of the human dummy head Bob and the media lab can be specified by measuring the Head Related Transfer Function (HRTF) for ITD and ILD. The localization of a source in the auditory scene is then approached by using the HRTF as a table look-up: The ITD and ILD of the left and right signal are computed and compared to the HRTF to find the dedicated position.

HRTFs are very sensitive to changes in the setup and the performance severely degrades, if the recording setup and the application scenario differ [114]. Even for different head positions in the same environment differences in the HRTFs occur (see figure 5.5).

The HRTF of the human dummy head Bob is measured by playing back white noise from each of the seven loudspeakers (positions depicted in figure 5.2). For each loudspeaker position, Bob turns from 0° (looking straight to the loudspeaker) to 360° (again looking straight to the speaker) in 1°-steps and records one second of white noise in each case. For each loudspeaker the incoming signal is filtered with a gammatone filterbank of 512 channels. For each of the 512 frequency channels of the resulting cochleagram and each of the 360 source positions, the ITDs and ILDs are computed. The complete HRTFs for ITDs for each loudspeaker position are plotted in figure 5.5.

To smooth out local variations and to make the HRTF more robust against changes in the environment, a mean HRTF is computed by taking the mean of all seven measured HRTFs. Inspecting the mean ITD-HRTF yields that each channel can roughly be approximated by a sinusoid of a specific frequency and amplitude – as already mentioned by other researchers

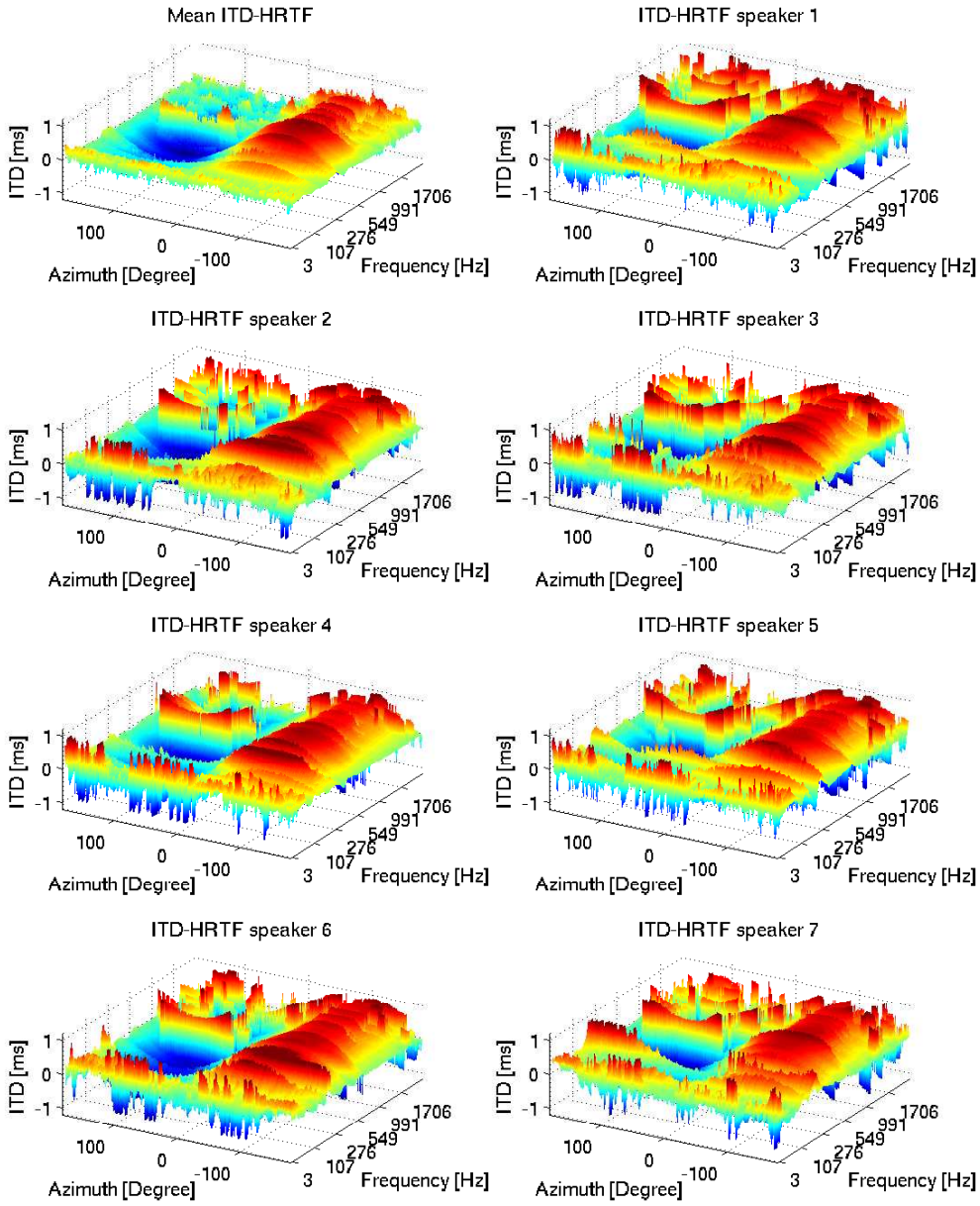

Figure 5.5: *HRTF for ITD and ILD for dummy head Bob residing in a normal office room.*

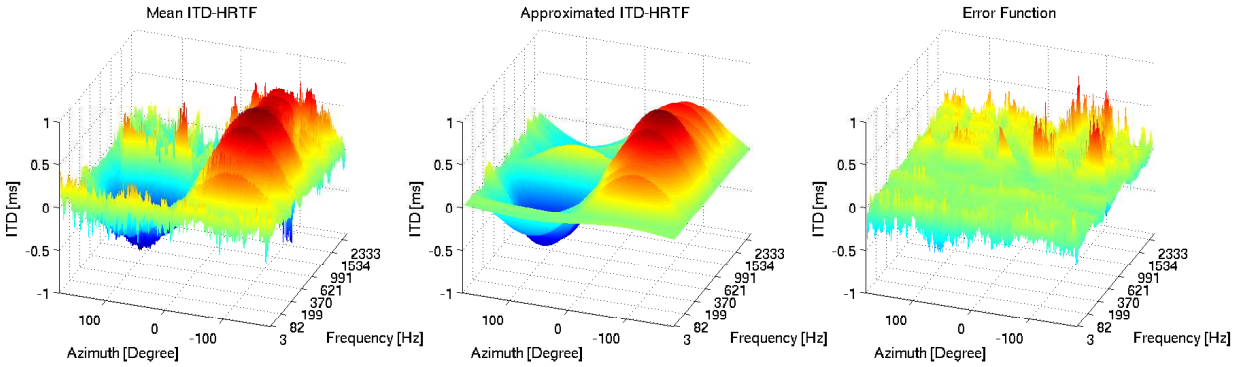

Figure 5.6: *Mean measured HRTF for ITD and approximated HRTF.*

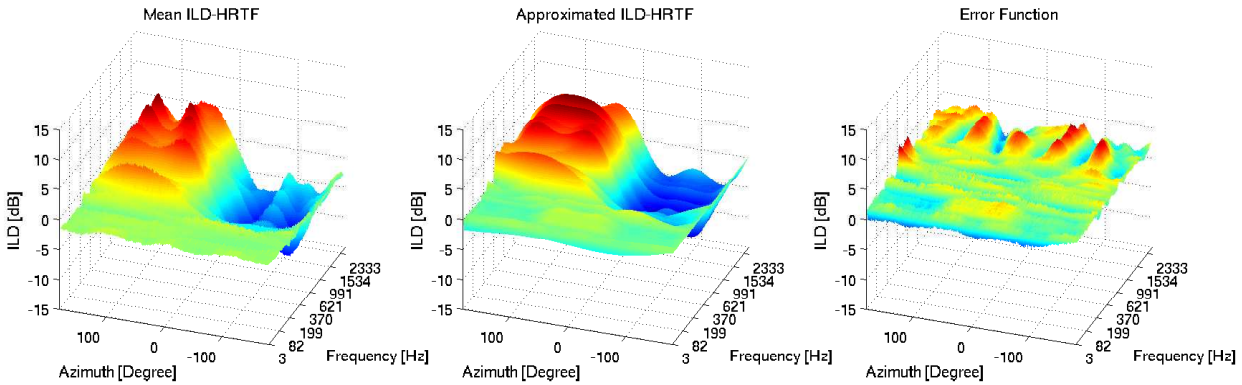

Figure 5.7: *Mean measured HRTF for ILD and approximated HRTF.*

(i.e. [113], [114]). An approximate HRTF is then found by fitting each channel to a sinusoidal model:

$$\Delta t(\phi, f) = \alpha_f \cdot \sin(\omega_f \phi),$$

where α_f denotes the frequency dependent scaling factor, and ω_f specifies the frequency dependent sinusoid frequency. The mean HRTF, the approximated HRTF and the resulting error function are plotted in figure 5.6. Especially in the low frequencies up to 1 kHz, the sinusoidal model fits very well and the error function shows only little deviations. This is consistent with the so called Duplex Theory which describes the human source localization based on the combined evaluation of the physical cues ITD and ILD and was first identified by Lord Rayleigh [86]. In frequencies larger than 1 kHz, the ITD starts to become ambiguous as the wavelength of the signal becomes comparable to the size of the head and the correlation function of the left and right ear signal starts to exhibit several peaks. For frequencies greater than 1 kHz, the ILD cues become a reliable measure of the incidence direction as the signals of short wavelength are not refracted by the human head – they either are reflected completely or pass with little refraction.

In analogy to the approximation of the ITD-HRTF, the ILD-HRTF can be approximated by a sinusoidal model. Viste [113] for example also approximates the ILD with sinusoids of different frequency and amplitude. Inspecting the mean ILD-HRTF however shows additional peaks and valleys in the curves, especially in frequencies higher than 1 kHz. In the case of the dummy head Bob, a model consisting of only one sinusoid does not adequately model the ILD-HRTF. Fitting harmonic Fourier series of length two however already gives a good and simple approximation of the function.

$$\Delta l(\phi, f) = \alpha_f \cdot \sin(\omega\phi) + \beta_f \cdot \sin(3\omega\phi) \tag{5.14}$$

This model is analog to a model used by Duda et. al [29]. Duda et. al noted that the ILD is periodic in ϕ and approximate the ILD by complete Fourier series expansions.

The result of the approximation is plotted in figure 5.7. In the low frequencies up to 800 Hz, there are only little ILDs as the waves pass through the head without reflection. Above 800 Hz the used model approximates the ILD well up to approximately 1.5 kHz. Above 1.5 kHz additional sinusoidal vibrations occur. Adding further harmonics to the model can eliminate these peaks and valleys and leads in the end case to the model used by Duda et. al [29].

At the current time there are no simple models known that describe the characteristics of the ILD in general [113] as in the case of the ITD, where the Freefield-Formula, the Woodworth-Formula and the Formula for a circular arrangement of the sound sources already give easy and good approximations for the estimation of the incidence direction.

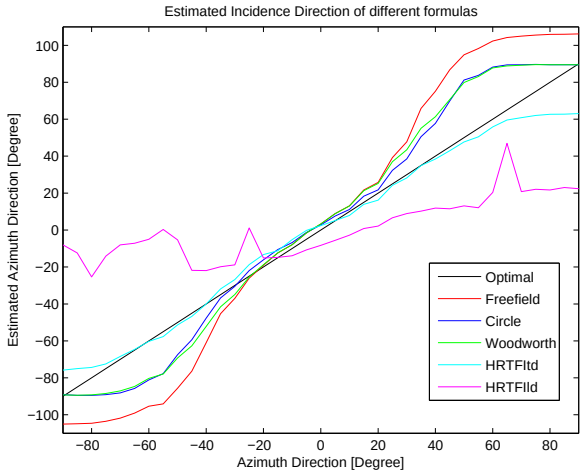

Figure 5.8: *Comparison of the five described formulas for incidence direction estimation based on the interaural time and level difference.*

Results

To estimate the incidence direction of a sound source in the auditory scene recorded by the human dummy head Bob, the five described formulas are compared to each other to find the most suitable and computational manageable algorithm for source localization. Figure 5.8 shows the results of the five described estimation formulas for sound source locations of $-90°$ to $90°$. The values are obtained by taking the average values for 240 speech sources of one second length taken from the speech database CMU Arctic [58] and played back from the specified directions. Each sound source is recorded with a sampling rate of 44.1 kHz. The estimation of the location by the HRTF formulas is performed by computing the average estimated direction of several frequency channels. For the ITD-HRTF the ITD value is used as table look-up to find the most probable incidence direction for all channels between 200 Hz and 1000 Hz, the range where the error function of the approximated ITD-HRTF is almost zero. The ILD-HRTF computes the location of the source analogously but uses only channels from 800 Hz to 1400 Hz as the error function of the ILD-HRTF is small in this range.

The black line shows the optimal results, where each sound source is assigned to its correct incidence direction. The Freefield formula (the red line) performs well for incidence directions between $-20°$ and $+20°$ and shows a localization blur of approximately three degree, which is comparable to the human localization blur [10]. For values greater than $\pm 20°$, the algorithm overestimates the directions by up to $40°$. The Woodworth formula (the green line) performs similarly and comparable to the Freefield formula, but the overestimation of incidence directions greater than $\pm 20°$ is only up to $20°$. The formula for the circular arrangement (the blue line) of the loudspeakers performs about 4 to $5°$ better than the Woodworth and the Freefield formula

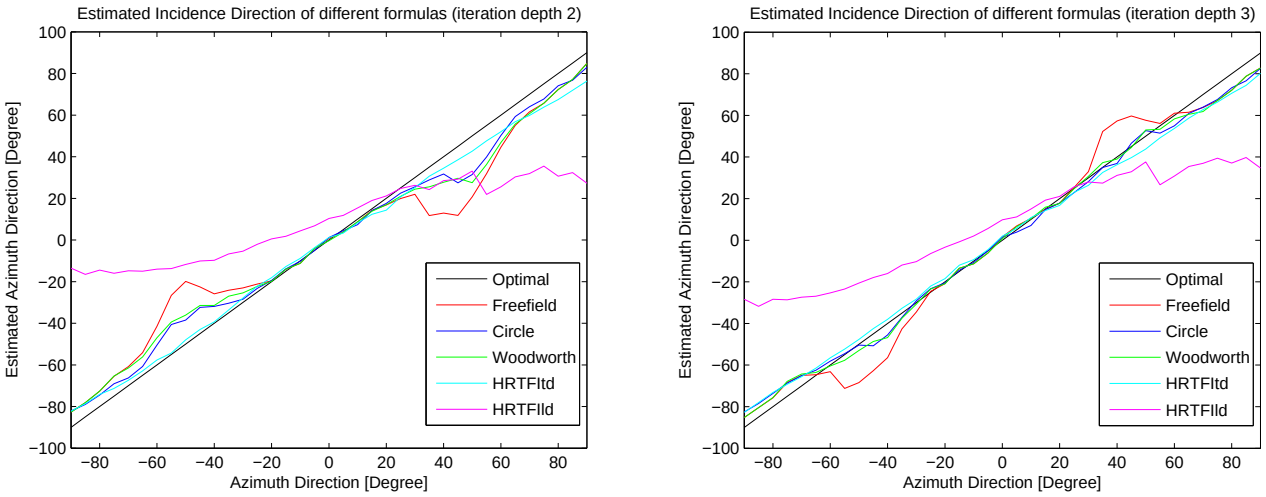

Figure 5.9: *Results of the iterative localization with two (left plot) and three (right plot) iterations.*

especially in the range between $\pm 20 - -40°$ and therefore should be preferred to cover a higher reliability range. The HRTF of the interaural time differences outperforms the other formulas and estimates the location of the sound source reliably up to $\pm 50°$ with less than three degree localization blur. The HRTF for the ILD performs only poorly and can only be used to derive a coarse direction like "on the left side" or "on the right side".

The presented algorithms for localization deliver reliable incidence direction estimations for source locations between $\pm 20°$ for the formulas and $\pm 50°$ for the HRTFs. The human dummy head Bob should be able to localize sources in the complete horizontal plane. To estimate also incidence directions greater than $\pm 50°$ reliably, the location estimation is used iteratively. In each iteration Bob turns to the estimated direction and refines his computation until a stable result occurs.

Figure 5.9 shows the results for the different formulas for two and three iterations. The ITD-HRTF formula estimates the directions almost optimal with less than three degree localization blur even in the two iteration case (see left plot), but the accuracy slightly increases in the three iteration case (see right plot).

For two iterations, the ITD formulas refine their accuracy in the range up to $\pm 30°$ and for incidence directions greater than 60° and achieve a localization blur of less than three degree for these directions. In the range between 30° and 60° on both sides of the head, the estimation suffers. While the Circle and the Woodworth formula misjudge the incidence directions by up to 10° in this range, the Freefield formula evaluates falsely up to 30°. The fall-off of the performance for the Circle, Woodworth and Freefield formula in this range can be traced back to the moderate performance of these formulas in the one-iteration case. Assuming a source position of 45°, the Freefield formula estimates the position to approximately 90° (see graph in figure 5.8). The

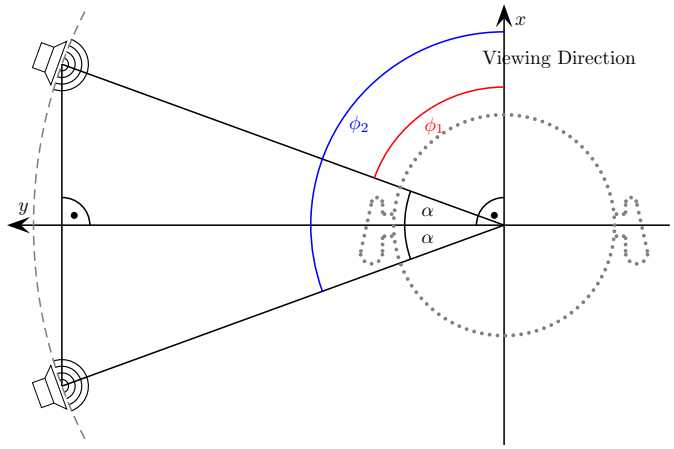

Figure 5.10: *Front-Back Confusion of single sound source.*

head is then moving to position 90° and the new relative position of the source is at −45°, which in turn leads to a false estimation of −90° and so on and so forth. The performance of the Woodworth and the Circle formula in this range can be explained analogously, but the first estimation is approximately 70° and the next iterations can resolve this error as can been seen in the three iteration case. The third iteration increases the accuracy of the estimation and the Woodworth, the Circle and the HRTF-ITD formula perform almost equally.

The results of the ILD-HRTF have been improved compared to the non-iterative case, but cannot keep up with the incidence estimation based on ITD. The performance of the ILD estimation has increased compared to the two iterations case, but is still not comparable to the incidence estimation based on ITD. By using more iteration steps, the ILD estimation can be further refined and an acceptable accuracy is achieved by approximately eight iterations.

These results show, that a detailed HRTF is not necessarily needed to perform a reliable source localization. By using a moving dummy head – just like a real human head – the iterative variants of the simple formulas like the Circle or Woodworth formula, perform equally to the HRTF, but are far more easier to compute compared to the HRTF, which is also heavily dependent on the used head and the environment. Especially in CASA systems which have to work under real-time conditions the easy to compute iterative variants of the formulas are advantageous.

To localize also sources in the back of Bob, a front-back discrimination is implemented as described in the next section.

5.1.3 Front-Back Confusion

All algorithms used for location estimation (except the HRTF) assume a spherical head and a symmetrical setup regarding the front and back direction. These formulas always assume that the sound source is located in front and return an estimated incidence between ±90° as they

are only invertible in this range. They are not able to distinguish, if a source is coming from the front or the back direction. Signals coming from the back are localized erroneously at the mirrored frontal position.

Assuming a constant distance of the sound sources, there always exist two incidence directions in the horizontal plane that exhibit the same time and level differences as depicted in figure 5.10. Denoting these directions with ϕ_1 and ϕ_2, the two possible source locations are related to each other by

$$\phi_2 = \phi_1 + 2 \cdot \alpha \tag{5.15}$$

All formulas – including the HRTF – approximate the ITD and ILD by a kind of sinusoidal function. The inverse table look-up of the time and level differences return degree values between $\pm 90°$, which always corresponds to position ϕ_1. It is therefore not possible to infer with these formulas, if the source is coming from position ϕ_1 or position ϕ_2.

If the distance of the sound source is not constant, the two possible locations of the sound source expand to half-lines emanating in directions ϕ_1 and ϕ_2. If the elevation dimension is additionally regarded, the set of possible locations further expand to the rotational solid of the two half lines, which results in the so called cone of confusion [10]. Experiments with human subjects revealed that for a real – not completely spherical – human head the half lines look more like hyperbolas, which corresponds to an hyperboloid in the three dimensional case [9].

Humans use the frequency and direction dependent filtering of the outer ear to determine if the sound source is coming from the front or the back direction [10]. Additionally humans use head motions to resolve front-back ambiguities [10]. A slight movement of the head yields a specific change in the ITDs and ILDs and is used to estimate, if the source is located in the front or the back direction (compare to figure 5.11).

The iterative source localization algorithm described in the last section resolves most of the front-back confusion errors by the separate incidence direction estimation in each direction. Figure 5.12 shows the results of the iterative localization for a variable number of iterations. The algorithm stops, when a stable result is reached or a maximum of seven iterations has occurred. All formulas perform comparable to the only front-estimation case. The Freefield formula further exhibits the peaks at $\pm 45°$ for the same reason described above. For incidence directions with absolute values greater than $170°$, the algorithms sometimes erroneously localize the mirrored location in front, which leads to the decreased mean value in the figure.

For incidence directions with absolute values greater than $170°$ an additional front-back estimation therefore evaluates if the source is coming from the front or the back. The front-back confusion is resolved by using the moving ability of Bob. During the iterative detection of the location of the source, Bob notices the changes in the ITD and the ILD in every iteration according to the corresponding positions of the head in each iteration.

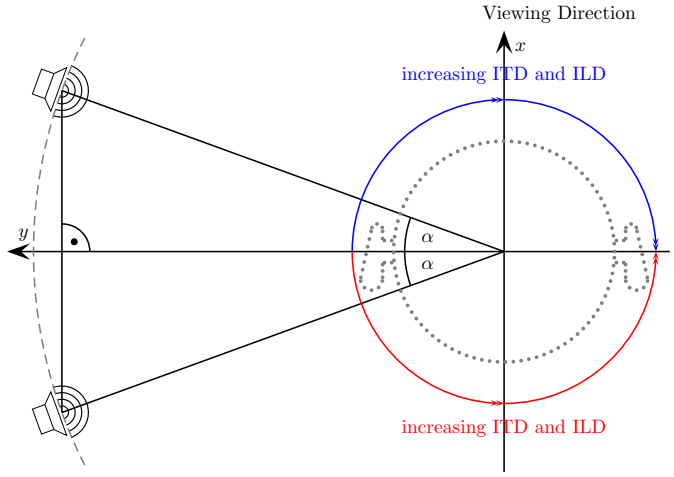

Figure 5.11: *Change of ITD according to the incidence direction of the sound source.*

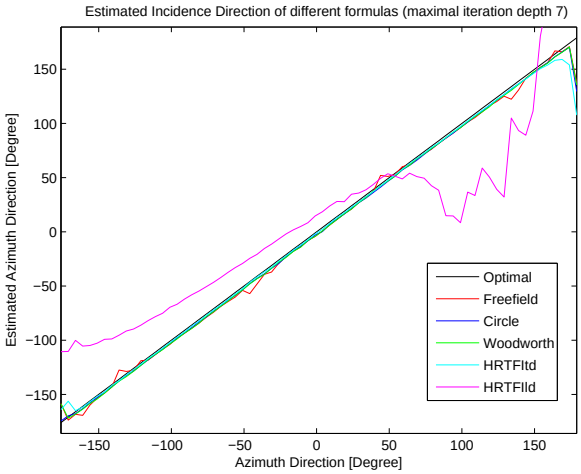

Figure 5.12: *Results of the iterative localization with front-back confusion for a variable number of iteration (maximal seven iterations).*

Algorithm	Correct Estimated [%]
Freefield	98.24
Circle	97.22
Woodworth	98.43
HRTF-ITD	94.35
HRTF-ILD	96.39

Table 5.1: *Results of the front-back confusion algorithm for the five described algorithms.*

It has turned out that the ILD changes are more reliable than the ITD changes to estimate the front-back direction. The algorithm first examines the positions of the localization track – the absolute head positions during each iteration. Then for each two iterations, the change of the ILD is assigned to the front or the back direction according to the sign and a final direction is judged based on a major vote.

Problems arise, when there is only one iteration. If the source is i.e. located at 180° and the localization algorithm estimates the incidence direction as 0° in the first iteration, no change in the ILD respectively ITD can be measured. In these cases, where only one valid position is available, the head is moved by a specific amount – i.e. 10° – to one side, the ILD change is measured and the front-back direction is judged.

Table 5.1.3 shows the results of the front-back-confusion algorithm for the five localization formulas, when using only the positions and the ILD changes of the first two iterations of each localization. The values are obtained by evaluating 80 speech signals of one-second length, played back from all positions between −20° to 20° and 160° to 200°, which leads to 6400 evaluations of the algorithm for each formula. For the Woodworth formula, the algorithm estimates the front-back direction correctly in 98.43 percent of the cases.

5.1.4 Localization of Several Sources

The localization of several sources in the auditory scene is implemented in many parts analog to the one source case. Bob evaluates the interaural time differences based on the summed correlation function described by formula 5.3, which performs clearly better in the multisource scenario than the normal correlation function as shown in section 5.1.1. The iterative head motion is used to verify and refine the results of the incidence direction estimation.

In each iteration the algorithm computes the complete correlation function and estimates the positions of the highest points which are assigned to the single sources in the auditory scene. Then Bob turns to the position of the highest point, repeats the estimation of the incidence directions and compares the results with the outcome of the last iteration shifted by the new relative head position. The final directions are estimated based on the summed correlation function of each iteration.

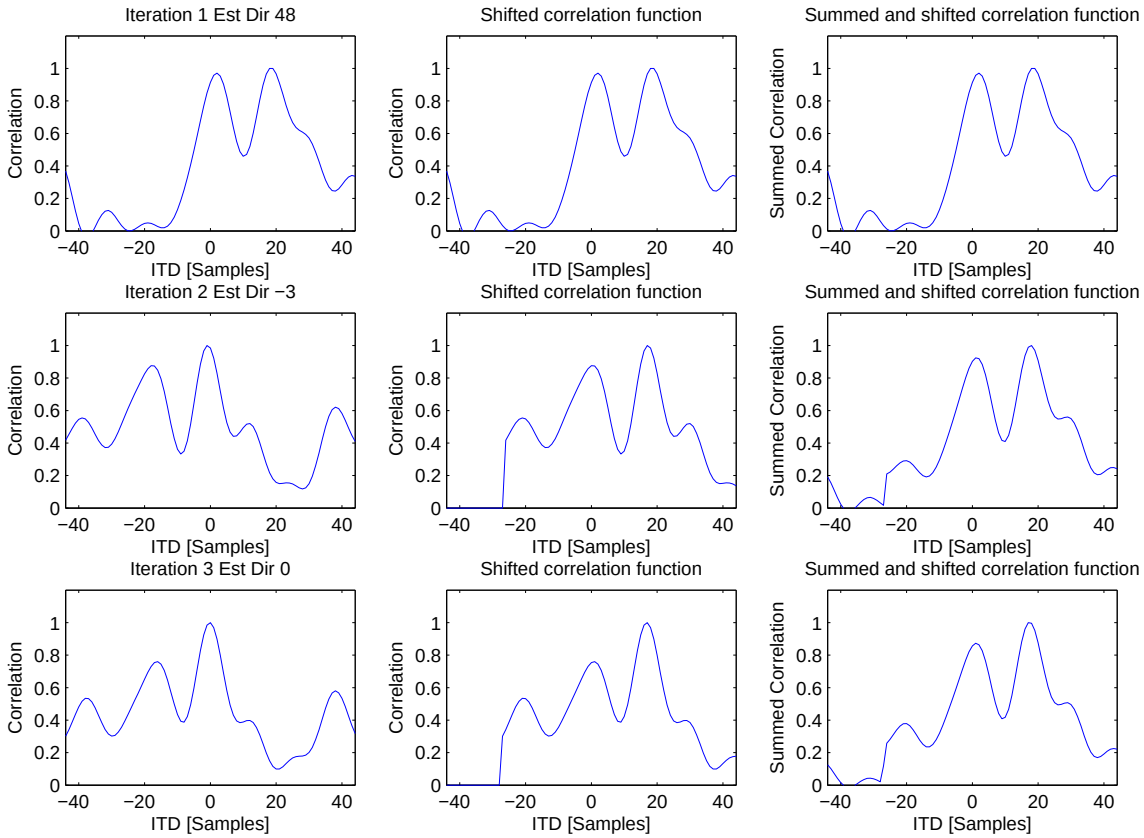

Figure 5.13: *Localization of multiple sources by summing correlation functions of different directions.*

Original Positions		Estimated Positions		Original Positions		Estimated Positions	
Source 1	Source 2	Source 1	Source 2	Source 1	Source 2	Source 1	Source 2
-45	0	-43.68	0.60	-90	0	-89.68	-0.14
-40	5	-37.48	5.07	-85	5	-85.23	4.55
-35	10	-32.81	10.38	-80	10	-84.18	10.04
-30	15	-27.52	15.67	-75	15	-79.59	15.00
-25	20	-23.41	20.26	-70	20	-75.33	20.76
-20	25	-18.31	25.54	-65	25	-67.55	27.20
-15	30	-14.04	30.46	-60	30	-60.90	31.75
-10	35	-8.96	35.96	-55	35	-54.90	37.65
-5	40	-4.11	39.78	-50	40	-50.10	42.45
0	45	2.00	45.24	-45	45	-44.95	47.70
5	50	6.00	53.88	-40	50	-38.70	52.15
10	55	10.00	59.50	-35	55	-34.45	58.10
15	60	15.31	62.57	-30	60	-26.21	64.94
20	65	18.96	67.08	-25	65	-25.29	66.82
25	70	23.71	74.46	-20	70	-18.29	73.06
30	75	27.39	82.43	-15	75	-14.17	80.47
35	80	30.87	86.50	-10	80	-9.47	86.06
40	85	37.65	87.96	-5	85	-5.53	89.18
45	90	44.96	86.25	0	90	-2.25	89.93

Table 5.2: *Results of the iterative algorithm for two sources of* 45° *(left table) and* 90° *(right table) distance for an iteration depth of 3.*

Figure 5.13 shows the principle of the algorithm for three iterations in a two source scenario, where source 1 is positioned at 0 degree and source 2 is positioned at 45 degree. The left plots show the estimated correlation functions for each iteration, that show clear peaks at the source positions. The middle plots depict the correlation function shifted to the absolute head position (the position of the head in the first iteration). The right figures show the summed correlation function for each iteration consisting of the summed correlation function of the last iteration and the shifted correlation function of this iteration. All shown correlation functions are normalized. In the first iteration the algorithm estimates the incidence direction of the highest point with 48° based on the Circle formula. Then the head moves to position 48° and recomputes the correlation function. The new correlation function is shifted by 48° and added to the summed correlation function. The new position of the highest peak is estimated as $-3°$ and the head moves to that position, adds the correlation function to the summed correlation function and so on and so forth. The final estimates of the source positions are taken from the final summed correlation function by extracting the positions of the highest peaks in the graph.

Table 5.2 shows the results of the multiple source localization algorithm for two source scenarios consisting of a male and a female speech source, which are spatially separated by 45° respectively 90°. The iteration depth of the algorithm has been fixed to three iterations and to convert the

Original Positions			Estimated Positions		
Source 1	Source 2	Source 3	Source 1	Source 2	Source 3
-45	0	-90	-66.76	1.06	-78.35
-40	5	-85	-57.59	6.23	-87.17
-35	10	-80	-47.74	14.84	-67.79
-30	15	-75	-41.61	19.50	-84.00
-25	20	-70	-34.52	25.47	-80.41
-20	25	-65	-25.14	30.28	-66.38
-15	30	-60	-21.42	37.36	-64.36
-10	35	-55	-20.50	40.16	-61.72
-5	40	-50	-16.37	47.52	-41.31
0	45	-45	-8.58	52.00	-47.84
5	50	-40	-5.28	58.77	-43.66
10	55	-35	-0.57	59.85	-36.95
15	60	-30	-1.12	65.37	-22.18
20	65	-25	8.38	69.76	-19.28
25	70	-20	14.19	56.57	-23.66
30	75	-15	14.45	71.63	-22.22
35	80	-10	22.90	81.63	-13.31
40	85	-5	24.13	85.36	-16.13
45	90	0	24.71	86.50	12.67

Table 5.3: *Results of the iterative algorithm for three sources of* 45° *distance for an iteration depth of 3.*

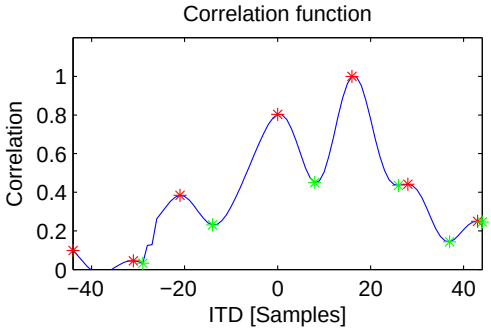

Figure 5.14: *Detection of the valid peaks of the correlation function.*

interaural time difference to incidence direction, the Circle formula has been used. For each direction, the results are averaged over thirty one-second recordings. For scenarios where the sources are separated by 45°, the algorithm estimates the source locations reliably with a maximal deviation of approximately $2 - 5°$. Analog results are obtained in the 90° difference case.

Table 5.3 shows the same evaluation for a three source scenario, where the sources are separated by 45°. The iteration depth is again fixed to three iterations and the results are obtained by averaging over thirty one-second recordings of three speech sources played back at the specified directions. Overall the average results turn out to be inferior to the two source scenarios. Some values however are distorted by outliers which are introduced by detecting a false peak in the correlation function which leads to a complete false incidence direction of one of the three sources. This outlier then distorts the mean of all detected source locations. Removing such outliers is not easy to manage as it is not clear to which of the three sources this outlier has to be assigned. For this evaluation the estimated incidence directions have been assigned to the optimal locations by the minimum absolute distance.

5.1.5 Estimating the Number of Sources

The described algorithm estimates the source positions reliably if the number of sources in the auditory scene is known in advance. In each iteration the algorithm picks the highest peaks of the correlation function according to the number of sources. If the number of sources is not known, the algorithm falsely estimates more source positions as the correlation function usually exhibits more peaks than valid source positions in the auditory scene (see i.e. figure 5.13) and so potentially moves to false positions in later iterations.

Valid peaks in the correlation function usually are the highest peaks in the function and the height difference to the neighboring dips of the function is large. Figure 5.14 shows a typical correlation function for two sources located at 0° and 45°. The valid peaks are located at sample positions 2 and 18 on the ITD axis that specifies the interaural time difference in samples. Further false peaks are located at sample positions -44, -30, -20, 28 and 44.

Considering only normalized correlation functions, peaks with a small height difference to neighboring dips are eliminated by using a threshold t_1. Denoting the correlation function as R, the i-th peak as p_i and the left and the right neighboring dip of p_i as d_l and d_r, the absolute height difference of p_i to its neighboring dips is specified as

$$h_l(p_i) = R(p_i) - R(d_l) \tag{5.16}$$

$$h_r(p_i) = R(p_i) - R(d_r) \tag{5.17}$$

Invalid peaks which are only marginally higher than their neighboring dips – such as i.e. the peak at ITD position 28 – are removed in this way. Remaining false peaks, like the peak at sample position -20 are removed by comparing the height to the highest peak in the function. A threshold t_2 admits only those peaks with a low absolute height difference to the highest peak.

$$h_h(p_i) = R(p_h) - R(p_i), \tag{5.18}$$

where p_h denotes the highest peak in the correlation function. The number of sources n in the auditory scene is obtained by extracting and counting those peaks that fulfill the two threshold conditions:

$$n = |\{p_i \in P, |\ h_l(p_i) > t_1 \wedge h_r(p_i) > t_1 \wedge h_h(p_i) < t_2\}|, \tag{5.19}$$

where P includes all peaks of the correlation function and $0 \leq t_1, t_2 \leq 1$. In limiting cases, where i.e. the left or the right neighboring dip does not exist, only the remaining existing conditions are evaluated.

The performance of the number of source estimation algorithm is evaluated for different threshold values t_1 and t_2 in figure 5.15 for two source scenarios. For each threshold pair, the algorithm is tested on 20 speech mixtures of one-second length, which are spatially separated by 45° and are played back from 60 different incidence directions. The correlation function is computed as the normalized correlation after three iterations as described above. For threshold values $t_1 = 0.15$ and $t_2 = 0.45$, the algorithm estimates the number of sources in the scenario correctly in approximately 94% of the cases. For three source scenarios, the performance of the algorithm decreases to approximately 69% correct estimated number of sources.

The accuracy of the number of source estimation can be increased by tracking the locations of the peaks in the correlation function over several iterations. Peaks that are visible in one iteration, but diminish in the next iteration, can be removed in this way.

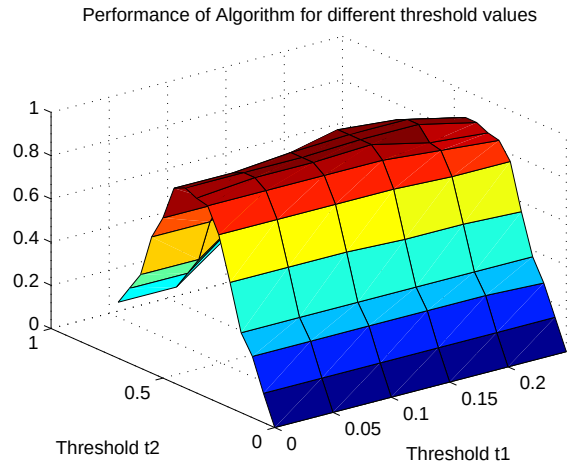

Figure 5.15: *Performance of algorithm for number of sources estimation for different threshold values.*

5.1.6 Conclusions

The incidence directions of the sources in the auditory scene can be estimated by the described localization algorithms. Although the source locations are estimated relative reliably, there is room for improvements especially in the multiple source scenarios.

One possible enhancement is to increase the sampling rate of the recorded mixtures. All files used in the previous experiments are sampled with 44.1 kHz. The accuracy resolution of the incidence direction estimation is limited to one sample time difference in the correlation function. Using i.e. the Woodworth formula for incidence direction estimation leads to an accuracy resolution of approximately 2.5°. Increasing the sampling rate also increases the localization accuracy. The processing of higher sampled audio files (i.e. 192 kHz) however is computationally very demanding for the described algorithms. The increasing of the sampling rate promises only marginal improvements and is therefore not investigated further in this thesis.

The human mind rarely estimates the incidence directions of sound emanating sources only based on the auditory system. In almost all cases humans additionally use their visual system to estimate the locations of the sound sources. The source localization of this project can surely be enhanced by mode fusion with a visual system. The human dummy head Bob is equipped with two camera eyes, searches the scene for the sound emanating sources and estimates the incidence directions based on visual characteristics of the sources. A mode fusion of these results with the results of the auditory system then gives final and valid estimates. Especially the algorithm for the estimation of the number of sources is easier to implement in the visual domain than in the auditory domain. Sound sources that are not directly inside the auditory scene such as i.e. music coming from the neighboring room can be identified as invalid source for the current

auditory scene and can be filtered out. Such a described visual analysis of the room around Bob to enhance the auditory scene analysis is currently investigated by Haschke [46].

5.2 Fundamental Frequency Estimation

Humans tend to use frequencies that are an integer multiple of their own fundamental frequency (F0) as described in detail in chapter 2.2.3. Especially voiced parts of speech contain most of the energy in the harmonics of the F0. Source separation approaches can use the F0 to determine those frequencies that are mainly used by a speaker.

5.2.1 Harmonicity of Human Speech

Most algorithms for fundamental frequency estimation rely on the assumption that the underlying signal $x(t)$ is periodic:

$$\exists T \neq 0 : \forall t : x(t) = x(t+T) \tag{5.20}$$

This periodicity implies that the harmonics F_i of the fundamental frequency $F0 = \frac{1}{T}$ are also harmonic and are integer multiples of the F0:

$$\forall i \in \mathbb{N}^+ : F_i = (i+1) \cdot F_0 \tag{5.21}$$

Simple string instruments like guitars or violins exhibit partials that are not an exact integer multiple of the fundamental frequency [97]. Xue and Sandler [126] for example describe the harmonicity of simple string instruments with the following formula that approximates the ratio of the partials to the fundamental frequency, if the stiffness B of the string is known:

$$F_i = (i+1) \cdot F_0 \cdot \sqrt{1 + B((i+1)^2 - 1)} \tag{5.22}$$

The field of speech analysis commonly assumes that the human voice works strictly harmonic for voiced speech parts (i.e. [105] [13], [108]). But if even simple string instruments exhibit a specific describable inharmonicity, the human voice as a far more complex architecture cannot be assumed to be strictly harmonic.

As almost all fundamental frequency estimation algorithms for human speech rely on the underlying harmonicity of human speech, Krämer [59] examines if the human voice also shows such a describable inharmonicity or if it can be regarded as strictly harmonic. Speech signals of the speech database CMU Arctic [58] are divided in approximately 55000 windows of length $\approx 185,8\ ms$ by applying the Short-Time-Fourier-Transform to the speech signals. For the harmonicity analysis only those windows are admitted that exhibit clear and stable spectral peaks

and show a spectral peak at the fundamental frequency[1] (see [59] for details). It turned out that only about 1.6 percent of all windows are suitable for the analysis. Of the admitted windows 92.6 percent are considered to be strictly harmonic. The remaining windows have been inspected by hand, but no common inharmonicity has been recognizable. The study therefore concluded that an inharmonicity as in the case of simple string instruments is not detectable for voiced human speech. This result is consistent with the common assumption used in the literature.

5.2.2 Fundamental Frequency Estimation

The algorithm used to estimate the fundamental frequency track of the speech sources in the auditory scene is implemented in many parts analog to the YIN-method [22]. The input signals are divided in time windows of 50 ms length, such that two periods of a 40 Hz wave are captured. For each of these windows a fundamental frequency is estimated to construct a complete F0-track. A postprocessing stage smoothes the F0-track by removing outliers and applying specific assumptions of the human voice. The following paragraphs outline the concept of the F0 algorithm. For a detailed description of the algorithm, its parallels to YIN and the specific implementation and parameter estimation, the reader is referred to the work of Krämer [59].

F0 estimation based on SDF In a first step the fundamental frequency of the windowed signal is estimated by searching for the first global minimum of the Squared Difference Function (SDF)

$$D_t(l) = \sum_n (x(n) - x(n+l))^2 \tag{5.23}$$

where l specifies the lag, n denotes the samples of the signal and t is the currently inspected window of the whole signal. An estimate of the fundamental frequency is computed by extracting the first global minimum of $D_t(l)$:

$$F_0^{\text{Step 1}}(t) = \frac{1}{\arg\min_l D_t(l)} \tag{5.24}$$

Thresholding When estimating the F0 with the SDF as described in step 1, dips that do not correspond to the F0 dip are sometimes slightly lower than the correct dip. In most cases these dips lie on an integer multiple of the fundamental period and harmonic errors occur in the estimation [59]. To avoid the estimation of higher harmonics, a threshold is implemented. Initially all dip positions P are estimated that are smaller than $2/3$ of the position of the global dip as the period of the harmonics of a dip with period T are expected to lie below $\approx T/2$ [59]. Assuming s is the threshold value, l_1 the position of the global

[1] Windows with so called missing fundamental frequencies are not regarded in this analysis.

dip and $\overline{D_t}$ denotes the mean of the SDF for the currently regarded window, the final F0 is estimated as follows:

$$F_0^{\text{step 2}} = \begin{cases} \frac{1}{\min(P_s)} & P_s \neq \emptyset \\ F_0^{\text{step 1}} & \text{else} \end{cases} \tag{5.25}$$

where

$$P_s = \{p \in P \mid \overline{D_t} - D_t(p) > (\overline{D_t} - D_t(l_1)) \cdot s\} \tag{5.26}$$

Polynome Interpolation The resolution of the F0 estimation of step 1 and 2 is limited to one sample. Especially in higher frequencies this leads to rounding errors. To enhance the resolution of the detected dip locations in the SDF, a parabolic interpolation is used that fits a polynome of degree two to the global dip and its direct neighbors. On clean test signals the resolution of the estimated F0 can be increased by up to two power of ten. For details see [59].

Postprocessing of F0-track The estimated F0-track $f_0(t)$ shows the coarse characteristics of the real F0-track, but also includes several errors. A postprocessing stage tries to eliminate these errors by applying specific assumptions about the human voice. The F0 region of human speech is limited to 50 - 600 Hz [22]. F0 estimates outside this range can be neglected. Additionally the human voice exhibits a continuity of the fundamental frequency. Under realistic conditions the F0 of human speech does not change faster than 1 oct/s [22].

The postprocessing of the estimated F0-track is implemented in several stages: First the reliability of an F0-estimate is evaluated by the following formula:

$$V(t) = \frac{\overline{D_t}}{D_t(T(t)) + \epsilon} \tag{5.27}$$

where $T(t)$ denotes the estimated fundamental period at time t and ϵ is a small and positive number. The higher the value $V(t)$, the more reliable is the result. Most of the remaining errors are harmonic errors, where the fundamental frequency is often estimated as twice or half the correct F0 – so called octave errors. Those octave errors are corrected by iterating over the complete F0-track and detecting for each F0-estimate if the current estimate, the next overtone or the next undertone fits better to the complete track. Remaining outliers have commonly no relation to the correct F0-track and are removed based on their distance to the stable mean of the whole track. The estimated F0-track is finally smoothed with a median filter that removes noise and remaining outliers from the track.

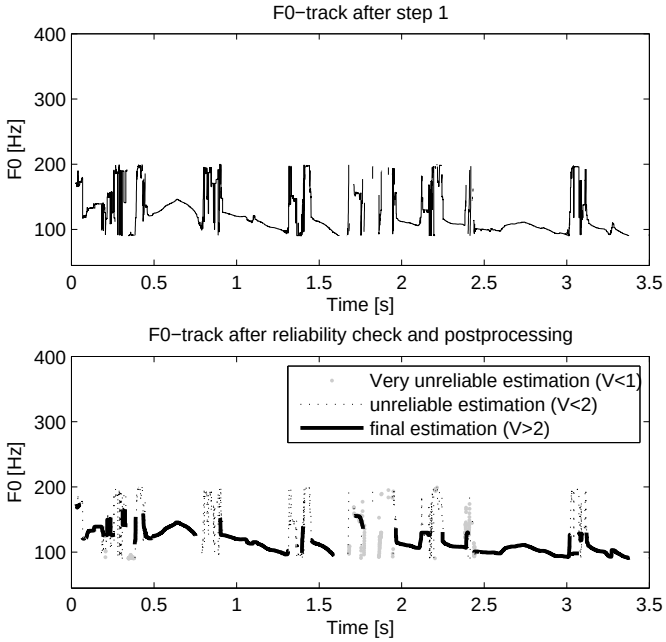

Figure 5.16: *Estimated F0-track of the described algorithm for a male speech source of 3.5 s duration. The upper plot shows the F0-track after step 1. The lower figure plots the result after the reliability checks and the postprocessing.*

Figure 5.16 shows the result of the described algorithm for a male speech source of 3.5 s duration. In the upper plot the result of step 1 – the F0-estimation based on the SDF – is depicted, which includes many false frequencies. The lower plot shows the resulting F0-track after the reliability check and the postprocessing steps: False candidates and harmonic errors that do not fit to the continuity of the track are eliminated and a continuous track remains.

The introduced algorithm works for single speech sources and estimates a fundamental frequency track for the signal. When several speech sources are present in a mixture, the quality of the results drastically decreases. To enable the estimation of several F0-tracks, the concept of the SDF is extended to the general case of the xDF – the difference function of x dimensions. The computation of the SDF is equivalent to the energy of the signal convolved with a flipped comb filter (see [59] or [21]):

$$SDF(l) = \sum_n (x(n) - x(n+l))^2 \tag{5.28}$$

$$= \sum_n (h_\tau(-n) * x(n))^2 \tag{5.29}$$

The general case of the xDF is then defined as

$$xDF(l_1, l_2, \cdots, l_N) = \sum_n (h_{\tau_1 \cdots \tau_N}(-n) * x(n))^2 \tag{5.30}$$

The estimation of the F0-track is similar to the estimation in the one source case, but instead of finding the minimum of the SDF, the minimum of the xDF is extracted. Finding the global minimum of a higher dimensional function is computationally very demanding. For the two source case the computational complexity of the first step of the described algorithm is already about 4000 times real time [59]. Additionally the postprocessing has to be adapted to identify the separate tracks. While this is still practicable for the two source case, already the three source case is very demanding (see [59] for explicit examples).

In the current scenario, where a human dummy head is listening to the auditory scene around, it would already be useful to extract the F0-track of the target source. This is accomplished by steering the head in the direction of the source. An easy directional beamformer (as described in section 2.4.1) is implemented by adding together the two ear signals and dividing by two. In this way the target source is enhanced in the resulting signal, while the other sources degrade. The described single source F0-algorithm then estimates the F0-track of the target source. Compared to the F0-estimation based on the single ear channels, this procedure enhances the algorithm by up 10% correct estimated F0s.

The research of multiple F0-estimation is beyond the main scope of this thesis and is therefore only regarded marginally. If the later presented algorithms for source separation need the single

F0-tracks of each source, the multiple F0-estimation algorithm is simulated by applying the single source F0-algorithm to each single recorded source. If only the track of the target voice is required, Bob uses the described beamforming approach to detect the F0-track.

6 Binaural Source Separation

The concept of source separation with ideal binary masks has been introduced in detail in the previous chapters. The goal of this chapter is to extract the ideal binary mask that includes the orthogonal time-frequency points of the speech mixtures. The binary masks are estimated based on the characteristics of the auditory scene, which has been analyzed by Bob in the previous step.

6.1 Architecture

The ideal binary mask is applied to a time-frequency spectrum to demix the target source from a mixture of several speech sources. The time-frequency spectrum can either be chosen to be the STFT or the cochleagram. In both time-frequency representations, speech sources exhibit an approximate orthogonality and the ideal mask demixing yields satisfactory speech quality and SIR values (i.e. [127], [17]).

The STFT is easy and lossless to compute, but the filter channels are positioned linear on the frequency scale which yields only a coarse frequency resolution in the important low frequencies of speech signals. Also the amplitude and phase information is averaged over the complete analysis window and so is not reliable in reverberant environments. The cochleagram on the other hand analyzes the signal with logarithmically spaced filter channels and allows a finer frequency resolution in the low frequencies, but the invertion of a given cochleagram to a time-domain signal is non-trivial and lossy.

The source separation framework presented in this thesis combines the positive features of the STFT with the positive features of the cochleagram while eliminating some of the negative features. The overall goal of the source separation is to find the ideal STFT-mask. The core source separation process however is based on the analysis of the corresponding region in an additionally computed cochleagram. This way the macroscopic STFT-transform is used to define the demixing masks and to finally demix the original sources. The core assignment of each STFT-bin to a specific source is based on the corresponding region in the microscopic cochleagram and is supported by the information gained from the STFT-spectrum.

This proceeding is analog to the approaches used in MPEG audio coders. For example MPEG Audio Layer 3 uses a FFT of 1024 samples to analyze the input signal and to apply the psychoacoustic models. The critical subsampling however is implemented using only 32 subbands [87].

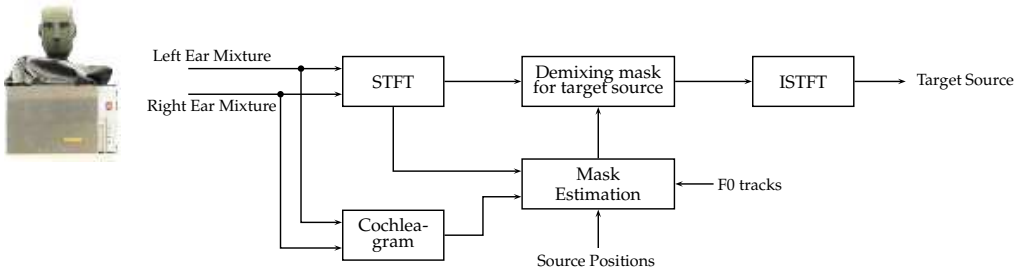

Figure 6.1: *Overall architecture for source separation framework.*

Figure 6.1 illustrates the system architecture of the source separation framework. The incoming signals of the left and right ear are STF-transformed and the respective cochleagram of each ear signal is computed. The mask estimation process computes a STFT-mask for the target source based on the information gathered from the detailed cochleagram and supported by coarse information of the STFT spectrum. Additionally the results of the previous scene analysis – the positions of the sources in the auditory scene and the F0 track of the target source – are input to the mask estimation algorithm. Finally the STFT spectrum is multiplied with the estimated STFT mask of the target source and the spectrum is transformed back to the time-domain, yielding the demixed time-domain signal.

The mask estimation stage tries to make use of all information that could be established using standard or sophisticated signal processing methods. In a first step Bob analyzes the auditory scene as described in detail in the last chapter. Bob estimates the positions of the sources in the azimuth plane and identifies estimates of the fundamental frequency track of the target speaker. In further steps this information is used to enhance the source separation, that uses both interaural and monaural cues to distinguish the TF-bins. Because of reverberation many of the cues used to separate bins are distorted and can only be used to some extend. To face the reflections and reverberations, several algorithms compute independent estimates of the binary masks. In a final stage the estimated masks of each algorithm are combined to find a best estimate.

6.2 Analysis of Interaural Differences

The interaural time and level differences of the single STFT bins of speech signals are analyzed in the following figures. The values are obtained by computing the mean values of the ITD and ILD of each STFT-bin of 400 one second mixtures of the CMU arctic database [58]. The two source sound mixtures are recorded with Bob under reverberant conditions. The signals are sampled with 44.1 kHz and the STFT uses an analysis window of 4096 samples and an overlap of 2048 samples between succeeding windows.

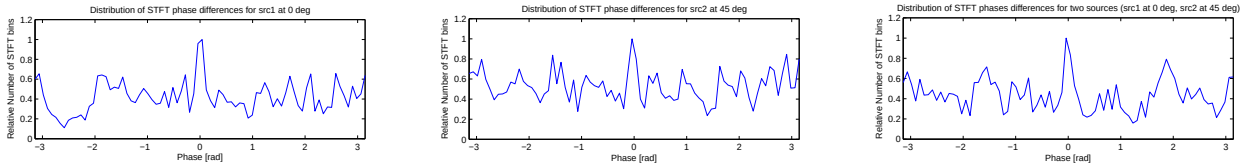

Figure 6.2: *Distribution of the estimated interaural time differences of the STFT bins for single source scenarios and two source scenarios based on STFT ITD estimation (all plots are normalized to one).*

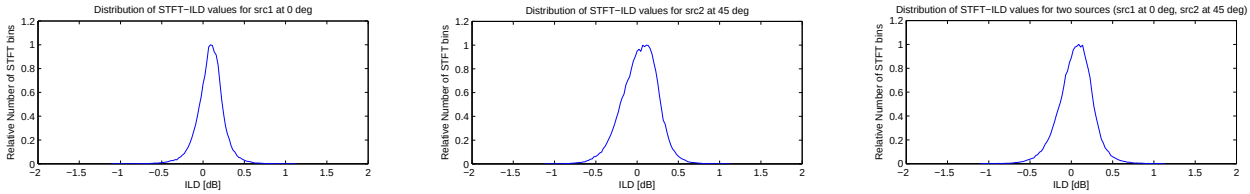

Figure 6.3: *Distribution of the estimated interaural level differences of the STFT bins for single source scenarios and two source scenarios based on STFT level difference (all plots are normalized to one).*

Figure 6.2 shows the distribution of the ITDs for the STFT bins for a two source scenario, where the sources are positioned at 0° and 45° and the distribution of the ITDs for the same sources played back individually. The ITD is computed based on the phase differences between the left and right STFT spectrum. Some source separation approaches for anechoic speech mixtures (i.e. [127]) use this ITD estimation to group TF-bins coming from the same direction. This works fine for anechoic and artificially mixed sound mixtures (see [127]), but for reverberant and HRTF-filtered speech signals, the STFT phase differences do not provide reliable ITD estimations as can be seen in figure 6.2. Even in the one source case, where the speech source is positioned at 0°, the STFT phase values do not show a clear peak. The peak at zero degree can be traced back to the fact that the STFT phases are referenced to the phase of the lowest frequency of the STFT and therefore high frequency bins all have phase values that tend to be zero.

Figure 6.3 analog analyzes the ILD of the STFT bins. It can be seen that also the ILDs do not show clear different peaks at different positions. The distribution of the ILD for source 2 – positioned at 45° – is broader compared to that of source 1, which can be traced back to the unsymmetrical incidence direction of the incoming sound waves regarding the two ears. The head shadow and the reverberation then disturb the signal more than in the case of source 1.

In contrast to the ITD estimation based on the STFT phase differences, this architecture estimates the ITD values by correlating the cochleagram windows corresponding to the currently regarded STFT-bin. The time-scale of the STFT consists of discrete time points at multiples of the window shifting distance. In this architecture a shifting distance of half a window length is used, so there is one Fourier spectrum available at multiples of 0.5× the window length.

The cochleagram filters the signal with its original time-scale, so there is one spectrum for each sampling point of the recorded signals. Let $X_{L_{stft}}$ and $X_{R_{stft}}$ denote the STFT-representation of the left and right ear signal and $X_{L_{co}}$ and $X_{R_{co}}$ the corresponding cochleagram representations. For each STFT-bin the corresponding left and right TF-windows $W_{L_{co}}$ and $W_{R_{co}}$ are cut out of the cochleagram and include the cochleagram-spectra for all time instances inside the STFT window. So for each single STFT bin a two-dimensional window of the cochleagram is regarded. The ITD estimates of $W_{L_{co}}$ and $W_{R_{co}}$ are computed using a running cross-correlation across the time-dimension of the time-frequency regions: $\forall l \in \{-\text{maxLag}, \text{maxLag}\}$

$$R_{W_{L_{co}}W_{R_{co}}}(l) = \sum_{t=t_s}^{t_e} \sum_{f=f_s}^{f_e} W_{L_{co}}(t+l, f) \cdot W_{R_{co}}(t, f) \tag{6.1}$$

The highest peak of $R_{W_{L_{co}}W_{R_{co}}}$ yields the best estimate of the ITD for this bin. As described in detail in chapter 5.1.1 reflections and reverberation can introduce further peaks in the correlation function that refer to the correct ITD and therefore should also be considered. According to Faller and Merimaa [32], the height of the peak in the correlation function is a measure of reliability: The higher the peak, the more reliable the ITD estimation.

Figure 6.4 analyzes the distribution of the ITD computed by the described algorithm under the same conditions as previously described. Opposed to the STFT phase differences there are clear peaks at the source positions. For one source positioned at zero degree, the ITDs are almost all positioned at zero degree. For the scenario, where the source is positioned at 45° also a clear peak at the ITD position corresponding to 45° shows up, but there are also a lot of bins with false ITDs. This can be traced back to the reflections and the head shadow that is existent for sources with an incidence direction of 45° and which disturbs the signals and the ITD between the two ears. In the two source scenario, two clear peaks show the locations of the sources in the auditory scene and divide the STFT bins in almost two parts: Those bins belonging to source 1 and those bins belonging to source 2.

The analysis of the ITD values shows that it is preferable to position the target source at position zero degree relative to the head, as the distribution-function of the ITDs then is more narrow than at other incidence directions. The human dummy head Bob has already analyzed the auditory scene in the previous stage and knows the positions of the source. Therefore Bob can easily align to the target source. Doing this way better separation results are expected.

The ILD can be computed in a similar manner as the estimation of the ITD, by computing the level differences between the corresponding cochleagram windows of each STFT bin. Unfortunately this proceeding does not result in the desired distribution of the ILDs with two separate peaks as in the ITD case as depicted in figure 6.5. As in the estimation of the ILDs based on the STFT level differences, no clear peaks at the corresponding positions occur. But again the distribution of the ILDs for source 2 is broader than in the distribution of source 1. The poor

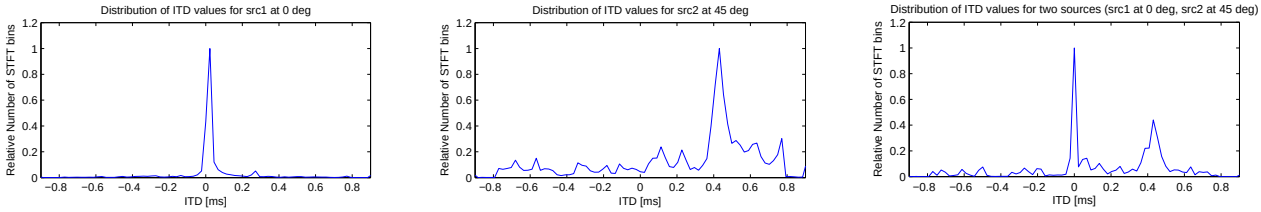

Figure 6.4: *Distribution of the estimated interaural time differences of the STFT bins for single source scenarios and two source scenarios based on cochleagram ITD estimation (all plots are normalized to one).*

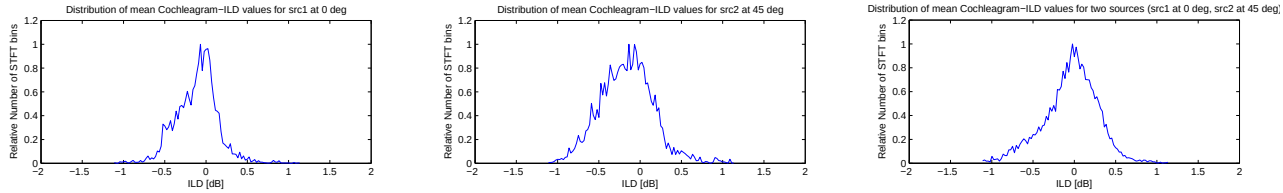

Figure 6.5: *Distribution of the estimated interaural level differences of the STFT bins for single source scenarios and two source scenarios based on cochleagram ILD estimation.*

performance of the ILD estimation can be traced back to the head filtering, which disturbs the signal levels by several dB (see i.e. HRTF-ILD in section 2.1.3).

The correlation function computed of the cochleagram windows and the resulting estimated ITDs are used to separate a target source out of a mixture of several speech sources by assigning those bins to the target source that emanate from the corresponding position. The ILDs between the left and the right ear for each STFT bin do not provide reliable information regarding the incidence direction of the dominant source in this bin and are therefore not used to estimate the demixing masks.

6.3 Evaluation Criteria

The ideal mask for the target speech source is computed as described in equation 2.30. To account for the noise in the reverberant recordings, a SNR is computed and only those bins of the STFT are considered that have an energy value higher than this threshold. For the recordings considered in this thesis a threshold of -20 dB is used. This threshold has been computed as the mean SNR of the used recordings.

To evaluate the following source separation algorithms based on the computation of the interaural time differences and the fundamental frequency track, five values are used to specify the capabilities of these algorithms:

Recovered energy of target source The percentage of recovered energy of the ideal mask for the target source specifies how much of the original energy of the target source is reconstructed. A value of 100% states that the estimated mask fully contains the ideal mask. This value is used to relativize the Signal-to-Interference-Ratio (SIR): If the estimated signal contains only very little of the target signal energy and almost no interfering energy, this leads to a high SIR value, but the resulting speech quality is very low because of the low recovered energy of the target source. The higher the percentage of recovered energy, the more energy of the original signal is recovered and the speech intelligibility of the target source increases.

False estimated bins The percentage of false estimated bins denotes the relative number of bins that are wrongly assigned to the target source. According to the ideal masks of the interfering sources, these bins should be assigned to one of the other sources of the auditory scene, as the absolute value of energy contribution to this bin of another source is larger than the energy contribution of the desired source. The number of false estimated bins however cannot be directly mapped to the speech quality of the separated source as each bin is weighted equally, but different bins have different influence on the speech quality (i.e. the amount of energy contributed to the whole signal). Nonetheless this value is interesting to evaluate the following source separation algorithms as it gives a coarse overview about the performance.

Gains in SIR, SDR and SAR In contrast to the values used in chapter 3 to analyze the window disjoint orthogonality of speech sources, the Signal-To-Interference-Ratio (SIR), Signal-To-Distortion-Ratio (SDR) and Signal-To-Artifacts-Ratio (SAR) are computed based on an approach by Vincent et. al. [112] (see section 2.7.2 for details), which is used for evaluation by several source separation algorithms (i.e. [121], [55], [35]). The SIR gain shows how well the estimated mask suppresses the interfering sources compared to the recorded mixture. The SDR and SAR gains evaluate in some sense the quality of the separated speech source.

To estimate the values, the freely available Matlab-Toolbox BSS-EVAL [36] is used. The values shown in the evaluations of the following source separation algorithms are the gains in SIR, SDR and SAR compared to the unprocessed mixture of the sources.

6.4 Separation based on Interaural Time Differences

Six algorithms based on the interaural time differences are evaluated regarding their source separation capabilities. Each algorithm is tested on 320 two source mixtures of one second length. Source 1 is located directly before the head at 0° as Bob has already geared to the source, while source 2 is located at 45° to the right. The values are obtained by computing the mean of all 320 separations. The signals are sampled with 44.1 kHz and the STFT uses an analysis window of 4096 samples and an overlap of 2048 samples between succeeding windows. The cochleagram

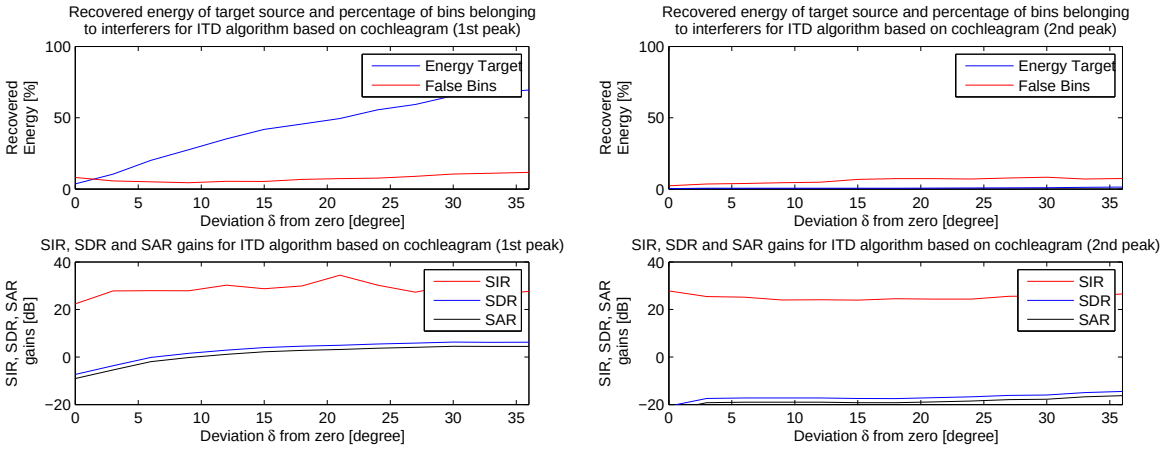

Figure 6.6: *Results of the separation algorithm based on the ITD computed by correlating the cochleagram windows corresponding to the STFT bins. The left figure (**algorithm 1**) shows the results when only the highest peak is considered while the right plot (**algorithm 2**) uses only the second highest peak.*

filters the signal in 512 channels. To convert the interaural time differences to corresponding incidence direction values in degree, the HRTF-ITD method is used (see section 5.1.2 for details). Each regarded correlation function is normalized to one at the highest peak.

Algorithm 1 Assign to the target source all TF-bins where the estimated ITD of the highest peak of the correlation function yields an angle of incidence that deviates not more than $\delta°$ from $0°$.

Algorithm 2 Assign to the target source all TF-bins where the estimated ITD of an existent second highest peak of the correlation function yields an angle of incidence that deviates not more than $\delta°$ from $0°$.

Algorithm 3 Assign to the target source all TF-bins where the correlation function at $0°$ is more than $h \in [0, 1]$ higher than the correlation function at the positions of the interfering sources.

Algorithm 4 Assign to the target source all TF-bins where the envelope of the correlation function at $0°$ is more than $h \in [0, 1]$ higher than the envelope of the correlation function at the positions of the interfering sources.

Algorithm 5 Assign to the target source all TF-bins where the distance to the nearest correlation peak is smaller than $\delta°$.

Figure 6.6 shows the results of algorithm 1 and 2 for different deviations δ. For algorithm 1 (left plot) one can clearly see that the percentage of recovered energy increases if the deviation

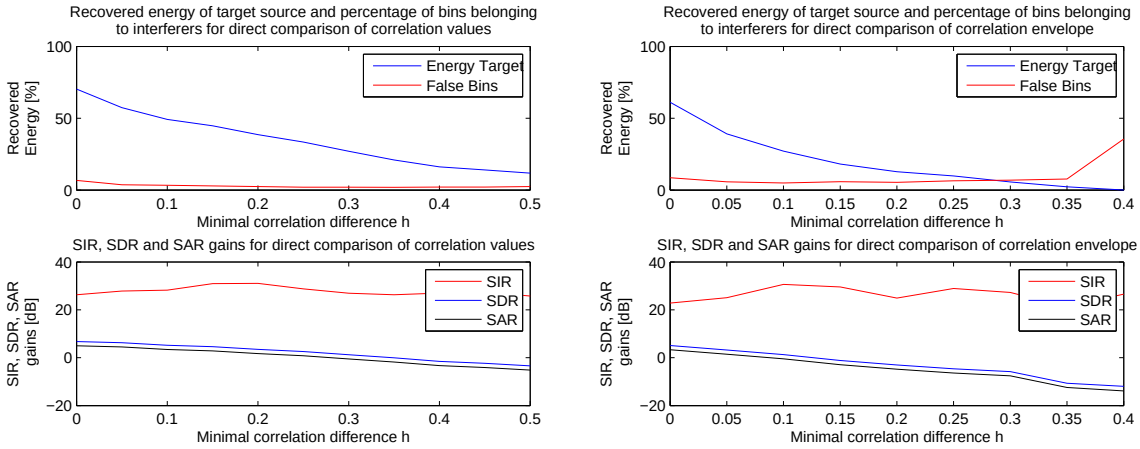

Figure 6.7: *Results of the separation algorithm based on the direct comparison of the correlation values at the previously known correct source positions. The left plot shows the results for the normal correlation function (**algorithm 3**), while the right plot compares the envelope of the correlation function (**algorithm 4**).*

from zero increases. The percentage of false estimated bins is very low (about 7 %) and grows only slowly to about 18 % at 35°. The SIR gains are high and range from 30 dB for small deviations to 37 dB for a deviation of 22°. The maximum SIR gain lies at the position directly in between the two sources, which also follows intuitively. The bins are then assigned to the nearer source based on the estimated incidence direction. A δ-value of 22.5° shows a good trade-off between the recovered energy and the achieved SIR gain. The SDR and SAR gains lie in the range of only 2-3 dB, which shows that the speech quality can be further enhanced.

The same results for algorithm 2 are illustrated in the right plot of figure 6.6. The percentage of reconstructed energy is much lower than in the case of algorithm 1, but almost all estimated bins are correct. This low percentage of recovered energy is due to the fact that a second peak in the correlation function in most cases only exists for TF-bins at high frequencies, where the correlation analysis window becomes bigger than the period of this bin. Those high frequency bins naturally include lower energy than low frequency bins, so the overall recovery is quite low. But these high frequency bins increase the final speech quality and should therefore be considered.

Algorithms 3 and 4 estimate the demixing mask based on the correlation value at the previously known positions of all sources in the auditory scene. Algorithm 3 uses the normal correlation function, while algorithm 4 uses the envelope of the correlation function to smooth out little variations in the correlation function. Figure 6.7 shows the results of the algorithms for different minimal correlation value differences h. If there is no minimal height difference ($h = 0$) the algorithm assigns the bin to those source with the maximal correlation value. Algorithm 3 then performs very good and recovers up to 75% of the total energy and achieves SIR gains of

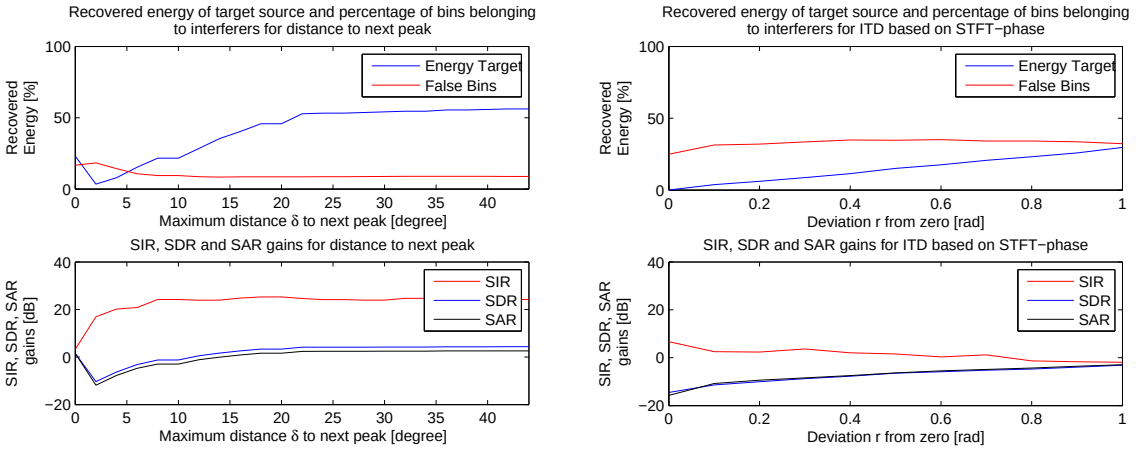

Figure 6.8: *Results of the separation algorithm based on the distance in degree to the nearest peak (**algorithm 5**) and the separation algorithm based on the ITD computed by the STFT phase differences.*

approximately 27 dB. Algorithm 4 in the same case recovers 70% total energy and achieves SIR gains of 22 dB. With increasing minimal height differences h the probability of the bins uniquely belonging to the target source increases (the correlation value of the target source position is more than h higher than that of the interfering sources). The total recovered energy decreases slowly with increasing height difference, but the percentage of false estimated bins decreases, which leads to maximal SIR gains of approximately 33 dB for both algorithms. The SDR and SAR gains for both algorithms lie in the range of up 7 dB.

The results of algorithm 5 are depicted in the left plot of figure 6.8. Algorithm 5 is similar to algorithm 1, but it regards all peaks and uses only the distance to the nearest peak in the correlation function. The recovered energy increases up to a deviation value of approximately 22°, where the SIR lies in the range of 22 dB. The SAR and SDR values are very low with approximate values of 2 dB, which can be traced back to overall low percentage of recovered energy of ca. 50%. To achieve high speech quality with high SAR and SDR value, also a high percentage of recovered energy has to be achieved.

For a comparison of the described algorithms, the right plot of figure 6.8 shows the performance of the separation based on the STFT phase (as i.e. the DUET algorithm [127] does in anechoic and artificially mixed scenarios). The algorithm assigns to the target source all TF-bins where the estimated phase difference of the STFT bin yields an angle of incidence that deviates not more than r radians from 0. As expected from the previously described distribution of the STFT phase differences between the left and right ear, the algorithm performs poorly with maximal SIR gains of 1-2 dB for low energy recovery and SDR and SAR losses of up to 15 dB. This

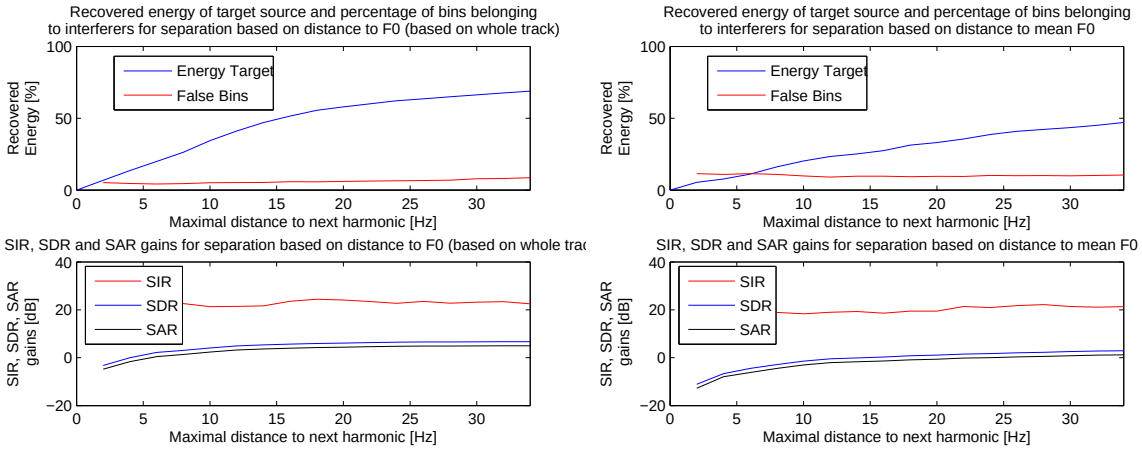

Figure 6.9: *Results of the separation algorithm based on the previously computed F0 of the target speaker. The left figure (algorithm 6) shows the results when using the complete F0-track, while the right plot (algorithm 7) uses only the mean F0.*

algorithm is hardly applicable for source separation and is only regarded here for a comparison to the other algorithms.

6.5 Separation based on Fundamental Frequency

If two persons speaking have a considerable different F0, their harmonics do not overlap in many frequencies. If the F0 of the target speaker is known in advance, this information can be used to assign the TF-bins. Each bin with a frequency value near to a multiple of the fundamental frequency is more probable to belong to the target source, than it is to belong to one of the other sources.

The following F0-based algorithms are examined for their source separation capabilities:

Algorithm 6 Assign to the target source all TF-bins where the frequency of the current STFT-bin deviates not more than Δf Hz from the nearest harmonic of the target source computed based on the complete F0-track.

Algorithm 7 Assign to the target source all TF-bins where the frequency of the current STFT-bin deviates not more than Δf Hz from the nearest harmonic of the target source computed based on the mean F0.

Figure 6.9 displays the results of algorithm 7 and 8. The percentage of recovered energy grows as the maximal distance of the nearest harmonic of the target speaker grows as more and more bins are considered. Algorithm 7 reconstructs up to 70 % of the total energy, while the percentage of false estimated bin is very low (approximately 7%). Compared to algorithm 7, algorithm 8

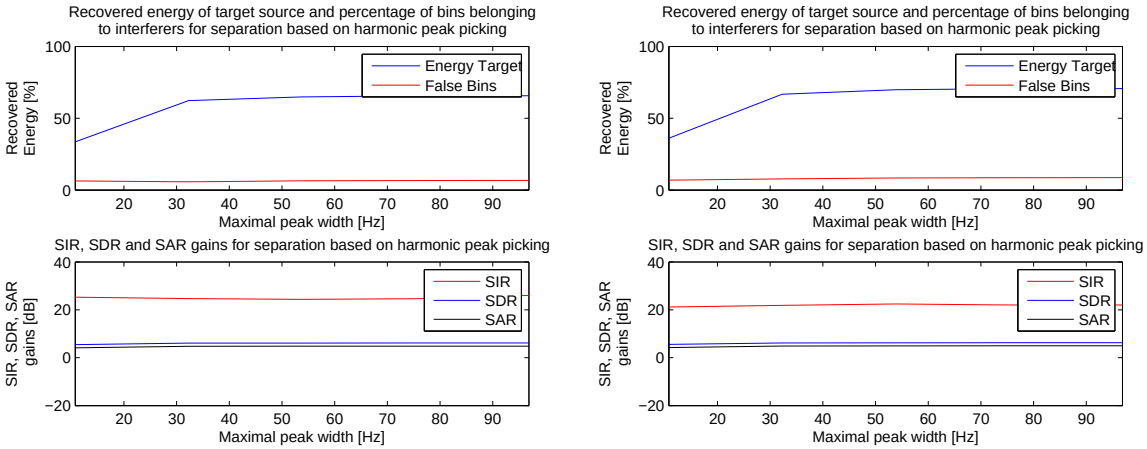

Figure 6.10: *Results of the separation algorithm that first extracts the bins that coincide with energy peaks in the frequency spectrum and lie on the harmonics of the target speaker and then extends the peaks up to the next left and right energy dips. The left figure shows the performance of the algorithm for a neighboring threshold of 20 Hz, while the right plot uses a neighboring distance of 30 Hz.*

uses only the mean F0 instead of the whole F0-track which leads to false local F0-estimates at many time instances. Nonetheless this algorithm then reconstructs up 45 % of the total energy of the target source. Both algorithms show SIR values in the range of 20 dB up to 25 dB. As the reconstructed energy of algorithm 7 is much larger than that of algorithm 8, positive SAR and SDR gains of up to 5 dB can be achieved. Overall algorithm 7 has the best performance of the algorithms considered so far.

Algorithm 8 uses a more intelligent way to extract the bins that are harmonically related to the target speaker. For each time instance of the STFT-spectrum the complete frequency spectrum is regarded and the energy peaks and dips are computed. Then those peaks are extracted that coincide with the harmonics of the target source up to a specific neighboring threshold. The remaining peaks are then extended up to the next left and right dip of the frequency spectrum. To avoid errors with non-existent left or right energy dips, a maximal peak width is used.

Figure 6.10 shows the results of the described algorithm 8 for neighboring thresholds of 20 Hz and 30 Hz and different peak widths. The left plot illustrates the performance for a neighboring distance of 20 Hz. The SIR gains lie in the range of 27 dB for all regarded peak widths. The percentage of recovered energy however increases to approximately 65 % for peak widths of 30 Hz. The respective SDR and SAR gains are quite high with 6 dB and 5 dB. Peak widths larger than 30 Hz do not noteworthy change the performance. Using a larger neighboring threshold (right plot) increases the percentage of recovered energy to up to 75 %, but decreases the SIR

gains by approximately 3 dB compared to the peak width of the left plot. The SDR and SAR gains stay constant, which shows that compared to the left algorithm the right algorithm does not add significant energy to the target source.

6.6 Combining of Algorithms

Each of the introduced algorithms yields a reconstructed target speech source with a low to intermediate intelligibility. To enhance the separation capabilities, the algorithms work together to combine the information regarding each TF-bin. In a first stage each discussed algorithm separately estimates a demixing mask for the target source. A second combining stage combines the single masks resulting in a final estimate of the ideal binary mask which is then used to demix the target source from the mixture.

One possibility to fuse the estimated masks of the different algorithms is to combine the masks in a sequential way similar to a chain of responsibility. The first algorithm in the chain assigns all bins according to its specification and passes the remainder of the bins to the second algorithm which in turn assigns those bins that match its specifications and passes the rest to the next algorithm and so on. This sequential combining of the algorithms is equivalent to computing the logical "or" of the estimated single masks.

Furthermore a parallel combining of the estimated masks is used to enhance the resulting speech quality. If several of the algorithms have assigned a specific bin to the target source, then this bin is more probable to belong to the source of interest than bins that are only assigned by a single mask. This parallel combining is implemented using the logical "and" of the single estimated masks.

For the following evaluation the parameters of the previously described algorithms are set to values that try to achieve the best SIR/recovered energy ratio. The maximum allowed deviation in degree from zero for algorithm 1 and 2 is set to 22.5°. For algorithm 3 and 4 a minimum correlation value difference of 0.18 is used, while algorithm 5 specifies a maximum deviation of 22.5°. Algorithm 6 and 7 use a maximum distance of 17 Hz and 25 Hz. The algorithm 8 uses a neighboring distance of 20 Hz and a peak width of 20 Hz to estimate the demixing mask.

To evaluate the performance of the algorithms, average values of 400 reverberant two source mixtures are used as described in detail in section 6.2. Table 6.6 summarizes the mean values of the best evaluation results of the combining of algorithms 1-8 for an auditory scene consisting of two sources.

In the mean all shown combinations achieve SIR gains of approximately 30 dB. They differ in the percentage of recovered energy and the SDR and SAR gains. Combinations with a high percentage of recovered energy also show high SAR and SDR gains, as much of the speech energy is reconstructed and therefore the speech quality is quite high. Combination $7 \cup 8 \cup 6 \cap 3 \cup 2 \cup 1$ for example achieves recovered energy values of 69.07% in the mean and SAR and SDR gains

Algorithm Order	Recovered Energy of ideal mask [%]	False estimated bins [%]	SIR gain [dB]	SAR gain [dB]	SDR gain [dB]
$7 \cup 1 \cup 6 \cap 3$	47.97	5.70	31.42	2.33	3.87
$2 \cup 6 \cap 8 \cap 3 \cup 7 \cup 1$	18.97	4.26	30.92	-3.15	-1.62
$1 \cap 8 \cup 7 \cap 3$	54.39	7.25	29.71	3.33	4.86
$6 \cup 3 \cup 8 \cap 7 \cap 1$	20.11	4.37	30.25	-2.86	-1.33
$6 \cap 2 \cap 8 \cup 1 \cap 7$	21.08	5.78	29.37	-2.47	-0.95
$6 \cup 2 \cap 8 \cup 7 \cap 1$	35.20	3.74	31.59	1.14	2.70
$3 \cup 6 \cap 8 \cup 7 \cap 1$	37.32	3.71	31.21	1.40	2.96
$8 \cup 7 \cap 1 \cap 6$	28.76	2.23	31.42	0.18	1.78
$8 \cup 2 \cup 7 \cap 3 \cap 1$	68.62	8.94	29.20	5.06	6.57
$7 \cup 8 \cup 6 \cap 3 \cup 2 \cup 1$	69.07	9.09	29.05	5.10	6.61
$6 \cap 8 \cup 7 \cap 3 \cup 1 \cup 2$	67.35	9.15	29.58	4.98	6.49
$8 \cup 6 \cap 7 \cup 2 \cap 3 \cup 1$	64.89	9.03	29.61	4.65	6.17
$2 \cup 7 \cap 1 \cap 8 \cup 3 \cap 6$	30.22	2.30	30.50	0.73	2.32
$6 \cap 2 \cup 8 \cup 7 \cap 1$	39.59	4.40	30.56	1.67	3.21
$8 \cap 3 \cap 2 \cup 1 \cap 7$	21.10	5.78	29.35	-2.46	-0.94

Table 6.1: *A selection of the best evaluation results of the combining of algorithms 1-8 for an auditory scene consisting of two sources (mean values).*

of 5.10 dB and 6.61 dB. But compared to other combinations that recover fewer total energy, the number of false estimated bins is quite high with 9.09%, which shows that there are a lot of interfering bins included. Combination $8 \cup 7 \cap 1 \cap 6$ on the other hand reconstructs only 28.76% of the energy, but only 2.23% of the estimated bins belong to the interferers. In this way the algorithm combination achieves mean SIR gains of 31.42 dB, but the speech quality parameters SAR and SDR are only improved marginally by 0.18 dB and 1.78 dB.

Table 6.6 shows the same results for auditory scenes consisting of three speech sources. In the mean of 240 three source speech mixtures at positions $-45°$, $0°$ and $45°$ SIR gains of approximately 16 dB can be achieved for the source at position $0°$, while up to 33 % of the ideal mask energy is recovered.

The strategy used for combining the masks estimated by the algorithms is dependent on the purpose of the separation infrastructure. If the target source is to be enhanced for better intelligibility by humans, combination $7 \cup 8 \cup 6 \cap 3 \cup 2 \cup 1$ is suitable as this strategy achieves a high percentage of recovered energy and therefore a high speech quality. However the interfering sources are not suppressed that well. Human source separation can take over the final step. If however the framework is used as input to an automatic speech recognizer – which in most cases is very sensitive to interfering speech sources – combination $8 \cup 7 \cap 1 \cap 6$ is adequate as this strategy maximizes the SIR gains. Other purposes could choose a combination which balances the percentage of recovered energy and SIR gains to achieve an intermediate quality.

Algorithm Order	Recovered Energy of ideal mask [%]	False estimated bins [%]	SIR gain [dB]	SAR gain [dB]	SDR gain [dB]
$7 \cup 1 \cup 6 \cap 3$	31.76	16.19	13.66	-0.60	1.17
$2 \cup 6 \cap 8 \cap 3 \cup 7 \cup 1$	33.72	16.07	13.33	-0.15	1.81
$1 \cap 8 \cup 7 \cap 3$	19.53	10.34	16.64	-3.40	-1.24
$6 \cup 3 \cup 8 \cap 7 \cap 1$	57.39	21.15	7.01	1.44	2.52
$6 \cap 2 \cap 8 \cup 1 \cap 7$	57.65	22.34	7.35	1.42	2.48
$6 \cup 2 \cap 8 \cup 7 \cap 1$	57.62	22.29	7.35	1.43	2.49
$3 \cup 6 \cap 8 \cup 7 \cap 1$	57.38	22.08	7.01	1.46	2.53

Table 6.2: *A selection of the best evaluation results of the combining of algorithms 1-8 for an auditory scene consisting of three sources (mean values).*

Algorithm	Recovered Energy of ideal mask [%]	False estimated bins [%]	SIR gain [dB]	SAR gain [dB]	SDR gain [dB]
Fixed Beamforming	99.95	41.73	0.70	0.47	0.56
DUET	96.48	40.26	0.28	0.19	0.12
Fair DUET	96.06	41.41	0.68	0.53	0.66

Table 6.3: *Results of state-of-the-art source separation approaches.*

6.7 Comparison to State-of-the-Art Techniques

The results of the separated speech sources are compared to Fixed Beamforming and the DUET [127] source separation algorithm. As the target source is assumed to be at zero degree, fixed beamforming can be implemented by adding the left and right ear signal and dividing by two. The DUET algorithm is originally designed to work in anechoic scenarios and heavily relies on the correct positions of the sources in terms of phase and amplitude differences between the two ear signals. The estimation of those values suffers a lot under reverberant conditions as shown in section 6.2. To be fair and comparable, the correct position values in terms of amplitude and phase differences of the sources in the auditory scene are provided to the algorithm by measured mean HRTF values of the head.

Fixed beamforming enhances the target signal only by about 0.7 dB, but preserves almost all of the target energy (99.95 %). The high recovered energy leads to a high percentage of false estimated bins of more than 40 % as the beamforming technique does not separate the sources on a TF-bin basis. But compared to other techniques based on the TF-masks beamforming does not manipulate the signal in non-linear ways as binary mask demixing does. For a stand alone source separation architecture the fixed beamforming technique is not appropriate, but it can

be used in front of another architecture to enhance the separation process. In the case of the current project, the fixed beamforming technique comes almost for free and can be implemented by simply adding the two ear inputs and dividing by two. It can either be used and integrated as a further "algorithm" in the previous described architecture or the beamforming can be applied to the final estimated left and right ear signals to achieve a final signal.

The DUET algorithm performs poorly also in the fair variant. It reconstructs much energy of the target source, but it only achieves SIR gains of 0.28 dB respectively 0.68 dB. The resulting masks include almost all possible bins of the masks, which leads to the high percentage of false estimated bins of over 40%. The original DUET algorithm additionally estimates completely false positions in about 30 % of all cases, which leads to a total reconstructed energy of about 0 % and sometimes quite high SIR values of about 20 dB. These values are not useful and are not considered in the mean computation as these would manipulate the total results.

6.8 Ideal Head Position

During communication humans move their head to perfectly align to the signal of interest. To find out if the straightforward approach to directly face the target source is ideal, the presented source separation algorithms are evaluated for different head steering directions between $-90°$ and $90°$. The sources are further located at $0°$ and $45°$.

Figure 6.11 shows the evaluation of the left (continuous line) and the right ear (dashed line) for a selection of the previously presented algorithms.

Algorithm 1 (separation based on ITD) performs best regarding the percentage of reconstructed energy, when the head is positioned between the two sources as in this position the incidence angle of all sources is minimized and the direction of arrival estimation is most reliable. When turning away from both sources the reconstructed energy suffers as the geometric free field assumption used to estimate the incidence direction isn't satisfied anymore due to reflections and resonances of the human ear and head. The SIR however is drastically better for positions on the left side of the sources as the interfering source is then in the head shadow. A good trade-off between all evaluation values is achieved when aligning the head directly to the target source. The other separation algorithms based on ITD perform similar.

Algorithm 6 (separation based on F0) shows a dependency between the position and the reconstructed energy of the left and right ear. This contributes to the monaural beamforming effect of the human ear: The more direct the incidence direction of the target source, the more energy of this source reaches the ear and the better the algorithm can separate this source. The SIR shows no clear dependency regarding the head steering direction. Again a head position directly steering to the target source is a good trade-off for the separation.

The results of the combinations $7 \cup 8 \cup 6 \cap 3 \cup 2 \cup 1$ and $7 \cup 1 \cup 6 \cap 3$ are clearly affected by the performance of the single constituent algorithms. Combination $7 \cup 8 \cup 6 \cap 3 \cup 2 \cup 1$ is strongly

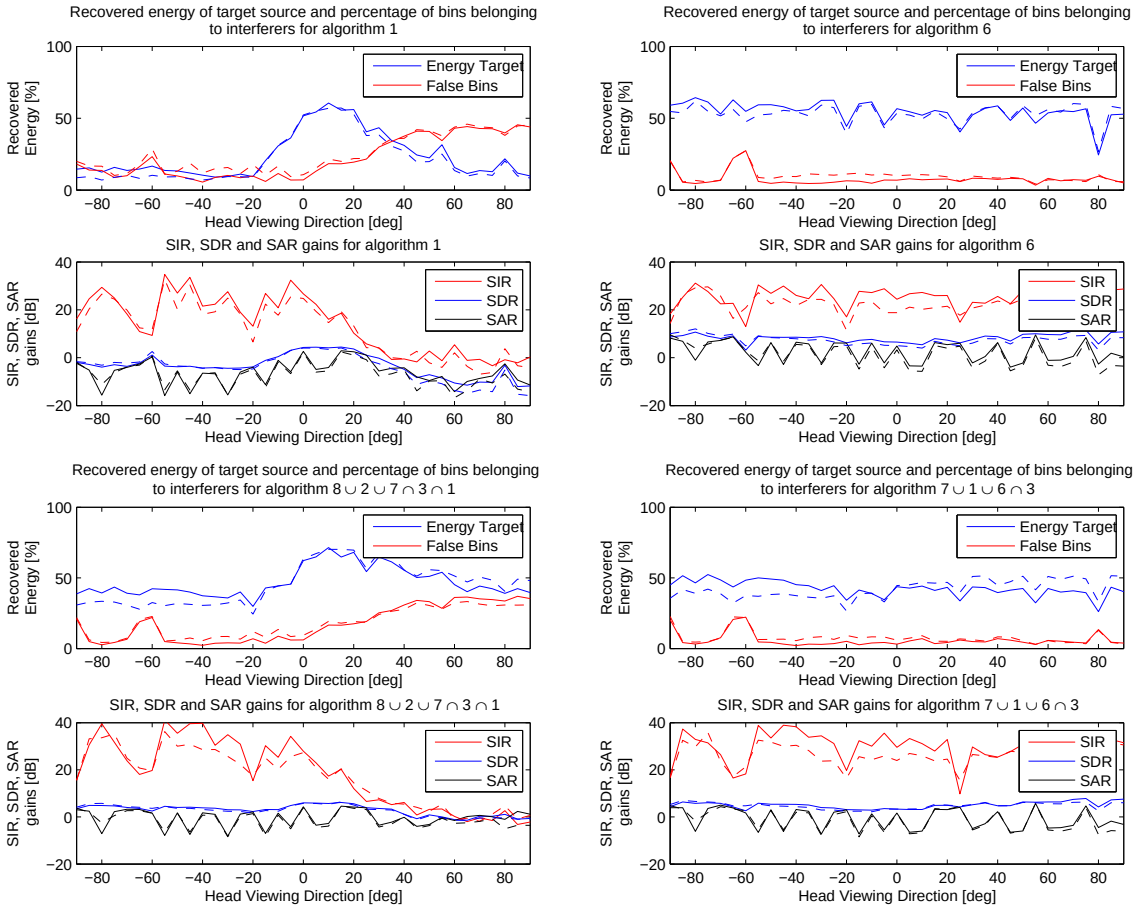

Figure 6.11: *Evaluation of selected source separation algorithms subject to the head steering direction for the left (continuous line) and the right ear (dashed line).*

influenced by algorithm 1, which leads to performance curves similar to that of algorithm 1, but with a higher percentage of recovered energy. Combination $7 \cup 1 \cup 6 \cap 3$ shows no clear best head direction, but directly steering to the source seems to be a good choice. Using the turned away ear to demix the sources is advantageous regarding the percentage of reconstructed energy as the interfering source is then positioned in the head shadow and is attenuated.

Overall the ideal head position depends on the setup of the auditory scene – such as the number and the positions of the sources – and the used combination scheme. For all separation algorithms and combinations the position where Bob is directly steering towards the target source is a good trade-off.

7 Conclusions

7.1 Summary

The goal of this thesis was to enhance the separation results of standard binaural signal processing approaches for source separation by making use of cognitive signal processing mechanisms. This cognitive separation approach is tested in a realistic non-ideal experiment setup by using a human dummy head that is able to move in three dimensions. The dummy head resides in a normal reverberant office room and listens to auditory scenes played back by a high quality 7.1 loudspeaker system.

Orthogonality of Speech Sources in Reverberant Humanoid Scenarios

An often used final goal of source separation architectures is finding the ideal binary time-frequency mask: Each entry of the T-F-mask is set to one if the target energy in this T-F-bin is greater than the interfering energy. This concept is based on the approximate orthogonality of speech sources in the time-frequency domain, which has been investigated in detail for anechoic speech mixtures. To be applicable to real world scenarios, such as a robotic human dummy head in the case of this research, this thesis investigated how the orthogonality of speech sources in the time-frequency domain drops with different reverberation times of the environment and indicated that separation schemes based on ideal binary T-F-masks are suitable to perform source separation under humanoid reverberant conditions. Reverberation and the HRTF filtering of the human head influence the orthogonality of speech sources in the time-frequency domain and the SIR decreases by approximately 5 dB for a two-source humanoid reverberant scenario compared to the anechoic case.

Auditory Scene Analysis

When humans enter an auditory scene, they automatically analyze the environment around them and estimate parameters like the positions and the fundamental frequency of the preferred target speaker. The source separation architecture presented in this thesis tries to mimic these cognitive abilities of the human brain. Prior to separating the speech sources, the human dummy head Bob analyzes the auditory scene and estimates the number, the positions and the fundamental

frequency tracks of the sources in the auditory scene, which are then used to enhance the following separation approaches.

A new localization approach which assumes that the sources are positioned on a circle around the listener is introduced and performs better than standard approaches for humanoid source localization like the Woodworth formula and the Freefield formula. Furthermore a localization approach based on approximated HRTFs is introduced and evaluated. Iterative variants of the algorithms enhance the localization accuracy and resolve specific localization ambiguities. In this way a localization blur of approximately three degrees is achieved which is comparable to the human localization blur. A front-back confusion allows a reliable localization of the sources in the whole azimuth plane in up to 98.43 % of the cases.

The algorithm used to estimate the fundamental frequency track of the speech sources in the auditory scene is implemented in many parts analog to the YIN-method [22]. The input signals are divided in time windows of 50 ms length, such that two periods of a 40 Hz wave are captured. For each of these windows a fundamental frequency is estimated to construct a complete F0-track. A postprocessing stage smoothes the F0-track by removing outliers and applying specific assumptions of the human voice.

Source Separation

The presented source separation framework combines the positive features of the STFT with the positive features of the cochleagram. The overall goal of the source separation is to find the ideal STFT-mask. The core source separation process however is based on the analysis of the corresponding region in an additionally computed cochleagram. In this way a much more reliable ITD estimation is realized and used for separation.

Several algorithms based on the ITD and the fundamental frequency of the target source are evaluated for their source separation capabilities. To enhance the separation capabilities of the single algorithms, the results of the different algorithms are combined to compute a final estimate. In this way SIR gains of about 30 dB for two source scenarios are achieved. For three source scenarios SIR gains of up to 16 dB are attained. Compared to the standard binaural signal processing approaches like DUET and Fixed Beamforming the presented approach achieves up to 29 dB SIR gain.

7.2 Future Work

Future work in the presented project especially includes an improvement of the core source separation by a multimodal extension of the whole project to the visual domain and a postprocessing of the estimated binary masks.

Extension to Visual Domain

An extension of the cognitive source separation project to the visual domain seems to be natural. Adding two eyes to the human dummy head Bob makes the setup more realistic. Humans rarely separate auditory sources without additionally using their visual system. This project therefore examines in the next step, if additional visual information is useful to enhance the source separation capabilities.

Haschke [46] realizes the extension of Bob's eyes, which are implemented by two cameras, that are movable as human eyes are able to do. When visual information is available to the source separation architecture, several approaches are to be examined for the capabilities in enhancing the source separation.

On/Offset-Detection With the help of the cameras, on- and offsets of the speakers can be reliably estimated i.e. by noting the motion of the mouth and lips. If the correct on- and offsets are known, the speech quality of those parts where only one speaker is audible can be drastically enhanced. At all other times, the source separation architecture can adapt to the number of currently speaking persons.

Face Recognition The visual system enables the source separation architecture to include further speaker-specific information in the separation process. A gender detection can support the estimation process of the fundamental frequency track. Women i.e. usually have a higher fundamental frequency as men, so that the correct assignment of the available F0s to the target speaker is enhanced.

If the visual system implements a complete face recognition system, several prior information which has been learned in the past can be made available. When a specific person is recognized, its fundamental frequency, tone, accent and its language habit are recalled from memory and can be used for separation.

Enhanced Localization The input of the visual system allows the implementation of a better and faster localization. Visual information is robust against reflections and reverberation and so is more reliable than acoustic information. The acoustic and the visual localization can complement each other in the localization task. Both can compute independent estimations of the number of sources and their positions in the auditory scene. In limiting cases i.e. when a speaker is not in the visible field or when two speakers are localized nearby, the fusion of the auditory and visual system surely enhances the accuracy and stability of the localization.

Lip Reading In future steps of the project, a lip reading ability can be implemented. When the target speaker is focused, lip gestures can be estimated and the recognized vowels and words can be reconstructed by applying filters that refill the resulting separation spectrum according to the recognized values.

Postprocessing of separated sources

The presented source separation architecture delivers the separated target source by applying binary masks to the mixture spectrum. This binary demixing yields spectra with hard edges and many wholes in the spectrum, which degrades the speech quality of the separated sources.

Future work in the project should analyze human speech spectra in detail and should try to adapt the characteristics of the target spectra to those of real speech spectra by non-binary postprocessing of the binary masks. For example the energy distribution over the frequency bins around harmonics and on- and offsets can be estimated. The speech quality of the binary target spectrum can then be enhanced by a non-binary refilling of the spectrum according to the estimated energy distributions.

Additionally whole models of human speech production – i.e. models which are used to perform speech synthesis – can improve the speech quality drastically. However by applying such models, the resulting voice does not necessarily be similar to the underlying voice of the target source in the mixture. Nonetheless the speech intelligibility can be much higher in this way. The use of such models is therefore dependent on the intended use of the source separation architecture.

Bibliography

[1] J. Karhunen A. Hyvärinen and E. Oja. *Independent Component Analysis (Adaptive and learning systems for signal processing, communications, and control)*. New York: Wiley, 2001.

[2] Ralph Algazi, Richard O Duda, Dennis M. Thompson, and Carlos Avendano. The CIPIC HRTF Database. In *WASSAP '01*. 2001 IEEE ASSP Workshop on Application of Signal Processing to Audio and Acoustics, 2001.

[3] S. Araki, S. Makino, A. Blin, and and H. Sawada R. Mukai. Blind separation of more speech than sensors with less distortion by combining sparseness and ICA. In *Proceedings of Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 271–274, September 2003.

[4] L. Atlas. Modulation spectral filtering of speech. *Proceedings of Eurospeech*, pages 2577 – 2580, 2003.

[5] Dirk Bechler and Kristian Kroschel. Three different reliability criteria for time delay estimates. *12 th European Signal Processing Conference*, 2004.

[6] Anthony J. Bell and Terrence J. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[7] F. Berthommier and G. Meyer. Improving of amplitude modulation maps for F0-dependent segregation of harmonic sounds. *Proceedings of Eurospeech*, 1997.

[8] Stanley T. Birchfield and Rajitha Gangishetty. Acoustic Localization by Interaural Level Difference. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2005.

[9] Jens Blauert. *Untersuchungen zum Richtungshören in der Medianebene bei fixiertem Kopf*. PhD thesis, Rheinisch-Westfälische Technische Hochschule Aachen, 1969.

[10] Jens Blauert. *Spatial Hearing (Revised Edition)*. MIT Press, 1997.

[11] M. Bodden. Modelling human sound-source localization and the cocktail party effect. *Acta Acoustica*, 1:43 – 55, 1993.

[12] J. Braasch. Localization in the presence of a distractor and reverberation in the frontal horizontal plane: II. Model algorithms. *Acta Acustica united with Acustica*, 88:956 – 969, 2002.

[13] Albert S. Bregman. *Auditory Scene Analysis – The Perceptual Organization Of Sound.* MIT Press, 1990.

[14] G.J Brown and M. P. Cooke. Computational auditory scene analysis. *Computer speech and language*, 8:297 – 336, 1994.

[15] G.J. Brown, S. Harding, and J. P. Barker. Speech separation based on the statistics of binaural auditory features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[16] Guy J. Brown and D. L. Wang. Separation of Speech by Computational Auditory Scene Analysis. In *Benesty J., Makino S., and Chen J. (ed.), Speech Enhancement*, pages 371 – 402. Springer, 2005.

[17] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, 120:4007 – 4018, 2006.

[18] M. D. Burkhard and R. M. Sachs. Anthropometric manikin for acoustic research. *Journal of the Acoustical Society of America*, 58 (1):214 – 222, 1974.

[19] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:679 – 698, October 1986.

[20] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of Acoustical Society of America*, 25(5):975 – 979, 1953.

[21] Alain de Cheveigne. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model for auditory processing. *Journal of the Acoustical Society of America*, 93(6):3271 – 3290, 1993.

[22] Alain de Cheveigne and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111:1917–1930, 2002.

[23] L. Cohen. Time-Frequency Distributions – A review. *in Proceedings of the IEEE*, 77(7):941 – 980, July 1989.

[24] Martin Cooke, Daniel P. W. Ellis, and Dan Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35 (3-4):141 – 177, 2001.

[25] N. Cowan. The magic number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioural Brain Science*, 24:87 – 185, 2001.

[26] J. F. Culling, Q. Summerfield, and D.H. Marshall. Effects of simulated reverberation on the use of binaural cues and fundamental frequency differences for separating concurrent vowels. *Speech Communication*, 14:71 – 95, 1994.

[27] S. Devore and B. G. Shinn-Cunningham. Perceptual consequences of including reverberation in spatial auditory displays. In *Proceedings of the International Conference on Auditory Display (ICAD 2003)*, pages 75 – 78, 2003.

[28] R. H. Domnitz and H. S. Colburn. Lateral position and interaural discrimination. *Journal of the Acoustical Society of America*, 61:1586–1598, 1977.

[29] R. O. Duda and W. L. Martens. Range-dependence of the HRTF for a spherical head. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 5 – 10, october 1997.

[30] N. I. Durlach. Equalization and cancellation theory of binaural masking level differences. *Journal of the Acoustical Society of America*, 35 (8):1206 – 1218, 1963.

[31] Dan Ellis. *Prediction-driven computational auditory scene analysis.* PhD thesis, MIT Department of Electrical Engineering and Computer Science, 1996.

[32] Christof Faller and Juha Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *The Journal of the Acoustical Society of America*, 116(5):3075–3089, 2004.

[33] G. D. Forney. The Viterbi algorithm. In *Proceedings of the IEEE*, volume 61 (3), pages 268 – 278, March 1973.

[34] O.L. Frost. An algorithm for linearly constrained adaptive array processing. In *Proceedings of the IEEE*, volume 60 (8), pages 926 – 935, August 1972.

[35] C. Févotte. Bayesian blind separation of audio mixtures with structured priors. *14th European Signal Processing Conference (EUSIPCO'06)*, 2006.

[36] C. Févotte, R. Gribonval, and E. Vincent. *BSS-EVAL toolbox user guide*, 2005.

[37] Mark Gales. *Model-Based Techniques for Noise Robust Speech Recognition.* PhD thesis, University of Cambridge, 1996.

[38] Mark Gales and Steve Young. The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing*, 1:195 – 304, 2008.

[39] S. A. Gelfand and S. Silman. Effects of small room reverberation upon the recognition of some consonant features . *Journal of the Acoustical Society of America*, 66:22–29, July 1979.

[40] L. Godara. Application of Antenna Arrays to Mobile Communications, Part 1: Performance Improvement, Feasability and System Considerations. In *Proceedings of the IEEE*, volume 85 (7), pages 1031 – 1060, June 1997.

[41] L. Godara. Application of Antenna Arrays to Mobile Communications, Part 2: Beam-Forming and Direction-of-Arrival Considerations. In *Proceedings of the IEEE*, volume 85 (8), pages 1195 – 1245, August 1997.

[42] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30:27–34, January 1982.

[43] W. M. Hartmann and B. Rakered. Localization of sound in rooms: The effect of a single reflecting surface. *Journal of the Acoustical Society of America*, 79 (2):524 – 533, 1985.

[44] William M. Hartmann. How we localize sound. In *Physics Today*, November 1999.

[45] Eric Haschke. Sound Source Localization using a movable human dummy head. Master's thesis, Saarland University, 2007.

[46] Eric Haschke. *Auditory Source Separation – New multimodal approaches in the time-frequency domain.* PhD thesis, Saarland University, to pe published.

[47] Simon Haykin. *Adaptive Filter Theory (Third Edition).* Prentice Hall, 1996.

[48] G. Hu and D. L. Wang. Auditory segmentation based on event detection. *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

[49] G. Hu and D. L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, 15:1135 – 1150, 2004.

[50] G. Hu and D. L. Wang. Auditory segmentation based on onset and offset analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:396 – 405, 2007.

[51] Guoning Hu. *Monaural speech organization and segregation.* PhD thesis, The Ohio State University Biophysics Program, 2006.

[52] L. A. Jeffress. A place theory of sound localization. *Journal of Comparitive Psychology*, 41:35 – 39, 1948.

[53] Peter I.M. Johannesma. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In *Symposium on Hearing Theory*, pages 58–69, Eindhoven, Holland, June 1972. Institute for Perception Research (IPO).

[54] W. E. Kock. Binaural localization and masking. *Journal of the Acoustical Society of America*, 22:801 – 804, 1950.

[55] Z. Koldovsky and P. Tichavsky. Time-Domain Blind Audio Source Separation Using Advanced Component Clustering and Reconstruction. *Hands-Free Speech Communication and Microphone Arrays*, pages 216 – 219, 2008.

[56] B. Kollmeier and R. Koch. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *Journal of the Acoustical Society of America*, 95:1593–1602, 1994.

[57] B. Kollmeier, J. Peissig, and V. Hohmann. Real-time multiband dynamic compression and noise reduction for binaural hearing aids. *Journal of Rehabilitation Research and Development*, 30 (1):82–94, 1993.

[58] John Kominek and Alan W Black. CMU ARCTIC databases for speech synthesis. In *5th ISCA Speech Synthesis Workshop – Pittsburgh*, pages 223 – 224, 2004.

[59] Jochen Krämer. Multiple Fundamental Frequency Estimation for Cognitive Source Separation. Bachelor's thesis, Saarland University, 2008.

[60] Sylvia Kümmel, Eric Haschke, and Thorsten Herfet. Human Inspired Auditory Source Localization. In *Proceedings of 12th International Conference on Digital Audio Effects (DAFx-09), Como, Italy*, September 2009.

[61] Russ Lambert. *Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures*. PhD thesis, University of Southern California, Department of Electrical Engineering, 1996.

[62] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 1998.

[63] B. Libbey and P. H. Rogers. The effect of overlap-masking on binaural reverberant word intelligibility. *Journal of the Acoustical Society of America*, 115:3141 –3151, 2004.

[64] J. C. R. Licklider. A Duplex Theory of Pitch Perception. *Experientia*, 7:128 –134, 1951.

[65] C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng. Localization of Multiple Sound Sources with Two Microphones. *Journal of the Acoustical Society of America*, 108:1888–1905, 2000.

[66] C. Liu, B. C. Wheeler, W. D. O'Brien, C. R. Lansing, R. C. Bilger, D. Jones, and A. S. Feng. A Two-microphone Dual Delay-line Approach for Extraction of a Speech Sound in the Presence of Multiple Interferers. *Journal of the Acoustical Society of America*, 110:3218–3231, 2001.

[67] M. E. Lockwood, D. L. Jones, R. C. Bilger, C. R. Lansing, W. D. O'Brien, B. C. Wheeler, and A. S. Feng. Performance of Time- and Frequency-domain Binaural Beamformers Based on Recorded Signals from Real Rooms. *Journal of the Acoustical Society of America*, 115:379–391, 2004.

[68] Robert Lorenz and Stephen P. Boyd. Robust minimum variance beamforming. In *IEEE Transactions on signal processing*, volume 53, May 2005.

[69] R. F. Lyon. A computational model of binaural localization and separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1148 – 1151, 1983.

[70] David Marr. *Vision.* Freeman, New York, 1982.

[71] Aaron S. Master. *Stereo Music Source Separation via Bayesian Modeling.* PhD thesis, Stanford University, June 2006.

[72] R. B. Masterton and T. J. Imig. Neural mechanisms for sound localization. *Annual Review of Physiology*, 46:275–287, 1984.

[73] Stephen G. McGovern. A Model for Room Acoustics. http://www.2pi.us/rir.html, 2004.

[74] T. Melia and S. Rickard. Extending the DUET blind source separation technique. In *Proceedings of Signal Processing with Adaptative Sparse Structured Representations Workshop (SPARS '05)*, 2005.

[75] D. L. Wang N. Roman and G. Brown. Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114:2236 – 2252, 2003.

[76] Tomohiro Nakatani, Masato Miyoshi, and Keisuke Kinoshita. One Microphone Blind Dereverberation Based On Quasi-Periodicity Of Speech Signals. *In Advances in Neural Information Processing Systems*, pages 1417 – 1424, 2004.

[77] Tomohiro Nakatani, Masato Miyoshi, Keisuke Kinoshita, and P. Zolfaghari. Harmonicity based monaural speech dereverberation with time warping and F0 adaptive window. In *Proceedings of the International Conference of Spoken Language Processing*, pages 873 – 876, 2004.

[78] Alan Oppenheim and Alan Willsky. *Signals and Systems, 2nd Edition.* Prentice Hall, 1997.

[79] S. E. Palmer. *Vision Science.* MIT Press, Cambridge, MA, 1999.

[80] Kalle J. Palomäki, Guy J. Brown, and D. L. Wang. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication*, 43(4):361–378, 2004.

[81] Alex Park. Using the Gammachirp Filter for Auditory Analysis of Speech, 2003. Wavelets and Filterbanks.

[82] T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *Acoustical Society of America Journal*, 60:911–918, October 1976.

[83] I. Pollack and W. J. Trittipoe. Binaural listening and interaural cross correlation. *Journal of the Acoustical Society of America*, 31:1250–1252, 1959.

[84] Lawrence R. Rabiner. On the Use of Autocorrelation Analysis for Pitch Detection. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25:24–33, 1977.

[85] Vikas C. Raykar. A Study of a various Beamforming Techniques And Implementation of the Constrained Least Mean Squares (LMS) algorithm for Beamforming, 2001. Project Report ENEE 624.

[86] J. W. Strutt (Lord Rayleigh). On our perception of sound direction. *Philosophical Magazine*, 13:214 – 232, 1907.

[87] Ulrich Reimers. *Digital Video Broadcasting - The Family of International Standards for Digital Video Broadcasting*. Springer, 2005.

[88] Scott Rickard, Radu Balan, and Justinian Rosca. Real-time time-frequency based blind source separation. *3rd International Conference on Independent Component Analysis and Blind Source Separation, San Diego, CA*, 2001.

[89] M. D. Riley. *Speech Time-Frequency Representations*. Kluwer Academic, Boston, MA, 1989.

[90] N. Roman. *Auditory-Based Algorithms for Sound Segregation in Multisource and Reverberant Environments*. PhD thesis, Ohio State University, Department of Computer Science and Engineering, 2005.

[91] N. Roman, D. L.Wang, and G. Brown. Localization-based sound segregation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1013 – 1016, 2002.

[92] N. Roman, S. Srinivasan, and D. L. Wang. Binaural Segregation in Multisource Reverberant Environments, 2005. Technical Report OSU-CISRC-9/05-TR60.

[93] N. Roman, S. Srinivasan, and D. L. Wang. Binaural segregation in multisource reverberant environments. *Journal of the Acoustical Society of America*, 120:4040–4051, 2006.

[94] N. Roman and D. L. Wang. Binaural tracking of multiple moving sources. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 149 – 152, 2003.

[95] N. Roman and D. L Wang. Pitch-based monaural segregation of reverberant speech. *Journal of the Acoustical Society of America*, 120:458 – 469, 2006.

[96] N. Roman, D. L. Wang, and G. Brown. Speech segregation based on sound localization. In *Proceedings of IJCNN-01*, pages 2861 – 2866, 2001.

[97] Thomas D. Rossing. *Science of Sound*. Addison-Wesley, 1983.

[98] R. Roy and T. Kailath. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):984 – 995, 1989.

[99] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276 – 280, 1986.

[100] M. R. Schroeder. New method of measuring reverberation time. *Journal of the Acoustical Society of America*, 37:409 – 412, 1965.

[101] M. R. Schroeder. Integrated-impulse method measuring sound decay without using impulses. *Acoustical Society of America Journal*, 66:497–500, August 1979.

[102] Sylvia Schulz and Thorsten Herfet. Binaural Source Separation in Non-Ideal Reverberant Environments. In *Proceedings of 10th International Conference on Digital Audio Effects (DAFx-07), Bordeaux, France*, September 2007.

[103] Sylvia Schulz and Thorsten Herfet. Humanoid Separation of Speech Sources In Reverberant Environments. In *Proceedings of 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP 2008), St. Julians, Malta*, March 2008.

[104] Sylvia Schulz and Thorsten Herfet. On the Window-Disjoint-Orthogonality of Speech Sources in Reverberant Humanoid Scenarios. In *Proceedings of 11th International Conference on Digital Audio Effects (DAFx-08), Espoo, Finland*, September 2008.

[105] Yoshinori Shiga and Simon King. Estimating the Spectral Envelope of Voiced Speech Using Multi-frame Analysis. In *Proc. Eurospeech-2003*, volume 3, pages 1737–1740, September 2003.

[106] M. Slaney. Auditory Toolbox, 1998. Technical Report, Interval Research Corporation.

[107] Robert L. Solso. *Cognitive Psychology*. Allyn and Bacon, 1998.

[108] David Talkin. A Robust Algorithm For Pitch Tracking. *Speech Coding and Synthesis*, pages 495 – 518, 1995.

[109] Thilo Thiede, William C. Treurniet, Roland Bitto, Thomas Sporer, Karlheinz Brandenburg, Christian Schmidmer, Michael Keyhl, John G. Beerends, Catherine Colomes, Gerhard Stoll, and Bernhard Feiten. PEAQ - der künftige ITU-Standard zur objektiven Messung der wahrgenommenen Audioqualität. In *20. Tonmeistertagung*, Karlsruhe, 1998.

[110] K. Torkkola. Blind Separation of Convolved Sources Based on Information Maximization. In *IEEE Workshop on Neural Networks for Signal Processing, Kyoto, Japan*, pages 423–432, September 4-6 1996.

[111] A. Varga and R. Moore. Hidden Markov Model decomposition of speech and noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 845 – 848, 1990.

[112] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in Blind Audio Source Separation. *IEEE Trans. Audio, Speech and Language Processing*, pages 462–1469, 2006.

[113] Harald Viste. *Binaural Localization and Separation Techniques*. PhD thesis, Swiss Federal Institute of Technology, Lausanne, June 2004.

[114] Harald Viste and Gianpaolo Evangelista. On the Use of Spatial Cues to Improve Binaural Source Separation. In *Proceedings of 6th International Conference on Digital Audio Effects (DAFx-03), London, UK*, September 2003.

[115] Harald Viste and Gianpaolo Evangelista. Binaural Source Localization. In *Proceedings of 7th International Conference on Digital Audio Effects (DAFx-04), Napoli, Italy*, October 2004.

[116] D. L. Wang. Primitive Auditory Segregation Based on Oscillatory Correlation. *Cognitive Science*, 20(3):409–456, 1996.

[117] D. L. Wang. On Ideal Binary Masks as the Computational Goal of Auditory Scene Analysis. *In Divenyi P. (ed.), Speech Separation by Humans and Machines*, pages 181 – 197, 2005.

[118] D. L. Wang and G. J. Brown. Separation of Speech from Interfering Sounds Based on Oscillatory Correlation. *IEEE-NN*, 10(3):684, May 1999.

[119] D. L. Wang and Guy J. Brown. *Computational Auditory Scene Analysis - Principles, Algorithms, Applications*. IEEE Press, Wiley Interscience, 2006.

[120] M. Weintraub. *A Theory and Computational Model of Auditory Monaural Sound Separation*. PhD thesis, Stanford University, August 1985.

[121] Stefan Winter, Walter Kellermann, Hiroshi Sawada, and Shoji Makino. MAP-Based Underdetermined Blind Source Separation of Convolutive Mixtures by Hierarchical Clustering and 1-Norm Minimization. *EURASIP Journal on Advances in Signal Processing*, 2007.

[122] R. S. Woodworth and H. Schlosberg. *Experimental Psychology.* Holt, New York, 1954.

[123] M. Wu and D. L. Wang. A one-microphone algorithm for reverberant speech enhancement. In *Proceedings of ICASSP-03*, pages 844 – 847, 2003.

[124] M. Wu and D. L. Wang. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:774 – 784, 2006.

[125] M. Wu, D. L. Wang, and G. J. Brown. A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, 11:229 – 241, 2003.

[126] Wen Xue and Mark Sandler. Sinuoid Modeling in a Harmonic Context. *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*, pages 33–40, 2007.

[127] Ö. Yilmaz and S. Rickard. Blind Separation of Speech Mixtures via Time-Frequency Masking. *IEEE Transactions on Signal Processing*, 52(7):1830 – 1847, July 2004.

List of Figures

List of Tables