Localization and Separation of Concurrent Talkers Based on Principles of Auditory Scene Analysis and Multi-Dimensional Statistical Methods

Localization and Separation of Concurrent Talkers Based on Principles of Auditory Scene Analysis and Multi-Dimensional Statistical Methods

Von der Fakultät für Mathematik und Naturwissenschaften
der Carl von Ossietzky Universität Oldenburg zur
Erlangung des Grades und Titels eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
angenommene Dissertation

von Johannes Nix
geboren am 20. Mai 1967 in Köln

Gutachter: Prof. Dr. Dr. Birger Kollmeier
Korreferent: Prof. Dr. Alfred Mertins
Tag der Disputation: 7. 7. 2005

Instructions for Binding this Document

The layout of this document is designed for ISO B5 Format (176 x 250 millimeter). This electronic version is prepared for printing on ISO A4 paper and cutting it to the right size. Observe the following instructions to print and bind it correctly:

- Print this document with Adobe Acrobat Reader or equivalent software on paper with ISO A4 format (210 x 297 millimeter). Use double-sided printing with long edge binding. Do not use any scaling. If paper size and scaling are correct, the layout of odd and even pages will match.
- Remove this sheet.
- Crop the paper stack so that the crop marks are removed.
- Bind the sheets at the left.

Contents

List of Figures **viii**

List of Tables **ix**

Abbreviations **xi**

Symbols **xv**

Abstract **xix**

Zusammenfassung **xxi**

1. General Introduction **1**
1.1. Tackling the Cocktail Party Problem 1
1.2. Overview Of Existing Methods 2
1.2.1. Approaches Based on Spatial Filtering 2
1.2.2. Spectral Estimation Approaches 4
1.2.3. Further Methods 4
1.3. Scope of This Work 6
1.4. Pursued Strategies 6
1.4.1. Usage of Principles of Auditory Processing 6
1.4.2. The Probabilistic Approach: Feature Integration and Competing Hypotheses 10
1.4.3. Robust Feature Integration by Multidimensional State-Space Methods 12
1.5. Structure of the Thesis 17

2. Localization Based on Statistics of Interaural Parameters **19**
2.1. Introduction 20
2.2. Methods 22

2.2.1. Assumptions . . . 22
2.2.2. Computation of Interaural Parameters in the Frequency Domain . . . 24
2.2.3. Description of Interaural Parameters as Random Variables 25
2.2.4. Data Acquisition . . . 26
2.2.5. $\mathbf{d}'$ Analysis of Differences in Interaural Parameters . . . 31
2.2.6. Moment Analysis . . . 31
2.2.7. Bayesian Analysis for Sound Localization . . . 32
2.3. Results . . . 33
2.3.1. Distributions of Interaural Parameters . . . 33
2.3.2. Simulation Results . . . 42
2.4. Discussion . . . 50
2.4.1. Consequences of Parameter Distributions for Sound Localization . . . 50
2.4.2. Detectability of Differences in Interaural Parameters . . . 51
2.4.3. Simulation of Absolute Localization . . . 52
2.4.4. Frequency Integration of Probabilities . . . 54
2.4.5. Statistical Representation of Interaural Timing . . . 54
2.4.6. Possible Physiological Representations of Interaural Timing 55
2.5. Conclusions . . . 57

3. Noise-Robust Sound Localization for the Separation of Voices 59
3.1. Introduction . . . 60
3.2. Methods . . . 62
3.2.1. Algorithm for Sound Localization . . . 62
3.2.2. Signal Recordings . . . 65
3.2.3. Generation of *A Priori* Statistics . . . 67
3.2.4. Error Measure for Direction Estimates . . . 67
3.2.5. Demixing of Sound Sources . . . 67
3.2.6. SNR Evaluation . . . 69
3.3. Results . . . 69
3.3.1. Localization of Single Sound Sources in Noise . . . 70
3.3.2. Localization of Concurrent Talkers . . . 74
3.3.3. Demixing Two Sound Sources . . . 79
3.4. Discussion . . . 80
3.4.1. Comparison to Other Approaches for Sound Localization 80
3.4.2. Use of Sound Localization for Directional Filtering . . . 82
3.5. Summary . . . 82

4. Estimating Envelopes of Concurrent Voices by Statistical Filtering **85**
4.1. Introduction 86
4.2. Methods 90
4.2.1. General Assumptions 90
4.2.2. Framework of State-Space Tracking of Voices 90
4.2.3. Bootstrap SMC Algorithm 92
4.2.4. Implementation of Statistical Algorithm 94
4.2.5. Technical Remarks 100
4.2.6. Algorithm Parameters 101
4.2.7. Experiments 102
4.3. Results 103
4.3.1. Experiment 1 (One Voice, Moving Azimuth) 103
4.3.2. Experiment 2 (Two Voices, Fixed Azimuth Variance) . . . 105
4.3.3. Experiment 3 (Two Voices, Decaying Azimuth Variance) . 106
4.4. Discussion 110
4.4.1. Integration of Principles used for CASA 110
4.4.2. Advantages and Potentials of Multidimensional Filtering . 113
4.4.3. Challenges and Improvement Strategies 115
4.5. Summary and Conclusions 117

5. General Summary and Conclusions **119**
5.1. Retrospect on the Goals 119
5.2. Summary of the Results 120
5.2.1. Bayesian Algorithm for Sound Localization 120
5.2.2. Application of Bayesian Sound Localization 121
5.2.3. CASA by Multidimensional Statistical Filtering 121
5.3. Suggestions for Future Work 123
5.3.1. Sound Localization 123
5.3.2. Separation of Speech and Nonstationary Noise by Statistical Methods 124
5.3.3. Integration of Bayesian Sound Localization and Multidimensional Statistical Filtering 125
5.4. Quintessence 126

A. Real-Time, Script-Based Development of Signal Processing Algorithms **127**
A.1. Current Real-Time Programming Environments 127
A.2. A Script-Based Real Time Processing Environment 129
A.2.1. Overview 129

A.2.2. Components 130
A.2.3. Characteristics of Implemented Solution 133
A.3. Summary 134

B. Moment Coefficients for Linear and Cyclic Random Variables 137
B.1. Moment Coefficients of Distributions of Linear Random Variables 137
B.2. Moment Coefficients and Parameters of Distributions of Cyclic Random Variables 138

C. Envelope Series Generated from Markov Statistics of Speech 139

D. Time Series of Interaural Timing Parameters 141

E. Distributions of Interaural Level and Phase Differences 145

Bibliography 170

Index of Citations 175

Index 191

Acknowledgments 191

Biographical Note 193

List of Figures

1.1. "Bregman's Bs" as an Illustration of Scene Analysis (I) 7
1.2. "Bregman's Bs" as an Illustration of Scene Analysis (II) 9
1.3. Illustration of the "Old Plus New" Principle 10
1.4. Illustration of Stochastic Tracking 13
1.5. Tracking of Signals in a Probable Subspace 14
1.6. Tracking of Objects in Multiple Dimensions 15
1.7. Representing a Multidimensional Function by Discrete Samples I 16
1.8. Representing a Multidimensional Function by Discrete Samples II 16

2.1. Fluctuations of ILD, Caused by Nonstationary Signals 23
2.2. Distribution Percentiles for the ILD 35
2.3. Histograms of ILD for different SNRs 36
2.4. Polar Plots of Histograms of the IPD for Different SNRs 38
2.5. d' Values for Interaural Parameters 44
2.6. Decision Histogram for Bayesian Localization at 5 dB SNR 46
2.7. Error Patterns in Decision Histograms 47
2.8. Localization Performance of Bayesian Estimation: Percentage of Front-Back Confusions 48

3.1. Block Diagram of Bayesian Real-Time Localization Algorithm . . 63
3.2. Effect of Frequency Integration 71
3.3. Decision Histogram of Localization Estimates for Cafeteria Noise, at an SNR of 15 dB 72
3.4. Histogram of Localization Estimates in Cafeteria Noise at an SNR of 15 db, with Reference Data for Train Station Concourse 73
3.5. Histogram of Localization Estimates, one Interfering Speaker . . 76
3.6. Contour Plot of Time-Dependent *A Posteriori* Probability Densities for the Azimuth 77
3.7. MAP Estimates of Azimuth for a Mixture of Three Talkers 78

4.1. Global Structure of Statistical Filtering Algorithm 95
4.2. Expected Value of Azimuth and true Azimuth for one Moving Voice 104
4.3. Estimated and true Short-Term Level for one Voice 104
4.4. Estimated and True Azimuths for two Voices 105
4.5. Estimated and true Azimuth for two Voices, decaying σ_α 106
4.6. Original and Estimated Spectrograms for two Voices 107
4.7. Inter-Band Cross Correlations Between Envelopes of Original and Estimated Signals 111

A.1. Global Structure of Real-Time Processing System 130

C.1. Spectral Series Generated from the First-Order Markov Statistics of Speech 140

D.1. Time Series of the Argument of the Maximum of the ICCF and of the IPD in Silence 142
D.2. Time Series of the Argument of the Maximum of the ICCF and of the IPD at an SNR of 20 dB 143

E.1. Mean Value and Standard Deviation of ILD as a Function of Frequency 146
E.2. Mean Value and Standard Deviation of IPD as a Function of Frequency 147
E.3. Histograms of ILD in Cafeteria Noise at Several SNRs 148
E.4. Histograms of IPD in Cafeteria Noise at Several SNRs 149

List of Tables

2.1. Noise Conditions for Analysis of Interaural Parameters 28
2.2. Moments of Distributions of the ILD at various SNRs 39
2.3. Moments of Distributions of the IPD at various SNRs 40
2.4. Moments of ILD and IPD in different Noise Environments 43
2.5. Percentage of Front-Back Confusions for Different Reference and Test SNR 49

3.1. Percentage of Front-Back Confusions for Different Noise Environments in Training and Operation Stage 74
3.2. SNR Improvements for Demixed Voices 80

4.1. Steps of the Bootstrap Algorithm 93
4.2. Parameters of the Statistical Filtering Algorithm 101
4.3. SNR Improvements by Statistical Envelope Filtering 109

List of Abbreviations

ALSA Advanced Linux Sound Architecture, collection of audio drivers and API

API Application Program Interface

ASR Automatic Speech Recognition

BSS Blind Source Separation

CASA Computational Auditory Scene Analysis

c.f. *confer*, compare

CIPIC Center for Image Processing and Integrated Computing

CPU Central Processing Unit

DAT Digital Audio Tape

DOA Direction Of Arrival

DSP Digital Signal Processor

d' **Analysis** Analyis of Detectability

e.g. *exempli gratia*, for example

ERB Equivalent Rectangular Bandwidth (of auditory filters)

FFT Fast Fourier Transform

fMRI functional Magnetic Resonance Imaging

HMM Hidden Markov Model

HRTF Head Related Transfer Function

IC (Neurophysiology) Colliculus Inferior

ICA Indenpedent Component Analysis

ICCF Interaural Cross Correlation Function

ICRA International Collegium of Rehabilitative Audiology

ILD Interaural Level Difference

IPD Interaural Phase Difference

ITD Interaural Time Difference

ITE In-the-ear Hearing Aid

ITF Interaural Transfer Function

JND (Psychoacoustics) Just Noticable Difference

LAM Local area multicomputer

LENS Location-Estimating, Null-Steering algorithm

LPC Linear Predictive Coding

LSF Line Spectrum Frequencies

LSO (Neurophysiology) Lateral Superior Olive

MAP Maximum A Posteriori

MIMO Multiple Input, Multiple Output

MPI Message Passing Interface

MRI Magnetic Resonance Imaging

MSO (Neurophysiology) Medial Superior Olive

OS Operating System

PAC (Neurophysiology) Posterior Auditory Cortex

PCA Principal components Analysis

PCI Peripheral Component Interconnect

PDA Personal Digital Assistant

PDF Probability Density Function

POSIX Portable Operating System Interface (API standard)

QNX (Commercial real-time operating system)

RMS Root Mean Square (of values)

SCSI Small Computer System Interface, bus standard

SIMD Single Instruction, Multiple Data

SMC Sequential Monte Carlo (Method)

SMP Symmetric Multiprocessing

SNR Signal-To-Noise Ratio

SRT (Speech audiometry) Speech Reception Treshold

SPL Sound Pressure Level

TASP Two Arc Source Positioning System, measurement device

TIMIT Speech database for automatic speech recognition, recorded at Texas Instruments (TI), and transcribed at the Massachusetts Institute of Technology (MIT)

TVAR Time-Variant Auto-Regressive Model

URL Uniform Resource Locator (internet address)

VAD Voice Activity Detection

VME Bus system (IEEE 1014 standard)

VQ Vector Quantization

Symbols

The numbers in parentheses indicate the page where the symbol is mentioned first.

$\sim$ 'is sampled from' (94)

α azimuth (24)

$\vec{\alpha}$ vector of azimuth values for different sound sources (68)

B number of frequency bands (25)

b frequency band index (29)

$c_{v,k}$ index of codebook entry (97)

D dimension of the state space (115)

d' detectability (31)

ΔT size of time steps between short-term spectra (30)

Δ_L interaural level difference (ILD) (25)

Δ_θ interaural phase difference (IPD) (25)

$\vec{\Delta}(k)$ vector of interaural parameters at time step k (25)

$\vec{\boldsymbol{\Delta}}(k)$ vector of interaural parameters (random variable) (25)

e Euler's number

$E[x]$ expected value of x

$F_r(f)$ and $F_l(f)$ spectrum of left and right ear signal (29)

$F_{ll}(b,k)$ intensity-weigthed level at left side for band b and time k (30)

$F_{rl}(b,k)$ intensity-weigthed cross spectrum for band b and time k (30)

$F_{rr}(b,k)$ intensity-weigthed level at right side for band b and time k (30)

$\mathcal{F}$ transformation from time series to short-term spectral coefficients (96)

f frequency (index of FFT coefficient) (24)

$G_{\alpha,\phi}(f)$ "Ear-independent Transfer Function" (68)

$g(\vec{x}_k)$ arbitrary function of system state (93)

γ coefficient of first order low pass filter (30)

$H_{\alpha,\phi,r}(f)$ and $H_{\alpha,\phi,l}(f)$ right and left HRTF for azimuth α, and elevation ϕ (24)

$\mathbf{H_f}(\vec{\alpha})$ mixing matrix for direction vector $\vec{\alpha}$ and frequency f (68)

$I_{\alpha,\phi}(f)$ Interaural Transfer Function (24)

i imaginary unit

$\Im[c]$ imaginary part of c

K kurtosis (also called kurtosis excess, zero for Gaussian PDF) (137,138)

k frame index (time step) (29)

l left ear, left side (24)

λ index of direction (26)

m_n n-th moment of a PDF (137)

μ expected value (99,138)

N_B number of frequency bands (97)

N_C number of codebook entries (101)

N_K number of bins (categories) in histogram (26)

N_λ number of discrete directions (26)

N_P number of particles (94)

N_V number of mixed voices (67)

n_{ppc} number of particles per coordinate (115)

$\mathcal{N}(\mu, \sigma)$ Gaussian PDF with expected value μ and standard deviation σ 100

$\mathbb{N}$ set of natural numbers (91)

ν circular variance (138)

$p(\alpha_k|\alpha_{k-1})$ PDF of the azimuth dynamics (97)

$P(\vec{\Delta}|\lambda)$ conditional PDF of interaural parameter vector for direction λ (26)

$\tilde{p}(\Delta_b|\lambda)$ marginal PDF of $\tilde{P}(\vec{\Delta}|\lambda)$ for frequency band b (26)

$\tilde{p}(\vec{\Delta}|\lambda)$ approximation of $P(\vec{\Delta}|\lambda)$ from marginal PDFs (26)

$p(\phi_k|\phi_{k-1})$ PDF of the elevation dynamics (97)

$p(\vec{S}_k|\vec{S}_{k-1})$ PDF of spectral transitions (97)

$p(\vec{x}_k|\vec{x}_{k-1})$ system dynamics PDF (91)

$p(\vec{z}_k|\vec{x}_k)$ observable statistics PDF (91)

ϕ elevation (24)

R_F frame rate (103)

$\Re[c]$ real part of c

r right ear, right side (24)

$\mathcal{R}$ state space (90)

ρ_{fg} cross correlation coefficient of series $f(k)$ and $g(k)$ (109)

ϱ vector strength (138)

$S(f)$ spectrum of free-field signal emitted from one sound source (24)

$\vec{S}_{v,k}$ short-term spectrum (97)

$(s_{b,v,k})$ short-term spectral coefficients (97)

s skew (137, 138)

σ standard deviation of spectral coefficients (99)

$\vec{\sigma}_{\mathrm{vq}}$ standard deviation of codebook coefficients (100)

T transition matrix (98)

t time (24)

t_g interaural group delay (24)

t_p interaural phase delay (24)

τ_a time constant of first order low pass filter (30)

θ phase angle (25)

$\mathcal{U}(a, b)$ uniform PDF in the interval $[a, b]$ (98)

v voice index (67)

w_k^i weight of particle number i at time step k (94)

$\vec{\mathbf{x}}_k$ state at time k (random variable) (91)

$\vec{x}_k$ state at time k (realization of random variable) (90)

x^* complex conjugate of x

$x \sim p(x)$ sampling x from the PDF $p(x)$ (93, 94)

$Y(f)$ spectral values of vector of independent free-field signals (67)

$\hat{\vec{z}}_k^i$ hypothetical observation for particle i (99)

$\vec{\mathbf{z}}_k$ observation at time k (random variable) (91)

$\vec{z}_k$ observation at time k (realization of random variable) (91)

Abstract

The topic of this work is "*statistical cocktail party processing*": The exploration of principles of statistical signal processing which are useful for noise reduction and speaker separation at high levels of non-stationary background noise and in multi-talker situations. Furthermore, an improved understanding of the accomplishments of the auditory system in multi-source situations is desired.

The approach chosen here starts from known spatial filtering concepts, which can provide high noise suppression when their assumptions are met, and statistical estimation techniques, which aim to deal with uncertain and incomplete data. So far, single-channel statistical estimation methods were not able to increase intelligibility at low signal-to-noise ratios; adaptive spatial filtering concepts are less applicable in multi-talker situations.

In this thesis, interaural parameters, as they are available in binaural hearing, are combined statistically to achieve spatial filtering. Additionally, important principles used in auditory scene analysis are taken into account, specifically the integration of across-frequency processing of "common fate cues" like common onset and common amplitude modulation. The study links them to the domain of multidimensional statistical estimation by using a probabilistic interpretation of neural processing.

To pursue these approaches, binaural recordings of speech signals and recordings from real-world noise environments are used, and their statistical properties are evaluated. The spectro-temporal dynamics of speech is measured and evaluated in a stochastic state-space framework, which allows the use of information that is unavailable to deterministic approaches, and that helps to reconstruct and estimate masked segments of the involved signals.

Based on this, three algorithms for on-line estimation of sound source direction and separation of voices are developed. In the first, the directional information described by interaural parameters is evaluated by a multidimensional, maximum *a posteriori* approach. The second employs the estimated directions, combined with head related transfer functions, to separate concurrent voices by a steered beamforming method. The third algorithm evaluates the combination

of dynamical, spectro-temporal features of speech and directional information. This is done by a non-linear, non-Gaussian, multidimensional state-space framework. The application of sequential Monte Carlo methods allows an implementation of this framework for high-dimensional spectro-temporal data.

The measurement of the statistics of interaural parameters in noise shows high levels of fluctuation. Based on *a priori* knowledge about these fluctuations, the first algorithm achieves robust estimation of the sound source azimuth and elevation, even at low signal-to-noise ratios. The second algorithm, which performs steered beamforming, has convergence times of about 0.2 s, it causes only few distortions for most of the time, and it yields signal-to-noise ratio improvements of up to 30 dB. The third algorithm, which implements multidimensional tracking using sequential Monte Carlo methods, achieves reliable tracking of the azimuth of one single talker, even if its azimuth changes rapidly. For a mixture of two concurrent voices, this algorithm achieves precise azimuth estimation, as well as tracking and on-line separation of the frequency-specific envelopes, yielding improvements of the signal-to-noise ratio of up to 7.8 dB at an initial signal-to-noise ratio of zero dB. The convergence times are between 50 ms and 200 ms. The algorithm has a high computational complexity; however, strategies for large reductions are pointed out.

As a main conclusion, knowledge of spectro-temporal properties of speech and statistics of interaural parameters supports robust and fast localization, as well as separation of the envelopes. Further, environmental noise should be taken into account for understanding binaural auditory processing of sounds. The steered beamforming algorithm can be executed in real time and is applicable to binaural hearing aids and other small microphone arrays. Sequential Monte Carlo methods can integrate principles of peripheral auditory processing with source-coding approaches, as those used in telephony. Efficient algorithms for low-dimensional state coding need to be developed to reduce the computational complexity, and to make these methods applicable for the reduction of non-stationary noises. Finally, these methods allow us to perform robust statistical feature integration, and they open new possibilities to understand and simulate the grouping of multiple features in the auditory system.

Zusammenfassung

Zielsetzung dieser Arbeit ist die statistische Signalverarbeitung in Cocktail-Party-Situationen: Die Untersuchung von erfolgversprechenden Prinzipien für die Trennung von mehreren Sprechern und die Störgeräuschunterdrückung bei hohen Pegeln von nichtstationären Störgeräuschen. Weiterhin wird ein verbessertes Verständnis der Leistungen des auditorischen Systems in Situationen mit mehreren Schallquellen angestrebt.

Die Untersuchung geht einerseits von bekannten Verfahren zur Richtungsfilterung aus. Diese können Störgeräusche wirkungsvoll unterdrücken, sofern die Annahmen, auf denen sie basieren, erfüllt sind. Andererseits werden statistische Schätzmethoden hinzugezogen, welche die Zielsetzung haben, verrauschte und unvollständige Eingangsdaten zu berücksichtigen. Einkanalige, statistisch basierte Störgeräuschunterdrückungsmethoden waren bisher nicht in der Lage, die Verständlichkeit bei niedrigen Signal-Rausch-Abständen zu erhöhen; Verfahren basierend auf adaptiver Richtungsfilterung verlieren dagegen einen Teil ihrer Wirksamkeit in Situationen mit mehreren Quellen.

In dieser Arbeit werden interaurale Parameter, wie sie beim binauralen Hören verfügbar sind, statistisch ausgewertet, um eine räumliche Filterung zu erreichen. Außerdem werden wesentliche Prinzipen der Auditorischen Szenenanalyse berücksichtigt, insbesondere die Einbeziehung von frequenzübergreifender Information über "Merkmale gemeinsamen Schicksals", wie zeitgleiches Einsetzen von Signalen, oder zeitsynchrone Amplitudenmodulationen. Die Untersuchung verbindet beide Ansätze mit Verfahren der multidimensionalen statistischen Filterung, indem eine statistische Interpretation neuronaler Verarbeitung zugrunde gelegt wird.

Um die genannten Ansätze zu verfolgen, werden binaurale Aufnahmen von Sprachsignalen und realen Störgeräuschumgebungen verwendet, und ihre statistischen Eigenschaften werden ausgewertet. Die spektro-temporale Dynamik von Sprachsignalen wird erfaßt und auf der theoretischen Grundlage eines stochastischen Zustandsraum-Modells untersucht. Dieses Vorgehen ermöglicht es, Informationen zu nutzen, die für deterministische Ansätze nicht verfügbar sind,

und hilft, maskierte Zeitabschnitte der beteiligten Signale zu schätzen und zu rekonstruieren.

Hierauf basierend werden drei on-line Algorithmen zur Schätzung der Richtung von Schallquellen, und zu ihrer Trennung entwickelt. Beim ersten wird die Richtungsinformation, welche durch interaurale Parameter erfaßt wurde, mit einem Bayes'schen maximum *a posteriori* Verfahren ausgewertet. Der zweite Algorithmus verwendet die geschätzten Richtungen unter Hinzuziehung von Außenohrübertragungsfunktionen, um überlagerte Stimmen mittels eines gesteuerten Richtungsfilters zu trennen. Der dritte Algorithmus untersucht die Kombination von spektrotemporalen, dynamischen Merkmalen von Sprache mit Richtungsinformation. Dies geschieht mittels eines nichtlinearen, nicht-Gaußschen, mehrdimensionalen Zustandsraum-Ansatzes. Die Anwendung von Sequentiellen Monte Carlo Methoden erlaubt es, diesen Ansatz für hochdimensionale spektro-temporale Daten umzusetzen.

Die Messungen der Statistik interauraler Parameter ergeben ein hohes Niveau an Fluktuationen. Basierend auf *a priori* Information über diese Fluktuationen, erreicht der erste Algorithmus eine robuste Schätzung des Azimuts und der Elevation der Schallquellen selbst bei niedrigen Signal-Rausch-Abständen. Der zweite Algorithmus, welcher den gesteuerten Richtungsfilter implementiert, weist Konvergenzzeiten von etwa 0.2 s auf, verursacht für die Mehrheit der Zeit nur geringe Verzerrungen, und erreicht Verbesserungen des Signal-Rausch-Abstands von bis zu 30 dB. Der dritte Algorithmus, welcher multidimensionales Verfolgen der Schallquellen mit Sequentiellen Monte Carlo Methoden verwirklicht, erbringt eine zuverlässige Verfolgung des Azimuts einer einzelnen Quelle, selbst wenn dieser sich zeitlich schnell verändert. Für eine Mischung von zwei überlagerten Stimmen ergibt der Algorithmus eine präzise Richtungsschätzung sowie eine Verfolgung und on-line Trennung der frequenzspezifischen Einhüllenden. Verbesserungen des Signal-Rausch-Abstands bis zu 7.8 dB werden bei einem ursprünglichem Signal-Rausch-Abstand von null dB erreicht. Die Konvergenzzeiten liegen zwischen 50 ms und 200 ms. Der dritte Algorithmus erfordert einen hohen Rechenaufwand; Strategien für erhebliche Verringerungen desselben werden aufgezeigt.

Eine wesentliche Schlußfolgerung ist, daß Vorwissen über spektro-temporale Eigenschaften von Schallen und über die Statistik interauraler Parameter eine schnelle und robuste Lokalisation von Schallquellen, und eine Trennung der Einhüllenden ermöglicht. Weiterhin sollten Umgebungsgeräusche berücksichtigt werden, um die binaurale auditorische Verarbeitung zu verstehen. Der Algorithmus zur gesteuerten Richtungsfilterung ist in Echtzeit ausführbar und anwendbar in binauralen Hörgeräten und anderen kleinen Mikrophonarrays. Sequentielle Monte Carlo Methoden können Prinzipien der peripheren auditorischen Verarbeitung integrieren mit Ansätzen zur Quellenkodierung, wie sie in der Telefonie verwendet werden. Effiziente Verfahren zur niedrigdimensionalen Darstellung des Quellenzustands müssen entwickelt werden, um den Rechenaufwand zu verringern und diese Methoden für die Störgeräuschreduktion anwendbar zu machen. Schließlich erlaubt dieser Ansatz es, eine robuste statistische Merkmalsintegration durchzuführen, und eröffnet so neue Möglichkeiten, die Gruppierung von mehreren Merkmalen im auditorischen System zu simulieren und zu verstehen.

1. General Introduction

> How do we recognize what one person is saying when others are speaking at the same time ("the cocktail party problem")? On what logical basis could one design a machine ("filter") for carrying out such an operation? *Colin E. Cherry (1953)*

1.1. Tackling the Cocktail Party Problem

Many people with hearing impairment have great difficulties in understanding speech in situations with several interfering talkers, other background noise, and reverberation, like a party, a family celebration, or a conference. They often report that this is the most important limitation they experience. The importance of managing such hearing situations originates from the fact that communication in groups is decisive for fully taking part in a community. Normally, the human auditory system has a remarkable ability to distinguish and separate the different voices in such situations. This is based on redundancies of the signals generated by the acoustic sources, and extraction and combination of several features pertaining to each source. Auditory research has coined this accomplishment as the "Cocktail Party Effect" (Cherry, 1959; Cherry and Wiley, 1967; Culling and Summerfield, 1995; Yost, 1997; Bronkhorst, 2000; Hawley *et al.*, 2004). Persons with cochlear hearing impairment, however, gather far less information from their auditory periphery, and frequently do not benefit much from such a capability of the higher auditory system (Bronkhorst and Plomp, 1989; Peissig, 1992; Hawley *et al.*, 1998; Beutelmann and Brand, 2005). For this reason, hearing impairment often affects participation in social life.

Research continues to examine in which way hearing aids can support listening in noise and reverberation. Based on the insight that the cocktail party effect is rather an accomplishment of the brain, than of the ears (see Yost, 1991, for example), noise reduction algorithms could use the signal processing capacity of modern digital hearing aids to eliminate the unwanted sounds and extract the desired ones (Levitt, 2001).

To contribute to the goal that hearing aids improve understanding in difficult listening situations, noise reduction methods for high levels of nonstationary noise require further development. Successful general solutions for this task are still not known. The investigation of innovative signal processing strategies which are designed to work in multi-talker situations and in presence of nonstationary noise sources is the main objective of this thesis.

1.2. Overview Of Existing Methods

So far, a number of different strategies have been reported in the literature. According to their underlying assumptions, most algorithms proposed can be classified into two main categories, i.e. spatial filtering and envelope estimation. The properties of these algorithms are closely related to their respective assumptions.

1.2.1. Approaches Based on Spatial Filtering

Spatial filtering algorithms aim to enhance the recorded signal based on the spatial arrangement of different sound sources. They can be further divided into four sub-categories, which have common characteristics, namely directional microphones, fixed microphone arrays, adaptive beamformers, and approaches of blind source separation (BSS).

Methods using *directional microphones* combine two or more microphones to enhance sounds from a certain, fixed direction (Soede *et al.*, 1993; Kompis and Dillier, 1994). Pairs of directional microphones can provide a small but robust increase in speech intelligibility. In favorable conditions, they are able to increase the speech reception threshold (SRT), the signal-to-noise ratio at which 50% of the words are intelligible by 3 to 5 dB (Marzinzik, 2000; Greenberg and Zurek, 2001). Today, most advanced hearing aids include them.

Microphone arrays, *adaptive beamformers*, and approaches of *blind source separation* all use several microphones and rely typically upon the fact that the sound sources are superimposed linearly. In practice, microphone arrays are so far the most successful concept (Griffiths and Jim, 1982; Schwander and Levitt, 1987; Soede, 1990; Van Compernolle *et al.*, 1990; Soede *et al.*, 1993; Compernolle, 2001; see Greenberg and Zurek, 2001, for a review). *Non-adaptive microphone arrays* enhance sounds coming from fixed (usually frontal) directions. Such arrays can achieve suppressions of interfering sources of up to 11 dB (Zurek *et al.*, 1996; Greenberg and Zurek, 2001), however their low-frequency directivity is limited by the dimensions of the microphone array. Large microphone arrays, as developed by Widrow (2000), provide a considerable suppression of lateral noise sources. Such arrays have been reported to be useful for persons with severe hearing loss, even though manufacturers as well as the majority of potential users do not consider them attractive, presumably for cosmetical and practical reasons.

Algorithms for *adaptive beamformers* and *BSS* assume that filtering based on a linear combination of the input signals, which has only slowly varying coefficients, separates the individual sound sources (Bell and Sejnowski, 1995; Hoffman and Buckley, 1995; Parra and Spence, 2000; van der Kouwe *et al.*, 2001; Liu *et al.*, 2001; Anemüller, 2001; Anemüller and Kollmeier, 2003; Greenberg *et al.*, 2003). Adaptive beamformers typically aim at minimizing some error signals; BSS approaches attempt to find these coefficients based on the assumption that the different sources are in some way independent from each other; the precise definition of independence differs from algorithm to algorithm. Typically, this assumption is used when minimizing a cost function, which depends on the linear coefficients, and is derived from higher order statistics.

In principle, they can provide a high noise source suppression; cancellation or degradation of the desired signal may be a problem, especially at high signal-to-noise ratios (SNRs). When more sound sources than sensors are present, however, a linear operation cannot separate the sources completely, because the problem becomes algebraically ill-posed. Furthermore, some of these approaches are likely to fail when a high amount of reverberation is present. In case of strong reverberation or a large number of interferers, adaptive beamformers generally do not exceed the performance of fixed microphone arrays (Greenberg and Zurek, 2001; Levitt, 2001). This is far from matching the performance of the auditory system, which is capable of improving intelligibility in the presence of several concurrent directional interferers (Hawley *et al.*, 2004). Moreover, the adaptation times of such algorithms are at least in the order of one second. Depending on the requirements on convergence and stability, even several seconds may be necessary. This makes their application difficult in rapidly changing spatial configurations, which are typical for multi-talker or outdoor environments (Edwards, 2004).

Adaptive beamformers controlled by location-estimating algorithms, as evaluated by Wittkop (2001), Liu *et al.* (2001), and Greenberg *et al.* (2003) in SRT measurements, or described by Roman *et al.* (2003), are attractive alternatives, because they may reach faster convergence and flexibility. Recent approaches employed hidden Markov Models (HMMs), which evaluate the statistics of the signals to control the beamforming operation, sometimes combined with approaches of independent component analysis (ICA) (Acero *et al.*, 2000; Reyes-Gomez *et al.*, 2003).

1.2.2. Spectral Estimation Approaches

The second large class of algorithms are *spectral estimation approaches*. Well-known approaches belonging to this category are the spectral subtraction algorithms by Lim (1978) and Boll (1979), Wiener filtering (Kailath, 1981), or the algorithms developed by Ephraim and Malah (1984, 1985). Rather than combining several channels, they aim to estimate the undisturbed speech signal by taking into account observations of the noise signal, which are captured during speech pauses. Typically, they are accompanied by a voice activity detection (VAD) algorithm. In general, spectral estimation approaches assume short-term stationarity of the noise spectrum. For example, the algorithm by Boll (1979) estimates the desired signal by subtracting the average magnitude spectrum of the background noise. For this class of algorithms, as well as for the VAD stage, numerous enhancement schemes strive to account for non-stationarity of the noise, for example by Xie and van Compernolle (1996); Cohen and Berdugo (2001). Under suitable conditions, they enhance speech quality and listening comfort without adversely affecting speech intelligibility (Marzinzik, 2000). In environments with high levels of nonstationary noise however, the assumption of stationarity leads to strong artifacts ('musical tones' or 'gurgling') which degrade speech quality and intelligibility (Cappé, 1994). So far, single-channel signal processing algorithms for noise reduction have not shown any measurable advantage in intelligibility, even though listening comfort is increased (Marzinzik, 2000; Levitt, 2001; Edwards, 2004).

1.2.3. Further Methods

Summarizing the results of both algorithm categories, spectral estimation approaches do not have a beneficial effect in adverse situations with nonstationary noise and multiple noise sources because they cannot recover the original sound signals from the disturbed input in a deterministic, unambiguous way. The root of the problem is that a stationary noise model is too weak. Because of their design and underlying assumptions, the classical algorithms are only applicable to suppress stationary noises.

Adaptive spatial filtering algorithms can separate a number of noise sources, which is limited to the number of microphones, with slow varying directions. For both spatial filtering and spectral estimation approaches, time-varying additional noise sources, or interfering speech, will mask or hide the desired signal. Further approaches have been the separation or enhancement of sources based on their fundamental frequency, as proposed by Lim *et al.* (1978), Liu *et al.* (1997) and Summerfield and Stubbs (1990), or based on filtering in the modulation frequency domain (Kollmeier and Koch, 1994; Strube and Wilmers, 1999). For a discussion of additional approaches, the reader is referred to Edwards (2004).

Because of the masking of signal segments, it is sometimes necessary to *reconstruct* the signal according to known properties of the sound sources which generate it, such as in the restoration of musical recordings (Godsill and Rayner, 1998; Levitt, 2001). For this reconstruction, pattern-matching techniques, often adapted from methods of automatic speech recognition (Boll, 1992), methods based on disjoint orthogonality or binary masking (Roweis, 2000; Jourjine *et al.*, 2000), and methods based on hidden Markov models (Ephraim *et al.*, 1989b,a; Ephraim, 1992; Gales and Young, 1993; Sameti *et al.*, 1998; Sameti and Deng, 2002) and Kalman filtering (Lee and Jung, 2000) have been proposed. In spite of all efforts and improvements in certain domains, these approaches could not prove to increase speech intelligibility at low SNRs.

The practical relevance of speech enhancement for mixtures with nonstationary noises is not limited to hearing aids. First successful applications of spectral estimation methods emerged in the domain of automatic speech recognition, which is adversely affected by much lower levels of noise than listeners. Environments with multiple talkers, reverberation, and high levels of nonstationary noise sources represent also a difficult challenge for communication technologies like mobile telephony in cars, or automatic speech recognition in offices, meetings or other common real-world environments. While speech processing algorithms provide only small improvements of the recognition rate of disturbed speech, the auditory processing of normal-hearing persons handles such situations almost effortlessly; Normal-hearing users expect speech recognition algorithms to work in such everyday environments, rather than wanting to adapt their environment to them. For such applications, microphone arrays are possible solutions, but are still not widely applied (Compernolle, 2001); for reasons of cost effectiveness and ease of operation, development engineers desire to keep the size and number of microphones as small as possible.

1.3. Scope of This Work

Reducing nonstationary noises in complex environments and in presence of multiple talkers is a challenging task. To define the scope of this work, the examined situations were narrowed to up to three concurrent talkers, or one talker and additive noise from a spatial noise field. However, the algorithms investigated do not suffer theoretical limitations to work in more complex environments. The processed signals are always binaural recordings or synthesized equivalents, which carry about the same spatial information as is present at the ear canal entrance of human listeners. Further, the algorithms investigated are on-line algorithms, which process the input signals with buffer lengths of not more than 25 ms, and typically require convergence times of not more than about 50 to 200 ms. A focus has been placed on principles which may enable algorithms to cope successfully with nonstationary noises. Though hearing aids have limited computing capacity, no restriction was put on the computational complexity of the algorithms in the second part of the work, because the goal of an operative noise reduction for hearing aids will not be reached within a short period of time.

1.4. Pursued Strategies

The strategies chosen here to pursue the goal of noise reduction in multi-talker situations and nonstationary noise environments rest on three main ideas. The first is to apply principles of auditory processing known from physiology and psychoacoustics, especially binaural hearing and mechanisms of auditory scene analysis. Second, this thesis uses a probabilistic approach to emulate effective neural signal processing. The third idea is to combine sound source direction and statistics of spectro-temporal features using robust multidimensional statistical methods, and hereby taking advantage of recent developments in this field.

1.4.1. Usage of Principles of Auditory Processing

The first main idea has the rationale to mimic a selection of helpful properties of evolution's solution to the task, which up to now outperforms every technical attempt by far. Among the used principles are analysis of the signal in independent frequency channels, extraction of features which give information on sound source direction and temporal development of the sources, and tracking

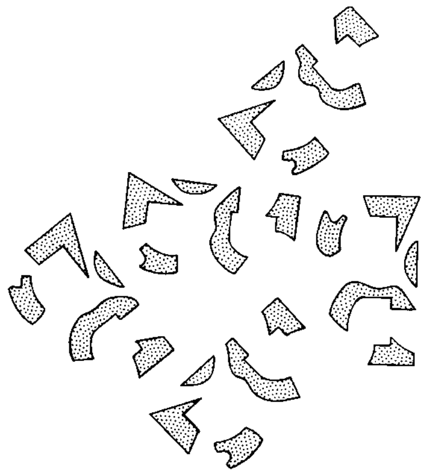


Figure 1.1.: Bregman's illustration of the task of scene analysis (I). Even if the observer knows that the depicted objects are letters, it is hard to recognize them because they are partially missing (Bregman, 1990, Fig. 1.13).

and identifying the sound-generating objects by their acoustical representation.

Principles of Auditory Scene Analysis

The process by which humans distinguish several sound sources is known as '*auditory scene analysis*', and its main principles have been described by Bregman (1993), McAdams (1993), and others (see Moore, 1989a, and Cooke and Ellis, 2001 for an overview). These and further investigations have shown up a number of important cues which humans use to separate sources (Bregman, 1990; Summerfield and Stubbs, 1990; Yost, 1991; Marin and McAdams, 1991; Carlyon and McAdams, 1992; McAdams and Bigand, 1993; Mellinger and Mont-Reynaud, 1996). A main result is that common onset and the dynamics of the short-term spectral envelope are among the most important cues; although a number of 'Gestalt' rules have been formulated, a fixed hierarchy of rules has not been found.

Bregman gave an example how knowledge about properties of objects and interferences can help to recognize them in the presence of clutter. Figure 1.1 shows the shape of several letters, which are partly missing. Even with the knowledge that letters are depicted, it is hard to recognize them. On the other hand, when a plausible shape, like an ink blot, hides their contours, the task of identifying them as "Bs" becomes easy, as Fig. 1.2 confirms. Bregman investigated analogous mechanisms and rules which allow to distinguish several sounds by their acoustical features.

Figure 1.3 shows an illustration of the "old plus new" principle, corresponding to the Gestalt law of "good continuation," which plays a role in the auditory continuity illusion or pulsation threshold (Bregman, 1990; Cooke and Ellis, 2001). The Figure shows a time-frequency representation of two sound sources. The gray shape indicates spectral energy from an ongoing sound source. The black shapes indicate time-frequency regions of suddenly increased spectral power density. Here, the auditory system will interpret the information in the way that some ongoing sound is temporally masked by a second sound source with sudden onsets. This guess results in the percept that the first source continues unchanged. The spectral subtraction approach developed by Boll (1979) mentioned before is a very simple application of this strategy; it takes the averaged spectrum of the instants of the first sound where no speech seems to be present, and subsequently subtracts it from the signal mixture to estimate the spectrum of the speech. Unfortunately, real-world signals are frequently too complicated for separation or reconstruction by a simple application of fixed rules, for example because both sounds involved will exhibit some amount of fluctuation.

Use of Binaural Information

An information source of special interest when examining the cocktail party problem is *binaural information*. Cherry and Wiley (1967) have examined the cues used by the auditory system in such situations, and found that binaural hearing can have a decisive importance. In difficult listening conditions and at a low SNR, binaural processing of directional information provides a significant increase of speech intelligibility (Durlach and Colburn, 1978; Colburn *et al.*, 1987; Peissig, 1992; Yost, 1997; Kidd *et al.*, 1998; Bronkhorst, 2000). Hearing loss partially impairs this binaural auditory processing (Tonning, 1973; Häusler *et al.*, 1983; Bronkhorst and Plomp, 1989; Colburn and Hawley, 1996). These results stimulated attempts to replace the impaired binaural processing by hearing aid algorithms (Allen *et al.*, 1977; Durlach and Pang, 1986; Bodden, 1992; Peissig, 1992; Bodden, 1993; Kollmeier and Koch, 1994; Bodden, 1996a; Liu *et al.*, 1997; Wittkop, 2001; Hohmann *et al.*, 2002b; Wittkop and Hohmann, 2003).

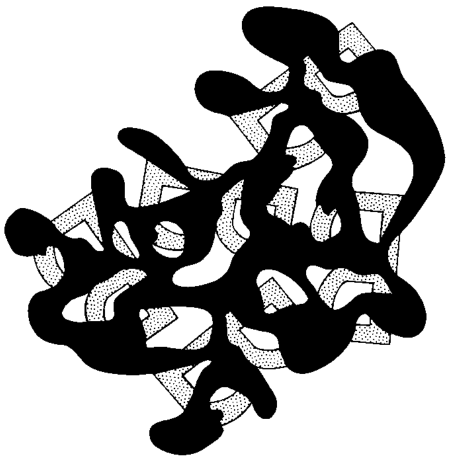


Figure 1.2.: "Bregman's Bs" as an illustration of the task of scene analysis (II). When the lacking visibility of part of the letters is explained by some other shape, it becomes easy to recognize them.

Motivated by Cherry's insight, taking into account the advantages of directional methods, and continuing the efforts made so far, this work uses binaural directional information as a main cue, extending the earlier work of Peissig (1992), Kollmeier and Koch (1994) and Albani *et al.* (1996). Their approach to sound localization was motivated by the physiological findings in the barn owl (Knudsen and Konishi, 1978; Brainard *et al.*, 1992; Wagner, 1991). In this work, sounds from recordings made with binaural hearing aids are evaluated. The incoming binaural signals are transformed with a short-term frequency analysis and are filtered with a bandwidth comparable to the bandwidths of auditory filters; thus, similar information is available as for the human auditory system. After this, phase and level differences are calculated, as established by Wittkop (2001). Then, the algorithm evaluates directional features of interaural phase and level differences.

Because earlier approaches to perform envelope estimation directly by these short-term binaural parameters reached improvements of the SRT in realistic environments (Wittkop, 2001; Wittkop and Hohmann, 2003), but partially suffered from the high amount of fluctuation of the interaural parameters, this study examines the statistical properties of these features systematically. Consequently, an important strategy chosen in this work was to take these fluctuations explic-

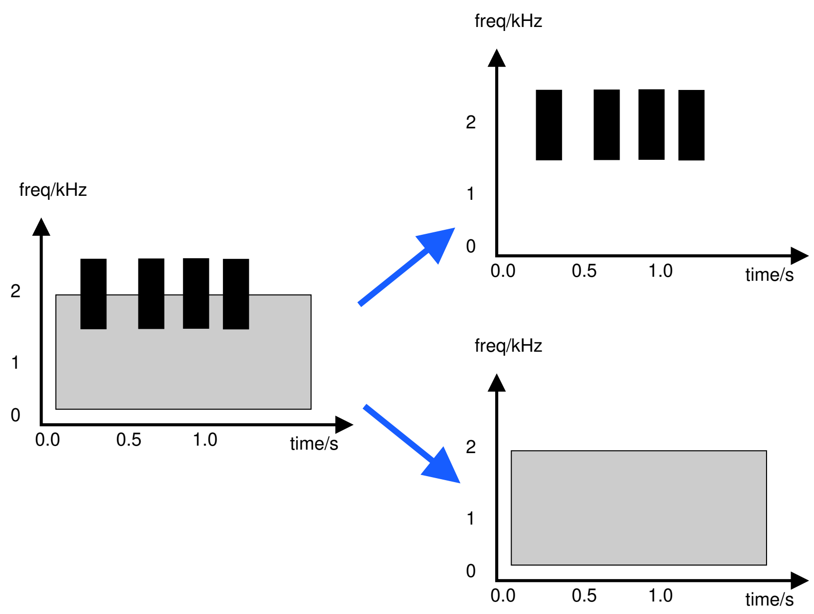


Figure 1.3.: Illustration of the "old plus new" principle (adapted from Cooke and Ellis, 2001). When a stimulus like in the left part of the figure is presented, the auditory system will make the interpretation that it is composed from two sounds, an ongoing, continuous sound (right side, bottom), and a sound with several sudden onsets (right side, top).

itly into account. This is performed by applying a multidimensional Bayesian method, which uses the statistical information as *a priori* information, to estimate the sound source directions simultaneously. Finally, the estimated directions serve to identify and separate the signals, by filtering them according to the known head-related transfer functions.

1.4.2. The Probabilistic Approach: Feature Integration and Competing Hypotheses

One result of the investigations on auditory scene analysis is that the auditory system usually integrates several cues to perform this task. In physiology, basic substrates of auditory feature integration have been discovered, for example, by Brainard *et al.* (1992) and Scott and Johnsrude (2003). Knudsen and Konishi (1978) and Brainard *et al.* demonstrated that barn owls (*tyto alba*) keep a neural map of auditory space, which is excited by interaural level and time differences. When narrow-band stimuli are presented, relatively large regions of this map become excited; when the stimuli have a high bandwidth, these excitations are combined in a manner that only regions excited by all frequency components are active in this case. Peña and Konishi (2001) showed that this integration can be explained by a model performing multiplication of neural excitations, similar to a logical AND relation, which is computed by 'fuzzy logic'. This conjunction is equivalent to a *multiplication of probabilities*.

Bushara *et al.* (2003) used functional magnetic resonance imaging (fMRI) experiments to examine the neural binding of modalities. They found that the neural binding of auditory and visual stimuli is correlated with reduced brain activity in some of the corresponding areas. This agrees with the concept that the brain explores competing hypotheses, and that their representations have mutually inhibitory interactions. This principle of "competing hypotheses" agrees with a Bayesian description of neural processing. The second main idea in this work is to use this probabilistic point of view to link the problem of scene analysis with methods of statistical signal processing. Statistical principles are also applied to describe dynamical properties of sources and signals. Cherry was probable the first who had the idea that the brain uses transition probabilities and *a priori* knowledge about speech to solve the cocktail party problem (Cherry, 1953, p. 976). Cherry's hypothesis about involvement of spectral dynamics and spatial information agrees with neurophysiological data reviewed by Scott and Johnsrude (2003), which indicate that spatial information is processed in the posterior parietal cortex in a 'where' processing stream, while the posterior auditory cortex (PAC) might subserve perception of the spectral dynamics of sound in a 'how' stream of processing (Belin and Zatorre, 2000).

Noise reduction algorithms could benefit from adapting the strategy of probabilistic feature integration. The need for feature integration arises from the fact that information about sound source direction is not sufficient to separate sound sources in situations containing more than one concurrent interfering talker. In practice, this problem is also an important drawback of beamformer algorithms and methods of blind source separation as mentioned before. To simulate the feature integration of the auditory system, the following questions have to be solved: First, how is it possible to perform an integration of these cues by computational methods? Second, how could they be combined in a way which is robust to the manifold and cluttered acoustical environments outside the laboratory? And third, which additional features used by auditory scene analysis can be utilized?

1.4.3. Robust Feature Integration by Multidimensional State-Space Methods

The third main idea, which distinguishes this work from the strategies pursued so far, is to emulate this combination of several perceptual relevant features by using multidimensional stochastic state-space methods, which not only can simulate partly the integration strategies of the auditory system, but also take into account that, in real-world situations, the incoming data carry a high degree of uncertainty and noise which would defeat purely rule-based solutions. So far, a number of different approaches has aimed at solving the task of feature integration (Strube, 1981; von der Malsburg and Schneider, 1986; Schneider, 1986; Cooke, 1993; Ellis, 1996; Unoki and Akagi, 1999; Acero *et al.*, 2000; Harding and Meyer, 2001; Reyes-Gomez *et al.*, 2003, 2004). With the goal to perform a *stochastic feature integration*, this study has chosen a stochastic state-space formalism as its theoretical framework. The use of *multidimensional, nonlinear, non-Gaussian state space models* for statistical filtering is what distinguishes the approach pursued here from earlier work.

State-space representations have three advantageous properties, which are exploited in this thesis. Figure 1.4 shows the first. The squares depict noisy observations of the state of an object, for example the azimuth of a sound source. Given that the manner of movement of the object is partially known – for example, we may know that it performs oscillatory movements – the combination of observations and knowledge allow to estimate the movement of the object in spite of the uncertainty of the observations - the object is *tracked*.

The second advantage is that stochastic tracking of objects can use knowledge about subspaces to which the states of the objects are restricted. Figure 1.5 shows symbolically the trajectory of a signal in a high-dimensional state space, which is restricted to a manifold of probable signals that has a lower dimension; the restriction may be due to limitations of energy or physiological feasibility, for example. Actually, signals occurring in nature belong to a small subset of the space of all possible signals, a property which is exploited not only for speech coding mobile telephony and model-based speech enhancement, but also by sensory coding in the visual and auditory system (Sameti and Deng, 2002; Lewicki, 2002; Olshausen and Field, 2004).

The third property that is helpful for separating voices is that the tracking

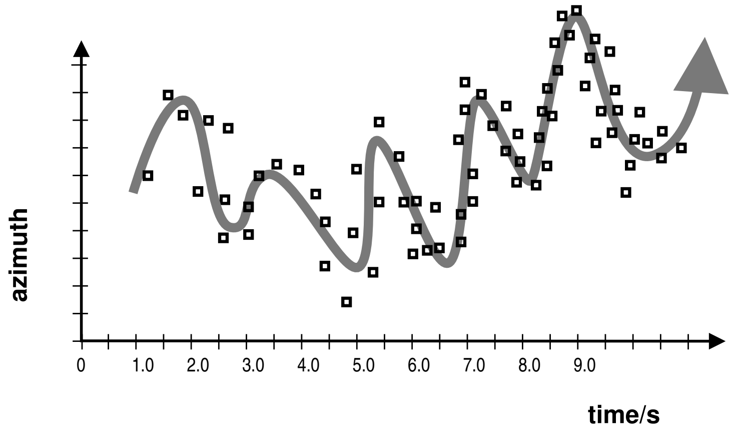


Figure 1.4.: Illustration of stochastic tracking. The squares indicate uncertain observations of the azimuth of an object as a function of time. Given some knowledge on the movement of the object, its true trajectory (solid line) can be estimated.

of objects in multiple dimensions allows to distinguish them even if their state coordinates are identical in a part of the dimensions, as Fig. 1.6 shows. Here, the time-dependent values of the coordinates of two objects moving in a two-dimensional state-space are drawn as two trajectories. The coordinates might be, for example, the azimuths and fundamental frequencies of two sound sources. The observation of each feature alone would not be sufficient to distinguish and track the two objects, because in some moments, they possess equal state coordinates. However, if their trajectory is followed in both dimensions, the objects can be distinguished most of the time.

To select a suitable computational strategy, this investigation compared properties of statistical methods which can perform an on-line estimation of the state of a dynamical system, based on noise-perturbed observations. Main requirements were the ability to process multidimensional data, efficiency, the capability to process signals with non-Gaussian statistics and a quasi-continuous state, and the possibility to account for nonlinear relationships between observed signals and the representation of the original sound. Traditional multidimensional statistical estimation algorithms, like the Kalman filter and its variants (Kalman, 1960; Kailath, 1981; Gelb, 1994), or hidden Markov models (HMMs), do not meet

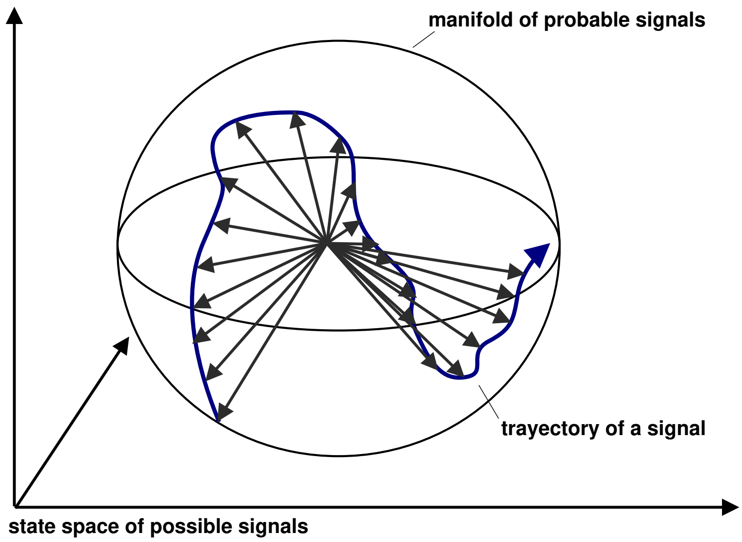


Figure 1.5.: Illustration of tracking signals in probable subspaces (adapted from Olshausen and Field, 2004). The Euclidian coordinate system denotes the high-dimensional space of all possible signals. The sphere, as a two-dimensional manifold, represents the space of probable signals, which has a lower dimensionality. Knowledge about the space of possible signals makes it easier to track a signal disturbed by noise.

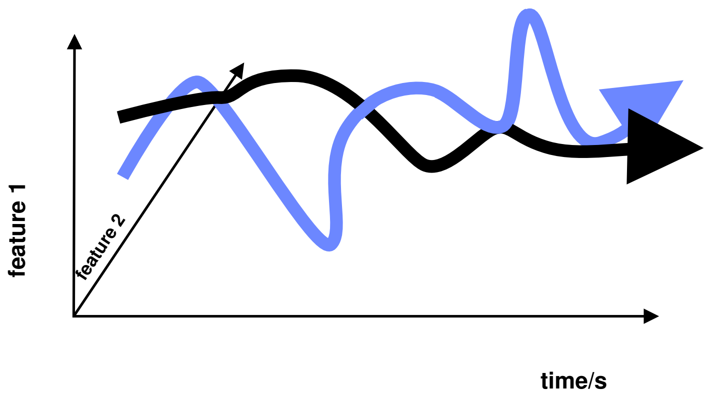


Figure 1.6.: Tracking of two objects in multiple dimensions. The figure illustrates the temporal movement of two objects in two dimensions. Because their coordinates do not become equal in both dimensions at the same time, their positions in the two-dimensional space can always be distinguished.

these requirements. However, a family of recently developed Monte Carlo methods matches them (Gordon *et al.*, 1993; Kitagawa, 1996; MacKay, 1999; Doucet *et al.*, 2001; Arulampalam *et al.*, 2002). Methods from this family, named "*sequential Monte Carlo methods*" (SMC methods), were developed and applied in recent years in scene analysis problems such as computer vision (Isard, 1998; Blake and Isard, 1998; Doucet *et al.*, 2001). In a few cases, they were applied successfully to acoustical tasks, like sound localization (Ward *et al.*, 2003) or formant tracking (Zheng and Hasegawa-Johnson, 2004). One basic principle of these methods is that they approximate the multidimensional probability density function (PDF) of the system state by a weighted set of discrete samples; this allows an efficient representation of non-Gaussian PDFs. This principle is illustrated in Figs. 1.7 and 1.8, which show the track of a goose in sand. In Fig.1.8, the two-dimensional gray values of Fig. 1.7 have been replaced by discrete black dots with different density. Note that for the representation of light areas (in our analogy, corresponding to regions of the PDF with low probability), very few dots are needed. The application of SMC methods to auditory scene analysis and to the separation of nonstationary sources poses considerable challenges, and is examined

Figure 1.7.: Representation of a multidimensional function by discrete samples I. The gray-scaled photo shows goose tracks at a river bank. The picture can be interpreted as a two-dimensional function of color density values. Reproduction with permission from Steven Sauter.

Figure 1.8.: Representation of a multidimensional function by discrete samples II. Here, black dots in variable densities represent the two-dimensional function of gray values.

here for the first time.

The application of SMC methods allows it to use spectro-temporal features and correlations of spectral envelopes to gain missing information on the sound sources. A detailed statistical description of spectro-temporal features of speech, based on statistical evaluation of a speech database, provided the additional information.

A notorious difficulty of SMC methods is the large computation time which they require in simplistic implementations, which, combined with their undeniable power, has led them to be called the "atomic bomb" of statistical estimation. Today, for the field of visual scene analysis, impressive real-time implementations exist (see Blake and Isard, 1998, and Isard, 1998 for an online resource). Therefore, the closing of this examination discusses strategies that can reduce the complexity in their application to auditory scene analysis and noise reduction, and compares them to previous work on speech enhancement based on statistical methods (Sheikhzadeh *et al.*, 1995; Boll, 1992; Vermaak *et al.*, 2002; Fong *et al.*, 2002).

1.5. Structure of the Thesis

The development of the mentioned key strategies throughout this thesis is divided into three major chapters, which are organized as research articles.

Chapter 2, presents an implementation and evaluation of a robust algorithm for binaural sound localization. Extending the concept of Peissig (1992) and Albani *et al.* (1996), the statistical properties of interaural parameters in several noise environments are examined in depth. Based on this *a priori* knowledge, a Bayesian approach evaluates and integrates the information about directions that may be present in a given binaural signal.

Chapter 3 evaluates the performance of the resulting algorithm more thoroughly for a number of realistic noise conditions. Six different types of real-life environments were recorded and used to test the binaural statistical algorithm. The chapter also examines how the algorithm performs with a mixture of several concurrent talkers as input. It continues with the demonstration that this information is sufficient for a rapid localization of three talkers and separation of two concurrent voices. For such mixtures, the algorithm provides good separation with SNR enhancements of up to 30 dB and an adaptation time of about 0.2 s. The chapter closes with a discussion of the strengths and limits of this approach.

Chapter 4 describes a basic implementation of a stochastic filtering algorithm based on SMC methods, delineating some solutions to the methodological and practical problems which emerged. The algorithm performs an on-line localization of sound sources and separation of the spectral envelopes of concurrent voices. First results show that the approach is in fact able to integrate several acoustical features, to localize sources, to track their short-term envelopes with delays of about 50 ms, and to provide an enhancement of the SNR. A specially interesting advantage of the method is that it can combine complementary strategies of representing speech, on one side strategies based on auditory principles, and source-modeling techniques on the other.

Chapter 5 summarizes and discusses the significance of the results, the integration of the investigated strategies, and their application. While many building blocks for improvements of the statistical approach on computational auditory scene analysis are well known, it becomes clear that integrating them into a practical algorithm is a long-term task which needs to combine joint efforts from many specialized fields, such as speech coding, acoustical source modeling, auditory models, and psychoacoustical research. The sequential Monte Carlo method family investigated here may well serve to join formerly rather disconnected fields of research.

Appendix A describes a technical solution for testing current hearing-aid algorithms in real-time implementations in an easy and flexible way. Although by this time, the complex on-line algorithm for envelope separation presented in the third chapter can not run in real-time, both algorithms developed here use this framework for their implementation.

2. Sound Source Localization in Real Sound Fields Based on Empirical Statistics of Interaural Parameters[1][2]

Abstract

The role of temporal fluctuations and systematic variations of interaural parameters in localization of sound sources in spatially distributed, nonstationary noise conditions was investigated. For this, Bayesian estimation was applied to interaural parameters calculated with physiologically plausible time and frequency resolution. Probability density functions (PDFs) of the interaural level differences (ILDs) and phase differences (IPDs) were estimated by measuring histograms for a directional sound source perturbed by several types of interfering noise at signal-to-noise ratios (SNRs) between -5 and +30 dB. A moment analysis of the PDFs reveals that the expected values shift and the standard deviations increase considerably with decreasing SNR, and that the PDFs have non-Gaussian shape at medium SNRs. A d' analysis of the PDFs indicates that elevation discrimination is possible even at low SNRs in the median plane by integrating information across frequency. Absolute sound localization was simulated by a Bayesian maximum a posteriori *(MAP) procedure. The simulation is based on frequency integration of broadly tuned 'detectors', which is consistent with recent physiological findings. Confusion patterns of real and estimated sound source directions are similar to those of human listeners. The results indicate that robust processing strategies are needed to exploit interaural parameters successfully in noise conditions due to their strong temporal fluctuations.*

[1]This chapter was submitted in modified and shortened form to the *Journal of the Acoustical Society of America*.

[2]Some earlier results on Bayesian methods for sound source localization have been presented in Nix and Hohmann (1999, 2000) and Hohmann *et al.* (1999).

2.1. Introduction

The filtering of acoustical signals by the human body, head, and pinna is characterized by the head-related transfer functions (HRTFs) and depends on both direction and distance of the sound source (Mehrgardt and Mellert, 1977; Blauert, 1983; Shaw, 1997). A set of HRTFs for the left and right ears characterizes the physical differences between signals recorded at the ear canal entrances. These differences are generally quantified by the frequency-dependent interaural parameters, i.e., the interaural level differences (ILDs) and the interaural time differences (ITDs). Interaural parameters measured from different directions exhibit a rich variety of possible features for sound localization (Wightman and Kistler, 1989b) and have therefore been considered in many physiological and psychoacoustical models of binaural processing. Jeffress (1948) proposed a 'place theory' of sound localization based on the ITDs, which suggests a physiological mechanism for coincidence detection. Influenced by Jeffress's hypothesis, ITDs have been used in many psychoacousticexperiments and models of binaural detection to characterize interaural timing (Durlach and Colburn, 1978; Colburn, 1996; Breebaart *et al.*, 1999). Responses of neurons to ITDs were also considered in physiological studies of binaural processing (Caird and Klinke, 1987; Kuwada and Yin, 1987; Brugge, 1992; Clarey *et al.*, 1992; Joris and Yin, 1996). Alternatively, interaural phase differences (IPDs), which are the frequency-domain representation of ITDs, have been used to quantify interaural timing cues (Kuwada and Yin, 1983; Spitzer and Semple, 1991). IPDs were used as well in recent quantitative physiological models (Borisyuk *et al.*, 2002). Whether IPD or ITD representations of interaural timing cues are more physiologically relevant for auditory processing in mammals is still an open question (McAlpine and Grothe, 2003). Regarding the processing of ILDs, there is a wide consensus that a combination of excitatory ipsilateral and inhibitory contralateral interactions takes place (Colburn, 1996). Interaural timing information and ILDs are then combined for sound localization (Brugge, 1992). In the barn owl (*tyto alba*) it has been shown that a topotopic map of auditory space exists, which performs such a combination (Knudsen, 1982).

Interaural parameters have also been used as basic parameters for sound source localization algorithms (Neti *et al.*, 1992; Albani *et al.*, 1996; Datum *et al.*, 1996; Duda, 1997; Janko *et al.*, 1997; Chung *et al.*, 2000; Liu *et al.*, 2000), "cocktail-party" processors (Bodden, 1993), and binaural directional filtering algorithms (Kollmeier *et al.*, 1993; Kollmeier and Koch, 1994; Wittkop *et al.*, 1997). The aim of such algorithms is to estimate the directions of the sound sources on a short-term basis and use this information for speech enhancement techniques like Wiener filtering (Bodden, 1996a). A short-term analysis of interaural parameters is commonly used in these approaches. It is also assumed that the auditory system initially evaluates interaural parameters on a short-term basis for exploiting binaural information.

For localization of signals in noise or of nonstationary signals, the information available for sound localization differs from the information in the HRTFs. In contrast to the stationary and anechoic conditions in which HRTFs are measured, the interaural parameters derived from subsequent windows in a short-term analysis fluctuate due to the statistics of the signal and due to the influence of noise, if present. As a consequence, the information about source location in the short-term values of the interaural parameters is likely to be degraded as compared to the information content of the set of HRTFs itself. These fluctuations have a close relationship with the properties of the HRTFs, e.g., spectral notches of the HRTFs for certain directions can lead to stronger fluctuations at these frequencies. Therefore, fluctuations can be used to retrieve additional information on the sound source direction.

Probability density functions of interaural parameters have been evaluated in several studies on binaural detection (Henning, 1973; Domnitz and Colburn, 1976; Zurek, 1991). However, these experiments do not allow us to draw conclusions about localization performance in noisy environments. Although fluctuations of interaural parameters have been explicitly considered in models of sound localization (Duda, 1997; Wittkop *et al.*, 1997), empirical data on the amount of fluctuations are still missing. The aim of the present study is to characterize fluctuations of interaural parameters in realistic listening conditions and to study their possible implications for sound localization.

The approach chosen here is to simulate the possible effective signal processing involved in binaural sound localization and to use *a priori* knowledge about statistical properties of the interaural parameters to estimate the sound source directions. This *a priori* knowledge is gathered empirically by observation of histograms of a large amount of short-term interaural parameter values. Consequences of the observed statistics for modeling sound source localization are discussed using the framework of detection theory and Bayesian analysis. It was not in the focus of these analyses to construct a detailed model of human sound localization, but to gain knowledge about the relevant properties of real-world signals.

2.2. Methods

2.2.1. Assumptions

Several general assumptions are made in this study: First, only binaural cues derived from the amplitude ratio and the phase difference of the left and right signals are used for simulating sound source localization.

Second, the sound to be localized is assumed to be speech coming from a constant direction without reverberation. The noise is assumed to be a noise field as found in real environments without preferred directions and to be incoherent between different frequencies. The long-term signal-to-noise ratio is known, however the shapes of the short-term magnitude spectra of the participating sound sources are not known.

The global structure of the processing is in line with current auditory peripheral modeling approaches: First, a short-term frequency analysis with resolution similar to the auditory critical bandwidth is performed. Interaural parameters are then calculated for the set of frequency bands as a function of time. Second, binaural information is integrated across the different frequency bands. In contrast to most models mentioned in the introduction of this chapter, this integration is defined as a *combination of probabilities*. Because time domain and frequency domain representation of a signal are equivalent in terms of information content, the IPDs as frequency-domain representation of the interaural interaural timing carry approximately the same information as the narrow-band ITDs. To represent interaural timing, the IPDs are used here in addition to the ILDs. Finally, a temporal integration with a longer time constant is performed. The time constants of the model are a 16-ms window for the analysis of interaural parameters, followed by a moving average with 8-ms time constant, and, subsequent to the computation of probabilities, a 100-ms window for the integration of statistical information, similar to the time constants found by Wagner (1991).The long-term integration consists of a simple moving average; statistical estimation procedures of a much higher complexity are possible and may be required to explain localization of several concurrent sound objects. However, they are disregarded here because this chapter focuses on localization of a single directional source.   Using short-term analysis of nonstationary signals to evaluate interaural parameters has important consequences on the amount of temporal fluctuations, especially if framing prior to ITD detection is assumed. As an example, Fig. 2.1 shows schematically the time course of the level at the left and right ear canal entrance. The source is assumed to be on the left side.

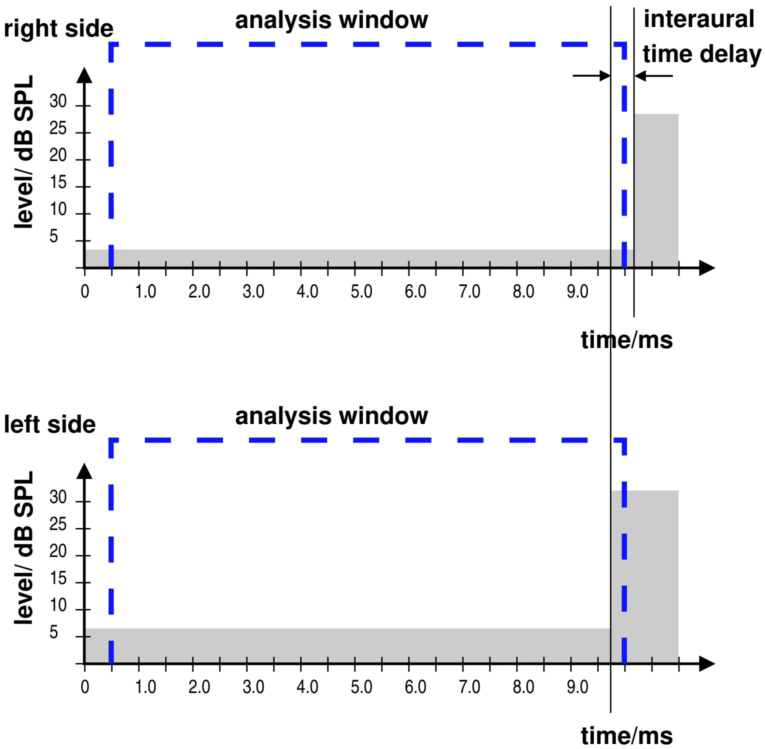


Figure 2.1.: The diagram shows the effect of fluctuations of a single signal with a fixed interaural time delay on the observed ILDs for the framework of a short-term analysis of both ear signals. As long as the signal is constant, an ILD of about 3 dB is observed. In the example, a sudden increase in level occurs close to the end of the analysis window. This increase is caught in the temporal window of the ipsilateral side, but not the one of the contralateral side. Because the ILD and IPD values are weighted according to their amplitude, the higher ILD near the end of the analysis window gains a high weight. A temporal fluctuation of the observed short-term ILDs is the result.

Therefore, we usually expect the level at the left side to be higher than the level at the right side, resulting in a stationary level difference of about 3 dB for our example. Now consider what happens if the signal has a sudden increase in level of about 25 dB near the end of the analysis window. At the left side, part of the additional energy enters the short-term analysis window, but not at the right side. Because the increase by 25 dB means a more than tenfold increase in the amplitude, the observed intensity-weighted level difference is much larger than one would expect because of the sound source direction. The same is valid for level-weighted estimates of interaural timing parameters, when analyzed by a windowed short-term analysis.

2.2.2. Computation of Interaural Parameters in the Frequency Domain

In this work, the interaural parameters are represented in the frequency domain by the interaural transfer function (ITF) (Duda, 1997). It is defined by the HRTFs, $H_r(\alpha, \phi, f)$ and $H_l(\alpha, \phi, f)$, which are functions of azimuth α, elevation ϕ and frequency f; r and l denote left and right side. As the elevation of a sound source, we define the angle to the horizontal plane; as the azimuth, we define the angle between the projection of the source onto the horizontal plane, and the median plane. The ITF is defined as the quotient of the left and right HRTFs, assuming both have non-zero gains:

$$I(\alpha, \phi, f) = \frac{H_r(\alpha, \phi, f)}{H_l(\alpha, \phi, f)}. \tag{2.1}$$

Interaural timing can be characterized by the phase of the ITF

$$\arg I(\alpha, \phi, f), \tag{2.2}$$

by the *interaural phase delay* t_p

$$t_p(\alpha, \phi, f) = -\frac{\arg I(\alpha, \phi, f)}{2\pi f}, \tag{2.3}$$

and by *the interaural group delay*, t_g (Duda, 1997):

$$t_g(\alpha, \phi, f) = -\frac{1}{2\pi} \frac{d \arg I(\alpha, \phi, f)}{d f}. \tag{2.4}$$

If one nonstationary sound source with spectrum $S(f, t)$ is present, t denoting the time variable, the spectra of the signals which arrive, filtered by the HRTFs, at the left and right ear canals are approximately:

$$F_r(\alpha,\phi,f,t) \approx S(f,t)\ H_r(\alpha,\phi,f), \tag{2.5}$$
$$F_l(\alpha,\phi,f,t) \approx S(f,t)\ H_l(\alpha,\phi,f). \tag{2.6}$$

We do not write this as an identity because of the possible windowing effects for nonstationary sources discussed above.

If we assume $F_l(\alpha,\phi,f,t)$ and $F_r(\alpha,\phi,f,t)$ to be non-zero, the time-dependent quotient transfer function $\hat{I}(\alpha,\phi,f)$ can be computed based on the binaural signals as an approximation to the ITF:

$$\hat{I}(\alpha,\phi,f,t) = \frac{F_r(\alpha,\phi,f,t)}{F_l(\alpha,\phi,f,t)}. \tag{2.7}$$

Here, the ILD Δ_L and the IPD Δ_θ are defined as amplitude and phase of $\hat{I}(\alpha,\phi,f)$.

Because, with X^* being the complex conjugate of X,

$$\hat{I}(\alpha,\phi,f,t) = \frac{F_r(\alpha,\phi,f,t)\ F_l(\alpha,\phi,f,t)^*}{|F_l(\alpha,\phi,f,t)|^2}, \tag{2.8}$$

the IPDs can be computed from the interaural cross-power spectrum without unwrapping. $F_r(\alpha,\phi,f,t)/F_l(\alpha,\phi,f,t)$ is, according to the cross-correlation theorem (a generalized form of the Wiener-Khintchine theorem), identical to the Fourier transform of the short-term interaural cross-correlation function.

2.2.3. Description of Interaural Parameters as Random Variables

It is assumed that short-term interaural parameters form a stochastic process. It is described by the vector random variable of interaural parameters $\vec{\boldsymbol{\Delta}}$ whose instances $\vec{\Delta}$ consist of the set of ILD and IPD variables in all frequency bands at a certain point of time. Writing $\Delta_{L,b}$ for the ILD of band b, and $\Delta_{\theta,b}$ the IPD of band b, $\vec{\Delta}$ is defined as

$$\vec{\Delta} = (\Delta_{L,1}, \Delta_{L,2}, \ldots, \Delta_{L,B}, \Delta_{\theta,1}, \Delta_{\theta,2}, \ldots, \Delta_{\theta,B}). \tag{2.9}$$

Assuming that B frequency bands are analyzed, $\vec{\Delta}$ is $2B$-dimensional. In the following, we disregard the temporal characteristics of $\vec{\boldsymbol{\Delta}}$, e.g. temporal correlations, and focus on the probability density function (PDF) of $\vec{\boldsymbol{\Delta}}$. The PDF of the random variable $\vec{\boldsymbol{\Delta}}$ is determined by the properties of the sound field at both ears. In principle, this PDF could be calculated analytically, requiring knowledge about the anechoic HRTFs, the power spectral density statistics of the sources, room acoustics, and the distribution of incidence directions of the noise sources. In practice, however, an analytical derivation is not feasible because the required statistics are not available (Slatky, 1993).

In this study, the PDF is therefore estimated empirically by measuring the time series of short-term interaural parameters calculated from actual binaural recordings of directional sound sources. The normalized histogram of the series is then regarded as being an estimate of the PDF of the underlying random process. We denote the PDF of $\vec{\mathbf{\Delta}}$ given the direction λ as $p(\vec{\Delta}|\lambda)$, where λ is one of N_λ possible discrete directions. As $p(\vec{\Delta}|\lambda)$ is a $2B$-dimensional PDF, and histograms with N_K categories require about $(N_K)^{2B}$ observations, its empirical estimation is not feasible for $B = 43$, because the number of observations required would be too high. Therefore, only the histograms of the components of $\vec{\Delta}$ are observed, and it is assumed that the components of $\vec{\mathbf{\Delta}}$ can be treated as statistically independent. In this case, the joint PDF is calculated by multiplication of the component PDFs (Papoulis, 1965),

$$p(\vec{\Delta}|\lambda) \approx \tilde{p}(\vec{\Delta}|\lambda) = \prod_{b=1}^{2B} p_b(\Delta_b|\lambda), \tag{2.10}$$

where the component or marginal PDFs $p_b(\Delta_b|\lambda)$ are estimated by the histograms of the respective variables Δ_b.

In addition to approximating the PDF of $\vec{\mathbf{\Delta}}$ by the product of its marginals, stationarity and ergodicity of the random process are assumed. Because the statistics of spectral power densities of speech cannot be assumed to be Gaussian, and interaural parameters are the result of a nonlinear operation on the spectral power densities (Slatky, 1993), it cannot be assumed, in general, that the resulting PDF has a Gaussian shape.

2.2.4. Data Acquisition

Signals

The signals used in this study are an additive superposition of one directional sound source recorded in an anechoic room (the target signal) and spatially distributed noises recorded in real listening environments. All signals were recorded binaurally in one individual male subject using hearing aid microphones mounted in in-the-ear (ITE) hearing aids (Siemens Cosmea M). The devices were fitted to the subject's ear canals with ear molds, allowing for a reproducible positioning of the microphones. They were positioned at the ear canal entrance, ensuring that most of the binaural information would be recorded; however, the effect of the concha may have been underestimated due to the size of the devices. The usable frequency range of the hearing aid microphones was between 100 Hz and about 8 kHz.

A nonstationary speech signal was generated as the target signal. Continuous speech from four different talkers was recorded at a sampling frequency of 25 kHz through a single microphone and added together to yield a speech signal with almost no pauses but still significant amplitude modulations. It included one female and three male talkers. One male read a German text and the other talkers read English texts. From this recording a segment of 20-s duration was selected as a target in which all talkers were present.

In the next step, directional information was generated by taking binaural recordings of the target signal in an anechoic room. The position of the source was set by means of a free-field apparatus (Otten, 2001). The apparatus consists of two loudspeakers attached to a movable arc of 3.6-m diameter, which can be positioned with stepping motors at all azimuths and at all elevations between -30° and 85°. Positions of the loudspeakers and generation of signals were controlled by a personal computer. The subject sat in the center of the arc on a stool and was instructed to move his head and arms as little as possible. The head was supported by a headrest. The error of direction is estimated to be about 3° in the azimuth and elevation, mostly the result of head movements and to a lesser extent due to position uncertainty. Recordings of the target signal on digital audio tape (DAT) were taken at 430 positions, ranging in azimuths from 0° to 355° in 5°-steps at elevations of -15°, 0°, 15°, and 30°. For azimuths in the range of -15° to 15° and 165° to 195°, additional elevations of -10°, -5°, 5°, 10°, and 20° were measured. Three breaks were scheduled during the session in order to change the tapes and to permit the subject to take a rest and move around. All elevations for each azimuth were recorded without a break.

In addition to the recordings of the target signal, spatially distributed noise signals were recorded in eight different real environments. By using the same subject, the same microphones, and the same recording equipment, it was assured that the anechoic HRTFs and the transfer functions of the equipment were equal in all recordings. The selected noise environments were a densely populated university cafeteria, an outdoor market, inside a moving car, a city street with traffic noise (two different recordings), a train station concourse, and a metal workshop with various machinery operating, such as a milling cutter, grinding wheel, and lathe (two different recordings). The goal was to record

target signal	noise	SNR / dB
speech	silence	—
	station concourse	15, 5
	cafeteria	30, 20, 15, 10, 12, 5, 3, 2, 1, 0, -1, -2, -5
	metal workshop 1	15, 5
	metal workshop 2	15
	car interior noise	15, 5
	outdoor market	15, 5
	traffic noise 1	15, 5
	traffic noise 2	15, 5

Table 2.1.: List of the 27 signal conditions chosen for the analysis of interaural parameters. Each condition includes a superposition of a directional target signal from 430 directions and a spatially distributed noise at a specific signal-to-noise ratio. The target signal in silence was included as well.

situations in which many nonstationary noise sources were impinging on the listener from different directions at the same time and in which no source was dominant for more than about 1 s. The level of the noise was not measured during the recording session but was ensured to be well above the noise floor associated with the recording equipment so that the interaural parameters were determined by the environmental noise rather than the recording noise. Segments of 25-s duration which met these criteria were selected as spatially distributed noise samples.

The DAT recordings of the target and all noise signals were sampled at 25 kHz with 16-bit resolution and stored on a computer. For further processing and analysis of the interaural parameters, 27 different target-noise conditions were selected. Each condition consists of a set of 430 signals covering the different directions of incidence. They were generated by digitally adding a target and a noise signal at various signal-to-noise ratios (SNRs, defined explicitly in the next paragraph) in the range between +30 dB and -5 dB as well as in silence. The conditions are listed in Table 2.1. Especially for the condition of a speech target mixed with cafeteria noise, a broad range of SNRs were selected to ensure good coverage of this important communication situation.

The SNR was defined as the difference in decibels of the level associated with the target and noise signals. The levels were calculated from the digitized recordings by averaging the overall RMS (root mean square) levels across both ears on a linear scale and transforming this average to decibels. For a specified SNR, the recorded signals were scaled appropriately for each direction, target, and noise type and then added. Using this definition, the SNR was controlled at ear level and did not vary with direction. As the SNR is expected to influence the distribution of interaural parameters, this definition seems more appropriate than defining the SNR at the source level which would have introduced SNR variations with direction due to the influence of the HRTFs. This issue will be discussed later on in this chapter.

Calculation of Interaural Parameters and their Distributions

Interaural parameters, i.e., ILDs and IPDs, were calculated using a windowed short-termshort-term fast Fourier transform (FFT) analysis (Allen and Rabiner, 1977; Allen, 1977) with a subsequent averaging across broader frequency bands.

Time segments with a length of 400 samples were taken from the left and right signals with an overlap of 200 samples. This corresponds to a total window duration of 16 ms and a window time shift of 8 ms, resulting in a frame rate of 125 Hz. The segments were multiplied by a Hann window, padded with zeros for a total length of 512 samples and transformed with an FFT. The short-term FFT spectra of left and right channels are denoted as $F_l(f,k)$ and $F_r(f,k)$, respectively. The indices f and k denote the frequency and time index of the spectrogram, respectively. The FFT bins were grouped to 43 adjacent frequency channels so that, according to the transformation from frequency to the equivalent rectangular bandwidth (ERB) of auditory filters given by Moore (1989b), a bandwidth of at least 0.57 ERB was reached. These frequency bands covered the range of center frequencies from 73 Hz to 7.5 kHz, i.e., 3.3 to 32.6 ERB. Because at low frequencies the frequency resolution of the FFT, 48.8 Hz, does not allow for a channel bandwidth of 0.57 ERB, the low-frequency channels have bandwidths of up to 1.37 ERB; the first band which includes more than one FFT bin has a center frequency of 634 Hz, and the average bandwidth is 0.76 ERB.[3]

Let $f_u(b)$ and $f_h(b)$ denote the lowest and highest FFT frequency index belonging to a frequency channel b, respectively. Frequency averaging was then performed by adding up the squared magnitude spectrum and the complex-valued

[3]Taking into account the effect of the 400-point Hann window, an effective average bandwidth of 0.96 ERB results.

cross-spectrum of left and right FFT results across the FFT-bins belonging to each channel:

$$F_{rr}(b,k) = \sum_{f=f_u(b)}^{f_h(b)} |F_r(f,k)|^2, \tag{2.11}$$

$$F_{ll}(b,k) = \sum_{f=f_u(b)}^{f_h(b)} |F_l(f,k)|^2, \tag{2.12}$$

$$F_{rl}(b,k) = \sum_{f=f_u(b)}^{f_h(b)} F_r(f,k)F_l(f,k)^*. \tag{2.13}$$

These parameters were filtered by a recursive first-order low-pass filter with the filter coefficient $\gamma = e^{-\Delta T/\tau_a}$, corresponding to a time constant of $\tau_a = 8\,\mathrm{ms}$, the frame shift being $\Delta T = 8\,\mathrm{ms}$:

$$\overline{F_{rr}(b,k)} = (1-\gamma)\, F_{rr}(b,k) + \gamma\, \overline{F_{rr}(b,k-1)}, \tag{2.14}$$

$$\overline{F_{ll}(b,k)} = (1-\gamma)\, F_{ll}(b,k) + \gamma\, \overline{F_{ll}(b,k-1)}, \tag{2.15}$$

$$\overline{F_{rl}(b,k)} = (1-\gamma)\, F_{rl}(b,k) + \gamma\, \overline{F_{rl}(b,k-1)}. \tag{2.16}$$

The binaural parameters, i.e., the ILD Δ_L and IPD Δ_θ, were then calculated as follows:

$$\Delta_L(b,k) = 10\log\left|\frac{\overline{F_{rr}(b,k)}}{\overline{F_{ll}(b,k)}}\right|, \tag{2.17}$$

$$\Delta_\theta(b,k) = \arg\overline{F_{rl}(b,k)}. \tag{2.18}$$

It should be noted that Eq. 2.13 describes an intensity-weighted average of the phase difference across the FFT-bins belonging to one channel. The cyclical nature of the IPDs is accounted for by taking the complex-valued average.

The time series of the binaural parameters from each signal were sorted into histograms employing 50 non-zero bins in an adjusted range from up to a -50 dB to 50 dB level difference and up to a -π to +π phase difference, respectively. No special effort was made to detect outliers, except speech pauses that were discarded using a threshold criterion. In each of the 27 different target-noise conditions described earlier, histograms of the ILD and IPD variables in each of the

43 frequency channels for each of the 430 directions were calculated, resulting in a total of 36 980 histograms per condition. The histograms were normalized so that they could be used as an estimate of the marginals of the PDF of the underlying process.

The processing described above was implemented on a VME-Bus-based digital signal processor (DSP) system, which consisted of five DSPs (Texas Instruments TMS320C40) that were attached to a SUN SPARC host system. The signals were read from hard disk and copied in blocks to the memory of the DSP system. The DSPs carried out the calculations and the results (i.e., the histograms of interaural parameters) were transferred back to the host system, and stored on hard disk. The system allows for taking input signals from either hard disk or A/D converters, ensuring real-time processing and calculation of binaural parameters and its distributions. A detailed description of the system can be found in Wittkop *et al.* (1997).

2.2.5. d′ Analysis of Differences in Interaural Parameters

We assume that a classification system discriminates between two directions with a Bayesian criterion on the basis of the observed values of short-term interaural parameters. No further information loss is assumed. In this case and with the further approximation of a Gaussian PDF, the detectability d' of differences in interaural parameters for two different directions can be calculated as the difference in mean values of the respective distributions divided by the geometric mean of their standard deviations. This measure was calculated from the first moments of the empirical univariate distributions of either ILDs or IPDs, neglecting the deviation from a Gaussian shape.

2.2.6. Moment Analysis

The marginal PDF estimates were further analyzed by means of descriptive statistics, (i.e., calculation of higher-order moments of the distributions). Following Tukey's rule (Sachs, 1992) for the estimation of the n-th moment, at least 5^n samples of a random variable should be taken. Here, at least 1 875 samples were evaluated (15 s duration at 125 Hz sampling frequency), so that the first four moments can be estimated according to Tukey's rule. Linear moments, namely expected value, standard deviation skew, and kurtosis (sometimes also called kurtosis excess), were used for the linear ILD variable, as defined in the Appendix B.1. For the distributions of the IPD variable, trigonometric moments as defined by Fisher (1993, p. 41) were used to calculate the mean phase angle, vector strength, circular standard deviation, circular skew, and circular kurtosis (see the Appendix B.2 for definitions). By using the trigonometric moments, no unwrapping of the IPDs is required.

2.2.7. Bayesian Analysis for Sound Localization

We describe the extraction of directional information from the noisy short-term interaural parameters as a Bayesian maximum *a posteriori* (MAP) estimate, which derives the most probable direction based on *a priori* knowledge of the PDF of the parameters for all directions. From Bayes's formula, the conditional probability for each direction λ out of N_λ possible directions is calculated given the parameter vector $\vec{\Delta}$ (compare Eq. 2.9) as

$$P(\lambda|\vec{\Delta}) = \frac{P(\vec{\Delta}|\lambda)P(\lambda)}{\sum_{\lambda=1}^{N_\lambda}[P(\vec{\Delta}|\lambda)P(\lambda)]}, \tag{2.19}$$

where $P(\vec{\Delta}|\lambda)$ is the conditional probability[4] of $\vec{\Delta}$ given the direction λ and $P(\lambda)$ the *a priori* probability for the occurrence of the direction λ. From Eq. 2.10 and under the assumption that all directions λ are equally probable, this formula yields

$$P(\lambda|\vec{\Delta}) = \frac{\prod_{b=1}^{2B} P_b(\Delta_b|\lambda)}{\sum_{\lambda=1}^{N_\lambda}\prod_{b=1}^{2B} P_b(\Delta_b|\lambda)}. \tag{2.20}$$

The right-hand side arguments are all known, as the $P_b(\Delta_b|\lambda)$ are estimated from the empirical analysis for all b and λ. The marginal distributions $P_b(\Delta_b|\lambda)$ form the *a priori* knowledge used to calculate the probability of all directions. The direction chosen as the most probable, given an observation of the parameter vector $\vec{\Delta}$ and assuming that one source is present in the given noise field, is then

$$\hat{\lambda} = \underset{\lambda\in[1\ldots N_\lambda]}{\operatorname{argmax}} P(\lambda|\vec{\Delta}). \tag{2.21}$$

Using Eqs. 2.20 and 2.21, the *a posteriori* probability of all directions can be calculated and the most probable direction can be selected from one observation of parameter $\vec{\Delta}$ and the known distributions of its components for the different

[4]Probabilities $P(\Delta)$ are given in capital letters here and can be calculated by multiplying the probability density $p(\Delta)$ with the parameter interval $\delta\Delta$. This factor is omitted here, because it is constant and does not change the results.

directions. A new set of probabilities can be calculated from every time frame so that an ongoing estimate of the most probable direction results.

The statistical localization model described above has been applied to several types of signals and spatially distributed noises. Results for a 25-s segment of speech from a single talker (different from the four-talker target signal) in cafeteria noise at 5 dB SNR are reported below. The distributions of the four-talker target and the cafeteria noise sample at 5 dB SNR were used as reference distributions[5] in Eq. 2.20.

The *a posteriori* probabilities were calculated for the total signal duration. This yields 3 200 samples of the probabilities for each of the 430 probed directions. The *a posteriori* probabilities were smoothed by a first-order, low-pass filter with 100-ms time constant, and the most probable direction was determined from the smoothed probabilities. The estimates for the most probable direction are plotted into a normalized histogram, which describes the probability that a specific direction is detected as the most probable direction, given the real direction. This plot can be described as a "decision histogram," displaying deviations and confusions in the estimates in one view (see Fig. 2.7).

2.3. Results

2.3.1. Distributions of Interaural Parameters

In this section, the empirical results on the statistics of short-term interaural parameters are described. Specifically, the dependence of the distributions on frequency, SNR, direction, and noise type is analyzed. Due to the large amount of data, the parameter space covered in this analysis had to be restricted. The signal conditions pertaining to speech in silence and speech in cafeteria noise at a moderate-to-low SNR of 5 dB were used as primary examples for one lateral direction and one direction near the median plane (0° elevation.). Results for low frequencies are reported for the IPD variable (340 and 540 Hz), and results for medium and high frequencies are reported for the ILD variable (830 and 2.88 kHz).

[5] Specifically, the reference distributions were clustered using the hierarchical Ward technique (Ward, 1963; Kopp, 1978) so that for each of the 36 980 histograms (43 frequencies, 430 directions and 2 parameters), 1 out of 550 samples of the marginal distributions was used. For simplicity, the influence of this data reduction technique on the localization accuracy is not discussed here. However, its application shows that a noticeable reduction of the *a priori* information is possible.

Dependence on Frequency

Figure 2.2 shows percentiles of the distribution of the ILD variable as a function of frequency for the different conditions pertaining to speech in silence (upper panel),speech in cafeteria noise at 5 dB SNR (middle panel), and car interior noise at 5 dB SNR (lower panel).

In each condition, the 5, 10, 25, 50, 75, 90, and 95 % percentiles of the distributions are plotted for -15° azimuth (lines) and +85° azimuth (symbols) and 0° elevation. The widths of the distributions can be assessed, e.g., by considering the vertical difference between the 95 % percentile line and the 5 % percentile line. For the case of silence, the widths of the distributions reveal that the parameter fluctuation is considerable, even though no additional noise is present, which is due to the statistics of the signal in short time frames. The width depends on both the frequency and the direction of the target. It is especially high in the lowest two frequency bands (50 and 100 Hz), because the band level of the speech is close to the recording noise level in these bands. The distributions are narrow as compared to the mean difference between these directions, except for the low frequencies of up to 8 ERB (310 Hz). In the noise conditions, however, the distributions are significantly broadened so that they overlap for the different directions. The broadening extends across the whole frequency range in the cafeteria-noise condition because the long-term frequency spectra of target and noise are nearly the same and the frequency-specific SNR is almost constant. In the car interior noise, the broadening is restricted to the lower frequency region, because the noise has an $1/f$ type of spectrum with primarily low-frequency content. The SNR of the remaining frequencies is higher than for the cafeteria noise at the same overall SNR.

Dependence on Signal-To-Noise Ratio (SNR)

In order to clarify the influence of noise on the distributions of interaural parameters, their dependence on SNR was studied. Figures 2.3 and 2.4 show the distributions of the ILDs and IPDs, respectively, for the speech target in cafeteria noise. Distributions are plotted for two directions (15° and 60° azimuth 0° elevation) and two frequencies (IPDs: 340 and 540 Hz; ILDs: 830 Hz and 2.88 kHz). For the ILDs, each of the four plots shows distributions for the SNR values of -5, -2, -1, 0, 1, 2, 3, 5, 10, 15, 20, and 30 dB, and in silence. The curves are separated for clarity by relative shifting in y-direction by 0.025 (ILDs), the uppermost curve in each plot showing the distributions in silence.

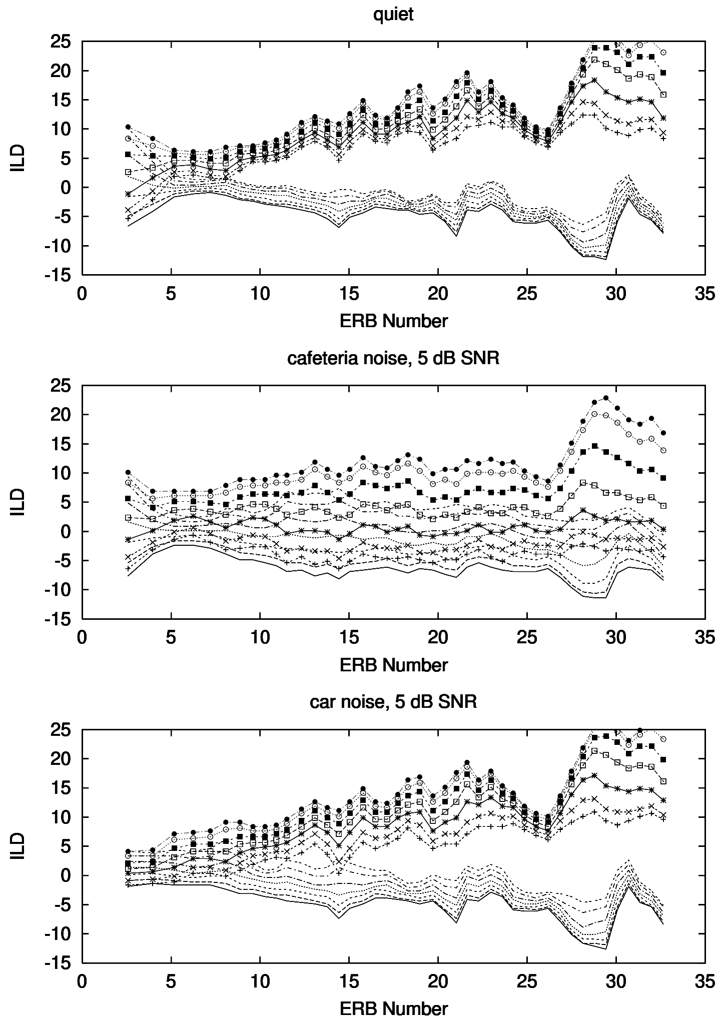


Figure 2.2.: Distribution percentiles for the interaural level difference (ILD) as a function of frequency for three different conditions: speech target in silence (upper panel), in cafeteria noise (5 dB SNR, mid panel), and in car interior noise (5 dB SNR, lower panel). Each panel shows 5, 10, 25, 50, 75, 90, and 95 %-percentiles for -15° azimuth (symbols) and +85° azimuth (lines) and 0° elevation. Percentiles are calculated by integration from the negative towards positive parameter values.

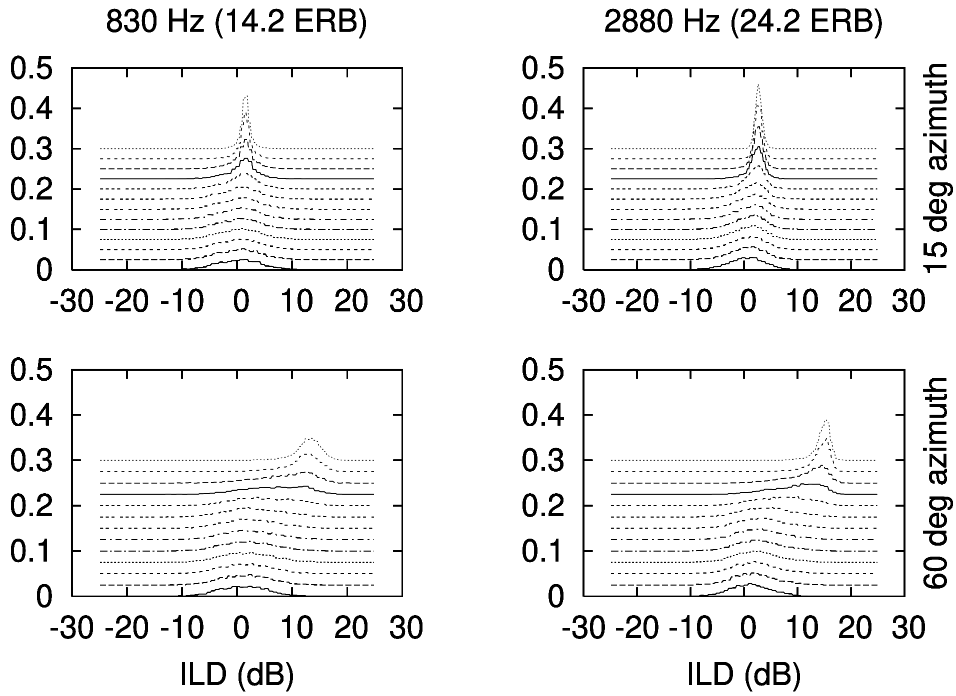


Figure 2.3.: Histograms of ILD values, i.e., number of observations of a specific ILD normalized to the total count, for the speech target in cafeteria noise. The left panels are for a frequency of 830 Hz and the right panels are for 2.88 kHz. The upper panels are for 15° azimuth and the lower panels are for 60° azimuth (0° elevation, respectively). Each panel shows the distributions for the SNR values of -5, -2, -1, 0, 1, 2, 3, 5, 10, 15, 20, and 30 dB, and in silence. The curves are shifted in this order successively by 0.025 in the y-direction for clarity.

In silence, the ILD-distributions (Fig. 2.3) are relatively narrow and show little overlap between directions at both frequencies (upper versus lower panels, respectively). The ILD increases significantly with frequency for both directions (left versus right panels, respectively). It is therefore a distinctive parameter for the direction. However, the distributions "decay" with decreasing SNR due to the influence of the noise. Specifically, the variance increases and the mean value is shifted towards zero. The distributions are skewed at medium SNRs. These effects are due to the nonlinear superposition of the PDFs of the target, which has a nonzero mean and a low variance due to its directionality, and of the noise, which has a zero mean and higher variance because of its diffusiveness. Both the increased variance and the systematic shift of the mean value towards zero lead to a reduction of the variation of the distributions with direction and frequency. The systematic shift is especially large for the lateral direction of 60° azimuth (lower panels). In this case, the observed level difference at low SNRs approaches the difference between the noise level at the contralateral ear and the source level at the ipsilateral ear rather than the large level difference expected from the anechoic HRTFs.

Figure 2.4 shows the corresponding histograms of the IPD variable; because the variable is cyclical, the histogram is plotted in a polar diagram. The angular parameter of the plot shows the IPD, in counter-clockwise orientation, and the IPD 0° corresponds to the half-axis with $y = 0, x > 0$. The radius of the curve, r, shows the relative frequency of occurrence h of this IPD value. To make the differences visible, the frequencies of occurrence are logarithmically transformed according to $r = 4.5 + log_{10} h$. The curves are directly superposed without offset. The maximum radii of each curve are ordered as the SNR values in which the histograms were measured. The figure shows that for the IPD variable the PDFs "decay" as well, converging to a uniform circular distribution with decreasing SNR. The variation of the mean value with decreasing SNR is much smaller than for the ILD variable.

This dependency is reflected in Table 2.2, which lists the higher-order moments of the ILD distributions at 830 Hz and 2.88 kHz and 60° azimuth (lower left and right panel in Fig. 2.3). Both frequencies behave similarly. The mean values and standard deviations show the broadening and shifting described above. Due to the nonlinear superposition of the PDFs of target and noise, the standard deviations show a non-monotonic behavior. They increase and then decrease slightly with decreasing SNRs. Skew and kurtosis are significantly different from 0 at high SNRs and converge towards zero with decreasing SNRs. Again, non-monotonic behavior is observed with decreasing SNRs.

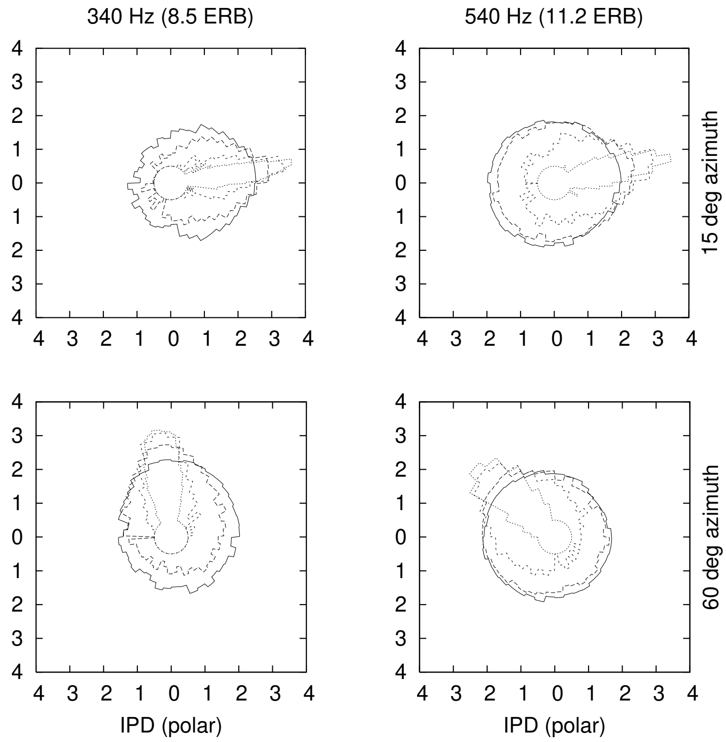


Figure 2.4.: Polar plots showing histograms of the interaural phase difference (IPD). The frequency subbands are 340 Hz (left panels) and 540 Hz (right panels), the azimuths 15° (upper panels) and 60° (lower panels); the elevation was 0°. The angle parameter of the plot shows the angle of the IPD, in counter-clockwise orientation, and the angle 0° corresponds to the half-axis with $y = 0, x > 0$. The radius parameter of the curve, r, shows the relative frequency of occurrence h of this IPD value. To make the differences visible, the frequencies of occurrence are logarithmically transformed according to $r = 4.5 + log_{10} h$. The maximum radii of each curve are ordered as the SNR values at which the histograms were measured. The SNRs are, in the order from the largest maximum to the smallest, silence, 20, 5, and -5 dB.

SNR	f = 830 Hz				f = 2880 Hz			
[dB]	$\bar{x}[dB]$	$\sigma[dB]$	$s[1]$	$K[1]$	$\bar{x}[dB]$	$\sigma[dB]$	$s[1]$	$K[1]$
silence	12.83	2.52	-0.92	1.78	14.63	1.68	-2.03	8.04
30	11.68	3.59	-1.26	2.35	14.09	2.15	-1.63	4.09
20	8.82	5.05	-0.70	0.12	11.78	3.81	-1.08	1.20
15	6.82	5.46	-0.31	-0.48	9.64	4.56	-0.58	-0.31
12	5.61	5.53	-0.12	-0.54	8.08	4.84	-0.33	-0.53
10	4.83	5.51	-0.02	-0.52	7.02	4.93	-0.17	-0.55
5	3.13	5.21	0.10	-0.29	4.49	4.83	0.13	-0.27
3	2.59	5.03	0.11	-0.21	3.63	4.66	0.21	-0.11
2	2.32	4.94	0.10	-0.17	3.22	4.55	0.24	-0.01
1	2.08	4.86	0.08	-0.15	2.82	4.50	0.17	0.24
0	1.87	4.77	0.08	-0.13	2.52	4.36	0.25	0.09
-1	1.67	4.69	0.05	-0.13	2.21	4.27	0.23	0.12
-2	1.49	4.61	0.04	-0.12	1.93	4.17	0.21	0.11
-5	1.04	4.42	-0.01	-0.11	1.24	3.93	0.11	0.05

Table 2.2.: Moments of the distributions of the ILD variable at 830 Hz (left columns) and 2.88 kHz (right columns) at various SNR. The target-noise condition was speech at 60° azimuth and 0° elevation in cafeteria noise. Moments are: $\bar{x}$ expected value, σ standard deviation, s skew, and K kurtosis.

SNR	$f = 340\,\mathrm{Hz}$				$f = 540\,\mathrm{Hz}$			
[dB]	$\phi[rad]$	$\sigma_z[rad]$	$s_z[1]$	$K_z[1]$	$\phi[rad]$	$\sigma_z[rad]$	$s_z[1]$	$K_z[1]$
silence	1.69	0.12	0.43	4.77	2.40	0.17	0.74	21.11
30	1.69	0.14	0.29	10.56	2.42	0.32	-0.05	23.74
20	1.67	0.26	2.00	30.92	2.46	0.60	-0.39	6.69
15	1.65	0.38	1.59	17.84	2.51	0.82	-0.23	2.99
12	1.63	0.48	1.24	11.79	2.55	0.96	-0.18	1.77
10	1.62	0.55	1.06	8.57	2.58	1.05	-0.16	1.27
5	1.54	0.77	0.62	3.64	2.71	1.28	-0.13	0.53
3	1.50	0.86	0.45	2.46	2.77	1.37	-0.10	0.35
2	1.47	0.91	0.44	1.99	2.79	1.41	-0.10	0.29
1	1.43	0.96	0.39	1.61	2.82	1.45	-0.09	0.23
0	1.39	1.02	0.36	1.27	2.85	1.48	-0.09	0.20
-1	1.35	1.07	0.32	1.02	2.88	1.51	-0.07	0.17
-2	1.29	1.13	0.30	0.80	2.92	1.54	-0.06	0.14
-5	1.08	1.28	0.24	0.36	3.03	1.61	-0.03	0.09

Table 2.3.: Same as Table 2.2, but for the IPD variable at 340 and 540 Hz.The moments are: ϕ expected value of phase angle, σ_z standard deviation, s_z skew, and K_z kurtosis (see Appendix B.2 for definitions).

The corresponding higher-order moments of the IPD distributions for the frequencies 340 and 540 Hz are shown in Table 2.3. The mean values depend less on the SNR, whereas the standard deviations show a monotonic increase with decreasing SNRs at both frequencies. Non-zero skew and kurtosis values are observed especially at the lower frequency. For a Gaussian PDF both parameters would be zero because a nonzero skew indicates an asymmetrical distribution, and a positive kurtosis indicates a distribution with a sharper maximum and wider shoulders than the normal distribution. Here, the skew varies between 0.3 and 2.0, and the kurtosis has values up to 30.9, which is a very large difference in shape from Gaussian PDFs.

The analysis of higher-order moments shows that non-Gaussian PDFs have to be assumed in general for the distributions of the interaural parameters. Whether the small deviations from Gaussian shape at low SNRs are still relevant for the retrieval of binaural information remains to be further analyzed.

Dependence on Direction

The interaural parameters of stationary, undisturbed signals, which are defined by the HRTFs, show a clear dependence on direction, and extensive sets of data exist on this in the literature (e.g., Wightman and Kistler, 1989a). Therefore, the direction dependence of the expected values of the short-term interaural parameters, which are associated with the HRTF-derived parameters, is not shown here. Instead, higher-order moments of the distributions are considered.

Data not shown here reveal that the ILD standard deviations depend strongly on the azimuth (for 15 dB SNR, they vary from about 2 to 5 dB at azimuth angles from 15° to 90°) and only moderately on the elevation (about 1 dB variation from -20° to 45° elevation at 15 dB SNR). Standard deviations are high for lateral directions, where the ILDs themselves are large. For the IPD variable, however, the vector strength is high for the lateral directions (i.e., for this frequency, the short-term phase vectors are more congruent in time than for the more central directions). For an SNR of 5 dB, the variation in the vector strength is from about 0.18 to 0.52 for a variation in azimuth from 15° to 90° and for a variation in elevation from -20° to 45°. It is clear from this analysis of the second moments that the fluctuation of the short-term interaural parameters depends strongly on the direction at a fixed SNR. IPDs and ILDs behave differently in this aspect.

Dependence on Noise Condition

The moments of the distribution of the ILD variable at 830 Hz and 2.88 kHz and of the IPD distributions at 340 and 540 Hz for various target-noise conditions are listed in Table 2.4. The target was always speech at 60° azimuth and 0° elevation and the SNR was 15 dB. The moments for the speech target in silence are included as a reference.

The data for the ILD show that the parameter values lie in a narrow range as compared to the deviation from the values in the silent condition. This shows that the influence of the noise type on the distributions is small relative to the influence of the SNR (compare Table 2.2). The only significant deviation is observed for the car interior noise, where the parameters resemble those of the silent condition. The mean valuefor the IPD variable is similar in all noise conditions and deviates little from the silent condition. For the standard deviation, an increase is observed, which is larger at the higher frequency. The skew and kurtosis, however, vary across noise conditions more than for the ILDs. Nevertheless, similar noise types have a similar impact on these parameters.

It can be concluded from the analysis of the moments of the distributions that different types of spatially distributed noise have a similar influence on the distribution of short-term interaural parameters. The SNR is therefore the most relevant parameter for the quantification of the noise's impact on meanand variance. However, the higher-order moments, (i.e., skew and kurtosis) vary with the noise condition especially for the IPD variable.

2.3.2. Simulation Results

A probabilistic approach of directional information extraction from short-term interaural parameters is studied in this section. Both discrimination of directions and absolute localization are considered.

d′ Analysis of Differences in Interaural Parameters

Figure 2.5 shows single-band d' derived from two different target directions as a function of SNR and in silence for the speech in cafeteria noise condition. Data are plotted for the ILD variable at 830 Hz (+) and 2.88 kHz (□) and for the IPD variable at 340 Hz ($\times$) and 540 Hz ($*$).

In the upper panel of Fig. 2.5, the two target directions are 0° and 5° azimuth in the horizontal plane. In silence, d' is greater than 1 in all cases except for the ILD

noise type	SNR				
ILD	[dB]	$\bar{x}[dB]$	$\sigma[dB]$	$s[1]$	$K[1]$
			$f = 830\,\text{Hz}$		
—	silence	12.83	2.52	-0.92	1.78
mean	15	6.64	5.55	-0.37	-0.34
(std)	15	(2.07)	(0.73)	(0.20)	(0.21)
inside car	15	12.43	2.95	-0.84	2.67
			$f = 2880\,\text{Hz}$		
—	silence	14.63	1.68	-2.03	8.04
mean	15	9.56	4.47	-0.65	0.24
(std)	15	(0.84)	(0.50)	(0.15)	(0.48)
inside car	15	14.58	1.83	-1.03	14.03
IPD	[dB]	$\bar{x}_z[rad]$	$\sigma_z[rad]$	$s_z[1]$	$K_z[1]$
			$f = 340\,\text{Hz}$		
—	silence	1.69	0.12	0.43	4.77
mean	15	1.67	0.29	1.18	20.91
(std)		(0.02)	(0.08)	(0.81)	(5.38)
inside car	15	1.68	0.18	2.10	25.14
			$f = 540\,\text{Hz}$		
—	silence	2.40	0.17	0.74	21.11
mean	15	2.45	0.59	-0.15	8.10
(std)		(0.07)	(0.15)	(0.88)	(4.57)
inside car	15	2.41	0.21	0.92	25.47

Table 2.4.: Distribution moments of the ILD variable at 830 Hz and 2.88 kHz and for the IPD variable at 340 and 540 Hz for various target-noise conditions. The target was always speech at 60° azimuth and 0° elevation and the SNR was 15 dB. The "speech in silence" condition is listed as a reference. The noise condition "inside car" is listed separately, the data for the noise conditions "station concourse," "cafeteria," "metal workshop," "outdoor market" and "traffic noise" were averaged and the mean and standard deviation are given.

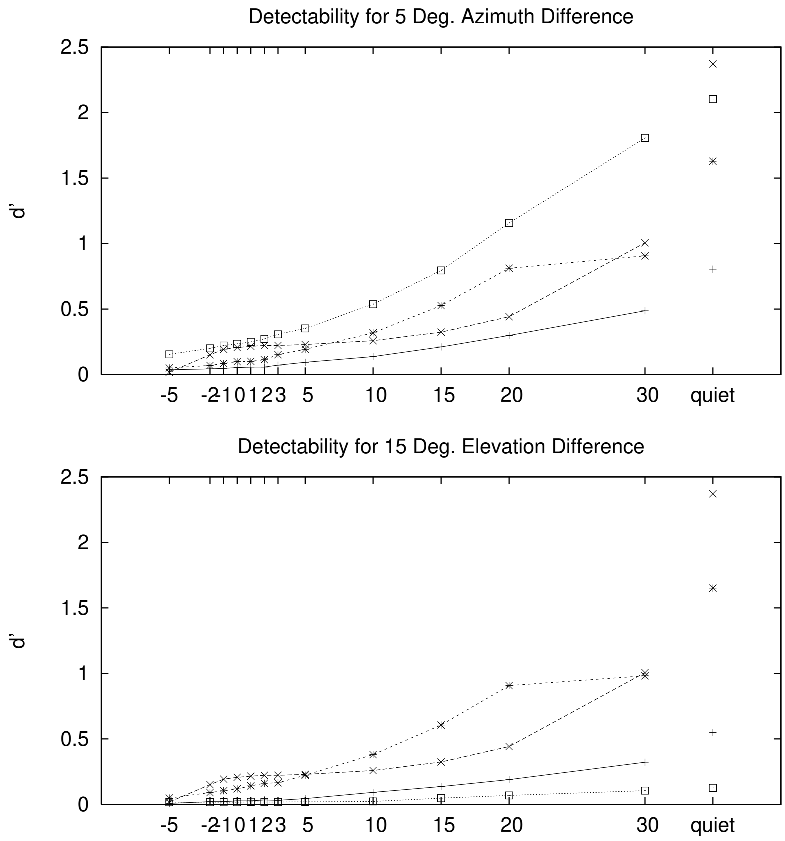


Figure 2.5.: d' of differences in interaural parameters derived from two different target directions as a function of SNR. In the upper panel, the two target directions were at 0° and 5° azimuth in the horizontal plane. In the lower panel, the differences in interaural parameters arise from a shift in elevation from 0° to 15° in the median plane. The target-noise condition was speech in cafeteria noise. Each plot shows data for the ILD variable at 830 Hz (+) and 2.88 kHz (□), and for the IPD variable at 340 (×) and 540 Hz (∗).

variable at 830 Hz. However, d' decreases significantly with decreasing SNRs and reaches a value of about 0.2 on average at an SNR of 5 dB. Assuming that the observations in different frequency bands are statistically independent, and that the ILDs vary systematically with small azimuth changes, the d' increases with a factor of the square root of the number of observations. The number of observations required to detect a difference in direction (i.e., $d' = 1$) is about 25 in this case.

In the lower panel of Fig. 2.5, the differences in interaural parameters arise from a shift in elevation from 0° to 15° in the median plane. The d' is lower than in the case of azimuth variation. At 5 dB SNR, d' is on average 0.1. In this case, about 100 independent observations, combined across time, frequency, or both, are needed to reach a d' of 1.

Simulation of Absolute Sound Localization

Figure 2.6 gives the decision histogram in columns for each direction as a gray-scale-coded frequency count. The condition is speech in cafeteria noise at 5 dB SNR. Each small rectangle in the plot represents a real and estimated azimuth for a given combination of real and estimated elevation. Real azimuths are varied along the x-axis of the subplots, estimated azimuths along the y-axis of the subplot. Each real elevation is represented in a column of subplots and each estimated elevation in a row of subplots. Possible error patterns are explained in Fig. 2.7. If all decisions are correct, a black diagonal intersecting the plot's origin should result. Decisions plotted on parallel lines intersecting the y-axis at different elevation boxes signify elevation confusions, whereas perpendicular diagonal lines indicate front-back confusions. Most pixels away from the diagonal are white, indicating that less than 2 % of the direction estimates were given for this real/estimated direction combination. If for a fixed "true" direction estimates would be evenly distributed across all possible directions, a white column across all subplots would result. The top row of the subplots is mostly white because the elevation value of 45° did not occur in the test data.

The percentages of front-back confusions, averaged across all target directions, were calculated from the decision histogram for the target-noise condition "speech in cafeteria noise" as a function of the SNR. A directional estimate was defined as a front-back confusion if the total angle of error was decreased by at least 15° when mirroring the azimuth coordinate at the frontal plane. Additionally, the direction estimates and the true directions were transformed from vertical-polar coordinates to the angle to the median plane, also known as the

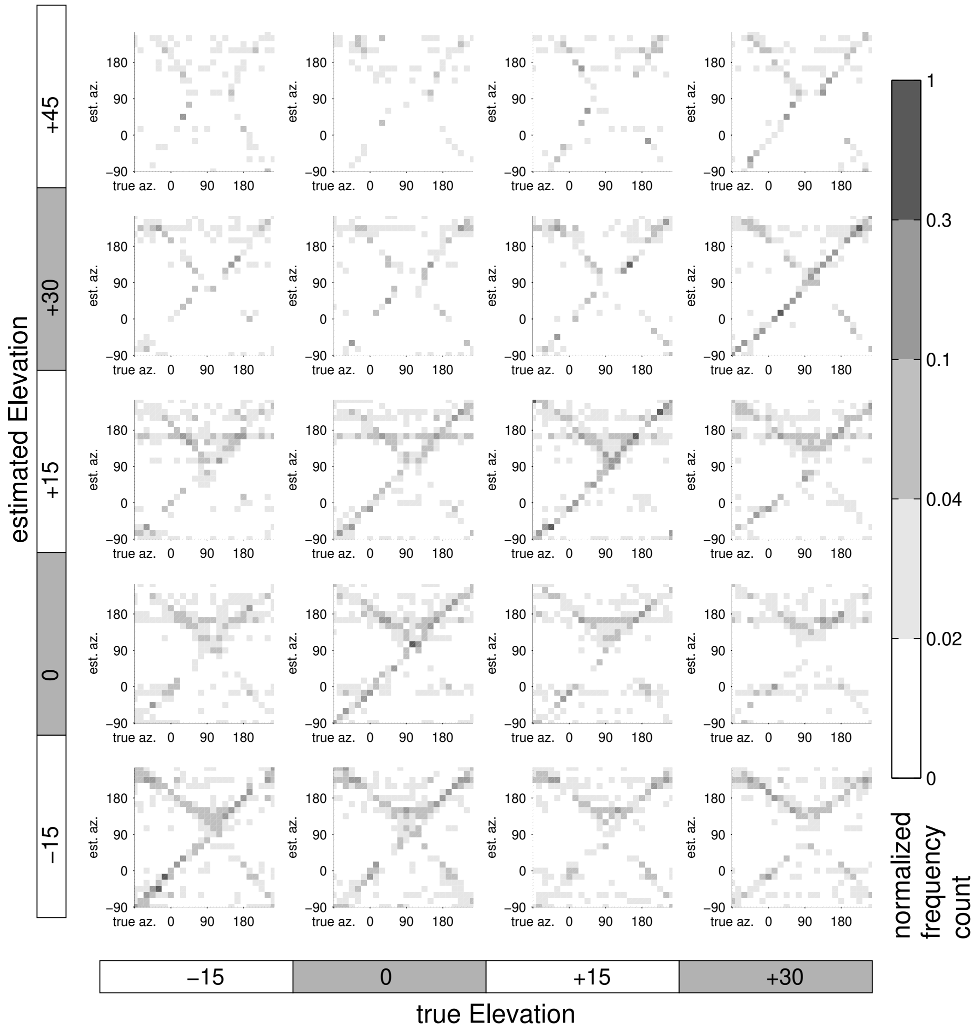
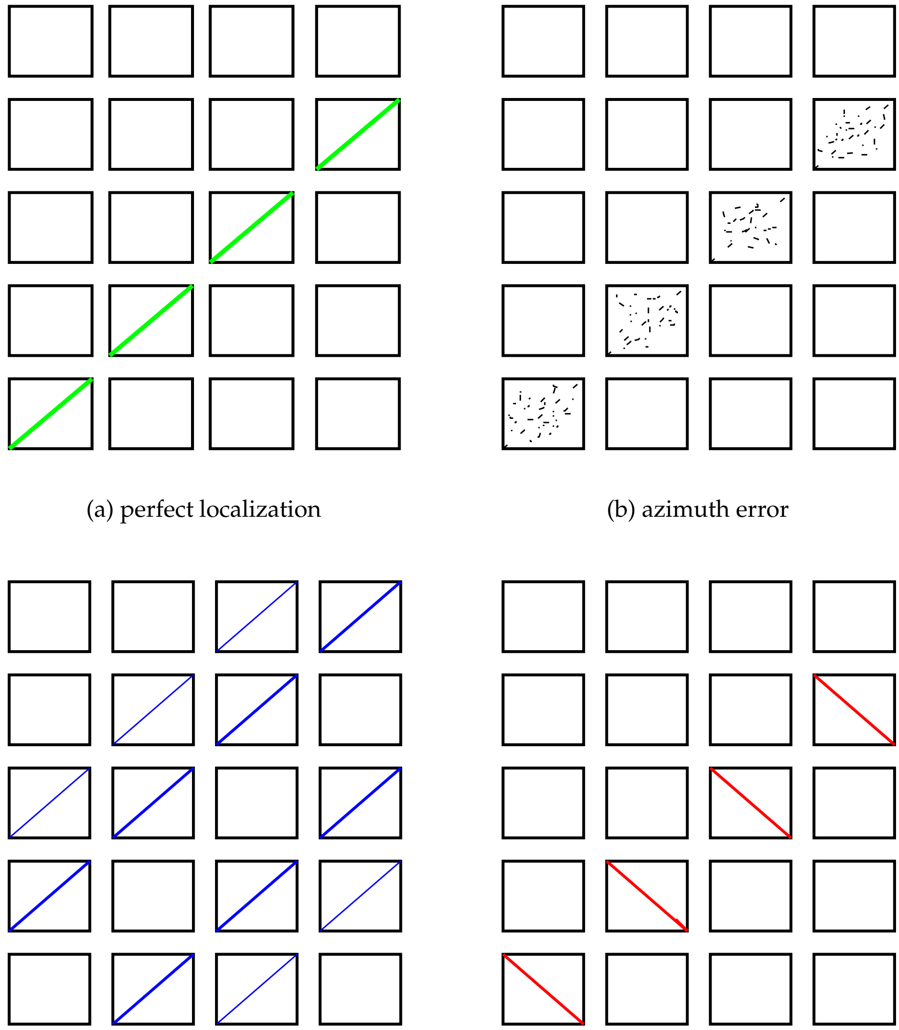


Figure 2.6.: Decision histogram for the target-noise condition “speech in cafeteria noise” at 5 dB SNR. The y-axis represents the detected direction and the x-axis the real direction of the target. Plotted is the normalized frequency count of localization decisions as an estimate of the probability that a specific direction is detected as the most probable direction, given the real direction. Each box in the plot gives real and estimated azimuth for a given combination of real and estimated elevation. If all decisions are correct, a diagonal intersecting the lower left plot’s origin should result.



(a) perfect localization

(b) azimuth error

(c) elevation error

(d) front-back confusion

Figure 2.7.: Error patterns in decision histograms: Panel (a) shows perfect localization, panel (b) correct elevation, but deviations in azimuth, panel (c) perfect azimuths, but deviations in elevation, panel (d) shows otherwise perfect estimates, but with front-back confusions. Note that the elevation +45° does not occur in the test data.

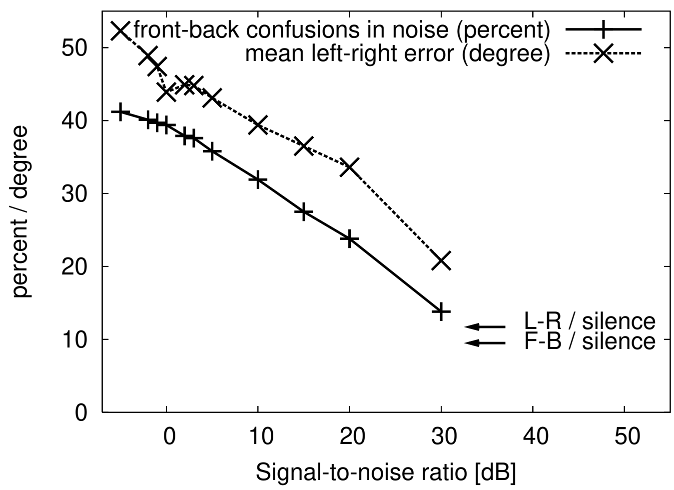


Figure 2.8.: Percentage of front-back confusions (+) and RMS error of the angle to the median plane (×) of the Bayes localization simulation as function of SNR. The target-noise condition is speech in cafeteria noise. Data were averaged for all target directions. An estimate was considered a front-back confusion if the total angle of error was decreased by at least 15° when mirroring the azimuth coordinate at the frontal plane.

	silence	30	20	15	10	5	3	2	0	-2	-5
silence	9.5			34.2							
30		13.8		11.4							
20			23.8	22.3							
15				27.5							
10	34.2	33.2	33.6	32.9	31.9	32.0	32.4	31.8	35.7	34.7	35.0
5				37.1		35.8					
3				37.9			37.6				
2				39.0				37.9			
0				39.6					39.4		
-2				40.1						40.1	
-5				41.4							41.2

Table 2.5.: Percentage of front-back confusions for the Bayes localization algorithm for the target-noise condition "speech in cafeteria noise." Rows show the data for the different tested SNR. The columns give the SNR of the reference distributions that were used as *a priori* knowledge. The conditions with optimal information form the diagonal of the matrix.

"left-right" coordinate (Wightman and Kistler, 1989b), and the RMS value of the error across all tested directions was evaluated. Figure 2.8 shows the percentage of confusions (+) and the RMS error of the angle to the median plane (×). The percentage of confusions increases from 9.5 % in silent conditions, to 41.2 % at the lowest SNR of -5 dB, while the RMS error of the left-right coordinate has a value of 11.7° in silent conditions, which rises to 52.3° at -5 dB SNR.

For a real-world application of the Bayes localization algorithm, its robustness against changes in the noise environment is crucial. The dependence of the distributions of interaural parameters on the SNR was found to be especially high so that the localization accuracy might be lowered, if the reference distributions used in Eq. 2.20 did not match the test stimuli in SNR. The localization accuracy for the target-noise condition "speech in cafeteria noise" has therefore been evaluated for different SNRs with *a priori* information measured at the same SNR ("matched" condition) and at different SNRs ("unmatched" condition). Data are shown in Table 2.3.2.

The rows show the percentage of front-back confusions for the different test SNRs. The columns give the SNR of the reference distributions that were used as *a priori* knowledge. The conditions with optimal information form the diagonal of the matrix (same data as in Fig. 2.8). For a test SNR of 10 dB (5th row), the percentage is lowest for the optimal condition, i.e., a reference SNR of 10 dB. However, the percentage is only slightly higher for an unmatched reference SNR. With a fixed reference SNR of 15 dB (fourth column), the dependence on the test SNR is similar to the dependence with matched references (entries on the diagonal).

Usually, very small differences among data in the same row are observed. However, if the reference condition which was measured in silence is tested at medium SNRs, the percentage of confusions increases from 9.5 to 34.2 % (first row).

2.4. Discussion

2.4.1. Consequences of Parameter Distributions for Sound Localization

The analysis described in the last section reveals significant variations of short-term interaural parameters in silent environments and especially in noise conditions, justifying their statistical description as random variables. The empirical approach of estimating probability distributions from observations of ILD and IPD time series from actual binaural recordings seems to be sufficient for the characterization of random variables, because it reveals the most relevant dependencies of their PDFs on the sound field properties.

The description of short-term interaural parameters as random variables has several consequences for the extraction of directional information from short-term interaural parameters. First, the parameters fluctuate because of the non-stationarity of the signals and their nonlinear combination. Therefore, information retrieval at low SNRs can only be performed in a statistical sense and requires a certain number of observations of short-term interaural parameter values. Second, directional information can be retrieved with *a priori* knowledge of the statistics of the interaural parameters. Restriction of the *a priori* knowledge to the interaural parameters derived from the anechoic HRTFs can lead to much larger errors, as Table 2.3.2 shows. Third, the detectability of systematic variations of interaural parameters with direction might be reduced due to their fluctuations. The amount of fluctuations is mainly a function of SNR and source direction and in the examined cases does not depend much on the noise type so that a general quantification of this effect as a function of SNR seems appropriate. Fourth, the systematic variation of the ILDs with direction itself is reduced due to the shift in mean values, i.e., the bias. Due to the SNR-dependent bias, it can be assumed that the large and significant ILD values observed in anechoic HRTFs from lateral directions cannot be fully exploited for localization or discrimination of directional sources in noise without taking into account the SNR. Finally, the direction dependence of higher-order moments possibly has to be taken into account. It could in principle be exploited for direction estimation, if an analysis of higher-order statistics is included.

2.4.2. Detectability of Differences in Interaural Parameters

The d' analysis shows that at moderate-to-low SNRs of about 15 dB or below, the physical difference in interaural parameters induced by significant variations of direction, according to our assumptions, is not detectable on the basis of the observation of either one of the variables alone. However, discrimination is possible by integration of the information across frequency, time, or both. The coarse estimate given above shows that the number of observations necessary for the detection of 5° azimuth or 15° elevation difference is in the range of the number of frequency bands in a critical band analysis of interaural parameters. In order to reach the same number of observations by temporal integration, time windows of about 400 to 800 ms would be needed, which are much larger than psychoacoustically derived time constants of localization in humans (Stern and Bachorski, 1983).Some authors discussed that interaural cues caused by asymmetries between the pinnae may be sufficient also for sound source localization along the ‘cones of confusion’ and the median plane (Searle *et al.*, 1975). However, as these asymmetries provide only small cues, it still has to be understood in which situations they are relevant.

If such an integration mechanism is used by the human hearing system, it can be assumed that the physically defined d' is perceptually relevant, because at moderate-to-low SNRs, the standard deviation of the IPDs at low frequencies and of the ILDs at medium-to-high frequencies is larger than the human just-noticeable differences (JNDs) in ILDs and IPDs in the respective frequency regions (compare Ito *et al.*, 1982; Stern *et al.*, 1983). It can thus be assumed that the external noise is the limiting factor in this range of SNRs rather than internal noise associated with neural processing.

The notion that frequency integration is needed to ensure detectability, although physically founded here, coincides with the physiological studies by Knudsen and Konishi (1978); Brainard *et al.* (1992); Peña and Konishi (2001) in barn owls that demonstrated binaural information processing in multiple frequency bands and subsequent frequency integration. Directional ambiguities have been shown to be resolved by this integration. The assumption of frequency integration is supported by psychophysical findings showing that the localization accuracy increases with increasing bandwidthof the stimulus (e.g., Butler, 1986).

2.4.3. Simulation of Absolute Localization

Figure 2.6 shows that the confusion pattern is qualitatively similar to the patterns of human performance found, e.g., by Good and Gilkey (1996). Confusions occur mainly as elevation errors or front-back confusions. Very few confusions occur that could not be explained by this typical pattern. At lower SNRs, localization of lateral directions becomes more blurred, which is not shown here.

Good and Gilkey obtained human data for a single noise masker in the front and a click train from different directions. The SNR was defined relative to the detection threshold of the target for the condition where the target was in the front, other directions were tested with the same free field sound level. That means that the SNR at the ear canal entrance varied for different directions of the sound source. Due to the different target-noise condition and the different definition of the SNR, the results are not directly comparable. The diffuse cafeteria condition with a single talker target is probably more difficult to localize at the same SNR than the single-noise source condition using a click train as a target employed for the psychophysical experiments. Also, the psychoacoustic experiment was performed at a constant free-field SNR for each trial block. Because human pinnae enhance the high-frequency part of the spectrum of sounds from frontal directions (Shaw, 1997), there is a systematic direction dependence of the SNR at the ear canal entrance, so that detectability may have been used by the subjects as a cue to estimate the sound source azimuth. A more thorough comparison of the confusion patterns of the algorithm and in humans using the same stimulus configuration is therefore indicated. However, it can be concluded from the data illustrated here that the localization accuracy of the model is qualitatively similar to the one in humans. The relatively good estimation of elevations and front-back directions in the median plane with binaural input only can be explained by the fact that the algorithm is able to exploit asymmetries of the pinnae by the across-frequency integration of probabilities.

Table 2.3.2 suggests that the performance of the algorithm depends mainly on the SNR of the test condition and only slightly on the SNR in the reference condition. This finding shows that the Bayes localization model is robust against changes of the SNR of the reference condition. Its performance decreases only slightly if the *a priori* knowledge does not match the actual target-noise condition to be analyzed. The results show that the distribution of interaural parameters as measured here could be one possible robust source of *a priori* information.

There are several reasons why human subjects probably are able to use binaural information in a more efficient way. First, the proposed MAP estimator disregards possible correlations between frequency channels, because in Eq. 2.10 the multidimensional PDF is approximated as a product of its marginals. Jenison (2000) has shown that a maximum-likelihood estimator with knowledge of the response covariance structure is able to perform better on a correlated population response than an estimator assuming independence. This might be relevant here, because the auditory system effectively possesses not only 43 frequency channels, but many thousands of nerve fibers with overlapping receptive fields. Second, the measured distributions include small variations of the interaural parameters due to head movements during the recording. These measurement errors decrease the localization performance at high SNRs. Third, humans use frequencies of at least up to 10 kHz for sound localization, and the high-frequency ILDs are probably particularly important, while the frequencies used here do not exceed 8 kHz. Fourth, the simulation does not include the interaural group delay, corresponding to time differences of the envelopes, which can be computed, e.g., according to Eq. 2.4. So far, the role of high-frequency ITDs has not been clarified completely (Macpherson and Middlebrooks, 2002); for high frequencies, envelope delays are probably more important. Therefore, it is possible that including some representation of interaural envelope delays improves localization performance. Fifth, humans can use monaural cues for sound localization in some circumstances. This can especially improve discrimination of directions on the median plane. Because monaural cues are not evaluated in the simulation, the distinction of front-back directions and elevations along the 'cone of confusion' is probably worse than the performance of humans at the same SNR. Sixth, the used resolution of the histograms of IPDs is in part of the cases coarser than psychoacoustically observed JNDs; this should only affect localization at high SNRs. On the other hand, there is one aspect which may improve the performance of the algorithm as compared to human subjects, i.e., that humans cannot extract the fine structure of waveforms beyond 2 kHz. However, at higher frequencies, the IPDs are not only strongly disturbed by noise, but also become highly ambiguous. Taking all preceding aspects into account, humans probably can use the binaural information in a more efficient way, especially for directions close to the median plane, and for higher frequencies. A preprocessing model which better matches human binaural processing including interaural envelope delays, and excluding IPDs for channels at 1.5 kHz and higher, can possibly explain most of the localization ability of humans by binaural parameters alone.

It should be noted that the Bayesian approach is equivalent to the one of Duda (1997) if one assumes that the distributions are Gaussian with constant variance.In this case, the MAP procedure reduces to a least-squares fit, which would not need the *a priori* knowledge of all distributions. In contrast, Eq. 2.20 allows a more general approach which is able to take noise explicitly into account.

2.4.4. Frequency Integration of Probabilities

Most models of lateralization or localization combine short-term cross-correlation values over frequency by a summation or multiplication (e.g., Stern *et al.*, 1988; Shackleton *et al.*, 1992; Stern and Trahiotis, 1997; Braasch and Hartung, 2002; Braasch, 2002b,a). Neurophysiological findings support that, for some species, after the detection of interaural parameters, a frequency integration is performed. This has been shown by Brainard *et al.* (1992) in the barn owl. Interaural cue detection followed by frequency integration has been used also successfully by frequency-domain models and technically motivated algorithms of sound localization (Duda, 1997; Wittkop *et al.*, 1997; Nakashima *et al.*, 2003). In difference to the models cited above, the quantities which are integrated across frequency in the model presented here are *probabilities*, which takes, according to the assumptions stated, the available information into account in an optimum way.

2.4.5. Statistical Representation of Interaural Timing

Because interaural parameters are considered in the frequency domain in narrow frequency bands, the interaural phase differences (IPDs) are used to describe timing differences. IPDs have several advantages over ITDs: For signals filtered by narrow-band auditory filters, the interaural cross-correlation function (ICCF) becomes nearly periodic. In the case that the signal is nonstationary or contains additional noise, the place of the maximum of the ICCF, which is used frequently to estimate the ITD, is not well-defined (Lyon, 1983; Stern *et al.*, 1988; Schauer *et al.*, 2000). Examples of time-series of ICCF maxima compared to IPD values are shown in Appendix D. This ambiguity of the maximum of the ICCF is directly related to the ambiguity of the phase in the frequency domain. This can be explained by the fact that, according to Eq. 2.8, both representations are linked by the combination of the generalized Wiener-Khintchine theorem (cross-correlation theorem) and the Wiener-Lee relation, and therefore have equal information content. Consequently, the probability density function (PDF) of ITD estimates based on the ICCF would have several maxima. While it is possible to describe such multimodal PDFs by histograms, there is no well-established approach to characterize it by a few parameters.

Contrarily, the statistics of the IPDs can be described neatly by statistics of cyclical data as defined by Fisher (1993); the expected value and variance can be calculated robustly, and empirically observed PDFs can be approximated well by the von Mises distribution. The phase difference is represented in the complex plane. Therefore, the error-prone operation of unwrapping the phase of noisy signals (Tribolet, 1977)is not necessary. This advantage of the IPDs has shown to become especially important in noise, as demonstrated by technical algorithms for robust sound localization (Liu *et al.*, 2000; Nakashima *et al.*, 2003).

2.4.6. Possible Physiological Representations of Interaural Timing

The processing structure sketched here aims to be a possible description of important features of the binaural auditory system. Clearly, the actual signal processing in binaural processing of interaural timing is still being discussed and may vary between different species. However, the explicit consideration of external noise, as proposed here, might be relevant for modeling physiological data.

Harper and McAlpine (2004), e.g., showed that when assuming a population code for distributions of IPDs for humans, as measured in indoor and outdoor sound fields, there are consequences for optimum distributions of best IPDs of auditory nerve fibers. Fitzpatrick *et al.* (1997) propose a population code based on the observation that localization of sounds is much more accurate than the spatial sensitivity of single neurons. Population codes have been proposed also, e.g., by Hancock and Delgutte (2004). The statistical data as well as the Bayesian approach described here could help to develop such models further, and eventually to decide which model matches neural data best.

The approach described here is solely based on the physical properties of the interaural parameters and subsequent Bayesian estimation. However, there is an interesting similarity with physiological models and data. When taking the logarithm of Eq. 2.20, the log probability $\log P(\lambda|\vec{\Delta})$ can be interpreted as an activity that is a sum of the activities (log probabilities) derived from the frequency bands. Frequency specific activities are generated by the log distributions $\log P_b(\Delta_b|\lambda)$ from the observed parameter Δ_b. In terms of neural processing, the log distributions can be interpreted as optimum tuning curves of neurons sensitive to single-channel ILDs and IPDs. These narrow-band tuning curves are rather broad due to the external noise. The tuning curve sharpens by summation or multiplication across frequency, which is equivalent to multiplication of probabilities, and a precise and robust localization is possible, although the basic tuning curves are unspecific. Measuring the "response" for ITD of narrow-band stimuli would yield periodic tuning curves. Also, with progressive frequency integration, ITD tuning curves would become less periodic and their shape should become more similar to a wide-band cross-correlation function, the bandwidth corresponding to the bandwidth of the receptive fields. Therefore, the width and shape of the tuning curves could be interpreted as useful to increase the robustness in noise. The observed deviations of the parameter distributions from Gaussian shapes suggest that properties of physiologically observed tuning curves for interaural parameters, such as asymmetry (skewedness), might be an adaptation to increase the robustness with real-world stimuli.

The interpretation of the log distributions as activation tuning curves and well-defined *a priori* information can be regarded as a major advantage of the Bayesian localization model compared to other approaches that use neural nets in combination with common training rules (e.g., Neti *et al.*, 1992; Datum *et al.*, 1996; Janko *et al.*, 1997; Chung *et al.*, 2000). In neural nets, the training procedure is less well-defined and sources of information used by the net are less clear than those in the approach described here.

The general approach employed here is not restricted to the specific ILD and IPD analysis carried out here, but is also applicable to more specific computational models of human binaural signal processing. Small nonlinearities in the extraction of binaural information by the models is acceptable for this type of analysis, as it has been shown here that the physically defined parameters are nonlinear functions of the sound field as well. Additional nonlinearities induced by the models, (e.g., level dependencies) could add some additional uncertainty, which is processed by the fuzzy information processing strategy proposed here in the same way as the nonlinearity in the physical parameters.

2.5. Conclusions

In noise conditions, the observed random variation of short-term, narrow-band interaural parameters (ILDs and IPDs) with time is large compared to the systematic variation induced by a change of direction of the sound source of several degrees in azimuth. Additionally, noise fields cause a systematic shift of the average values of these parameters. Because of the stochastic temporal variability, integration of information across frequency, or time, or both, is necessary to estimate directions from interaural parameters in such conditions.

A way to achieve this integration is the combination of statistical information across frequency. A Bayesian approach was used for this, which takes the estimated probability density functions (PDFs) of ILDs and IPDs from a reference noise condition as *a priori* information. These *a priori* PDFs were measured and evaluated for a large number of conditions. The shapes of the observed distributions depend mainly on the SNR, the azimuth, and the elevation. The noise environment has a smaller influence on the shape of the distributions. This influence is most notable at medium SNRs. Using the Bayesian approach, the azimuth and elevation can be estimated robustly. The elevation can be estimated even in the median plane at SNRs as low as 5 dB. The localization performance depends mainly on the SNR in the test condition.

The high level of external noise in combination with the hypothesis of integrating probabilities in the neural system could explain why tuning curves of neurons sensitive to interaural timing found in physiological measurements are broad, unspecific, and often asymmetric, while the behavioral localization performance is robust and accurate. External noise with realistic statistical properties should be explicitly considered in physiological measurements and models of binaural processing.

Acknowledgments

We are grateful to Birger Kollmeier for his substantial support and contribution to this work. We thank the members of the Oldenburg Medical Physics Group, especially Thomas Wittkop, Stephan Albani, and Jörn Otten, for providing technical support and for important discussions. Ronny Meyer prepared additional material which helped to discuss the results. Also, we are grateful to the staff of the Hearing Research Center at the Department of Biomedical Engineering, Boston University, for fruitful and motivating discussions during a visit from the second author. Thanks to Fred Wightman, two anonymous reviewers, Steve Greenberg, Hermann Wagner, and Jesko Verhey for helpful suggestions and comments on earlier versions of this manuscript.

This work was supported by DFG (European Graduate School Psychoacoustics) and BMBF (Center of Excellence on Hearing Technology, 01 EZ 02 12).

3. Application of noise-robust binaural sound localization for the identification and separation of concurrent voices[1]

Abstract

For the goal of noise reduction in hearing aids or by small microphone arrays, robust sound source localization in nonstationary noise is desirable. A real-time source localization algorithm based on interaural parameters is investigated here, which is based on the Bayesian maximum a posteriori *(MAP) approach described in Chapter 2. Now, we employ the approach to estimate the directions of concurrent voices. Further, we evaluate the robustness of the algorithm thoroughly in different noise environments, and with different numbers of sources. Results show that the Bayesian estimation resolves ambiguities caused by HRTFs, narrow-band filtering, and noise, yielding fast, noise-robust localization, even if the* a priori *data do not match the noise environment. The algorithm tracks the azimuths of up to three concurrent talkers reliably. The MAP estimates are used to separate two concurrent voices from binaural recordings by controlling a beamforming algorithm based on head-related transfer functions (HRTFs). An evaluation of SNRs of signals filtered by beamforming shows improvements of up to 30 dB and adaptation within 0.2 s. The algorithm can also provide microphone arrays with directional information. The probabilistic information about the directions of sound sources is applicable to extended statistical frameworks.*

[1] Preliminary results concerning the separation of concurrent voices have been presented in Nix and Hohmann (2001).

3.1. Introduction

In difficult listening situations, humans exploit spatial characteristics of sound signals by mechanisms of binaural hearing (Rayleigh, 1907). Part of the parameters which carry directional information are the interaural time delay (ITD), the frequency-specific interaural phase differences (IPDs), and interaural level differences (ILDs).

For the domain of noise reduction, e.g., in binaural hearing aids, these direction-indicating parameters have the interesting advantage that they do not require prior knowledge about the sound source spectrum. Also, compared to features like the fundamental frequency or amplitude modulations, the directions of sound sources change relatively slowly. Using information on sound source direction therefore allows us to identify sound sources, and to suppress them selectively with beamforming algorithms (Greenberg and Zurek, 2001).

Interaural parameters show characteristic patterns for each direction. These are caused by directional filtering of head and pinna, which is described by the head-related transfer functions (HRTF) (Mehrgardt and Mellert, 1977; Blauert, 1983; Middlebrooks *et al.*, 1989). In silent environments, they allow for the classification of sound source direction by neural networks and pattern-matching approaches (Neti *et al.*, 1992; Datum *et al.*, 1996; Isabelle *et al.*, 1998). Other approaches, based on physiological models, evaluate parameters based on the interaural cross correlation function, motivated by the place theory of sound localization (Jeffress, 1948; Lyon, 1983; Lindemann, 1986a,b). The direction estimates obtained from these algorithms are usable to control, for example, Wiener filtering for speech enhancement in situations with interfering talkers (Lyon, 1983; Bodden, 1996a; Slatky, 1993; Bodden, 1996b; Roman *et al.*, 2003). Albani *et al.* (1996) and Duda (1997) demonstrated estimation of azimuth and elevation based on a frequency-domain representation of binaural parameters. Liu *et al.* (2000) employed a representation based on the narrow-band ITD. Speech separation algorithms based on direction estimation followed by spatial filtering were developed, e.g., by Liu *et al.* (2001) and Greenberg *et al.* (2003). Directional filtering has been applied in a variety of other algorithms, e.g. those described by Peissig (1992); Soede *et al.* (1993); Wittkop and Hohmann (2003); see Greenberg and Zurek (2001) for a review.

There are two kinds of ambiguities that make correct localization by binaural parameters especially difficult: Ambiguities caused by narrow-band filtering of interaural timing cues, and ambiguities caused by the symmetry of the HRTFs.

The first ambiguity emerges when the interaural timing parameters are evaluated in narrow frequency bands. Frequently, this is accomplished by computing the place of the maximum of the interaural cross correlation function (ICCF) of the band-pass filtered signals. With decreasing bandwidths, this function approaches a cyclic shape. The consequence is that the maximum of the ICCF becomes ambiguous as soon as the period of this cyclic function – given by the center frequency of the band-pass filtering – becomes smaller than the time delay caused by the microphone distance[2], which in part of the cases leads to wrong direction estimates (Lyon, 1983; Liu *et al.*, 2000).

The second class of ambiguities corresponds to sets of directions with very similar HRTFs, especially the directions with equal angle to the interaural axis. The similarity is caused by the symmetry of the head, which results in very small differences in the patterns of the interaural parameters. Humans confound these directions easily, therefore resulting in the name "cones of confusion" (Damaske and Wagner, 1969; Searle *et al.*, 1975; Blauert, 1983; Shinn-Cunningham *et al.*, 2000). In consequence, most algorithms, with exception of the ones proposed by Duda (1997) and Albani *et al.* (1996), do not distinguish between sounds coming from the front or back.

The influence of noise augments both ambiguities, because it causes the patterns to acquire a high degree of uncertainty[3]. This degrades the performance of localization algorithms in complex noise situations. The problem of sound localization with high levels of background noise is so far unsolved.

The solution investigated here, motivated by findings on the mechanisms of sound localization in the barn owl (Brainard *et al.*, 1992; Albani *et al.*, 1996), evaluates interaural parameters in the frequency domain, and accounts for the ambiguities of the interaural timing parameters by using the interaural phase difference (IPD). As developed in Chapter 2, a multidimensional Bayesian estimation method tackles the localization task, by taking the influence of the noise field and the statistics of the target signal explicitly into account.

The first objective of this chapter is to examine further the use of binaural parameters for sound localization in real-world environments. The effect of noise environments which deviate from the *a priori* reference condition is evaluated more thoroughly. As a second objective, the chapter examines the localization of concurrent voices. The third objective is to investigate the separation of concurrent voices by using the estimated directions. Once the sound source directions are known, there are several possible strategies for enhancing a specific sound

[2]See Appendix D for examples of time series of ITD computed from ICCF maxima.
[3]See Figs. E.1 to E.4 in Appendix E.

source. If only two sources are present, both can be reconstructed by inverse filtering of the microphone signals using the known interaural transfer function (ITF) for each detected direction. This approach is evaluated here.

3.2. Methods

3.2.1. Algorithm for Sound Localization

Figure 3.1 shows a diagram of the algorithm. It shows the flow of data from top to bottom, and divides the algorithm into two stages, a 'training stage' and an 'operation stage'. The computation begins with a short-term frequency analysis. Subsequently following is the averaging across frequency bands, and the calculation of level and phase differences. The algorithm does this in the same way for the training stage and the operation stage. Below this preprocessing block, the right side depicts the training stage, which performs the acquisition of statistical references. The algorithm measures histograms of ILD and IPD. Then, they are compressed by a cluster analysis, and stored. The left side shows the operation stage which performs the on-line estimation of sound source directions. When this part of the algorithm starts, it loads the reference statistics produced in the training stage. For each time step, the probabilities for the observed ILD and IPD values are calculated, achieving a 'fuzzy' comparison between observed values and references. These probabilities are multiplied across frequency to yield *a posteriori* values for each direction. Subsequently, the resulting probabilities are low-pass filtered with a time constant of 100 ms, and the algorithm returns the directions with the largest *a posteriori* values as a result.

The operation stage of the algorithm was implemented to run on a dual processor workstation with AMD Opteron 246 CPUs and a 2 GHz clock. The implementation uses the Python script language, which provides efficient high-level vector operations, and is therefore able to run in real time. Appendix A describes this implementation.

Short-Term Frequency Analysis

An overlap-add technique processed the binaural signals (Allen, 1977), employing 25 000 Hz sampling frequency for the 18-bit digitized stereo signal, a Hann window of 16 ms length, 8 ms window shift, and a 512-point fast Fourier transform (FFT). This generates two short-term spectra $S_r(f,k)$ and $S_l(f,k)$ for time step k every 8 ms.

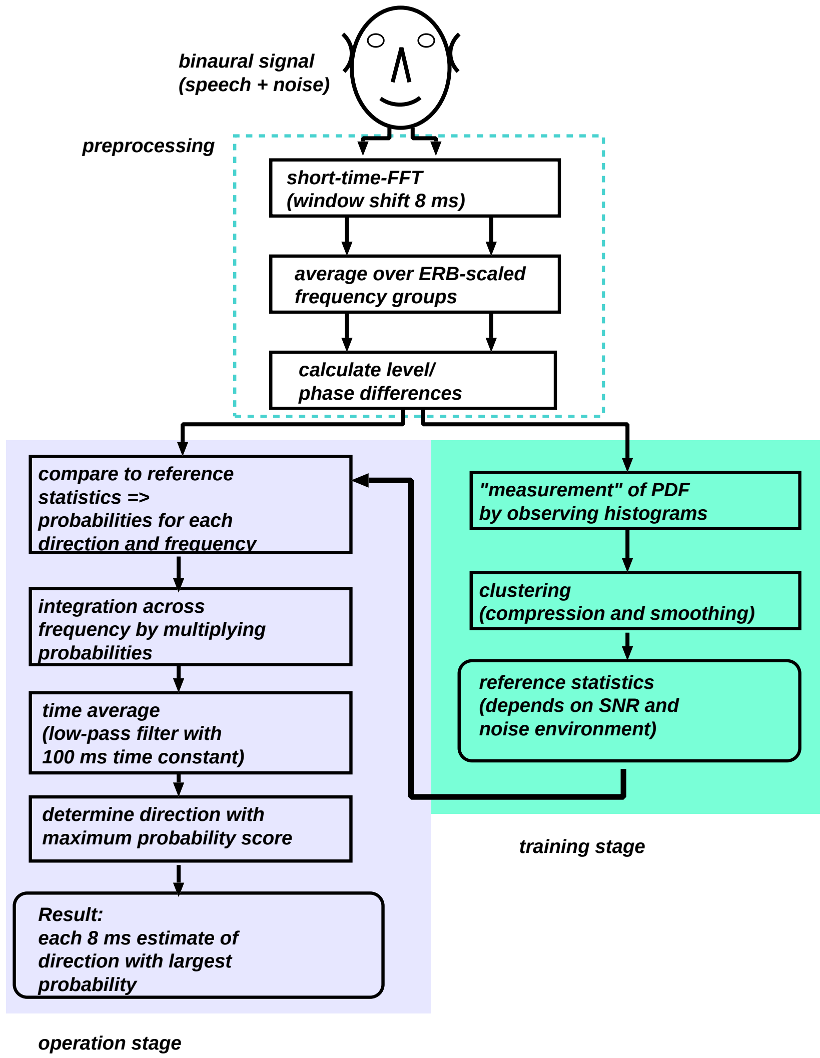


Figure 3.1.: Block diagram of Bayesian real-time localization algorithm

Let $H_{\alpha,\phi,r}(f)$ be the head-related transfer function (HRTF) for the direction with the azimuth α and the elevation ϕ for the right ear, and $H_{\alpha,\phi,l}(f)$ the HRTF for the left ear, both assumed to be non-zero. Assuming that one source is present, and denoting its original free-field spectrum as $S(f,k)$, and the signal spectra arriving at the left and right ear channel as $F_l(f,k)$ and $F_r(f,k)$, they relate to the HRTFs as

$$F_r(f,k) = H_{\alpha,\phi,r}(f)S(f,k) \tag{3.1}$$

$$F_l(f,k) = H_{\alpha,\phi,l}(f)S(f,k) \tag{3.2}$$

Computation of Short-Term Interaural Parameters

To compute short-term ILD and IPD, the algorithm averages the short-term spectra $F_r(f,k)$ and $F_l(f,k)$ over $B = 43$ adjacent frequency bands with a mean bandwidth of 0.96 ERB; b is used as the band index. As explained in Chapter 2, the short-term binaural parameters, i.e. the ILD Δ_L, and the IPD Δ_θ, were then computed from the band-averaged, smoothed short-term levels, $\overline{F_{rr}(b,k)}$ and $\overline{F_{rr}(b,k)}$, and the smoothed short-term cross spectrum $\overline{F_{rl}(b,n)}$ as follows:

$$\Delta_{L,b}(k) = 10\log\left|\frac{\overline{F_{rr}(b,k)}}{\overline{F_{ll}(b,k)}}\right| \tag{3.3}$$

$$\Delta_{\theta,b}(k) = \arg\overline{F_{rl}(b,n)} \tag{3.4}$$

Localization by Bayesian Estimation

This section summarizes the Bayesian estimation algorithm. As the extensive statistical evaluations in Chapter 2 have shown, superpositionof nonstationary signals from different directions will result in fluctuating phase and level differences. At low signal-to-noise ratio (SNR), the amount of fluctuation is higher than the direction-specific variation of ILD and IPD between similar azimuths (or same azimuth and different elevations, because of the "cones of confusions"); see Chapter 2, and Figs. E.1 to E.4 on pages 146 to 149. To account for this statistical variability, we consider $\Delta_\theta(b,k)$ and $\Delta_L(b,k)$ as realizations of random variables; the time index k is omitted for convenience. In the following, N_λ is the number of possible discrete sound source directions, with direction index λ, the azimuth value α_λ and elevation value ϕ_λ. Phase and level differences for each frequency band are grouped in the time-dependent feature vector $\vec{\Delta}$, defined as

$$\vec{\Delta} = (\Delta_b) = (\Delta_{L,1}, \Delta_{L,2}, \ldots \Delta_{L,B}, \Delta_{\theta,1}, \Delta_{\theta,2}, \ldots \Delta_{\theta,B}). \tag{3.5}$$

The conditional PDF of this compound random variable, given the direction λ, is $p(\vec{\Delta}|\lambda)$. Given that a feature vector $\vec{\Delta}$ from an unknown direction is observed, the probability for the presence of the direction λ is, according to Bayes's formula:

$$p(\lambda|\vec{\Delta}) = \frac{\tilde{p}(\vec{\Delta}|\lambda)p(\lambda)}{\sum_{\lambda=1}^{N_\lambda} p(\lambda)\tilde{p}(\vec{\Delta}|\lambda)} \tag{3.6}$$

We approximate $p(\vec{\Delta}|\lambda)$ as the product of marginal distributions $\tilde{p}(\Delta_b|\lambda)$ by assuming the components of $\vec{\Delta}$ to be statistically independent:

$$p(\vec{\Delta}|\lambda) \approx \tilde{p}(\vec{\Delta}|\lambda) = \prod_{b=1}^{2B} \tilde{p}(\Delta_b|\lambda) \tag{3.7}$$

Then, the resulting time-dependent probabilities $p(\lambda|\vec{\Delta})$ for each direction are smoothed by a first-order low-pass filter with a time constant of 100 ms. We refer to the result of this smoothing as *a posteriori* value for this direction. The direction $\hat{\lambda}$, defined by

$$\hat{\lambda}(k) = \underset{\lambda}{\operatorname{argmax}}\, p(\lambda|\vec{\Delta}_k),$$

and with coordinates $(\hat{\alpha}, \hat{\phi}) = (\alpha_{\hat{\lambda}}, \phi_{\hat{\lambda}})$ defines the maximum *a posteriori* (MAP) estimate of the direction of the most active sound source at time k.

The algorithm represents the required estimates of $\tilde{p}(\Delta_b|\lambda)$ by ordinary histograms for both ILD and IPD. These histograms are generated during the "training stage" of the algorithm, and can be viewed as "learned" references, the learning step being equivalent to the training of neural networks.

3.2.2. Signal Recordings

For operating and testing of the algorithm, four different sets of signals, named A, B, C, and D, were recorded. For the Bayesian algorithm, set A (REF) and set D (NOISE) served to generate *a priori* information; B (TEST) and C (TALK), mixed with samples from D (NOISE), were used to test the algorithm. All recordings were recorded binaurally by the same human subject using in-the-ear (ITE) hearing aids (Siemens Cosmea M)[4]. The signal sets A (REF), D (NOISE), and C (TALK) were spatial recordings of speech, made in an anechoic room; signal set D (NOISE) consisted of recordings of noise environments. For sampling of directional signals in the anechoic room, an automatic device, the 'two arc source positioning system' (TASP) positioned two loudspeakers. The systems allowed to position them with high precision at directions between -40° and 80° elevation and arbitrary azimuth (Otten, 2001). A portable digital audio tape (DAT) device recorded the two-channel microphone signals. Care was taken to conserve the original level differences between channels during recording and the subsequent A/D conversion.

For signal set *A (REF)*, reference signals of 25 s duration, recorded in the anechoic room, were generated as a single-channel mixture of four talkers. The recordings were spatially sampled from 430 directions; the angles were between 0° and 355° azimuth with a spacing of 5°. Elevations between $-15°$ and 45° were measured. The elevation spacing had the value of 15° for lateral directions. To improve the distinction of directions, for directions within 15° from the median plane, the elevation spacing was reduced to 5°.

Recording set *B (TEST)* had the purpose of testing the algorithm. Spatial recordings with 27 s duration, each containing the same sample of continuous speech from a single talker, were made in the anechoic room with the same technique as recording set A (REF). 96 directions were recorded: Elevations were $-15°$, 0°, 15°, and 30°, and the azimuth spacing was 15°.

We generated recording set *C (TALK)* from three different talkers which were reading newspaper articles in the anechoic room at different positions. The recordings had durations between 30 and 60 s, the azimuthal directions were 20°, 90°, 170°, 180°, and 225°.

For set *D (NOISE)*, we made recordings from six different noise environments: A cafeteria, a train station concourse, a metal workshop with several running machines, the interior of a car and two outdoor environments with traffic noise[5].

[4]The choice of ITE hearing aids has implications for the obtained data. Hearing aid microphones possess a smaller frequency range than normal-hearing humans, and have a relatively high level of microphone noise. The decisive advantage of this recording technique was, however, that the microphone position can be reproduced much more precisely than with conventional tube microphones.

[5]Further details about the recordings are described in Chapter 2, page 27.

3.2.3. Generation of *A Priori* Statistics

The *a priori* statistics, or reference data, were generated by mixing each of the 430 signals from set A (REF) with each of the noise recordings from set D (NOISE) at several SNRs. The algorithm computed IPD and ILD according to (3.3) and (3.4) and recorded their histograms as $\tilde{p}(\Delta_b|\lambda)$. In order to reduce the amount of data, the histograms were compressed and smoothed by a hierarchical cluster analysis using Ward's method (Kopp, 1978; Ward, 1963). By combining similar histograms into an average representant, this method allowed for a reduction of histogram data by a factor of 70 and reduced the computational complexity by the factor of 6.

3.2.4. Error Measure for Direction Estimates

Performance was tested with signals from recording set B (TEST), and 1250 estimates per direction. To evaluate the accuracy of the azimuth estimate, the standard deviation of the azimuth error was computed. To quantify the quality of elevation estimates, we evaluated the number of front-back confusions. The following criterion decided whether an estimate was a front-back confusion: The estimated direction was mirrored at the frontal plane, and the two angles between the estimated direction and the true direction, and the mirrored estimated direction and the true direction were compared. If the deviation of the mirrored estimate was at least 15° smaller, the estimate was counted as front-back confusion. In noise environments, a part of the time series of estimates correspond to noise sources that have at these instants a higher level than the target signal. Such estimates corresponding to the "wrong" source increase the error measure, although they may give the "right" position of this noise source; The test scheme used here for global performance assessment cannot distinguish them from "wrong" estimates of the "right" sound source.

3.2.5. Demixing of Sound Sources

In this subsection, we assume N_V directional sound signals (e.g., voices) to be present, which we will denote by the index v. The directional dependence of the HRTFs for several directional sources is indicated with the parameter $\lambda(v)$, and $Y_v(f)$ denotes the individual free-field spectrum of source v; $F_r(f)$ and $F_l(f)$ denote again the short-term spectra of the right and left ear signals for the frequency band f. With the estimated directions, we can write the filtering of the sound sources with directions $\lambda(v)$ by the corresponding left and right HRTF $H_{f,l,\lambda(v)}$ and $H_{f,r,\lambda(v)}$ as a linear operation:

$$\begin{pmatrix} F_r(f) \\ F_l(f) \end{pmatrix} = \begin{pmatrix} H_{f,r,\lambda(1)} & H_{f,r,\lambda(2)} & \dots & H_{f,r,\lambda(N_V)} \\ H_{f,l,\lambda(1)} & H_{f,l,\lambda(2)} & \dots & H_{f,l,\lambda(N_V)} \end{pmatrix} \begin{pmatrix} Y_1(f) \\ Y_2(f) \\ Y_3(f) \\ \vdots \\ Y_{N_V}(f) \end{pmatrix} \tag{3.8}$$

This, using $\vec{\lambda} = (\lambda(v))$, $\vec{Y}(f) = (Y_l(f), Y_r(f))$, and $\vec{F}(f) = (F_l(f), F_r(f))$, can be written as

$$\vec{F}(f) = \mathbf{H_f}(\vec{\lambda})\vec{Y}(f). \tag{3.9}$$

If $\vec{F}(f)$ and the sound source directions $\lambda(v)$ are known, the left and right HRTFs $H_{f,l,\lambda(v)}$ and $H_{f,r,\lambda(v)}$ can be looked up.

In the case that N_V equals 2, matrix inversion of $\mathbf{H_f}(\vec{\lambda})$ demixes the original sound sources. Because the sound source directions change only slowly in time, the first two maxima of the long-term average of the *a posteriori* values $p(\lambda|\vec{\Delta}_k)$ serve to estimate the directions. In this case, demixing the sources is equivalent to a two-microphone adaptive beamformer with perfect adaptation.

The inversion of $\mathbf{H_f}(\vec{\lambda}))$ poses two difficulties: First, the HRTFs themselves are not available from the measured data. Second, the coefficients of $\mathbf{H_f}(\vec{\lambda})^{-1}$ depend on the differences of the HRTF coefficients. The coefficients of different directions are frequently very similar, and for a correct matrix inversion they need to be known with relatively high precision. We solve both problems by separating the HRTF coefficients in an interaural component I_f, which is the interaural transfer function (ITF), and an "ear-independent" component G_f. With the ITF defined as

$$I_{f,\lambda(v)} = H_{f,r,\lambda(v)} / H_{f,l,\lambda(v)},$$

and

$$G_{f,\lambda(v)} = H_{f,r,\lambda(v)} / \sqrt{I_{f,\lambda(v)}},$$

the HRTF coefficients can be written as

$$H_{r,f,\lambda(v)} = G_{f,\lambda(v)} \sqrt{I_{f,\lambda(v)}} \tag{3.10}$$

$$H_{l,f,\lambda(v)} = \frac{G_{f,\lambda(v)}}{\sqrt{I_{f,\lambda(v)}}}, \tag{3.11}$$

and thus, for $N_V = 2$,

$$\vec{F}(f) = \begin{pmatrix} I^{\frac{1}{2}}_{f,\lambda(1)} G_{f,\lambda(1)} & I^{\frac{1}{2}}_{f,\lambda(2)} G_{f,\lambda(2)} \\ I^{-\frac{1}{2}}_{f,\lambda(1)} G_{f,\lambda(1)} & I^{-\frac{1}{2}}_{f,\lambda(2)} G_{f,\lambda(2)} \end{pmatrix} \vec{Y}(f) \tag{3.12}$$

$$= \begin{pmatrix} I^{\frac{1}{2}}_{f,\lambda(1)} & I^{\frac{1}{2}}_{f,\lambda(2)} \\ I^{-\frac{1}{2}}_{f,\lambda(1)} & I^{-\frac{1}{2}}_{f,\lambda(2)} \end{pmatrix} \begin{pmatrix} G_{f,\lambda(1)} Y_1(f) \\ G_{f,\lambda(2)} Y_2(f) \end{pmatrix}. \tag{3.13}$$

Eq. (3.13) means that apart from the interaural components, the HRTF cause a spectral distortion to the filtered signal. To apply the inverse filtering, this distortion component was simply neglected, assuming $G_{f,\lambda(v)} = 1$. Because the human ear is relatively insensitive to constant or slowly changing spectral distortions, this does not degrade the quality of the separated signal seriously. The expected values of the reference statistics of the interaural parameters which were measured without noise deliver the values for $I_{f,\lambda(v)}$. Expected values of the IPD were computed as defined in Appendix B.2.

3.2.6. SNR Evaluation

To evaluate the results of demixing of two sources, SNR improvements were computed. The SNR improvement is defined as the difference between the SNR of the processed, demixed signal, and the SNR of the unprocessed signal mixture. Because the coefficients in the mixing matrix $\mathbf{H_f}(\vec{\lambda})$ above are taken from the ITF in substitution for HRTF coefficients and $G_{f,\lambda(v)}$ are assumed to be 1, the demixed sources are linearly distorted in comparison to the signals recorded by the hearing aid microphones. This linear distortion is usually not audible, but would enlarge the error signal. Therefore, the signals, which served as "original signal" reference in the SNR evaluation, were obtained by filtering the separate originals, as they were recorded from the microphones, with the same coefficients as the mixture, but without the interfering voice.

3.3. Results

This section describes three series of experiments and their results. In the first series, the objective was to characterize the performance of the sound localization algorithm in several noise environments. The second series evaluated the performance and robustness of the algorithm in the presence of concurrent talkers. The third examined the separation of concurrent talkers.

3.3.1. Localization of Single Sound Sources in Noise

Experiment 1: Effect of Frequency Integration

We analyzed the effect of the frequency integration of *a posteriori* probabilities by combining probabilities for several numbers of frequency bands according to Eq.(3.7), summing the values corresponding to the same azimuth, and plotting the resulting *a posteriori* probabilities as a function of azimuth.

Figure 3.2 demonstrates the effect of the frequency integration of probabilities, showing in panels (a) to (e) multiplied probabilities for several degrees of frequency integration. The maxima are still ambiguous and do not match the direction precisely, but become much sharper with an increasing number of combined frequency bands. Panel (f) of Fig. 3.2 shows the resulting probability distribution for the combination of both ILD and IPD values across all frequency bands; a single maximum value very close to one results at the actual source position of -45° azimuth. For high SNRs, this is a typical result both for azimuth and elevation. For lower SNRs, more uncertainty for directions on the same cone of confusion remains, leading sometimes to front-back confusions in the estimates.

Experiment 2: Generality of *A Priori* Data

Because the Bayesian estimation procedure requires *a priori* knowledge of the statistics of the parameters in noise as "training" data, it could be affected adversely if the actual noise environment does not match the environment which served to generate the reference statistics. Therefore, the algorithm was tested in different noise environments with the signals from set B (TEST), each mixed with a signal from set D (NOISE). For each of the 96 directions, a test signal segment of 10 s duration was evaluated, resulting in 1250 estimates per direction.

Fig. 3.3 gives a decision histogram of tested and estimated directions as a gray-scale-coded frequency count. Here, the reference condition as well as the test condition was speech in cafeteria noise at 15 dB SNR. The layout of the plot is the same as in Fig. 2.6 in Chapter 2; Fig. 2.7 on page 47 explains the possible pattern of errors: Decisions plotted on lines parallel to the main diagonal indicate elevation confusions, whereas perpendicular diagonal lines indicate places of front-back confusions.

An example for the localization performance for a non-matching reference situation shows Fig. 3.4, where the reference condition was the train station noise at 15 dB SNR, while the test condition was cafeteria noise. While the pattern of

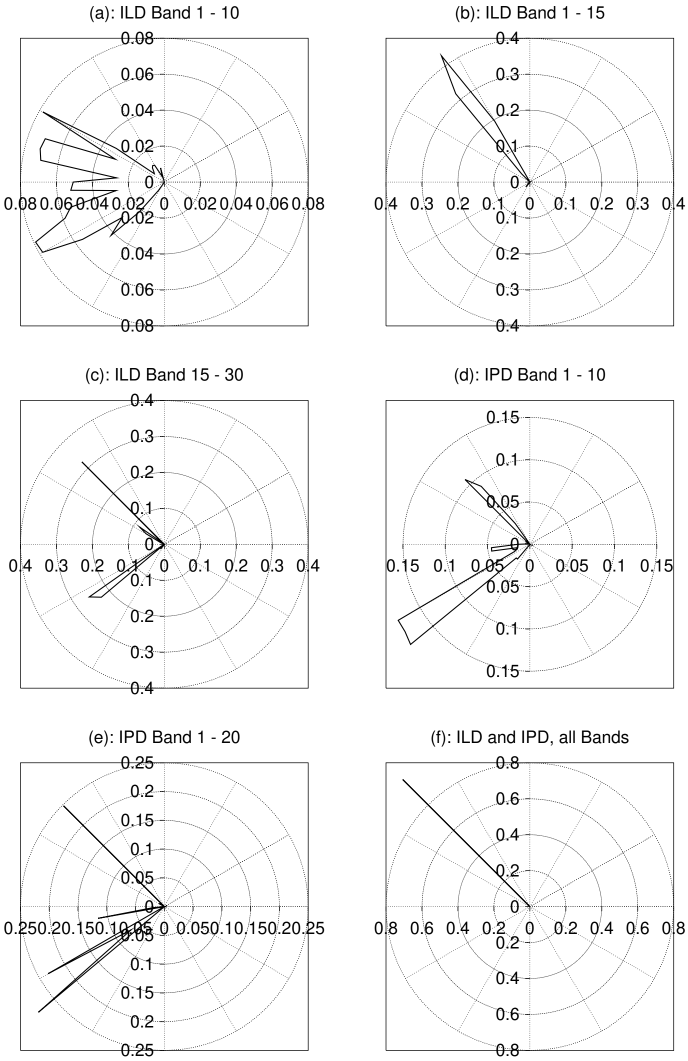


Figure 3.2.: Effect of frequency integration: *A posteriori* probabilities derived from ILD and IPD, integrated across selected frequency bands. From left to right and from bottom to top: (a) ILD, band 1-10 (48-537 Hz); (b) ILD, band 1-15 (48-878 Hz); (c) ILD, band 15-30 (830-2980 Hz); (d) IPD, band 1-10 (48-537 Hz); (e) IPD, band 1-20 (48-1367 Hz); (f) ILD and IPD, all bands (48-7763 Hz). Panel (f) shows a single maximum at the true azimuth position of -45°.

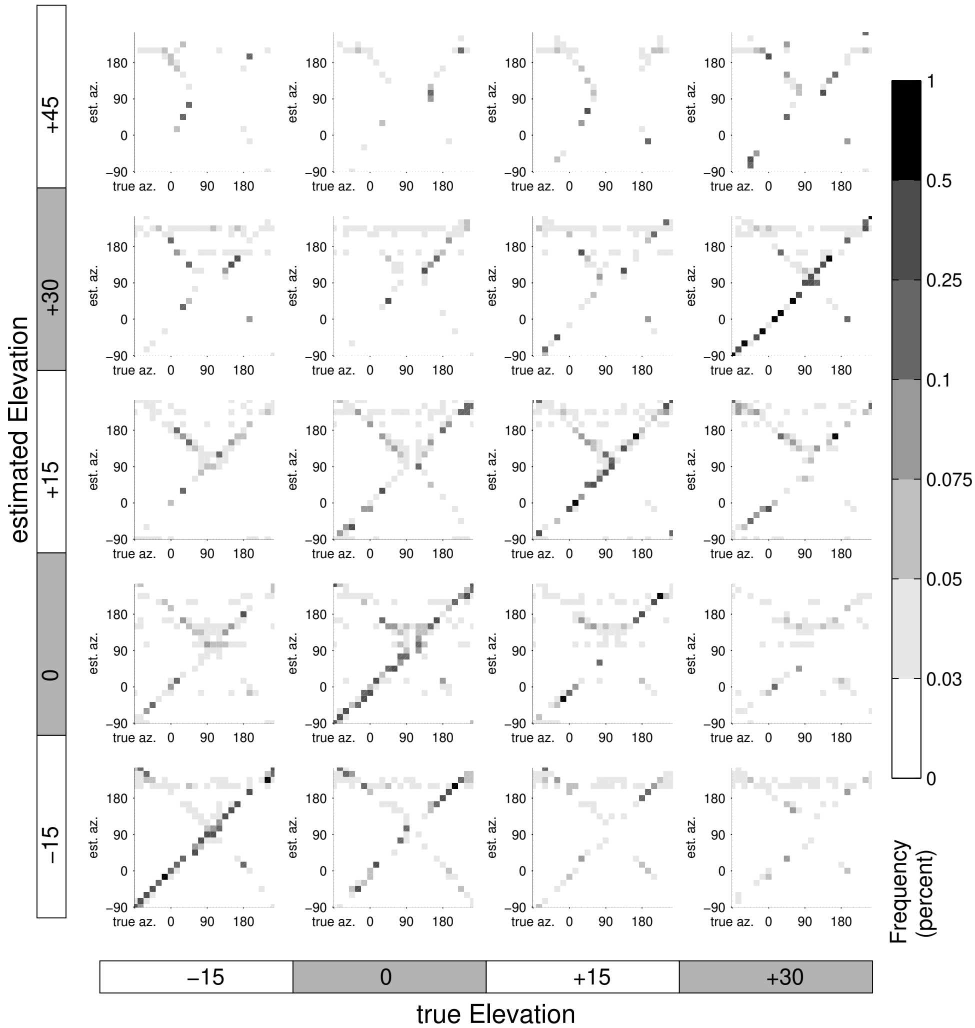


Figure 3.3.: Decision histogram of localization estimates for cafeteria noise, at an SNR of 15 dB. The y-axes represents the detected direction and the x-axes the real direction of the target. Plotted is the estimate of the probability that a specific direction is detected as the most likely direction, given the real direction, as a gray-scaled frequency count of localizations. Each box in the plot gives real and estimated azimuth for a given combination of real and estimated elevation. Explanations to possible error patterns see Fig. 2.7.

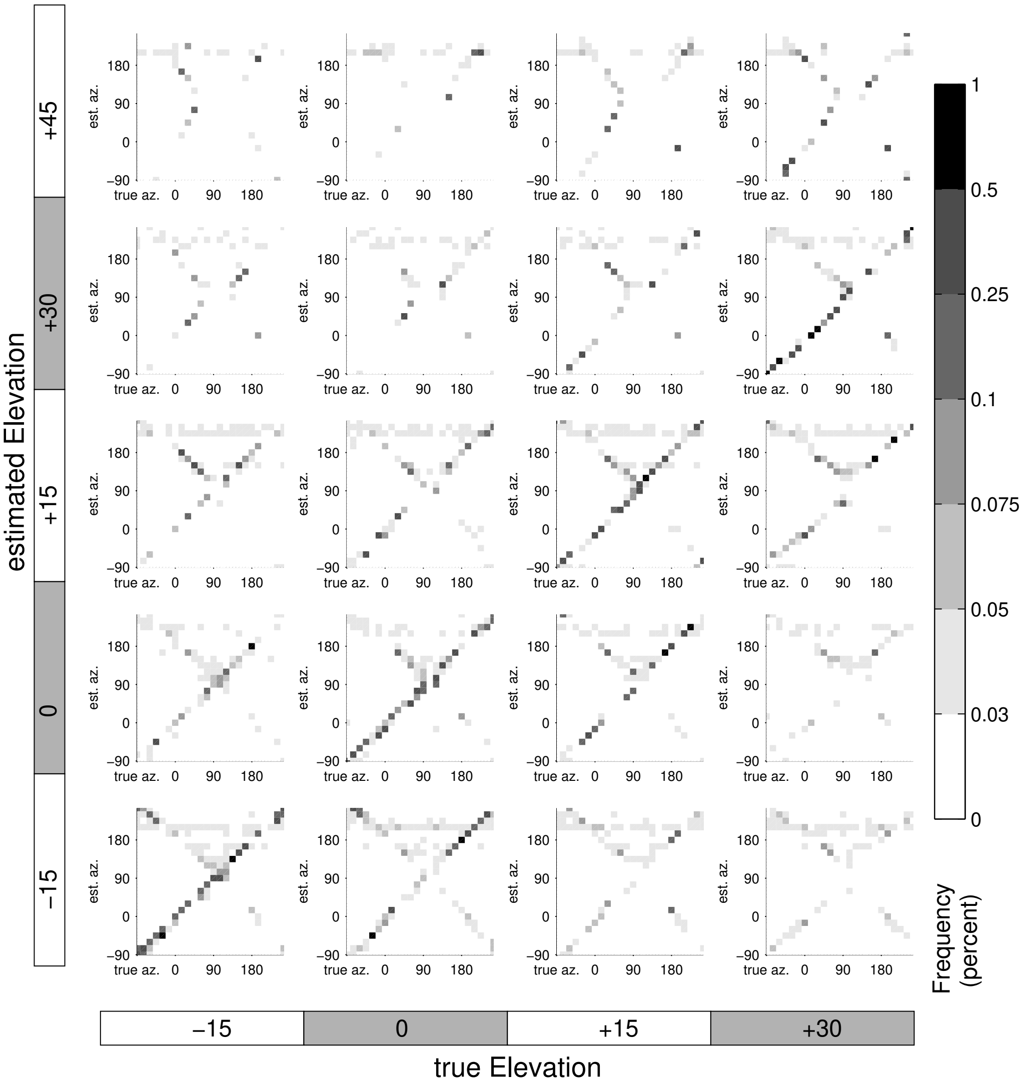


Figure 3.4.: Histogram of localization estimates in cafeteria noise, at an SNR of 15 dB. Histograms generated from the train station concourse noise environment at 15 dB SNR were used as 'trained' *a priori* data. Axes and grey scale code same as in Fig 3.3.

trained noise	C	S	M	T1	T2	A
tested noise						
Cafeteria	34.0	36.7	37.7	37.0	**43.7**	**47.1**
Station concourse	35.2	36.8	37.8	37.7	**42.0**	**49.8**
Metal workshop	36.3	38.4	37.8	36.7	**42.1**	**45.5**
Traffic 1	35.6	37.5	37.5	36.1	**43.5**	**46.5**
Traffic 2	**40.9**	**41.4**	**44.0**	**49.2**	34.0	**44.6**
Automobile (inside)	**33.3**	**35.0**	**35.3**	31.1	**35.3**	30.1

Table 3.1.: Percentage of front-back confusions for different noise environments in training and operation stage. The SNR is 5 dB in each case. Percentages which are higher than 5 % above the minimum of the row are printed in boldface.

direction estimates is very similar in both conditions, it can be seen that, e.g. for the elevation of +15°, front-back confusions occur slightly more frequently.

For a global quantitative comparison, the percentage of front-back confusions was calculated for each possible combination of noise and reference condition, for 96 tested directions and 1250 estimates. The result is given in Tab. 3.1, which contains a matrix of noise conditions containing speech and other noise sources, for 36 training / test condition pairs. Numbers in which the percentage of front-back confusions is more than 3 % higher than the minimum for this reference are printed in boldface. The environments including speech noise, as well as the metal workshop and one traffic noise environment, have references whose exchange causes only small increases in the percentage of front-back confusions. The relationship between the frequency of front-back confusions and the azimuth error is very similar to the results from Chapter 2, Fig. 2.8 on page 48, therefore it is not repeated here.

3.3.2. Localization of Concurrent Talkers

Experiment 3: Localization with a Single Interfering Talker

To examine the effect of an interfering talker, one of the recordings from set C (TALK) was mixed with the signals from recording set B (TEST), and localization was again tested for 96 directions and 1250 estimates per direction. The localization algorithm yields two distinct maxima for the *a posteriori* values for two interfering speakers, even if the levels of the speakers differ. Figure 3.5 illustrates an example of MAP estimates for a talker from 170° azimuth and 0° elevation. The azimuth of the interfering talker appears as a horizontal line at the correct azimuth value of 170° for the elevations of 0°, 15°, and 30°. Estimates of unrelated directions occur rarely, but front-back confusions and elevation errors have a higher frequency than without interference. In this series of tests, the voice with the highest level always reached a percentage of 'most likely direction' between 30 % and 50 %, and the percentage for the voice with the lowest level was typically about 10 % to 20 %. Excluding front-back confusions and elevation errors, *a posteriori* values for unrelated directions were almost always less than 2 %.

Experiment 4: *A Posteriori* Probabilities in Presence of Two Concurrent Talkers

The result of experiment 3 suggests that the algorithm can track several concurrent talkers. To examine multi-talker situations further, we evaluated *a posteriori* probabilities for all azimuths for mixtures of two voices from recording set C (TALK). The rationale for examining *a posteriori* probabilities was to examine if, and to which degree, interactions between the estimated sources occur, which might not be visible from MAP estimates. For each azimuth, the probabilities from Eq. (3.6) were smoothed with a first-order low pass filter with a time constant of 2.5 s and summed across elevations:

$$p(\alpha|\vec{\Delta}_k) = \sum_{\phi} p(\alpha, \phi|\vec{\Delta}_k)$$

Fig. 3.6 shows, as a typical example, the result as contour plot of log-scaled values of $p(\alpha|\vec{\Delta}_k)$ as a function of azimuth and time. A mixture of two talkers at 20° and 170° azimuth served as the input signal. Although the outermost contour line, corresponding to an overall probability of 0.03 % per 5° azimuth, occasionally extends across a fairly large azimuth range, for example between 4.4 and 5.2 s, the innermost contour line, which corresponds to a probability density of 3 % per 5° azimuth, remains always very close to the actual direction of the corresponding sound sources.

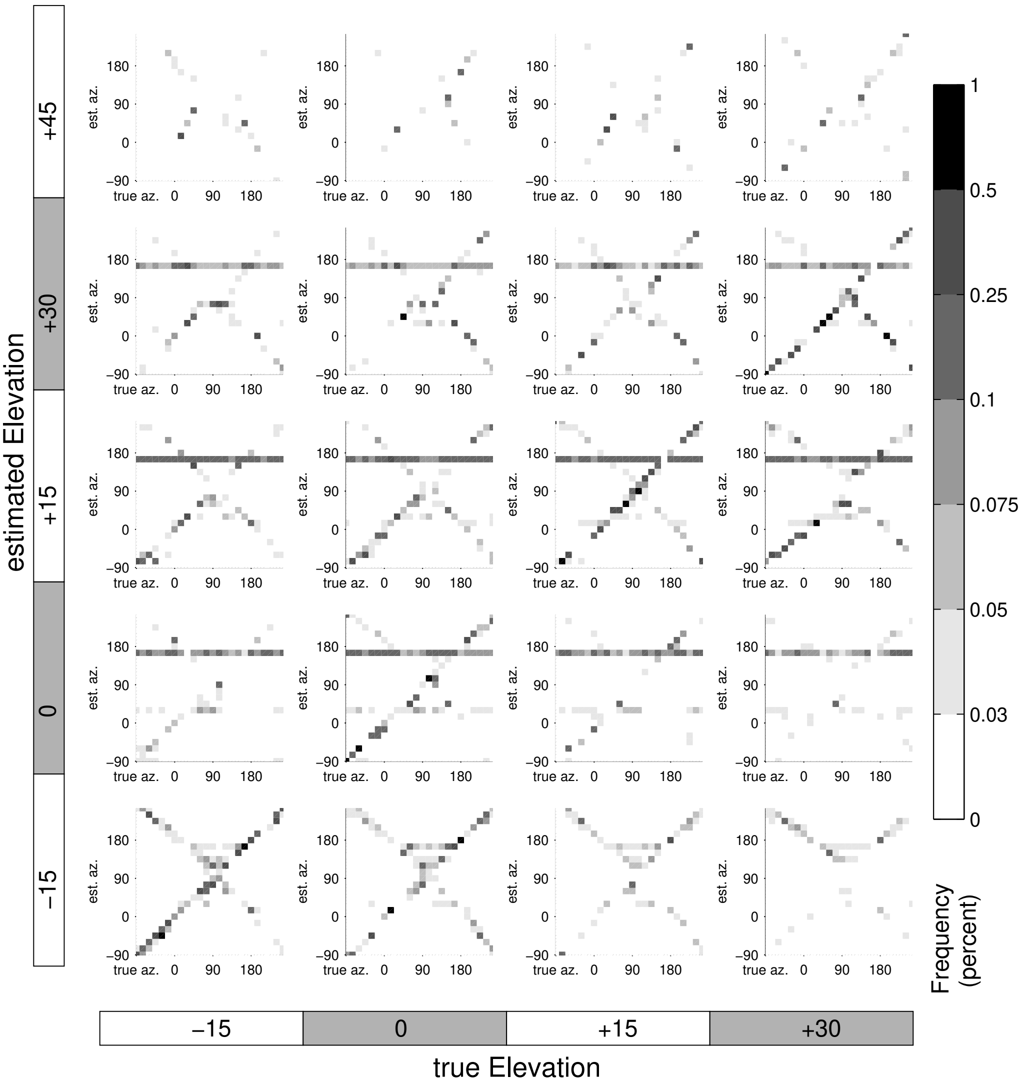


Figure 3.5.: Histogram of localization estimates for one interfering speaker at about 170° azimuth and approximately 0° elevation, and an SNR of 15 dB. Axes and gray scale code are the same as in figure 3.3. The dark horizontal lines at about 170° azimuth and elevations of 0°, 15°, and 30° correspond to the position of the interfering talker.

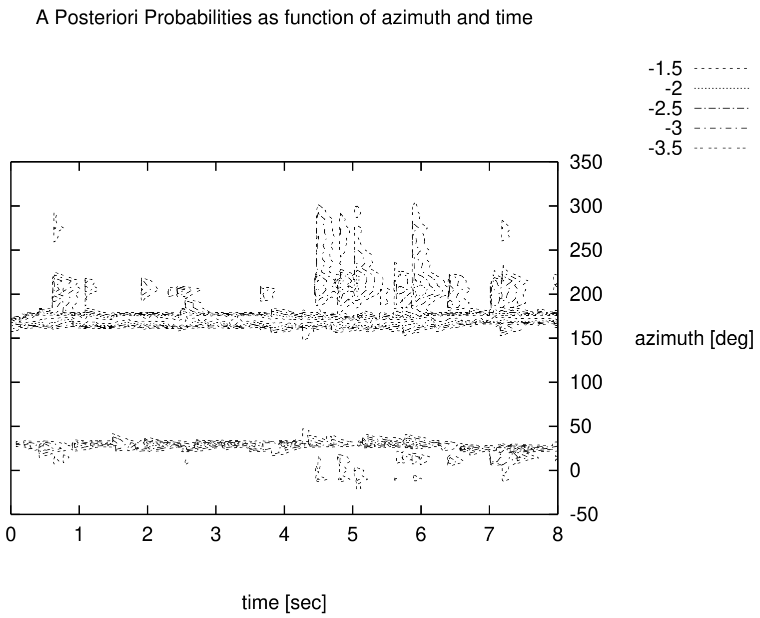


Figure 3.6.: Contour plot of logarithmically scaled *a posteriori* probability densities $p(\alpha|\vec{\Delta}_k(t))$ for the azimuth directions as a function of time. The signal is a mixture of two talkers with true azimuth angles of 20° and 170°, and an SNR of 0 dB. The contour lines indicate probability densities between 0.31 % and 3 % ($10^{-1.5}$ and $10^{-3,5}$) per 5° azimuth.

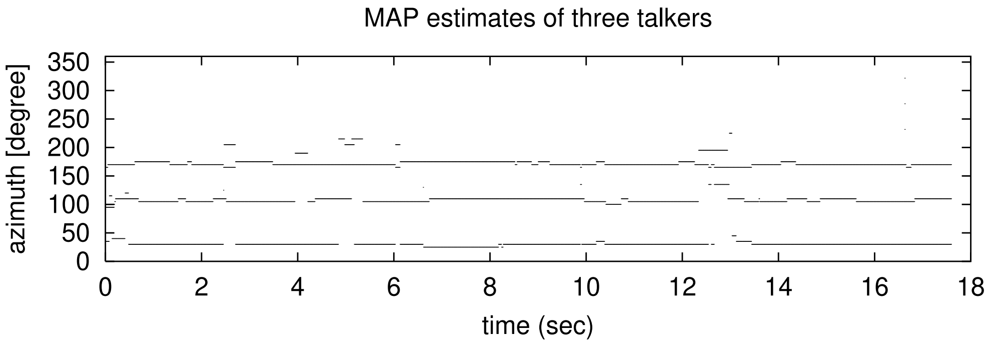


Figure 3.7.: Azimuths of the three directions with the three largest a posteriori probabilities per time step, as a function of time for a mixture of 3 talkers at about 20°, 100° and 170° azimuth.

Experiment 5: MAP Estimates of Sound Source Azimuth for Three Concurrent Talkers

Experiment 5 evaluated the usefulness of MAP estimates for sound source localization with a larger number of talkers. The algorithm was tested with a mixture of three talkers with azimuth coordinates of about 20°, 100° and 170°. For each time step, three maximum *a posteriori* (MAP) values were selected as an estimate for the azimuths of the incoming sources. Fig. 3.7 shows them as a function of time. The algorithm tracks three talkers quite reliably; for about 90 % of the time, the error of the estimate is in the order of 5° or less for all three sources. However, it does not reach the robustness demonstrated here in every case. When pronounced level differences between the sources exist, if fails frequently to identify the weaker sources. Further, from the seven recordings with concurrent speech (recording set C, TALK), two recordings could not be localized well, in contrast to the high reliability of localization for signal set B (TEST). A possible reason is that slightly changed adjustments to the recording equipment, or microphone placement, caused level differences of a few decibel, which would be enough to defeat the statistical pattern-matching procedure.

3.3.3. Experiment 6: Demixing Two Sound Sources

In this experiment, demixing of sound sources based on the on-line localization estimates and inversion of HRTF coefficients was applied to mixtures of two and three superimposed voice recordings (recording set C, TALK and B, TEST) with several spatial configurations and several SNRs.

To demonstrate the applicability of the localization algorithm for de-mixing sound sources, as discussed in section 3.2.5, the algorithm selected the ITF coefficients for the two estimated directions with the highest *a posteriori* probability, after filtering them with a time constant of 2.5 s. From these coefficients, the de-mixing matrix was calculated by algebraic inversion. Listening to the results reveals that for a mixture of two talkers, after a short initial convergence phase of about 0.2 s duration, the two voices are separated reliably and with high suppression. During the convergence time, frequently a short burst of noise is audible. For a mixture of three voices, the algorithm cancels out one of the three voices, and switches between the two remaining sources according to the highest level on a time scale of one to two seconds. In both cases, the processed speech signals are clearly intelligible. As an example for the results, sound files are available online[6] (Nix, 2005).

For mixtures of two voices, we computed SNR improvements from unmixed signals of 20 s duration as described in section 3.2.6. Table 3.2 lists them. The controlled parameters in the table were, in column two and three, the azimuths of the sound sources, and in column four the level ratio of the input mixture at the ear canal (left and right channels were added to compute this level). Because of the directionality of the HRTF, this level ratio does not correspond directly to the ratio between the mixture and the original, directionally filtered voices; these SNRS are tabulated in columns 5 and 6. Columns 7 and 8 give the SNR of the demixed voices, and column 9 and 10 the improvement between mixture and filtered voices.

The quality of the separated voices was assessed by informal listening. Usually, high SNR values indicate a high level of separation. However, despite the difference in the evaluated SNRs, the signals in trial 6 have about the same quality as in trial 7, for example. Sometimes, a large SNR improvement coincides with a poor audio quality in the demixed signals, such as blasting, ringing or aliasing effects. This can be explained by the fact that the filter coefficients obtained from $\mathbf{H_f}(\vec{\lambda})^{-1}$ (Eq. (3.9)) are not restricted. Therefore, for single frequencies, excessively high gains can occur, with the result that the impulse response

[6] http://medi.uni-oldenburg.de/demo/demo_voice_unmixing.html

trial	azimuth voice 1	azimuth voice 2	ear[a] canal	SNR (dB) SNR mixture[b] ch. A	SNR mixture[b] ch. B	SNR proc.[c] ch. A	SNR proc.[c] ch. B	improvement (dB) ch. A	improvement (dB) ch. B
1	-45	170	0	-15.0	10.6	-4.6	11.8	10.5	1.1
2	-45	170	10	-14.7	14.5	5.4	21.8	20.1	7.2
3	-45	170	20	-14.6	14.8	15.4	31.8	30.0	17.0
4	20	90	0	-2.3	1.0	11.8	8.6	14.1	7.5
5	20	90	-10	1.6	-2.1	17.3	18.7	15.7	20.6
6	-45	90	0	-16.4	9.0	-3.8	12.3	13.1	2.4
7	-45	20	0	-14.8	12.8	0.3	23.3	15.1	10.6

[a]Level ratio of the input signals at ear canal entrance (powers of left and right channel added).
[b]SNR of the mixed input relative to the filtered voices without mixing
[c]SNR of the filtered output relative to the filtered voices without mixing

Table 3.2.: SNR improvements for demixed voices for various configurations

of the filtering operation can become so long that aliasing occurs in the inverse FFT operation. This can be prevented by restricting the coefficients of $\mathbf{H_f}(\vec{\lambda})^{-1}$ by windowing in the temporal domain. In the audio examples, sometimes noise bursts are also audible when the direction estimates change suddenly, e.g., because of small movements of the talkers. In the scope of this work, further optimization of the beamforming operation was set aside.

3.4. Discussion

3.4.1. Comparison to Other Approaches for Sound Localization

The algorithm evaluated here has similarities with the algorithms which were proposed by Duda (1997) and Nakashima *et al.* (2003); both algorithm use binaural parameters, which are evaluated in the frequency domain, as well as a distance metric between the observed values of the interaural parameters and the reference values. However, different from Duda's algorithm, the approach proposed here takes the environmental noise explicitly into account by calculating a probabilistic distance metric from measured data. In contrast to this, many other approaches (Lyon, 1983; Lindemann, 1986a; Bodden, 1996b; Schauer *et al.*, 2000; Roman *et al.*, 2003) combine values derived from the narrow-band interaural cross-correlation function, or from related measures, motivated by the model from Jeffress (1948). If additional sound sources are present, or if the incoming sound is nonstationary, this can cause ambiguities in the time difference (see Lyon, 1983, p. 1151, and the examples in Appendix D).

Here, the interaural timing information is evaluated in the frequency domain in the form of the interaural phase difference. The approach takes the ambiguity of narrow-band interaural timing parameters explicitly into account. A similar approach to use ambiguous information on interaural timing by high-frequency integration has been proposed by Liu *et al.* (2000) in a 'stencil filter' method. In contrast to the approach described there, the integration of probabilities derived from ILD and IPD explicitly includes the effect of environmental noise.

Several concurrent talkers can be localized because voice signals do not overlap completely in a time-frequency representation. This is sufficient to generate local maxima in the estimated PDF $p(\lambda|\vec{\Delta}_k)$.

The strategy of estimation of the direction of arrival (DOA) and separation by linear operations is related to approaches of blind source separation (Parra and Spence, 2000; Anemüller and Kollmeier, 2003). However, the convergence times needed by the Bayesian approach, being in the range of about 0.2 s, are quicker than customary approaches for blind source separation, which often require convergence times of about one second (Parra and Spence, 2000). We explain the advantage by the fact that the algorithm is not completely 'blind', but exploits properties which are present similar in many common real-world signals (Chapter 2, Tab. 3.1). Many algorithms for BSS assume time delays and level differences which are constant across frequency, so that performance might detoriate under the influence of HRTFs. In contrast, the proposed algorithm uses HRTFs to resolve ambiguities of directions. Several authors proposed to perform sound localization by means of neural networks, which use 'trained' reference data as well (Datum *et al.*, 1996; Isabelle *et al.*, 1998; Schauer *et al.*, 2000). However, such networks seem to be difficult to train for robustness in high levels of noise.

The current realization of the algorithm still has two disadvantages. The first is that it requires individual HRTFs. A second shortcoming is that the algorithm does not take the fact into account that sound sources usually do not perform abrupt changes of directions. In the presence of multiple sources, implausible changes in the estimated directions can occur, see Fig. 3.7. Therefore, it would be a promising extension of the algorithm to include a statistical model of source movements. Statistical algorithms as, e.g., described in Chapter 4 of this work, or by Ward *et al.* (2003), would not use only the MAP values, but could make use of the complete probabilistic information contained in the estimated $p(\lambda|\vec{\Delta})$. Further improvements can be made by including models of head rotations as well. Wallach (1940) observed in psychophysical experiments that such movements help to resolve ambiguities along the cones of confusion. This is supported by more recent research (Perret and Noble, 1997; Hofman *et al.*, 1998; Wightman and Kistler, 1999).

3.4.2. Use of Sound Localization for Directional Filtering

The algorithm has several possible technical applications. One possible application is in future binaural hearing aids (Greenberg and Zurek, 2001; Soede *et al.*, 1993; Widrow, 2000). Kompis and Dillier (1994) and Korompis *et al.* (1995) evaluated adaptive beamforming methods for hearing aids; Shields and Campbell (1997) examined array processing in independent frequency subbands, and Hoffman and Buckley (1995) investigated robust methods for adaptive beamforming. Steerable microphone arrays controlled by an estimation of direction of arrival (DOA) were developed by Wang *et al.* (1996). Greenberg *et al.* (2003) found improvements of the speech reception threshold (SRT) of 7-9 dB compared to omnidirectional microphones as a reference condition. Approaches motivated by models of binaural processing have been proposed and evaluated by Peissig (1992); Bodden (1996a); Wittkop *et al.* (1997); Wittkop (2001); Liu *et al.* (2001); Roman *et al.* (2003).

Further, the Bayesian approach can be applied in small general-purpose microphone arrays. Such small microphone arrays are of interest for many areas of speech technology, e.g., automatic speech recognition for portable computers or mobile telephony in cars (Compernolle, 2001). They can also be combined with more sophisticated beamforming schemes, as, e.g., discussed in Brandstein and Ward (2001). The Bayesian approach used here is not restricted to two-channel signals, but can easily be adapted to three or more microphones which have different directional characteristics. In such a multi-microphone application, *a posteriori* probabilities derived from the cross spectra of each microphone pair, computed in analogy to Eqs. (3.3) and (3.4), can simply be multiplied to yield the overall probability for each direction, as defined by Eq.(3.7).

3.5. Summary

Bayesian classification of interaural phase and level differences can extract directional information from binaural recordings. Although these parameters have a high degree of fluctuations and possess ambiguities, they allow a robust sound localization by integrating probabilities across frequency (Fig. 3.2). The localization algorithm presented here works well down to an SNR of about 5 dB. It is also able to distinguish front and back directions, although this capability is more affected by noise. The performance of the algorithm for sound localization usually depends only weakly on the noise environment used to generate *a priori* parameters (Fig. 3.3 and Fig. 3.4, Tab. 3.1). Apart from localization in distributed noise, the algorithm is able to estimate the directions of concurrent talkers (Fig. 3.5); the *a posteriori* probabilities of the PDF $p(\alpha|\vec{\Delta})$ computed from the interaural parameters reflect well the probable directions of the talkers (Fig. 3.6). The computed MAP values can identify the directions of up to three concurrent talkers (Fig. 3.7). For a two-source situation, demixing of the sound sources is possible by inversion of a matrix composed of the quotients of the left and right HRTF. For two talkers, this allows to separate the sound sources with adaptation times in the order of 0.2 s, reaching SNR improvements of up to 30 dB (Tab. 3.2). Finally, the algorithm has been implemented as a computationally inexpensive on-line algorithm. The implementation used here, which is based on the Python script language, is able to run in real time on a PC with dual AMD Opteron 246 CPU with a 2 GHz clock frequency.

Acknowledgments

We are grateful to Birger Kollmeier for his substantial support and contribution to this work. We thank the members of the Medical Physics Group, especially Thomas Wittkop, Stephan Albani, and Jörn Otten, for providing technical support and for important discussions, and also Jürgen Peissig, who contributed important early ideas.

This work was supported by DFG (Graduate Programme Psychoacoustics), the German Ministry for Science and Education (BMBF, FDG 01 VJ 9305), and the HörTech Center of Excellence with funding from the German Ministry for Science and Education.

4. Combined Estimation of Spectral Envelopes and Sound Source Direction of Concurrent Voices by Multidimensional Statistical Filtering[1][2]

The logical principles involved in the recognition of speech seem to require that the brain have a vast "store" of probabilities, or at least of probability-rankings. *Colin E. Cherry (1953)*

Abstract

A key question for speech enhancement and simulations of auditory scene analysis in high levels of nonstationary noise is how to combine principles of auditory grouping, and integrate several noise-perturbed acoustical cues in a robust way. We present an application of recent on-line, nonlinear, non-Gaussian multidimensional statistical filtering methods which integrates tracking of sound-source direction and spectro-temporal dynamics of two mixed voices. The results show that the algorithm tracks sound source directions very precisely, separates the voice envelopes with algorithmic convergence times down to 50 ms, and enhances the signal-to-noise ratio (SNR) in adverse conditions, requiring high computational effort. The approach has a high potential for improvements of efficiency and could be applied for voice separation and reduction of nonstationary noises.

[1] A modified version of this chapter was submitted to the *IEEE Transactions on Speech and Audio Processing*.

[2] Preliminary results on sequential Monte Carlo Methods for envelope tracking have been presented in Nix *et al.* (2003a) and Nix *et al.* (2003b).

4.1. Introduction

A key challenge for computational models and simulations of auditory scene analysis is the question of how several acoustical features belonging to various mixed sound sources can be combined and grouped together to form "acoustical objects," according to their original sources. The problem becomes especially difficult when the sound sources mask each other partially and additional noise perturbation occurs. The task of feature grouping is a special case of the "binding problem" in neural information processing, and solving the task might provide general insights into how the brain combines several features from one or different sensory modalities. Simulating auditory grouping is an important intermediate step to extract useful information from everyday complex listening environments (Yost, 1991).

The particular acoustic features and cues used by the human auditory system have been investigated for about fifty years (Cherry, 1953; Cherry and Wiley, 1967; Bregman, 1990; Evans, 1992; McAdams and Bigand, 1993). The psychoacoustical evidence suggests that spectral dynamics of sound sources, especially common onsets of spectral components, as well as harmonicity and direction of incidence are among the most important cues for sound separation in difficult listening conditions. As a grouping cue, sound source direction also has the advantage that it helps to group sounds which are separated in time, because it changes more slowly than other features (Brown and Cooke, 1994). Bregman, Cherry and others summarized the use of these cues by the auditory system in abstract form to certain 'grouping rules', inspired by the 'Gestalt Laws' developed by psychologists (Cherry, 1959; Godsmark and Brown, 1999). One important rule concerns, for example, the significance of 'common onsets'; qualities such as high levels of spectral power densities emerging at the same time are likely to belong to the same sound source. Other important qualities are spatial location, common amplitude modulation, common frequency modulation, and harmonicity. If part of the spectral energy might be masked, an "old-plus-new principle" also applies, as auditory induction effects show (Bregman, 1990; Cooke and Ellis, 2001). This principle states that sounds that cease to be observable simultaneously with the onset of another more intense sound, probably continue to exist with little change. Many of these rules have in common that they apply to the combined spectral and temporal properties of sound sources, and that they correspond to the physical processes which generate the most sounds.

Early approaches to simulate this accomplishment of the auditory system by computational auditory scene analysis (CASA) successively implemented grouping rules by 'blackboard' systems, which attributed the acoustical features to sound sources by hierarchical schemes, that are characterized either as bottom-up processing or as top-down processing (Godsmark and Brown, 1999; Cooke and Ellis, 2001). Cooke *et al.* (1993) and Cooke (1993) demonstrated that these grouping rules are useful in tasks like separating mixtures of sounds. A similar approach was proposed by Unoki and Akagi (1999). Ellis (1996) proposed a prediction-driven scheme, which implements a layered structure that delivers successive abstractions of auditory features and expectations about their temporal development. Related to the prediction-driven schemes are model-driven approaches which originated from the extension of methods for automatic speech recognition (ASR), for example the "missing feature" approach (Cooke *et al.*, 2001b), parallel model combination (Gales and Young, 1993), and hidden Markov model (HMM)methods (Varga and Moore, 1990). A combination is the multisource decoder of Barker *et al.* (2005), which aims at integrating prediction-driven, and model-driven approaches in a statistical framework.

A central problem in models of grouping is how to select a plausible solution when the available cues allow several different interpretations of the mixed signals, leading to conflicting explanations of the state of the original sources. Grouping procedures based on a sequence of rules show the difficulty that the result will depend on the order of rule applications. Features that in principle are able to resolve the ambiguity might be considered too late. An alternative approach is 'competition of hypotheses'. Neurobiology, theory of neural networks, and theoretical physics have inspired solutions to the grouping problem which use networks of coupled oscillators (von der Malsburg and Schneider, 1986). Wang and Brown (1999) and Wang (2000) developed a similar system which segregates natural sounds successfully based on their fundamental frequency. As van der Kouwe *et al.* (2001) point out, these approaches possess the interesting property that they model a neural mechanism coined as "local excitation - global inhibition": Several hypotheses about grouping compete, and the hypothesis that matches the sensory input best will suppress other candidate solutions. Possibly, such strategies are more robust to input features of varying reliability than a scheme which is based on a fixed order of hierarchical rules.

The human auditory system is able to separate several concurrent sound sources at signal-to-noise ratios of 0 dB or below. Because of the importance of robust communication, the auditory system is also highly adapted to understand speech under such circumstances (Cherry and Wiley, 1967; Yost, 1997; Bronkhorst, 2000). Such conditions occur frequently in social gatherings, and the task to mimic this accomplishment has therefore been coined as the "cocktail party problem" (Bronkhorst, 2000). It poses extreme challenges to applications like robust automatic speech recognition or noise suppression in hearing aids. The non-stationarity of the interfering speech causes classical solutions for speech enhancement to fail (Marzinzik, 2000). In addition, deterministic, rule-based strategies of auditory scene analysis, like the ones mentioned above, lack the robustness necessary in such situations.

Technical approaches for speech enhancement, for example spectral subtraction (Lim, 1978; Boll, 1979), beamforming (Griffiths and Jim, 1982), blind source separation (BSS) and independent component analysis (ICA) methods (Bell and Sejnowski, 1995; Parra and Spence, 2000; van der Kouwe *et al.*, 2001; Anemüller and Kollmeier, 2003) have the common drawback that the assumptions necessary to estimate the noise and the target signal, which require frequently either that the noise is stationary, or that as many microphones are used as sound sources exist, are often not met. Thus, information for the reconstruction of the target signal is lacking.

Technical procedures have also applied principles corresponding to some of the grouping rules. For example, short-term envelopes of speech in different frequency bands have high cross-correlation values (Li *et al.*, 1969). Anemüller and Kollmeier (2003) proposed an algorithm to exploit this property in order to find solutions for BSS.

Summarizing the mentioned aspects, CASA in realistic environments faces the problem of robust binding of features and properties of the sound sources, the challenge of acoustical features being perturbed by highly nonstationary noise, the necessity to reconstruct masked parts of the spectral information of the sound sources, and the difficulty of estimating solutions for under-determined sets of equations (less microphones than sources).

To overcome these difficulties, alternative approaches for noise suppression employing pattern-matching, nonlinear estimation, modulation filtering, and HMM have been proposed (Ephraim *et al.*, 1989a; Boll, 1992; Kollmeier and Koch, 1994; Xie and van Compernolle, 1996; Sameti *et al.*, 1998; Strube and Wilmers, 1999). Some approaches also include aspects of model-driven scene analysis (Gales and Young, 1993; Reyes-Gomez *et al.*, 2003).

We propose to model the binding of auditory features as a multidimensional, nonlinear statistical estimation task. In this approach, the properties of undisturbed sound sources are represented by a stochastic feature vector in a multidimensional space, each coordinate representing some aspect of the sound sources. The original sound, transformed by the transfer functions of head, torso, and pinnae, leads to an input signal at the microphones. This signal is transformed to a perceptual representation in line with fundamental principles of peripheral auditory processing. Additionally to the desired sound source, we assume noise from a nonstationary source to be present. Stationary noise, for example from the microphones, may be present as well. The nonstationary noise masks partially the original sound. Therefore, the perceptual representation alone is not sufficient to derive the desired undisturbed original features unambiguously from it. As additional *a priori* knowledge, the statistics of the temporal evolution of sound source features, and the correlative dependencies between the different features are used. The task is then to estimate the time-dependent probability density function (PDF) of the sound source feature vector, and thereby the expected values of the true features of the sound sources, from the perturbed observation in a statistically optimal way. We show that the estimation task can be defined precisely in a Bayesian framework and that computational methods exist which are able to evaluate and combine several noisy features within such a framework.

To demonstrate this idea, we implemented an estimation algorithm for concurrent voices that is based on sequential Monte Carlo methods, also known as "particle filters" (Gordon *et al.*, 1993; Kitagawa, 1996; Arulampalam *et al.*, 2002). Such methods do not require assumptions regarding linearity or Gaussian PDFs, have been shown to be extremely flexible, and have been used for on-line tracking in visual scenes (Blake and Isard, 1998), sound source direction (Ward *et al.*, 2003), formant frequencies (Zheng and Hasegawa-Johnson, 2004), fundamental frequencies (Gandhi and Hasegawa-Johnson, 2004) and filtering applications (Vermaak *et al.*, 2002). We show here that these methods can integrate acoustical cues and source properties, and that they agree with the principle of "local excitation - global inhibition."

For the algorithm developed here, we assume the sounds that contribute to the mixture to be speech. Speech signals from two talkerswere used as mutually interfering original sound sources. The short-term magnitude spectra and the spatial positions of the sound sources were selected as features. The sound sources were filtered with head-related transfer functions (HRTFs) according to the sound source directions, and superimposed subsequently. The superimposed signals formed the input for the estimation algorithm, and the task was to estimate the individual magnitude spectra as well as the sound source azimuths. Throughout the following sections, vectors are denoted by arrows, like $\vec{a}$, and random variables are marked in bold face, like $\mathbf{b}$. Instances of random variables are written as b, for example.

4.2. Methods

4.2.1. General Assumptions

We summarize the assumptions used to apply the statistical method as follows:

- The incoming sound is a mixture of voices.
- Statistical *a priori* knowledge on spectro-temporal features of speech is available.
- The incoming sounds are free-field signals without reverberation. The sources can move with an angular velocity of up to 100° per second.
- The sounds are filtered by known head-related transfer functions (HRTF) (Mehrgardt and Mellert, 1977; Middlebrooks *et al.*, 1989; Algazi *et al.*, 2001), and recorded by two microphones. Thus, binaural features as present in the ear canal of humans (Wightman and Kistler, 1989c) are available.
- Characteristics of the individual voices are not known.
- We assume the number of voices to be known.
- In a mixture of several voices, each voice has about the same level.

Most assumptions have the purpose to keep the algorithm simple and to allow the examination of the main characteristics of the approach. The possible relaxation of these assumptions is discussed in section 4.4.

4.2.2. Framework of State-Space Tracking of Voices

The aim of exploiting stochastic spectro-temporal characteristics of speech requires a statistical description of the succession of speech spectra. An adequate formalism for this task is a general Markov-type state-space model, which will be used as follows.

Let $\vec{x}_k \in \mathcal{R}$ denote the *state* of all contributing sound sources at time k. This state vector describes the momentary condition of the sources in a number of coefficients, for example some representation of the spectra of two voice signals, their direction, and so on. $\mathcal{R}$ is called the state space, which is the set of all possible states of the system. In our statistical approach, we consider $\vec{x}_k$ to be an instance of the random variable $\vec{\mathbf{x}}_k$. Because of the statistical uncertainty, the PDF of the state vector $p(\vec{x}_k)$ describes all of the available information.

The PDF
$$p(\vec{x}_k|\vec{x}_{k-1}) \tag{4.1}$$
describes the *system dynamics* by quantifying the probability that a specific state $\vec{x}_k$ happens to succeed a state $\vec{x}_{k-1}$ at the previous time step. Further, the state cannot be observed directly, but only by means of an *observation* $\vec{z}_k$, which is a stochastic function of the state and is an instance of the random variable $\vec{\mathbf{z}}_k$. The observation $\vec{z}_k$ may contain some amount of uncertainty induced by the measurement or by additional environmental noise. This uncertainty is called *measurement noise*. State $\vec{x}_k$ and observation $\vec{z}_k$ are linked by the *observation statistics PDF*
$$p(\vec{z}_k|\vec{x}_k). \tag{4.2}$$
The measurement noise is reflected in the variance of $p(\vec{z}_k|\vec{x}_k)$.

The task is now to estimate the PDF of the state $p(\vec{x}_k)$ from the series of observations $\{\vec{z}_1, \vec{z}_2, \vec{z}_3, \dots \vec{z}_k\}$, $k \in \mathbb{N}$ with optimal consideration of $p(\vec{z}_k|\vec{x}_k)$ and $p(\vec{x}_k|\vec{x}_{k-1})$. Using the symbol $\vec{z}_{1:k}$ to express the series of observations, the goal is to calculate
$$p(\vec{x}_k|\vec{z}_{1:k}).$$
As, e.g., Doucet *et al.* (2001) and Arulampalam *et al.* (2002) show, the required PDF can be derived recursively by combining the Chapman-Kolmogorov equation with Bayes's Formula, which yields
$$p(\vec{x}_k|\vec{z}_{1:k}) = \frac{p(\vec{z}_k|\vec{x}_k)\, p(\vec{x}_k|\vec{z}_{1:k-1})}{p(\vec{z}_k|\vec{z}_{1:k-1})} \tag{4.3}$$
with
$$p(\vec{z}_k|\vec{z}_{1:k-1}) = \int_{\vec{x}_k \in \mathcal{R}} p(\vec{z}_k|\vec{x}_k)\, p(\vec{x}_k|\vec{z}_{1:k-1})\, d\vec{x}_k, \tag{4.4}$$
but in practice, the integral generally cannot be calculated directly because of the computational expense of the high-dimensional integration[3].

The PDF $p(\vec{x}_k|\vec{x}_{k-1})$ constitutes *a priori* knowledge on the dynamical characteristics of the system, for example how often certain combinations of succeeding

[3]Section 5.2.3 in Chapter 5 gives an assessment of the computational expense of a grid-based integration of all possible state sequences, whose number can be, for the parameters used here, as large as 10^{230}, and compares this to an evaluation using hidden Markov Models.

short-term spectra occur. $p(\vec{z}_k|\vec{x}_k)$ contains information about the connection between state and observation and about the precision with which a certain combination of source states will lead to a certain observation; for example, it will reflect microphone noise, background noise, and quantization effects.

In our application, the uncertainty of the state $\vec{\mathbf{x}}_k$ is caused first by the non-stationarity of the interfering voices, second by the stochastical uncertainty of the observation, and third by the fact that estimating the individual source signals with unknown directions from a mixture of sounds is an under-determined problem. The observation $\vec{z}_k$ is derived from the linear combination of the signals characterized by spectra and direction. For a number of voices N_V larger than one, the dimensionality of the observed superimposed signal which leads to the observation is smaller than the dimensionality of the state. In consequence, the transformation from $\vec{x}_k$ to $\vec{z}_k$ is not invertible.

A multidimensional statistical filtering algorithm for separating concurrent voices by approximating Eq. 4.3 has to meet several requirements. First, it should be able to integrate information across frequency, which is similar to the auditory processing involved in exploiting "common onset" cues or common amplitude modulation. Second, it should work for a high-dimensional state-space, because $\vec{\mathbf{x}}_k$ will have between about twenty and one hundred coefficients. Third, speech has complex non-deterministic properties. For example, certain speech sounds can be followed quickly by a number of other speech sounds, as long as the human vocal apparatus is able to produce the sound sequence. This means that a certain type of speech spectrum can succeed another spectrum with lower or higher probability. In consequence of this plurality of possibilities, $p(\vec{x}_k|\vec{x}_{k-1})$ is multimodal, and a procedure permitting non-Gaussian, multimodal PDFs needs to be applied. Fourth, the functional relationship between the state representation $\vec{x}_k$ and the observation $\vec{z}_k$ will be nonlinear. Last, for a reconstruction of the original signals, the representation of the state needs to be continuous or quasi-continuous.

While traditional methods for multidimensional tracking, like Kalman filtering, do not meet these requirements, sequential Monte Carlo (SMC) methods, also known as "Particle Filters," are in principle suitable and will be investigated here.

4.2.3. Bootstrap SMC Algorithm

For the voice separation task, we selected the 'bootstrap algorithm' or 'generic particle filter' as the SMC algorithm. While a derivation of the SMC framework

1. *Initialize* a number of N_P particles $\{(\vec{x}_k^i, w_k^i), \quad i \in [1 \ldots N_P]\}$ by drawing the state vector from the initial PDF

$$\vec{x}_k^i \sim p(\vec{x}_{k=0}), \tag{4.5}$$

   and setting, for all i, the weights w_k^i to $1/N_P$.

2. *Predict* the new state for each particle by drawing it from the system dynamics PDF:

$$\vec{x}_k^i \sim p(\vec{x}_k^i | \vec{x}_{k-1}^i) \tag{4.6}$$

3. For each particle, compute a hypothetical observable $\hat{\vec{z}}_k^i$ from the state vector:

$$\vec{x}_k^i \longmapsto \hat{\vec{z}}_k^i \tag{4.7}$$

4. *Update* the weights w_k^i recursively according to the observed value of $\vec{z}_k$, the anterior weights w_{k-1}^i and the PDF of the observation statistics PDF $p(\vec{z}_k|\vec{x}_k)$:

$$w_k^i = w_{k-1}^i \cdot p(\vec{z}_k|\vec{x}_k^i) = w_{k-1}^i \cdot p(\vec{z}_k|\hat{\vec{z}}_k^i) \tag{4.8}$$

5. *Normalize* the weights:

$$\tilde{w}_k^i = \frac{w_k^i}{\sum_{i'=1}^{N_P} w_k^{i'}} \tag{4.9}$$

6. Evaluate *expected values* of suitable functions $g(\vec{x}_k)$ of the state:

$$E[g(\vec{x}_k)] = \int_{\vec{x}_k \in \mathcal{R}} p(\vec{x}_k) g(\vec{x}_k)\, dx \approx \sum_{i=1}^{N_P} g(\vec{x}_k^i) \tilde{w}_k^i \tag{4.10}$$

7. *Resample* the approximated PDF: Eliminate particles with low weights $\tilde{w}_k^i$ and duplicate those with high weights.

8. Repeat from step 2

Table 4.1.: Steps of the filtering algorithm (adapted from the SIR particle filter algorithm, Arulampalam *et al.*, 2002, p. 180, Algorithm 3)

would exceed the scope of this article, the basic algorithm is sketched briefly in the following section. More detailed discussions are given by Kitagawa (1996); Doucet *et al.* (2001); Arulampalam *et al.* (2002), for example. Table 4.1 lists an overview of the successive steps of the bootstrap algorithm. The symbol "$\sim$" as in $z \sim p(z)$ means 'is sampled from'.

The basic principle of SMC methods is that the state PDF $p(\vec{x}_k)$ is represented by a set of N_P pairs of samples $\vec{x}_k^i$ and weights $\vec{w}_k^i$. These pairs are also called particles, hence the name "particle filter". SMC methods approximate the temporal evolution of the sampled PDF by drawing at each time step one possible, hypothetical evolution of the state for each particle $\vec{x}^i$ from $p(\vec{x}_k|\vec{x}_{k-1})$ (Eq. 4.6), and updating the weights recursively by multiplying them with $p(\vec{z}_k|\vec{x}_k)$ (Eq. 4.8), which quantifies the congruence between the predicted state and the actual observation $\vec{z}_k$. This recursion is analogous to the enumerator in Eq. 4.3; the denominator is computed by normalizing the sum of all weights to one (Eq. 4.9). With a sufficient number of particles N_P, the estimate of the PDF $p(\vec{x}_k|\vec{z}_{1:k})$ (the 'filtered PDF') approaches the true PDF under very general conditions (Doucet *et al.*, 2001).

The general design of the algorithm implemented in this study was kept simple. Nevertheless, it was necessary to simplify or omit many details to keep the following description concise. The main design choices for any implementation of the bootstrap algorithm involve the generation of the observable $\vec{z}_k$, the definition of the state vector $\vec{x}_k$, and the representation of the PDFs $p(\vec{x}_k|\vec{x}_{k-1})$ and $p(\vec{z}_k|\vec{x}_k)$. The choices made will be delineated in the next subsection.

4.2.4. Implementation of Statistical Algorithm

Overview on Algorithm Structure Figure 4.1 sketches the structure of the algorithm. The computations in the Eqs. 4.5 to 4.10 of Tab. 4.1 are mapped to the processing steps (g) to (n) in the figure. The numbering of the paragraphs of this subsection corresponds to the numbered steps in the figure. The six data entries are depicted by ovals. The first is a speech database which is used to generate statistical information. Four further data entries, a transition matrix, a codebook, the observation statistics, and a HRTF database, include the *a-priori* or "learned" knowledge of the algorithm; these data are generated in steps (b) - (d), (f), and (k) as preparation for the filtering operation. The sixth input is the binaural superimposed signal, which is processed on-line (dashed oval). Each input data set enters the estimation loop of the algorithm, which consists of a circular flow of information on particles (e). The particles are initialized in step (g). According

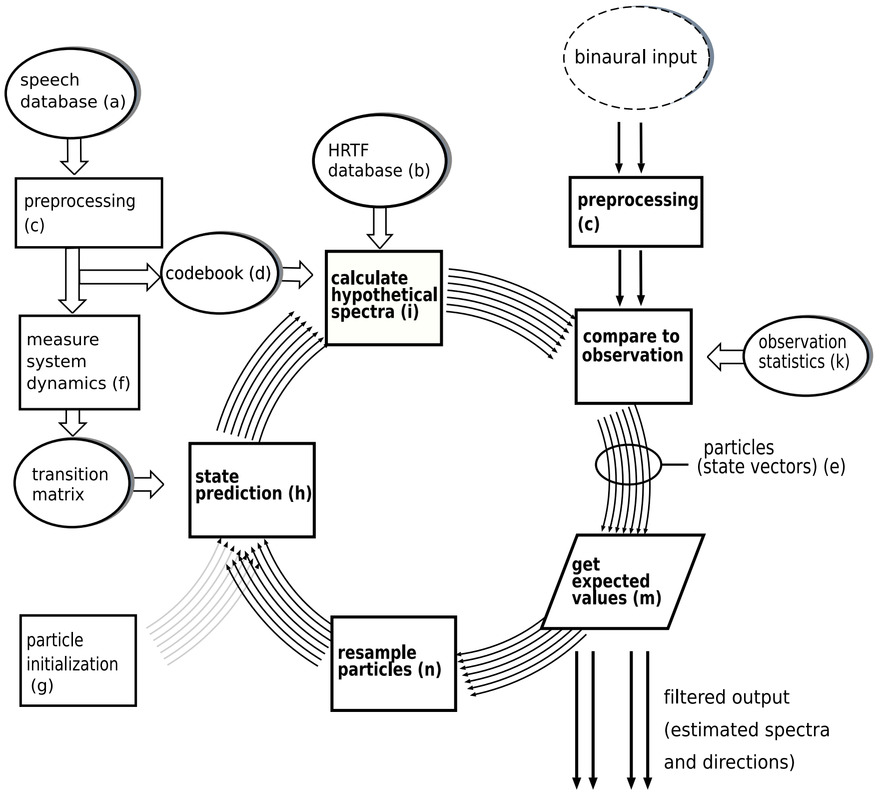


Figure 4.1.: Global structure of statistical filtering algorithm. For detailed explanations, see section 4.2.4 (numbering of the paragraphs of this subsection corresponds to the numbered steps in the figure).

to the rectangular blocks, the new positions of the state vectors are predicted in step (h), and their corresponding hypothetical spectra are evaluated (step i). The particle's weights are updated according to the result of the comparison between hypothetical spectra and actual observation, using the observation statistics (k). The particles simulate the evolving PDF of the state $p(\vec{x}_k)$; calculating expected values of the current magnitude spectrum of each voice and the sound source direction derives the output of the algorithm from the sampled PDF in step (m). Finally, in step (n), the particles are resampled.

(a) Speech database The initial time signals used to generate *a priori* data and test material were sentences from the PHONDAT (Schiel, 2003; Schiel *et al.*, 1997) and TIMIT (Garofolo, 1998) speech databases, which were sampled at the sampling frequency F_s. Tab. 4.2 lists the used values of F_s and the other relevant parameters.

(b) HRTF database To generate the binaural test signals, speech samples were filtered with head-related transfer functions (HRTFs) $\vec{H}(\alpha, \phi)$. The used transfer functions were selected according to the simulated directions of the input sound sources. HRTFs depend on the azimuth α and elevation ϕ of the sound source. A publicly available database (Algazi *et al.*, 2001) provided the HRTFs in the form of impulse responses. The signals were convolved with the impulse responses and then added to generate the mixture of superimposed sounds. Because the HRTF database has a finite spatial resolution of about 5° in azimuth, which is coarser than required for tracking and filtering, a two-dimensional linear interpolation generated the impulse responses for each continuous-valued desired direction.

(c) Preprocessing A preprocessing operation $\mathcal{F}$ transforms the superimposed binaural time signals into short-term spectral sub-band coefficients to generate the observable $\vec{z}_k$. The preprocessing aims to simulate the most basic functional aspects of the auditory system, without representing a detailed simulation of auditory periphery.

Specifically, $\mathcal{F}$ maps N_W buffered time-series values to short-term spectral coefficients as follows: The binaural signals are windowed using a N_W-point Hann Window ($N_W/2$ point window shift), zero-padded to N_F points and transformed by a discrete Fourier transform (Allen, 1977). The absolute values of the short-term Fourier transform coefficients are squared. Then, the power densities are averaged within N_B frequency bands, each having a width of at least B_f times the psychoacoustical equivalent rectangular bandwidth (ERB) (Moore, 1989b), or the frequency resolution of the FFT if this bandwidth is higher. Finally, the power densities in each frequency band are converted to the decibel scale. The symbol $\vec{z}_k$ denotes the resulting two-channel spectral data for the time step k.

(d) Codebook A codebook with a large number of short-term spectra was used to vector-quantize spectra from the speech databases. The codebook was generated from clean speech from the TIMIT and PHONDAT databases, by applying $\mathcal{F}$ to the complete signal set, leading to a number of N_G spectral vectors. A hierarchical clustering method (Ward, 1963; Kopp, 1978) computed N_C codebook elements from this set of vectors.

(e) State vectors The state vector $\vec{x}_k$ contains sound source direction and short-term spectra of a number of voices N_V. The symbol $v \in \{1, 2, \dots N_V\}$ denotes the voice index. Azimuth $\alpha_{v,k}$ and elevation $\phi_{v,k}$ characterize the sound source direction; the resulting spectral coefficients are denoted by $\vec{S}_{v,k} = (s_{b,v,k})$. Here, b denotes the frequency band. Thus the complete state vector $\vec{x}_k$ at time k is defined as

$$\vec{x}_k = (\alpha_{1,k}, \phi_{1,k}, \vec{S}_{1,k},\ \alpha_{2k}, \phi_{2,k}, \vec{S}_{2,k} \dots,\ \alpha_{N_V}, \phi_{N_V,k}, \vec{S}_{N_V,k})$$

Within the filtering algorithm, the short-term spectra $\vec{S}_{v,k}$ as part of the state vector $\vec{x}_k$ are represented in vector-quantized form by the index of a corresponding codebook entry $c_{v,k}$.

(f) Measurement of the System Dynamics The PDF $p(\vec{x}_k|\vec{x}_{k-1})$ governs the update equation (Eq. 4.6) in Tab. 4.1. Because of statistical independence, it factorizes into separate PDFs for the respective voices. Also, source spectral dynamics and sound source direction are independent. Additionally, we approximate azimuth and elevation to be independent Gaussian random variables, so that separate PDFs can be used to describe the azimuth dynamics $p(\alpha_k|\alpha_{k-1})$ and the elevation dynamics $p(\phi_k|\phi_{k-1})$. Finally, the PDF $p(\vec{S}_{v,k}|\vec{S}_{v,k-1})$ represents the Markov dynamics of the spectra. Thus,

$$p(\vec{x}_k|\vec{x}_{k-1}) = \prod_{v=1}^{N_V} p(\alpha_{v,k}|\alpha_{v,k-1}) p(\phi_{v,k}|\phi_{v,k-1}) p(\vec{S}_{v,k}|\vec{S}_{v,k-1}). \tag{4.11}$$

As explained in section 4.2.2, $p(\vec{S}_k|\vec{S}_{k-1})$ cannot be approximated by a Gaussian. Therefore, we measured this PDF by gathering information from the speech database. To represent $p(\vec{S}_k|\vec{S}_{k-1})$ in a discrete form, a transition matrix T was chosen as data structure to record the statistics of spectral transitions from the empirically observed spectra. T was generated as follows: N_T short-term spectra were taken from the speech database. They were neither mixed nor filtered by HRTFs. The vector-quantization algorithm for generating the *a priori* statistics looked up each short-term spectrum $\vec{S}_k$ in the codebook by searching for the entry with minimum Euclidian distance, and represented it by a single codebook index c_k; this transformed the time series of speech spectra $\vec{S}_k, k \in \{1, \ldots, N_T\}$ to a time series of scalar indices c_k, with $c_k \in \{1, \ldots, N_C\}, k \in \{1, \ldots, N_T\}$. The time series defined a sequence of index pairs $(l, m) = (c_{k-1}, c_k)$. The number $s_{l,m}$ of occurrences of each index pair (l, m) was counted, normalized to $\sum_{m=1}^{N_C} s_{l,m} = 1$, and assigned to

$$T_{l,m}. \tag{4.12}$$

The transition matrix together with the codebook used to derive it can be considered as ‘learned’ knowledge about the spectro-temporal dynamics of speech[4]. Because speech has the property of quasi-stationarity of short-term spectra for time scales of about 10 – 20 ms, and because of the overlap of the subsequent analysis frames, the coefficients $T_{l,m}$ usually have the highest values for $l = m$. As a result of this, the transition matrix is a relatively precise first-order representation of transitions between speech spectra, amplitude modulations, onset sounds, and other spectro-dynamical characteristics of speech signals.

(g) Particle initialization At first, the algorithm assigns starting values to the coordinates of the state vector of each particle according to Eq. 4.5. The starting values are assigned this way: For the azimuth of voice v of particle i at time $k = 0$, $\alpha^i_{v,k=0}$ is drawn from $\mathcal{U}(0, 360)$, where $\mathcal{U}(a, b)$ denotes the uniform distribution in the interval $[a \ldots b]$. The elevation $\phi^i_{v,k=0}$ is drawn from $\mathcal{U}(-45, 45)$. The initial codebook index of the voice spectrum $c_{v,k=0}$ is drawn in accordance to the total frequency of occurrence of each index in T. For $k = 0$, all expected values are set to zero.

[4] Because of the multidimensionality, the structure of these data is difficult to represent. However, it is possible to use the transition matrix to generate Markov series of short-term spectra. An example is shown in Appendix C.

(h) State Prediction The state prediction step, corresponding to Eq. 4.6, requires to draw new hypothetical states for each particle x_k^i based on the anterior state x_{k-1}^i and $p(x_k|x_{k-1})$. According to Eq. 4.11, the azimuth, elevation, and spectrum for each voice are updated separately. The azimuth and elevation are drawn from Gaussian distributions $\mathcal{N}(\mu, \sigma)$ with mean μ and standard deviation σ:

$$\alpha_{v,k}^i \sim \mathcal{N}(\alpha_{v,k-1}^i, \sigma_\alpha) \tag{4.13}$$

$$\phi_{v,k}^i \sim \mathcal{N}(\phi_{v,k-1}^i, \sigma_\phi) \tag{4.14}$$

The values σ_α and σ_ϕ influence the possible tracking speed of moving sound sources. State updates for the quantized spectra, represented by codebook indices $c_{v,k}$, are drawn from $p(c_k|c_{k-1})$, which is represented by T:

$$c_{v,k} \sim p(c_{v,k}|c_{v,k-1}) \tag{4.15}$$

(i) Calculation of hypothetical Spectra For each particle $\vec{x}_k^i$, hypothetical binaural spectra $\hat{\vec{z}}_k^i$ resulting from the spatial superposition of voices are calculated (Eq. 4.7). For a particle with index i and voice indices v, the codebook provides the hypothetical ERB-averaged short-term spectra $\vec{S}_{v,k}^i$. Then, the spectra are multiplied with ERB-averaged HRTF coefficients $\vec{H}(\alpha, \phi)$, corresponding to the hypothetical sound source azimuth $\alpha_{v,k}^i$ and elevation $\phi_{v,k}^i$ (indices for the left and right side are omitted):

$$\hat{\vec{S}}_{v,k}^i = \vec{H}(\alpha_{v,k}^i, \phi_{v,k}^i) \cdot 10^{\frac{\vec{S}_{v,k}^i}{20}}$$

The sum of the individual spectra defines the resulting hypothetical binaural input magnitude spectra, according to the principle of linear superposition of sounds:

$$\hat{\vec{z}}_k^i = 20 \log_{10} \left(\sum_{v=1}^{N_V} \hat{\vec{S}}_{v,k}^i \right) \tag{4.16}$$

Possible effects because of different phase values are neglected in Eq. 4.16. This introduces some error if both sources have a similar frequency-specific short-term level.

It should be noted that the retrieval of spectra from the codebook in step (i) requires almost no computation as compared to the lookup operation in step (f), which generates $p(c_k|c_{k-1})$. . During the filtering operation itself, no vector-quantization operation is necessary, but only the retrieval operation.

(k) Observation Statistics $p(\vec{z}_k|\vec{x}_k)$ compares state and observation (Eq. 4.8). First, the hypothetical voice spectrum for each particle is evaluated to a hypothetical observation $\hat{\vec{z}}_k^i$ as described above. Thus, $p(\vec{z}_k|\vec{x}_k) = p(z_k|\hat{\vec{z}}_k^i)$.

A multivariate Gaussian $\mathcal{N}(\hat{\vec{z}}_k^i, \vec{\sigma}_{\mathrm{vq}})$ was chosen to approximate $p(z_k|\hat{\vec{z}}_k^i)$. The component-wise standard deviation of the average codebook lookup error is used as an estimate of the value of $\vec{\sigma}_{\mathrm{vq}}$. It was computed additionally during the generation of the transition matrix T. Because the Gaussian depends on $|\vec{z}_k - \hat{\vec{z}}_k^i|$, which is evaluated in the log-spectral domain, different spectral vectors with small Euclidian distances can be assumed to be perceptually similar, and acquire consequently similar weights.

(m) Computation of Expected Values Finally, the algorithm evaluates the *expected values* from the sampled PDF of the state, as defined by Eq. 4.10. In spite of the fact that the system dynamics $p(\vec{x}_k|\vec{x}_{k-1})$ is multimodal, we assume the filtered PDF $p(\vec{x}_k|\vec{z}_{1:k})$ to have almost always only one maximum to simplify evaluation of the result by the expected value of the spectra. Expected values of the azimuth, elevation, and the magnitude spectrum of each voice are computed, which form the desired filtered output of the system.

(n) Resampling of particles For the resampling operation, the generic resampling algorithm described by Arulampalam *et al.* (2002) is used; ordering of particles is not employed.

4.2.5. Technical Remarks

Technically, the Python script language was used to implement the algorithm. The Numpy vector library provided most of the numerical operations; a few speed-critical sections were coded in the C programming language. The clustering algorithm used for the codebook generation was implemented in C++. A number of particles N_P of up to 16e6 was employed, depending on the number of voices. To reduce the computation time, the algorithm was parallelized to execute on a workstation cluster with up to 32 nodes. Parallel programming was done using the LAM implementation of the MPI standard (Squyres and Lums-

F_s	16 kHz	sample rate
N_W	400	window length
N_F	512	FFT length
N_B	53	number of frequency bands
B_f	0.5 ERB	approximate bandwidth of frequency bands
N_G	$2.5 \cdot 10^6$	number of spectral vectors used to generate codebook
N_C	$10 \cdot 10^3$	number of codebook entries
N_T	$2.35 \cdot 10^6$	number of spectral vectors used to generate the transition matrix

	1	2	3	experiment number
N_V	1	2	2	number of mixed voices
N_P	$70 \cdot 10^3$	$12 \cdot 10^6$	$12 \cdot 10^6$	number of particles
σ_α	1.0°	1.0°	variable	standard dev. of azimuth
σ_ϕ	0.05°	1.0°	variable	standard dev. of elevation
N_M	4	16	15	number of Monte Carlo trials

Table 4.2.: Parameters of the statistical filtering algorithm

daine, 2003; Burns *et al.*, 1994). The software environment is described further in Appendix A and by Langtangen (2004).

The computation time is roughly linear with $N_P \cdot N_V$; for 12e6 particles and $N_V = 2$, it amounts to 52 hours for the processing of one second of signal when run on a PC cluster with five AMD Opteron CPU with 2.0 GHz clock frequency. This is equivalent to 0.2 s of computation for performing steps (h) to (n) of Fig. 4.1 once with 1000 particles representing two voices.

4.2.6. Algorithm Parameters

Table 4.2 lists the parameters of the algorithm and their used values. For F_s, N_W, and N_F, values were chosen which allow to encode a speech signal of good qualityas well as resynthesis of the individual signals from their estimated envelope. N_B and B_f were chosen so that low-frequency spectral structures can be partially resolved. N_C was set high enough that the codebook can represent a speech signal with little degradation in speech quality; this was confirmed in informal listening tests. Whether tracking was achieved depended strongly on N_P. The number of particles was increased for a larger number of voices N_V in order to account for the higher dimensionality of the state-space. N_V had to be constrained to two because of limitations of computation time. The interesting case of estimating envelopes and directions of three simultaneous talkers, which because of the additive structure of Eq. 4.16 does not require any changes to the algorithm, could not be tested in the current implementation. In exploratory experiments, smaller values for N_B and N_C, and a larger B_F did not lead to an enhanced tracking of spectra. Further tests revealed that a small standard deviation of the assumed dynamics of the azimuth coordinate σ_α, used in Eq. 4.13, requires a larger number of particles to make convergence of the algorithm likely, especially if two voices are tracked. On the other hand, in the case that the algorithm converges, a smaller value of σ_α results in more precise azimuth tracking and a better separation quality, because the particles spread less along the spatial dimensions of the state-space. Therefore, several values of σ_α were tested.

4.2.7. Experiments

The performance of the algorithm was tested in three experiments with different values for the number of particles N_p, and number of voices N_V. For the case of one voice, a source with rapidly oscillating azimuth movements, and a small value for σ_α was tested (see Tab. 4.2). For the two-voice case, the value of σ_α was set to be small and constant in the second experiment, and exponentially decaying in the third. The true elevation of the sources was set to be constantly at 0°. To test the behavior of the algorithm with two voices, two sentences from the TIMIT database were mixed after adjustment to the same overall root mean square (RMS) level. The remaining parameters were as follows:

1. One voice with an oscillating movement with increasing oscillation frequency along the azimuth axis, according to

$$\alpha(t) = A\sin(\pi/8 + 2\pi \cdot t(a + tb)).$$

   with A = 60°, $a = 0.025/\mathrm{s}$, and $b = 0.01/\mathrm{s}^2$. The elevation of the source was constantly at 0°.

2. Two voices, fixed azimuths of -60° and 20°.

3. Two voices, fixed azimuths of -60° and 20°, σ_α and σ_ϕ decaying exponentially according to

$$\sigma_\alpha(k) = \sigma_\phi(k) = \sigma_{\mathrm{start}}[(1-d) + de^{-k/(\tau R_F)}]$$

with $\sigma_{\text{start}} = 1.0°$, $d = 0.98$, $\tau = 70\,\text{ms}$, and the frame rate $R_F = F_s \cdot 2/N_W$.

To assess the influence of the stochastical initialization of the particles, each condition was tested in a number of N_M Monte Carlo trials with different durations. Non-convergent trials were terminated after about 2 s of processed signals. The expected values of the magnitude spectra and sound source directions were calculated according to Eq. 4.10 and saved for each time step for later processing and visualization. Also, magnitude spectra of the original input signals and the input mixture were saved.

4.3. Results

4.3.1. Experiment 1 (One Voice, Moving Azimuth)

Directional Tracking

Figure 4.2 shows the result of azimuth tracking for one voice moving in azimuth with up to 120° per second (experiment 1). The dashed line shows the true position, and the solid line the expected value of the azimuth. The jump from the (arbitrarily assigned) initialization value of 0° to the start position of the source of 23° occurs within the first few time steps during the initialization of the algorithm. The largest error in tracking in this trial is about 20°. In this one-voice task, the algorithm converges rapidly at each start of the algorithm. Movements of the sound source were followed reliably. The extent of the deviations from the true azimuth position depends on the standard deviation in Eq. 4.13. The higher the standard deviation is, the higher is the uncertainty of the azimuth tracking.

Tracking of Short-Term Level

Figure 4.3 shows the expected value of the short-term level (solid line), and the true level (dashed line) as a function of time for one voice (experiment 1, trial #01). Most of the time, the two curves match each other closely; the estimate follows the true short-term level rapidly, with an RMS error of 4.3 dB. In some cases, there are convergence times of about 50 ms in the adaptation of the tracked level.

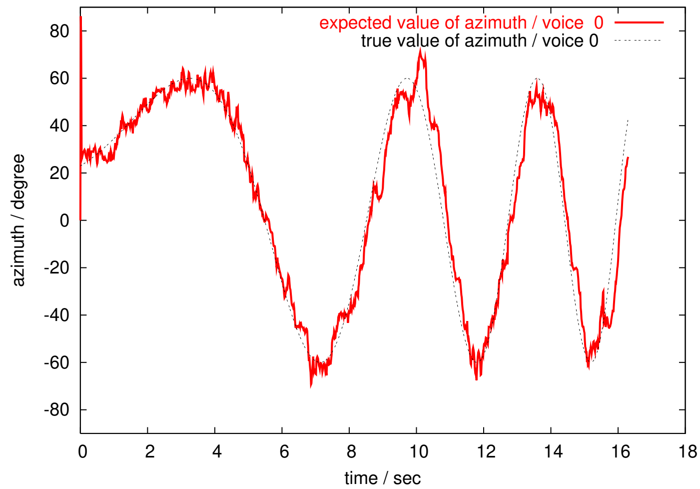


Figure 4.2.: Expected value of azimuth and true azimuth for one moving voice (experiment 1, trial #01). The time value zero corresponds to the algorithm at $k = 0$; all expected values were set to zero for this point.

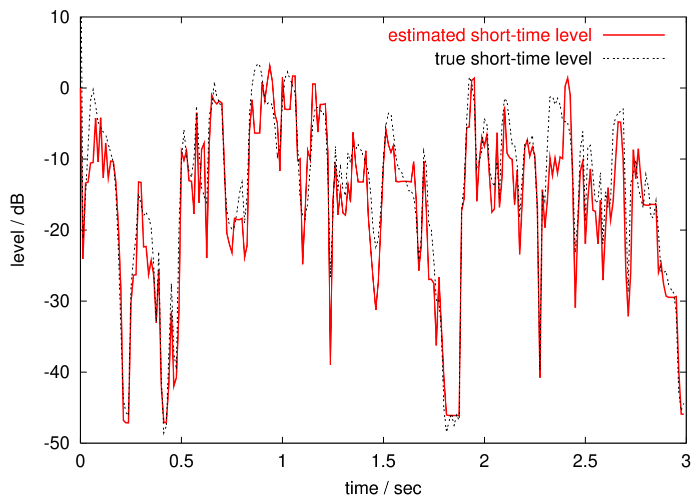


Figure 4.3.: Short-term level for one voice, expected value (solid line) and true level (dotted line) (experiment 1, first 3 s of trial #01).

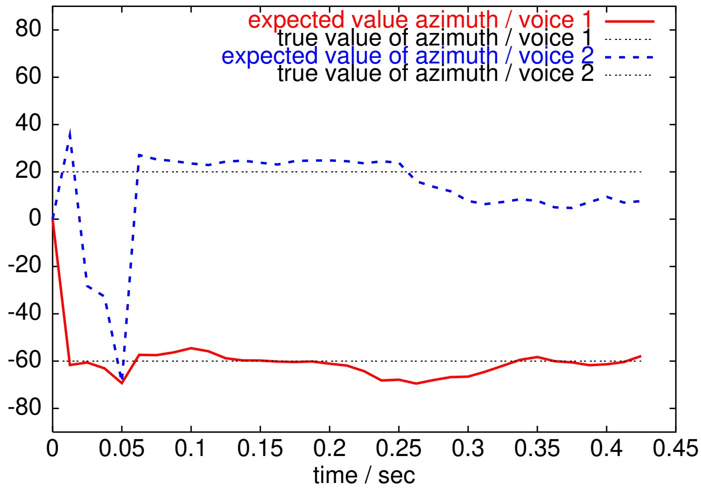


Figure 4.4.: Expected value of azimuth and true azimuth for two voices at -60° and 20°, experiment 2, trial #06.

4.3.2. Experiment 2 (Two Voices, Fixed Azimuth Variance)

Directional Tracking

When two voices are tracked, the algorithm does not always converge. The higher the number of particles N_P was chosen, and the larger σ_α is, the more probable convergence is reached. On the other hand, with increasing σ_α the tracking error becomes larger. With a large σ_α, it happens occasionally that both estimated azimuths converge to the same original voice, and the other voice is neglected. Figure 4.4 shows the expected values of the azimuths and the true azimuth of two sound sources as a function of time (trial #06 of experiment 2). The algorithm converges within 60 ms. After convergence, the error of the estimated azimuth is usually 10° or less, even if one source has at that moment an intensity which is more than 40 dB smaller than the other. In experiment 2, the algorithm converged in 3 out of 9 trials to an azimuth error of less than 10°.

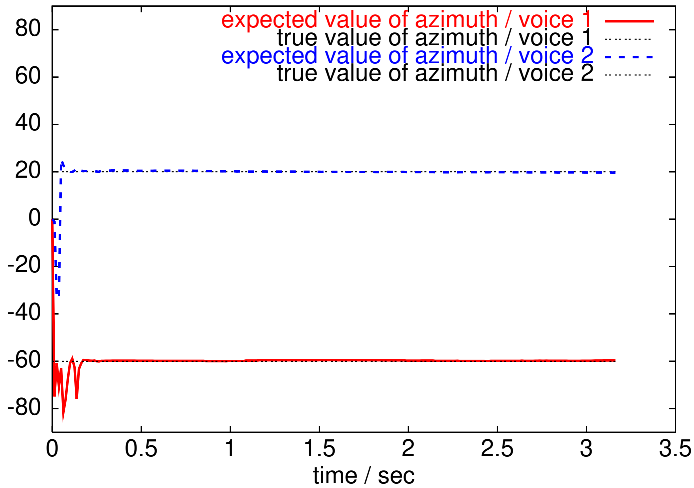


Figure 4.5.: Expected value of azimuth and true azimuth for two voices at -60° and 20°, σ_α decaying exponentially, experiment 3, trial #05, entire length of trial.

4.3.3. Experiment 3 (Two Voices, Decaying Azimuth Variance)

Directional Tracking

If exponentially decreasing values for the standard deviations σ_α and σ_ϕ from Eqs. 4.13 and 4.14 are chosen, the error in the ongoing direction estimate reduces down to a maximum of 1°, as shown in Fig. 4.5 (trial #05 of experiment 3). The algorithm converged to an azimuth error of less than 5° in four out of 15 cases, and an error of less than 10° in three cases. In the remaining eight cases, the algorithm failed to converge.

Separation of Envelopes

Figure 4.6 shows the first 0.5 s of the filtered spectra of two voices for the same signals as used in Fig. 4.5, and trial #07. The spectral estimate is not very precise in the first 50 ms, in which the algorithm is converging and the azimuth estimate is still not estimated correctly. After 50 ms, the envelope of the first source is tracked with high precision even with most of the harmonic structure around 0.2 s. Onsets as visible around 0.4 s are recovered correctly. The first formants

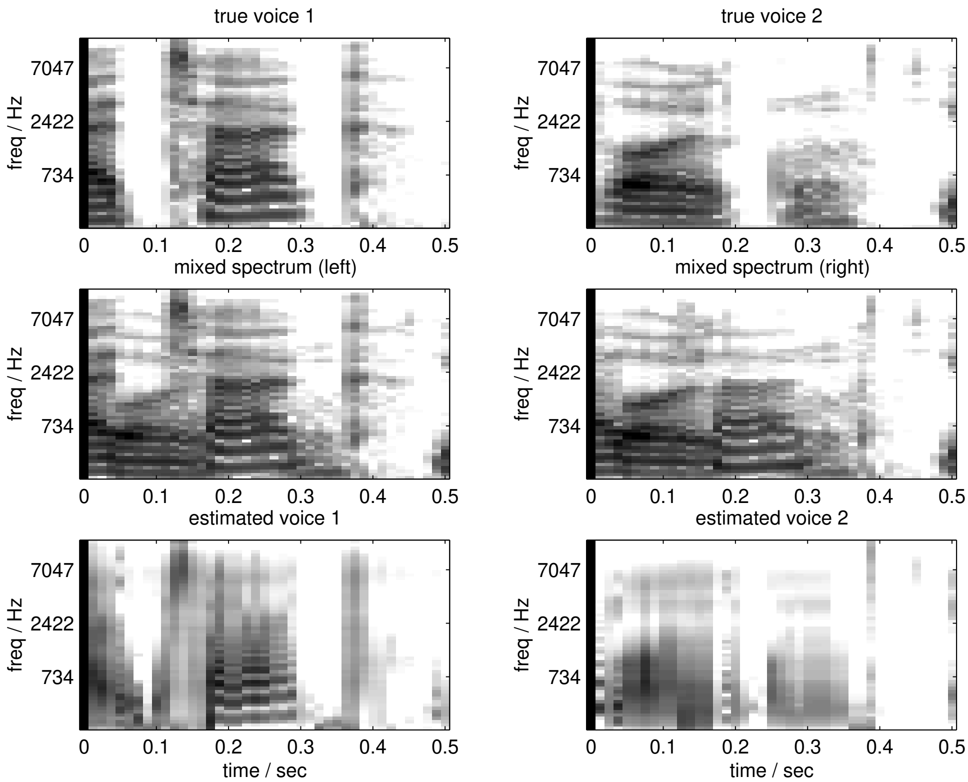


Figure 4.6.: Original, mixed and estimated spectrograms for two voices, experiment 3, trial #07, first 0.5 seconds. The abscissas display the time in seconds, the ordinates the frequency in Hz. From top to bottom: true spectra (left panel: first voice at -60°, right panel: second voice at 20°), mixed spectra (left and right channel), and expected value of filtered spectra (left panel: first voice, right panel: second voice). At time step zero, estimated, mixed, and original spectra were set to 0 dB, resulting in the vertical black bar at the time value zero.

and broadband onsets of the second voice are recovered well. The harmonic structure of the second voice is not recovered, but a large part of the overall envelope and the formant structure are reconstructed.

Two kinds of quantization effects are discernible in the spectrograms: First, along the time axis, effects from the temporal discretization with a windows shift of 12.5 ms (200 samples at 16 kHz sample rate), and second, along the spectral axis, the effect of the averaging into 53 sub-bands is visible. Moreover, the effect of the quantization of the spectra can be recognized. Sometimes, e.g. at 0.25 s for the left voice, the expected value computed from the vector-quantized representations of $\vec{S}_{v,k}$ remains the same in spite of small changes in the actual short-term spectra, resulting in short regions with horizontal lines of constant power density. This is notable especially for the left voice, which has a clear harmonic structure for the time interval between 0.2 and 0.3 s. For the right voice, the harmonic structure is not resolved. The suboptimal tracking of the right voice might be explained by the hypothesis that the sampling of $p(x_k)$ is suboptimal due to an insufficient number of particles. Another possible reason is that the vector-quantized codebook entries, from which the hypothetical envelope is composed, are not able to represent the harmonic structure with sufficient detail. At some moments, e.g. at 0.07 s and 0.5 s for the left voice, or 0.22 s for the right voice, spectral energy from the right voice is assigned to the left voice, or vice versa. This "crosstalk" occurs occasionally when the algorithm seems to 'lose track' of the envelopes for a short period.

Comparison of Signal-To-Noise Ratio

To evaluate the improvement of the signal-to-noise ratio (SNR) for the estimated voices, the SNRs of the mixed and estimated versions are calculated as follows from the envelope estimates of trial #05: The SNR is calculated as the power ratio of the original speech and error signals. The error signal is calculated in the frequency domain as the difference between the original and either the estimated, or the noise-perturbed, envelopes. The envelopes are linearly scaled magnitude spectrograms. Both channels (left and right ear) are evaluated separately. In the unprocessed case, the error signal is defined as the difference between the undisturbed envelopes before mixing, and the mixed envelopes. In the processed case, it is defined as the difference between the undisturbed voice envelopes before mixing, and the estimated envelopes.

From these SNR values, the difference value was calculated between the mixed and processed cases (Tab. 4.3). The change in SNR depends on the channel (left

condition	unprocessed		processed		improvement	
side	right	left	right	left	right	left
voice A	5.7	0.38	2.24	2.32	-3.49	1.95
voice B	-4.3	1.4	3.51	4.00	7.85	2.51

Table 4.3.: Left: SNRs of voices relative to unprocessed mixture for the left and right channel, experiment 3, trial #05. Mid: SNRs of estimated voices. Right: SNR difference of the estimated versus the unprocessed voice (positive means improvement)

or right ear) and is between an improvement of +2.5 dB and +7.8 dB for the channel with lower initial SNR. For the other channel, the SNR difference is between an improvement of +1.95 dB and a decrease by 3.49 dB. The decrease can be explained as an quantization effect which becomes dominant at high initial SNR.

Correlations Between Estimated and true Envelopes

To evaluate the separation of envelopes further, a correlation analysis of the original and the estimated envelope time series was performed with the components of the estimated two-voice mixture from trial #05. In each subplot of Fig. 4.7, for two envelope time series, $f_b(k)$ and $g_{b'}(k)$, the cross correlation coefficients $\rho_{f,g}$ have been calculated for each combination of frequency bands b and b', and are displayed in three rows and two columns of matrices. The abscissa and ordinate of each matrix correspond to b and b'. The duration of the evaluated signals and the estimated envelopes is 3.2 s.

In all columns of plots, the envelope time series $f_b(k)$ belongs to the original voice one (left column of plots), or to the original voice two (right column of plots). In the upper row of matrices, the time series $g_{b'}(k)$ belongs to the same voice as an $f_b(k)$, but may refer to a different frequency. The diagonal elements of both matrixes always have the value of one, because b and b' are equal there, and $\rho_{f,g}$ is the value of the autocorrelation function at lag zero. In the mid row of plots, the time series $g_{b'}(k)$ belongs to the estimated voice with the smaller azimuth estimate; in the bottom row, it belongs to the estimated voice with the larger estimated azimuth. The mid and lower rows of Fig. 4.7 show that for one estimated voice, the cross correlation values are for one original voice close to the values in the top row, and small for the other voice, while for the other estimated voice, the opposite is true. This shows that each envelope estimate follows closely one of the original voices while they are largely uncorrelated to the other voice.

4.4. Discussion

4.4.1. Integration of Principles used for CASA

The approach described here to integrate acoustical features with prior knowledge on sound sources and their spectral and temporal dynamics implements several principles that have been suggested for CASA in the literature.

The first principle is parallel processing and extraction and abstraction of features in different successive stages, as applied by Marr (1980) for computational visual scene analysis. Ellis (1996) and Cooke and Ellis (2001) demonstrated that such a separation of stages is viable for the auditory domain as well. Their models include stages which perform feature extraction and abstraction and several grouping stages in separate steps. This corresponds to physiological results, as, for example summarized by Scott and Johnsrude (2003), who show that in speech perception, several differently processed representational streams converge at some later point. The statistical approach presented here follows this principle. It performs an extraction of features from the incoming sounds and the evaluation of the hypotheses mostly in parallel, with the exception of the steps of normalization, resampling and part of the evaluation of expected values.

The second important principle that is realized here is the competition of hypotheses. Physiological results suggest that binding in neural systems is performed by some representation of competing hypotheses. For example, Bushara *et al.* (2003) performed functional magnetic resonance imaging (fMRI) experiments on neural binding between hearing and vision. They found not only areas of *increased* neural activity, but also regions of *reduced* activity when binding occurred. As mentioned in the introduction, similar structures have been implemented in several CASA models, e.g. by Wang and Brown (1999), which have shown that segmentation can be performed by a two-layered network of coupled oscillators. A possible segmentation, which matches the input from a feature-extracting layer well, tends to suppress other candidate segmentations. This is similar to “local excitation and global inhibition” interactions in the neural system (van der Kouwe *et al.*, 2001). Hypotheses formulated by a state-space

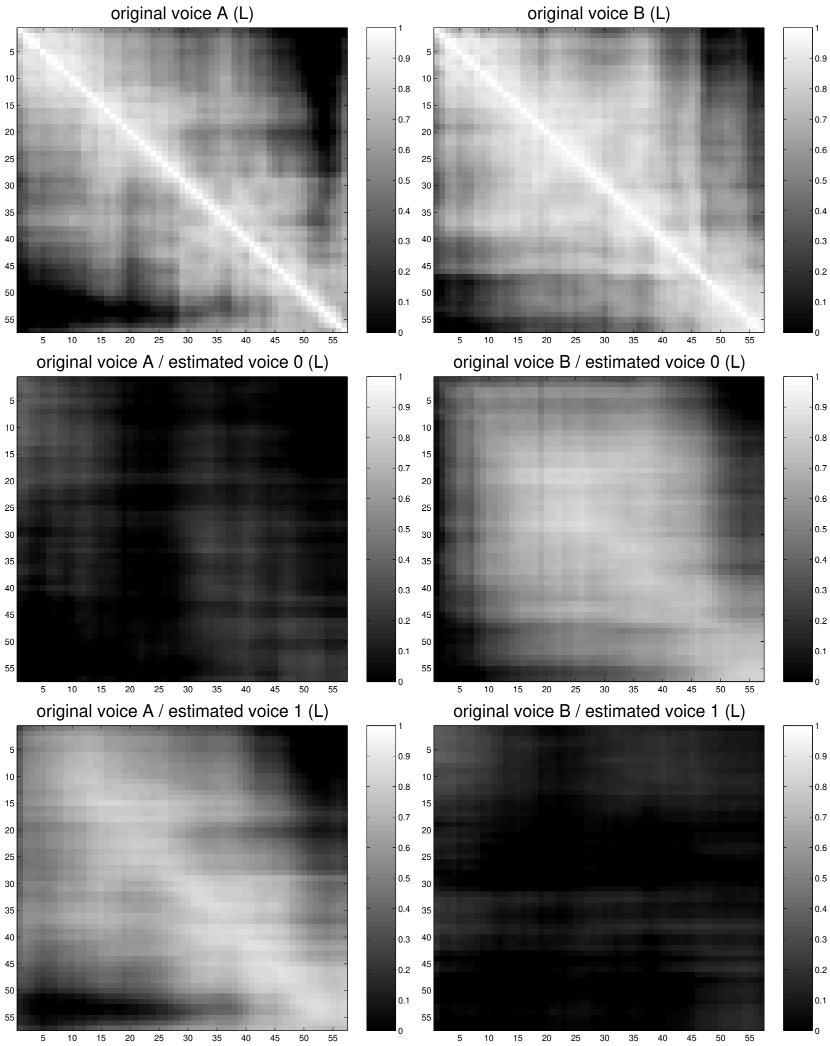


Figure 4.7.: Inter-band cross correlations of envelopes of original signals and original signals (upper row), original signals and estimated first voice (mid row), and original signals and estimated second voice (lower row). Signals are from experiment 3, trial #05 (more details see text in section 4.3.3).

approach will include all desired features in the state description. The "best explanation" of the incoming sound, which is derived from these features, will win the competition. The algorithm proposed here realizes a 'competitive' structure in steps 4, 5 and 7 in Tab. 4.1, which implement weighting of the hypotheses, normalization to a total probability of one, and elimination of hypotheses with small weight. If a state hypothesis matches well to the observation, it will gain a high weight. Because the sum of weights is normalized to a total probability of one, the weights of the other hypotheses are reduced in turn. The resampling step may eliminate those hypotheses whose weights are already small.

The third principle incorporated in the algorithm is the 'perceptual restoration' or 'continuity' principle (Cooke and Ellis, 2001) which assumes that parts of signals that are masked for a time up to 200 ms, most probably remain approximately the same. The PDF $p(\vec{S}_{v,k}|\vec{S}_{v,k-1})$ used in Eq. 4.11 incorporates the quasi-stationary property of speech which has the effect that short-term spectra in subsequent frames usually do not change much. This property is similar to the technique developed by Godsill and Rayner (1998); Fong *et al.* (2002), and others for the restoration of music recordings degraded by nonstationary noise. Redundancy of speech is also used by the "missing features" approaches for application in automatic speech recognition (Cooke *et al.*, 2001a,b).

In addition to the principles outlined above, the concept of multidimensional statistical filtering is innovative in that it allows to model a non-hierarchical organization of auditory grouping. As the implementation shows, for simulating binding of spectro-dynamical features and sound source direction, no explicit hierarchy of cues needs to be assumed. Consequently, the result of grouping does not depend on the order of application of grouping rules. This is important when features are partially masked, or several explanations are possible for the same observed data. The 'strength' of cues is considered intrinsically by the SMC framework when using them for grouping; the reliability of the individual features is quantified by the variance of the corresponding factors of the observation statistics PDF $p(\vec{\mathbf{z}}_k|\vec{\mathbf{x}}_k)$.

The statistical procedure integrates features in a single step, by multiplying marginal PDFs to construct the PDF of the system dynamics $p(\vec{\mathbf{x}}_k|\vec{\mathbf{x}}_{k-1})$ and the observation statistics PDF $p(\vec{\mathbf{z}}_k|\vec{\mathbf{x}}_k)$. Additional perceptual features can be integrated by extending the observation variable $\vec{z}_k$. They could, for example, correspond to source properties such as the formant frequencies (Zheng and Hasegawa-Johnson, 2004) or the fundamental frequency (Summerfield and Stubbs, 1990; Gandhi and Hasegawa-Johnson, 2004). Beyond contributing to a product of marginal PDFs, estimated source features may have interdependent effects on the expected observation. In this case, some factors of the PDF $p(\vec{\mathbf{z}}_k|\vec{\mathbf{x}}_k)$ will depend on several coordinates of the state-space.

4.4.2. Advantages and Potentials of Multidimensional Filtering

The algorithm developed here has several properties favorable for speech enhancement by combining advantages which are mutually exclusive in traditional approaches.

First, the proposed approach is able to integrate complementary strategies for *speech coding*: Evaluation of the observable $\vec{z}_k$ can be performed by taking the perceptually most relevant features of the sound streams into account, as perceptual coding techniques do. In addition, strategies from source coding, like linear predictive coding, can be applied in the representation of the source states. The description of the system dynamics by a discrete Markov transition matrix models the non-Gaussian statistics of spectro-temporal features of speech in a detailed, highly-resolved way, incorporating a large amount of *a priori* knowledge. As discussed by Li *et al.* (1969) and Anemüller and Kollmeier (2000), and shown similarly in Fig. 4.7, high correlations of the envelope time series exist between adjacent frequency channels of speech, and spectral sequences which contain such correlations occur frequently. They are recorded in the transition matrix T, and therefore they are predicted with a higher *a priori* probability. By using a spectral prediction based on spectro-temporal statistics of speech, the implemented algorithm is designed to exploit these across-frequency correlations. The incorporation of *a priori* knowledge about spectro-temporal properties is probably applicable to other types of sound sources, because according to the generating principle, each source type has specific spectro-temporal properties. To take advantage of such modeling of source properties, techniques for feature extraction still need to be improved.

Second, the stochastical framework combines important properties from classical noise suppression methods. On the one hand, it acquires virtues from multi-channel methods, like blind source separation and beamforming. It uses spatial information and symmetrical treatment of desired and undesired sound sources, reaching SNR enhancements even in highly nonstationary noises at low SNRs. On the other hand, the approach includes, similar to envelope filtering algorithms, an Bayesian solution to the problem of separating spectral power densities of several sound sources, which partially overlap in time-frequency representations. In order to estimate short-term envelope spectra of overlapping sound sources, it is not necessary to recognize speech pauses to estimate the noise signal. This allows it to estimate envelopes of highly nonstationary sound sources at SNRs as low as 0 or -4 dB, a range in which spectral subtraction algorithms have not shown success. The convergence times of the implementation are, depending on the parameter σ_α, in the order of 50 ms, which is fast as compared to many algorithms for BSS (Parra and Spence, 2000). Because of this fast convergence, the algorithm can follow rapidly moving or changing sources in a multi-talker environment.

The third advantage is that multidimensional estimation is in principle able to separate more than two sound sources from two microphones in an optimum way, because the calculation of the hypothetical binaural signals in Eq. 4.16 is sufficient to compare state and observation; an inverse transformation from observation to state is not necessary.

Fourth, the algorithm also includes advantages of hidden Markov models (HMMs) and pattern-matching techniques, as discussed, for example, by Boll (1992) and Ephraim *et al.* (1989a). Additionally, the implementation of a quasi-continuous state-space description allows for a high-resolved description of the voices, making 10 000 or more discrete states available for each voice. This high resolution is one requirement necessary to reach a reconstruction of the acoustical signal which is needed for speech enhancement.

A last advantage is that the proposed approach presents an on-line algorithm, which might allow for real-time applications as soon as more efficient schemes for state coding and feature extraction are implemented, and fast enough hardware exists.

The assumptions listed in section 4.2.1 still limit an application in realistic environments. They mostly had the rationale to make the implementation simpler and easier to examine, but do not reflect an inherent limitation of the SMC framework.

First, the restriction that all participating sounds are a mixture of voices can be expanded to that the types of the participating sounds are known, e.g. a voice in car noise. In this case, the statistical knowledge of speech would need to be supplemented with knowledge of the spectro-temporal statistics of the involved sound types. Second, the restriction of the angular velocity of the sources does not seems to be critical; if necessary, the algorithm can be extended by a more precise kinematic model of the sources and head rotations. Third, the assumption of knowledge of the number of voices or sources can be relieved by incorporating this number into the Markov Model (Berzuini and Gilks, 2001; Larocque *et al.*, 2002). Fourth, the assumption that the levels of the individual voices are about equal is not critical; the levels can be included into the state description in a straightforward way.

Last, the restriction to a non-reverberant environment can in principle be lifted if some model of reverberation is included into the state description, like the positions of the closest reflecting surfaces and the reverberation time of the room. Because these parameters change about as slowly as the sound source direction, there are realistic chances that they can be estimated along with the tracking operation. Summarizing these considerations, the high flexibility of particle filtering methods allow in principle to lift most of the restrictions used here.

4.4.3. Challenges and Improvement Strategies

The presented implementation of the concept still has important drawbacks. The most important one is the very high computational effort, which prevents practical applications. The reason for the excessive computational complexity is the high number of particles needed to represent the PDF of source spectra sufficiently; the reason for this in turn is the high dimensionality of the state-space, which originates from the simple spectral representation, that requires a large number of coefficients to describe the spectra.

An additional drawback is that the the speech quality of the reconstruction is still low. The imperfect reconstruction of envelopes can be interpreted as a partial removement of the interfering voice, with a simultaneous addition of a high level of quantization noise from the sampled PDF. Further, convergence was reached at only about 50% of trials for two voices. We explain both problems by the hypothesis that the used number of particles is still not sufficient to sample the state-space adequately. Moreover, the high dimensionality of the state space promotes an uneven distribution of particle weights, called 'degeneration' (Doucet *et al.*, 2001), which lowers the efficiency of the representation of $p(\vec{x}_k)$. In our experiments, we observed that any decrease of the number of particles reduced the performance of the algorithm in a nonlinear manner (the "track is lost"), supporting this explanation. A saturation of performance at some number of particles, as for example observed by Vermaak *et al.* (2002) for a one-voice estimation task, did not occur.

In fact, for typical applications of SMC algorithms, the expression

$$n_{ppc} = \sqrt[D]{N_P}$$

has usually values between 10 and 100 (Blake and Isard, 1998; Ward *et al.*, 2003); here, D is the dimension of the state-space, and N_P the number of particles. n_{ppc}

can be considered as "number of particles per coordinate." For the algorithm implemented here, with N_P =12e6 and $D = 118$, results an n_{ppc} of only 1.14. Both the exponential relationship between the dimensionality of the state-space, and the nonlinear relationship between tracking performance and number of particles, suggest strongly to reduce D as much as possible, because, e.g., reducing the dimensionality from $B = 53$ to $B = 20$ will lead to a much better sampling of the state space, and a much improved performance even if the number of particles, and thereby the complexity, is reduced.

Efficient representations of speech spectra as used in modern telephony codes achieve high-quality speech codingwith as few as eight coefficients. Similarly, Fong and Godsill achieved enhancement of speech in stationary noise with a sixth-order time-variant autoregressive model (Fong *et al.*, 2002) and an $N_P \leq 1000$. Vermaak *et al.* (2002) observed no improvement for a time-variant autoregressive model when the order of the model was increased to more than 4. They found that to model a single speech signal disturbed by Gaussian noise, between 10 and 1000 particles are necessary, depending on the SNR. Ephraim *et al.* (1989a) and Acero *et al.* (2000), e.g., reported to have used vector quantization approaches for speech enhancement with codebooks based on autoregressive coefficients, including only 256 entries.

A Rao-blackwellised particle filtering method combined with Kalman filtering has been described to separate mixtures of autoregressive signals with convolutive channels, using a number of only 50 particles for a source representation consisting of four coefficients (Daly *et al.*, 2004). However, in the reported simulations, besides that the order of the autoregressive and the convolutive models may not be sufficient to represent speech and transmission channel well, the process noise was very small, and dynamic properties of speech are not taken into account. So far, the approach has not been applied to speech, which exhibits a large amount of process noise. In any case, this example illustrates the relationship between the dimensionality of the state space and the required number of particles.

Use of such features leads to a state-space with a dimension D of about 20 for two voices. Consequently, a much smaller volume of the state-space needs to be explored during the initial convergence of directional estimates, leading to a faster computation and more reliable convergence behavior. Also, advanced representations of speech spectra can represent the physical state of the source better, which leads to a better description of source dynamics. Additionally, special variants of SMC methods, like partitioned sampling, exist for tracking multiple independent objects (MacCormick and Blake, 2000; MacCormick and Isard, 2000), and yield large improvements in applications like person tracking (Choo and Fleet, 2001). The combination of these strategies might lead to efficiency improvements by factors between 10^3 and 10^6.

To summarize the preceding points, the approach to model binding by multidimensional statistical filtering based on an Bayesian approach integrates important properties of earlier models for CASA. Additionally, grouping can be formulated in a non-hierarchical structure. The computational expense of the resulting algorithm is still impractical but we expect large improvements to be possible. Efficient abstract representations of sound sources need to be included, and suitable procedures for extraction of acoustical features in complex environments still need to be developed.

4.5. Summary and Conclusions

Statistical state-space methods are able to integrate different cues and model the 'binding' process in the cognitive system. The general properties of SMC methods (particle filters) allow to use them for statistical filtering of disturbed speech.

We present an algorithm which combines knowledge on spectro-temporal dynamics of speech with directional information to track one or two concurrent voices. This statistical algorithm is designed to exploit statistical interdependencies between different time variables, such as the correlations of spectral power densities at adjacent frequency bands. The algorithm tracks the direction of two concurrent voices very precisely. It can localize and track one moving talker reliably after an algorithmic convergence time of 50 ms or less.

As evaluations of the SNR show, the presented algorithm improves the SNR in a mixture of two concurrent talkers, at initial SNR around zero or below. This is a principal improvement in comparison to existing algorithms for envelope filtering which use a stationary noise model and estimate the noise spectrum in speech pauses.

To compute the estimate, only a forward transformation from the hypothetical spectra of sound sources to the actually observed spectra is needed. Therefore, the algorithm is in principle able to estimate the envelopes of more than two talkers, and to include reverberation.

The computational complexity of the algorithm is too high for a direct application. An efficient, low-dimensional representation of speech, a precise model of its dynamics, and more adapted SMC methods are considered key requirements for reducing the complexity of the algorithm. Their combination can achieve a reduction of the computational complexity by many orders of magnitude.

Acknowledgments

We are grateful to Birger Kollmeier for his substantial support and contribution to this work. We thank Michael Kleinschmidt for contributing to this work with numerous dedicated discussions and important ideas, and Jörn Anemüller and Ronny Meyer for revising and commenting the manuscript. Also, we thank the computing center of the University of Oldenburg for maintaining the Linux cluster. Thanks to the members of the CIPIC interface laboratory (Richard O. Duda) for providing the HRTF database. And finally, we say thanks to the members of the Oldenburg Medical Physics Group for their encouragement and technical support. We also thank to three anonymous reviewers for helpful comments and suggestions.

This work was supported by DFG via the International Graduate School for Neurosensory Science and Systems, and the HörTech Center of Excellence for Hearing Aid Technology with funding from the German Ministry for Science and Education (BMBF).

5. General Summary and Conclusions

5.1. Retrospect on the Goals

Summarizing the goals and questions mentioned in the introduction, the primary objective was to explore strategies for speech enhancement in high levels of nonstationary noise. As earlier investigations have shown, spatial filtering can achieve suppression of interfering sources of up to about 10 dB for environments of modest complexity, but does not reach the capabilities of the auditory system in the presence of multiple interferers (Greenberg and Zurek, 2001; Hawley *et al.*, 1999). Steering spatial filtering by estimation of the sound source direction was demonstrated to work (Liu *et al.*, 2001; Greenberg *et al.*, 2003), but sound localization methods adapted to high levels of distributed noise are still lacking. On the other hand, single-channel spectral estimation approaches so far have not been able to improve intelligibility in nonstationary noise situations (Marzinzik, 2000; Edwards, 2004). Because of the relative success of spatial filtering methods in the past, a secondary goal was to find robust ways to extract spatial information from binaural parameters which are strongly disturbed by additional noise.

A third goal was to incorporate knowledge about psychoacoustics and physiology of auditory scene analysis into processing schemes for noise reduction and spatial filtering. This raised three further questions. The first was, how to simulate the mechanisms of frequency integration and the feature integration performed by the neural binding process, as observed e.g., in barn owls, and how to take environmental noise into account. Second, it was explored how spatial information can be combined with an evaluation of spectro-temporal properties of sound sources by novel multidimensional statistical filtering methods, called sequential Monte Carlo (SMC) methods, and if this combination can provide separation of sources. The temporal evolution of the sources was modeled by a state-space approach. The approach mimics neural binding by following competing hypotheses, and weighting these hypotheses depending on their accordance to signals observed by two microphones. One of the most intriguing questions was if this strategy is applicable at all for separation of sound sources, given the high complexity of the task.

Last, a long-term goal was to explore if use of SMC methods in situations, where sufficient information about the sound sources is not available, can reconstruct the envelopes of the signals in a sense of optimum multidimensional estimation.

5.2. Summary of the Results

5.2.1. Bayesian Algorithm for Sound Localization

Chapter 2 introduced the Bayesian algorithm for sound localization, which offers several innovative properties. It is able to use binaural information to localize sound sources in the median plane, in spite of the almost symmetrical head related transfer functions (HRTFs) in respect to the interaural axis. It uses a frequency domain approach to represent interaural timing by interaural phase differences. Fluctuations of interaural level differences (ILD) and interaural phase differences (IPD) in real-world environments were assessed explicitly by measuring their probability density functions (PDFs). A moment analysis of these PDFs revealed that their shape depends mainly on the signal-to-noise ratio (SNR) and, to a much lesser degree, on the type of the noise environment. The Bayesian approach allowed to take the statistics of the parameters into account, integrating information across frequency by combination of probabilities. Probabilistic information was evaluated across a grid of azimuth and elevation values, similar to topotopic neural maps found in the barn owl.

The results with the sound localization simulations based only on binaural cues allow to gain several insights for physiological and psychoacoustical models. The remarkable good performance of the simulation gives a possible explanation for the broadness of neural tuning curves actually found in mammals. The interpretation of the PDFs of interaural parameters as optimized neural tuning curves is in line with physiological hypotheses favoring a population code for the representation of interaural timing information, as an alternative to Jeffress's place code hypothesis (Harper and McAlpine, 2004; McAlpine and Grothe, 2003). The model shows that by integration of information across frequencies, it is to a certain degree possible to use pinna asymmetries for sound localization, stressing the significance of binaural parameters for this task.

A main result is that environmental noise has to be taken into account for understanding how sound localization works in realistic environments, and how binaural parameters might be processed in the auditory system.

5.2.2. Application of Bayesian Sound Localization

Chapter 3 examined the application of the Bayesian sound localization scheme for concurrent talkers and for suppression of interfering voices. For the evaluation of the algorithm, implementation based on a script language was established, which is able to perform the computation in real time. Extensive tests measured the performance of the algorithm. The results show that the Bayesian approach provides an efficient and simple processing strategy for noise-robust sound localization, and that the integration of probabilities across frequency is a powerful concept which allows to distinguish sounds source directions even in high levels of noise, and to localize three concurrent talkers robustly. The spatial information served to steer a beamforming algorithm; an improvement in SNR was reached of up to 30 dB.

5.2.3. CASA by Multidimensional Statistical Filtering

Chapter 4 developed a new approach for tracking and separation of concurrent voices mixed with additive, highly nonstationary noises. Motivated by findings on auditory scene analysis, this is done by integration of spatial and spectro-temporal cues and use of correlations of spectral power densities across frequency, which are characteristical for speech. To tackle this difficult task, the algorithm developed here applies sequential Monte Carlo (SMC) methods in a multidimensional stochastical state-space framework. Further, it applies an extensive description of spectro-temporal dynamics of speech, gathered by evaluating a vector-quantized speech database. The two-channel input signals, carrying directional information from the HRTFs, were transformed into a "perceptive space," using a feature extraction which emulates the most basic properties of peripheral auditory processing, by applying short-term frequency analysis and use of frequency sub-bands oriented at the bandwidth of auditory filters.

The results demonstrate that on-line sound source localization and envelope separation is in fact possible by a such an approach. Because high-dimensional statistical knowledge on the sound sources is used, which existing algorithms do not exploit so far, the algorithm is able to localize two concurrent voices within a few time steps, often within 50 ms, which is much faster than current approaches for blind source separation achieve. It can track a single sound source performing rapid movements reliably, and is able separate the envelopes of two concurrent voices.

Evaluations of improvement in SNR by the separation found enhancements of up to 7 dB; considering the fact that many technical details of the new approach are still not optimized, this is a very encouraging result. It can be compared to results from Vermaak *et al.* (2002) and Fong *et al.* (2002), which applied SMC methods to single voices in white Gaussian noise and achieved (for an input SNR of 0 dB) improvements of up to 8.5 dB.

A further goal that could not be reached because of constraints of computation time, was to simulate simultaneous tracking and separation of more than two voices by use of two microphones. In the scope of this thesis, it was not possible to explore alternate strategies of state coding, like linear or time-varying autoregressive models, which are commonly used for speech recognition tasks. There are convincing reasons that by use of such coding strategies, computational complexity can be reduced drastically.

To relate the computational complexity of the sequential Monte Carlo (SMC) algorithm to other solutions of the task, it is compared here with a grid-based Bayesian evaluation of all possible paths through a discrete state-space. In this case, the complexity depends exponentially on the number of possible states, which we call N_T. Given the parameters for spatial resolution and representation of spectra used in Chapter 4, or their discrete equivalents, namely $v = 2$ concurrent voices, $N_C = 10\,000$ codebook entries for spectral representation, and $N_a = 360$ different azimuths, combined with $N_e = 5$ elevations, N_T is equal to $(N_C * N_a * N_e)^v = 3.24 \cdot 10^{14}$. For a sequence of $n = 16$ states (roughly equivalent to the average length of a syllable), the number of possible state sequences which would require evaluation in such a scheme, is in the order of $(N_T)^n$, which is about $1.47 \cdot 10^{232}$ – much more than the estimated number of particles in the universe, about 10^{78}. This is the reason why classical grid-based approaches, which explore all different possibilities in parallel, and were used successfully in a real-time implementation for the task of localization described in Chapter 2, are completely unsuitable for tracking in high-dimensional state spaces. Common hidden Markov models, e.g. based on the forward-backward algorithm, still have to evaluate a number of transition probabilities in the order of nN_T^2, which would be in this case approximately $1.6 \cdot 10^{30}$, because they need to evaluate even highly improbable state transitions at first. The advantage of SMC methods is that the number of hypotheses which need to be evaluated is only a very small fraction of the number of possible states; "uninteresting" regions of the state space are sampled less densely. Here, $1.2 \cdot 10^7$ hypotheses were sufficient for a rough representation of the envelopes of two voices and very exact azimuth tracking. This advantage can be expanded, e.g., by reducing the codebook size, and thereby N_T. For HMMs, codebook sizes of $N_C = 256$ have been used successfully to represent single voices for speech enhancement (Ephraim *et al.*, 1989a). The advantage of SMC methods has stimulated further successful attempts to apply them for the separation of voices (Gandhi and Hasegawa-Johnson, 2004).

5.3. Suggestions for Future Work

5.3.1. Sound Localization

The outcomes of the strategies pursued here lead to several suggestions concerning future research. Models of binaural processing and sound localization in the auditory system should take robustness to environmental noise into account. Physiological research should examine neural responses from a statistical point of view, considering the properties of real-world stimuli.

Simulations of auditory processing can apply the Bayesian approach developed here to alternative representations of interaural timing, or a more precise model of peripheral signal processing. The influence of such a modification on localization performance should be verified experimentally. Further, the contribution of information on envelope time differences, as expressed by the interaural group delay in the frequency domain, may be investigated.

Given that real-world signals exhibit correlations of spectral power densities at adjacent frequency bands (Li *et al.*, 1969; Anemüller, 2001, and Fig. 4.7), the neural system may also use correlations of interaural parameters between adjacent frequencies for sound localization. The possible advantage of inclusion of such information for the localization in real-world environments, especially at medium levels of noise, should be investigated further.

Technical applications of the Bayesian approach will probably use an increased frequency resolution of the short-term spectral analysis, because increasing the number of observations will improve the robustness in noise further. Also, they may compute the necessary parameter statistics numerically, based on average HRTF data, instead of a direct measurement. The future application in hearing aids and speech processors, for example cochlear implant processors, depends on the availability of some information-transmitting link between both channels.

Further, the technical advantages of the Bayesian approach are not limited to signals recorded at the ear canal entrance. It can also be applied to microphone arrays with a small number of microphones and modest dimensions. Such microphone arrays can serve, e.g., to provide applications of automatic speech recognition with enhanced speech input, or to assist mobile telephony in cars. The algorithm can provide such setups with exact information on the position of the sound sources as well as with hints about their relative SNR.

5.3.2. Separation of Speech and Nonstationary Noise by Statistical Methods

The envelope-tracking algorithm in the present form has the shortcoming that it is computationally too expensive to be applied practically. As explained in Chapter 4, a key point for improvement of the SMC algorithm is a more efficient representation of the state space resulting in a lower dimensionality of the state vector. Many techniques provide efficient representations for speech; a large part of them have been developed for telephony. To name a few, linear predictive coding (LPC), or line spectrum frequencies (LSF) coefficients, jointly with advanced vector quantization techniques, are well suitable to represent voices with as few as eight coefficients for each time step (Vermaak *et al.*, 2002). Tree-based vector quantization methods can speed up the generation of the reference statistics (Ephraim *et al.*, 1989a). Further improvements might be gained from combination with well-known methods for dimensionality reduction, e.g. principal component analysis (PCA) (Li *et al.*, 1969; Zahorian and Rothenberg, 1981). Alternatively, approaches describing the *physical* state of the voice-producing system, i.e. the vocal apparatus, may be explored; their advantage is that they can easily model the physical constraints which determine the set of possible states and state dynamics. In Chapter 4, page 110, the large possible reduction in computational complexity of such low-dimensional approaches has been discussed. In short, the idea is to take advantage of the fact that natural sound sources can only produce signals which pertain to a small subspace of the space of possible signals. Such principles, which agree with state-space descriptions, seem to be used also in the neural system in sparse neural representations of sensory information (Lewicki, 2002; Olshausen and Field, 2004).

A further domain of possible improvements is to incorporate important aspects of present models of peripheral auditory processing into the representation of the observation variable. For example, some representation of fundamental frequency or timbre (Cheveigné, 1993), an additional representation of onsets, or information from the modulation spectrogram (Kollmeier and Koch, 1994; Strube and Wilmers, 1999) could be added. The idea behind this is that humans might keep some model of acoustical sources and voices in the brain to facilitate stream separation, but that on the other hand the peripheral auditory system is probably highly adapted for delivering, for a wide variety of inputs, the information needed to decide which model, and which hypothesis, is given most credibility. The most fundamental feature, the frequency analysis performed by the human inner ear, is already included in the approach. It is a great advantage of the SMC scheme that it uses clearly separated domains for the acoustical observable, and the model of the processes generating the observation. In future, the SMC-based scheme may help to join the auditory approach with approaches of source coding. To develop this idea further and to turn it eventually into applications, it will probably be necessary to join efforts and knowledge from several fields. Although based on well-established concepts, their integration into a compound algorithm may be a major task.

5.3.3. Integration of Bayesian Sound Localization and Multidimensional Statistical Filtering

A final suggestion is to pursue the integration of the Bayesian approach examined in Chapter 3 with the SMC algorithm deployed in Chapter 4. Because the localization algorithm provides not only quite exact estimates of azimuth and elevation with small computational effort, but also a spatial map of probability of presence of sounds from each direction, it can provide initialization values for the SMC algorithm, and speed up convergence. Spatial filtering techniques can provide additional observables, in which in each case one of the interfering voices is nulled out according to their hypothetical direction. The SMC framework in turn is able to combine observations for several directions and to compute estimates of the spectra of the participating voices even if more voices than microphones are present. Also, the SMC method is able to remedy the weakness of the Bayesian approach which does not incorporates knowledge about plausible source movements and their possible extend. Additionally, the fundamental frequency of the voices should be incorporated into the state vector (Meyer *et al.*, 2005). First steps in this direction were taken; the results showed that they actually accelerate convergence, but that a robust integration of the different approaches is by no means a trivial task.

In the long run, extensions of the SMC algorithm could prove useful to model aspects of auditory scene analysis, because it provides a statistical framework which can integrate several feature extraction procedures based on auditory models. In 1953, Cherry described the idea that processing of transition probabilities is involved in the understanding of concurrent talkers and speech in noise; it took about thirty years until hidden Markov models were adapted to the task of recognition of undisturbed speech, and twenty years more until it became possible to buy software based on such models for automatic speech recognition for use in offices.

5.4. Quintessence

To condense the above in two sentences, this work has shown multidimensional statistical approaches to be principally applicable for the difficult, and so far unsolved, task of tracking sound source directions and separating nonstationary sound sources in complex environments by combining spatial and spectro-temporal information. This opens new possibilities for noise reduction in high levels of nonstationary noise, and may help to develop solutions which are useful in daily life.

A. A real-time, script-based, multiprocessing Solution for experimental Development of Signal Processing Algorithms

Abstract

Evaluation of audio signal processing algorithms on real-time platforms has unique advantages. However, such environments also used to have the disadvantage of requiring expensive hardware, and tedious work to set them up, while providing only a short useful life. This report proposes to exploit advances in hardware and software development by integrating real-time processing with script-based explorative development and use of multiprocessing hardware. The concept was implemented based on standard hardware and open source software, and its realization and characteristics are presented here. Applications of the system for algorithm development and evaluation are described briefly.

A.1. Current Real-Time Signal Processing Programming Environments

When developing speech signal processing algorithms, real-time processing environments allow rapid testing and adjustment of numerous parameters with interdependent effects. After development, such environments allow to evaluate algorithms in extensive tests, for example objective measurements of speech intelligibility. In the next paragraphs, three techniques, namely DSP platforms, use of specialized script languages, and multiprocessor hardware, are discussed, which are useful for development of algorithms. The first and the third are centered on hardware, while the second is centered on programming techniques.

Traditional real-time *digital signal processor (DSP) platforms* usually consist of a host computer and several processors which are mutually connected by a high-bandwidth bus, supporting a rapid data exchange. These environments provide high speed, but have several disadvantages: First, they are expensive. Second, they are tedious and difficult to program; the programming languages available are usually C and assembler code. Furthermore, programming such hardware requires an increasing amount of expert knowledge, e.g., on the CPU architecture. Data transfer to and from the host system, for example of *a priori* statistics in statistical algorithms, is often complicated, and the task to extract and visualize data from the running program may have to be programmed individually. Often, such systems are also difficult to debug. As the run-time environment of such DSP systems is generally not protected, a single error in the memory access of the DSP program can crash the entire system, requiring even the host computer to reboot. To access the capacity of multiple processors, the program has to be divided carefully among them, and changes to the algorithm may require to start this task from the beginning. Furthermore, because of their limited distribution, the platform-specific numerical libraries tend to contain significantly more hidden errors than widely used standard software. As of today, signal processor systems have lost most of the speed advantage compared to general-purpose workstations, and their expected useful life is short.

A second, widely used alternative for development of signal processing algorithms are *script languages* with specific numerical extensions. Typically, they provide commands which perform complex operations like matrix multiplication, vector addition, and handling of multidimensional data structures, thus allowing a short and compact notation.

Being interpreted languages, errors can be identified immediately, and complex scripts and functions can be composed from commands entered line by line. In contrast to languages like C, they provide a safe run-time environment, automatically managing the memory for complex data structures. This protection of the run-time environment has the big advantage that violations of the memory address space are not possible, and the programs can be warranted to either run as specified and to deliver correct results, or to be aborted, e.g., when a program attempts to retrieve array values with indices outside the array limits.

Further, some of these script languages support highly modular programming and re-use of code, and contain easy and powerful tools for data plotting and visualization. Traditionally, script languages, and the operating systems which support them, do not meet real-time requirements, and for evaluation purposes, algorithms developed in script languages are still often ported to DSP platforms, sometimes with the help of integrated development environments.

A third development which is becoming increasingly important is signal processing on *workstations with symmetric multiprocessing (SMP) architecture and multicomputers*, like workstation clusters. They provide an enormous amount of CPU power while requiring low specific costs. Several algorithms can run in parallel on the same data, allowing sophisticated processing schemes, for example evaluating the same time series on different time scales, or in multichannel filterbanks with different rates. Multiprocessor systems provide many services which DSP systems lack or provide only in rudimentary form, for example file access, graphical user interfaces, audio device drivers, and network services. In the last years, such workstations became suitable for low-latency, full duplex processing, which is frequently required for real-time evaluation of audio processing algorithms.

An integration of the three principles explained above, real-time processing, algorithm development based on script languages, and parallel computing on SMP systems, workstation clusters, or multicomputers, has the potential to combine several advantages of each approach. The most important advantage is that exploratory programming will be much easier and faster, because the result of small changes made to a script can be heard seconds later with real-time audio signals. A solution which provides this integration is presented and discussed in the next sections.

A.2. A Script-Based Real Time Processing Environment

A.2.1. Overview

Am overview over the system is depicted in Figure A.1. The system is layered in several levels. Each layer depends only on layers of a lower level. The lowest level is the hardware of the system, and the highest level the script which implements the signal processing algorithm. Each layer shields upper layers as much as possible from the peculiarities of the layers below, and the interfaces of each layer allow to replace them by different implementations. In the next paragraphs, the system is described from bottom to top.

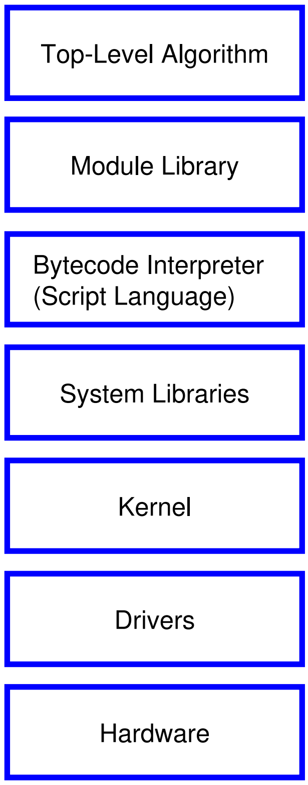


Figure A.1.: Global structure of real-time processing system

A.2.2. Components

Hardware

The system consists of a workstation with one or multiple processors. For the first implementation, an UP2000 workstation with dual Alpha 21264 CPU running at 800 MHz clock frequency was used. Later implementations included IBM PC architectures with Intel Pentium IV CPU with 2.4 GHz clock frequency, and SMP workstations with dual Opteron 246 CPUs produced by AMD, running at 2 GHz. For computing-intensive operations, the system was set up on a commercially available workstation cluster (Hewlett Packard High Performance Line) with 16 nodes and 32 Intel Pentium IV CPUs with 1.6 GHz clock frequency.

As audio hardware, the RME Digi96 and Hammerfall 9652 sound cards with optical I/O were used. They provide sampling rates between 44.1 and 96 kHz and up to eight channels. For purposes with lower demands on audio quality, or a lower sampling rate, cheaper soundblaster PCI 128 cards were also used. Hard disks with SCSI interface served to access data during real-time processing. Gigabit ethernet cards with PCI interface provided the network access. The remaining parts of the hardware consists of standard IBM PC components.

Driver Layer

To reach low-latency duplex processing and hardware independence, the audio drivers are of special importance. To access the sound cards, ALSA (Advanced Linux Sound Architecture) drivers were used. In combination with the RME cards mentioned above they allow very low latencies. At the same time they provide complete hardware abstraction for more than 300 commercially available sound cards (Kysela and many others, 2000-2005).

Kernel and Operation System

The system kernel consists of a standard Linux system, whose source code was modified with the "low-latency patches" provided by Morton (2001). These modifications allow worst-case interrupt latencies of about 500 μs for processes which use the real-time scheduling capabilities of the system. Among other services, this kernel layer also provides inter-process communication for exchanging data via shared memory, and scheduling of real-time processes. Both are necessary because on a typical workstation with graphical user interface, about 80 processes are active after startup, and even when using a system tailored for audio processing, several additional tasks still will be needed for operation system services. Real-time scheduling ensures that the signal processing task receives always as much CPU time as necessary, and the memory locking function protects the memory area of the process from being paged out to disk.

System Libraries

The system libraries provide, in addition to common operating system services and basic math functions, application program interfaces (APIs) to access the audio hardware, real-time scheduling, and POSIX IV threads and inter-process communication (Nichols *et al.*, 1996; Gallmeister, 1995). For parallel computation on workstation clusters, the local area multicomputer (LAM) library is used, an open-source implementation of the message passing interface (MPI) standard (Squyres and Lumsdaine, 2003). This library enables programs, that consist mainly in vector operations, to run with small modifications as a single instruction, multiple data (SIMD) program on workstation clusters, and SMP workstations.

Script Language and Virtual Machine

The script-language used is Python (van Rossum *et al.*, 1998-2005), a general-purpose script language. Among its characteristics are a compact syntax, a small core of instructions, a large general-purpose library, a design supporting re-use and exchange of source code and libraries, support for object-oriented programming paradigms, and extensive documentation of high quality (Lutz and Ascher, 2003; Lutz, 2001; Langtangen, 2004). Beneficial for real-time processing, all data are passed by reference, reducing out-of-cache memory accesses when passing large numerical data structures in object-oriented programming paradigms. So-called "generator functions" allow layered processing of (potentially infinite) data sequences, while avoiding much of the overhead and complexity of object-oriented programming paradigms. Also, functions coded in C or Fortran can be integrated easily. The run-time environment for the language is a virtual machine similar to the Java virtual machine, and requires only a few megabytes of memory.

Python Library Modules

Among many other facilities, Python library modules, which are mostly written in C, provide vector math, and convenient handling of numerical arrays with an arbitrary number of indices, linear algebra routines, numerical analysis functions, high-quality data plotting, access to POSIX shared memory functions, and bindings to the LAM library. The standard library also provides platform-independent file access, and process management. For real-time requirements, additional modules had to be developed by the author to access the ALSA audio library, as well as special functions of the operating system, like setting real-time scheduling priorities, and memory locking.

Signal Processing Application

The signal processing algorithm consists of a top-level script, which is written in the Python language, and a number of additional modules which can be implemented as Python scripts, or programmed in C or Fortran as well. The implementation of these modules is in turn based on a module library shared among applications, which provides, e.g., mixing of several audio streams from files and sound cards, short-term window analysis and synthesis, access to commonly used databases and advanced vector operations. Because Python fosters to collect frequently used operations in modules, such libraries are created and extended naturally during algorithm development.

A.2.3. Characteristics of Implemented Solution

A short summary of the properties of the implemented system follows. First, it is able to execute signal processing algorithms in full duplex mode (processing of sound card input to sound card output). Signal processing algorithms typical for hearing aids can be executed at latency times below a threshold of about 25 ms, which is necessary to avoid confusing echo effects in hearing aid simulations. To a large part, this latency is caused by the buffering of the short-term frequency analysis processing, which can be further reduced by special filter banks. The system kernel provides worst-case latencies of about 0.5 ms, and the sound card hardware requires additional buffering of at least $3 \cdot 64$ samples at 44.1 kHz. Therefore, stable latencies of about 5 ms may be achieved. Although the standard Linux kernel is not designed to meet very low latencies without any exception ('hard' real-time), the value of 25 ms is met very reliably even when other applications without real-time priority scheduling are started during processing. Second, parameters of the algorithms can be set and changed at run-time via a network connection or a graphical user interface. Third, the Python programs can run on SMP workstations or workstation clusters supporting the MPI standard[1]. Fourth, the language promotes a highly modular design and re-use of code while emphasizing simplicity. Fifth, because it is easy to profile programs and replace time-critical routines which are called often, by C or C++ code, development can be performed by first quickly writing a correct program completely in script language, and subsequently implementing just the

[1] Real-time audio processing was not tested on clusters. Provided that the cluster is connected by fast network hardware, real-time processing can be expected to work on such platforms as well.

sufficient amount of low-level code in C as a replacement. This approach has been called 'tip-of-the-iceberg strategy' (Lutz, 2001, p. 708). Sixth, since the script language is freely availably as source code and runs on a large number of platforms, like the variants of MS Windows, MacOS, QNX, Be-OS, and others, the algorithms are highly portable. It was easy to re-compile the C modules initially developed for the Alpha workstation hardware on different hardware with 32-bit-Intel (IA32) or 64-Bit Opteron architecture, requiring only changes to a few lines of code. The supported platforms include small, portable hardware. The Linux operating system runs on very small devices, like personal digital assistants (PDAs), and the Python interpreter has been implemented on real-time operating systems like QNX and even on mobile phones with Symbian OS (Nokia Corporation, 2005). Because of this, once such devices achieve the necessary processing capability, it will be possible to port algorithms developed on workstations almost effortless to smaller portable devices for testing in everyday environments. Due to the layered structure of the solution, porting to other platforms is possible as long as low-latency audio drivers and real-time process scheduling functions are available. This high degree of portability ensures the long-term availability of the environment and is especially attractive for scientific purposes.

The described system was used for the evaluation of the binaural sound localization algorithm, described in chapter 3 of this thesis, and also for the development and evaluation of the particle filtering algorithm described in chapter 4. Additional applications were development and extensive real-time comparisons of hearing aid algorithms, and teaching signal processing and binaural audio processing strategies in one-week lab courses.

A.3. Summary

Real-time audio processing as well as exploratory development based on script languages with numerical extensions provide important advantages for developing audio processing algorithms. A portable environment that is able to combine both advantages was developed, and its hardware and software components were described briefly.

Acknowledgments

I want to express my gratitude to Birger Kollmeier for his substantial support to this work. Thanks to Volker Hohmann and Giso Grimm for many fruitful discussions. Many thanks also to Paul Barton-Davies for developing the RME audio drivers and friendly personal support, and to Kai Vehmanen for helping to solve problems with dynamical loading. Jürgen Kahrs and Stefan Münkner gave important first inspirations. Also, Andrew Morton, Benno Senoner, Dave Beazly, Ingo Molnar, Jaroslav Kysela, Konrad Hinsen, Travis Oliphant, and Winfried Ritsch contributed decisive work to this project.

B. Moment Coefficients for Linear and Cyclic Random Variables

B.1. Moment Coefficients of Distributions of Linear Random Variables

The n-th statistical moment of a distribution $f(x)$ is defined as

$$\bar{x} = \int_{-\infty}^{\infty} x f(x)\, dx, \tag{B.1}$$

$$m_n = \int_{-\infty}^{\infty} (x - \bar{x})^n f(x)\, dx. \tag{B.2}$$

Using expected value $\bar{x}$ and standard deviation $\sigma = \sqrt{m_2}$, skew s and the kurtosis K (also called kurtosis excess) are calculated as follows (Sachs, 1992, p.169):

$$s = \frac{m_3}{\sigma^3} \tag{B.3}$$

$$K = \frac{m_4}{\sigma^4} - 3 \tag{B.4}$$

The skew is a measure for the asymmetry of a distribution. The kurtosis is a measure of the width of the shoulders of a distribution relative to the standard deviation. Both values are zero in case of a Gaussian distribution. Distributions with wider shoulders than a Gaussian PDF with same standard deviation have a positive kurtosis.

B.2. Moment Coefficients and Parameters of Distributions of Cyclic Random Variables

The first moment μ_1 of the PDF $f(\theta)$ of a circular variable θ is:

$$\mu_1 = \int_{-\pi}^{\pi} e^{i\theta} f(\theta)\, d\theta \tag{B.5}$$

$$= \varrho e^{i\phi} \tag{B.6}$$

The argument ϕ of μ_1 denotes the *mean phase angle* and the absolute value ϱ denotes the *vector strength* or *resultant length*. The *circular standard deviation* σ_z is defined as

$$\sigma_z = \sqrt{-2\log\varrho} \tag{B.7}$$

(Fisher, 1993). The *circular variance* ν is defined as $\nu = 1 - \varrho$.

The vector strength can assume values between 0 and 1. If ϱ equals 1, the distribution has the shape of a delta function and all phase values are coincident. By way of contrast, $\varrho = 0$ could mean that the random variable is uniformly distributed at all phase values, or that the distribution has two peaks at an angular difference of π, for example.

Analogous to the central moments for the linear case, trigonometric central moments μ_p of order p can be defined as

$$\mu_p = \int_{-\pi}^{\pi} e^{ip(\theta-\phi)} f(\theta)\, d\theta \tag{B.8}$$

The imaginary part of the second central trigonometric moment $\Im[\mu_2]$ can be used to calculate the *circular skew*

$$s_z = \frac{\Im[\mu_2]}{\nu^{\frac{3}{2}}} \tag{B.9}$$

and the real part $\Re[\mu_2]$ defines the *circular kurtosis*:

$$K_z = \frac{\Re[\mu_2] - \varrho^4}{\nu^2} \tag{B.10}$$

Using these quantities it is possible to describe distributions of circular variables by a few descriptive parameters as in the linear case.

C. Envelope Series Generated from Markov Statistics of Speech

This Appendix shows Markov series of short-term spectra generated from the statistical information, which was captured in the spectral transition matrix $T_{l,m}$ defined in chapter 4, Eq. 4.12. To select the first short-term spectrum, an initial codebook index $c_{k=0}$ was drawn randomly from the cumulated sum of all codebook entries stored in $T_{l,m}$. The following indices of the sequence were generated by drawing for each time step k one succeeding codebook index c_k from the row of the transition matrix given by $T_{c_{k-1},m}$. Each spectrum belonging to the resulting series of codebook indices was looked up in the codebook. This generated a series of spectra with the same first-order transition statistics as was measured from the speech database.

Figure C.1 shows an spectrogram of these series, which looks similar to a speech spectrogram. Time signals were generated by multiplying this spectro-temporal envelope with phase values from ICRA 5 noise (Dreschler *et al.*, 2001). These signals sound rougher than actual speech signals, but have a very similar temporal structure. Uninformed listeners may confound them with distorted speech, trying to understand words. Such signals might be useful for investigating spectro-dynamical and across-frequency processing in the auditory system.

Acknowledgments

Ronny Meyer synthesized the time signals.

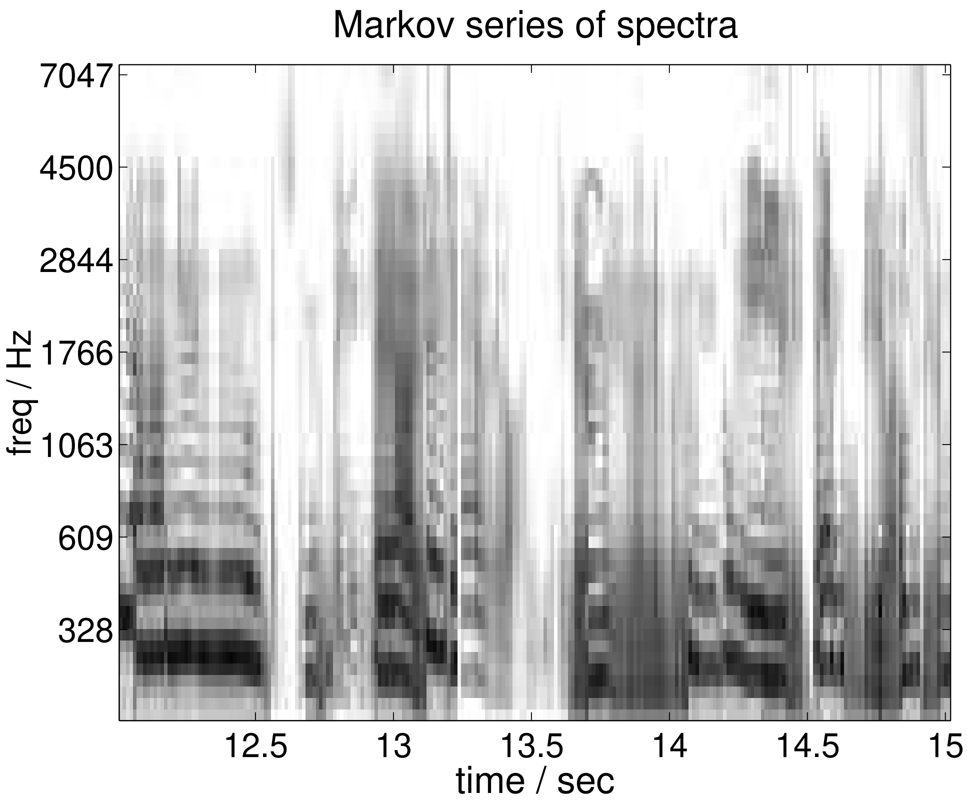


Figure C.1.: Series of short-term spectra, generated from the first-order statistics of spectral transitions of speech, as described in chapter 4. The ordinate represents the time in seconds; the abscissa the ERB-scaled frequency in Hz.

D. Time Series of Interaural Time Differences and Interaural Phase Differences

This section compares the representation of interaural timing by the interaural phase difference (IPD) of narrow-band signals with the interaural time difference (ITD), computed as the argument of the maximum of the interaural cross correlation function (ICCF). The ICCF was computed from the windowed short-term time series, which were used to calculate the IPD in chapter 2, Eqs. (2.13) to (2.18); The time series were filtered in the frequency domain with rectangular bandpass filters, defined by the subband parameters $m_u(b)$ and $f_h(b)$. Different from the computation of the IPD, no averaging across the subbands was applied in this case. The cross correlation theorem was applied to the band-pass filtered signals to compute the ICCF. The ITD was estimated from the maximum of the ICCF in a search range of ± 3 ms time delay. Figures D.1 and D.2 show in the left panels the estimated ITD, and in the right panels the IPD, computed as defined in Eq. 2.18, both as a function of the number of the time step. For frequencies above 1200 Hz, the time series of the estimated ITD show several stripes with frequent values within the physiological range of ITDs (about 600 μs, determined mostly by the head diameter), illustrating the ambiguity of the ICCF for narrow-band signals.

Acknowledgments

Thanks to Ronny Meyer for help with preparing the time series.

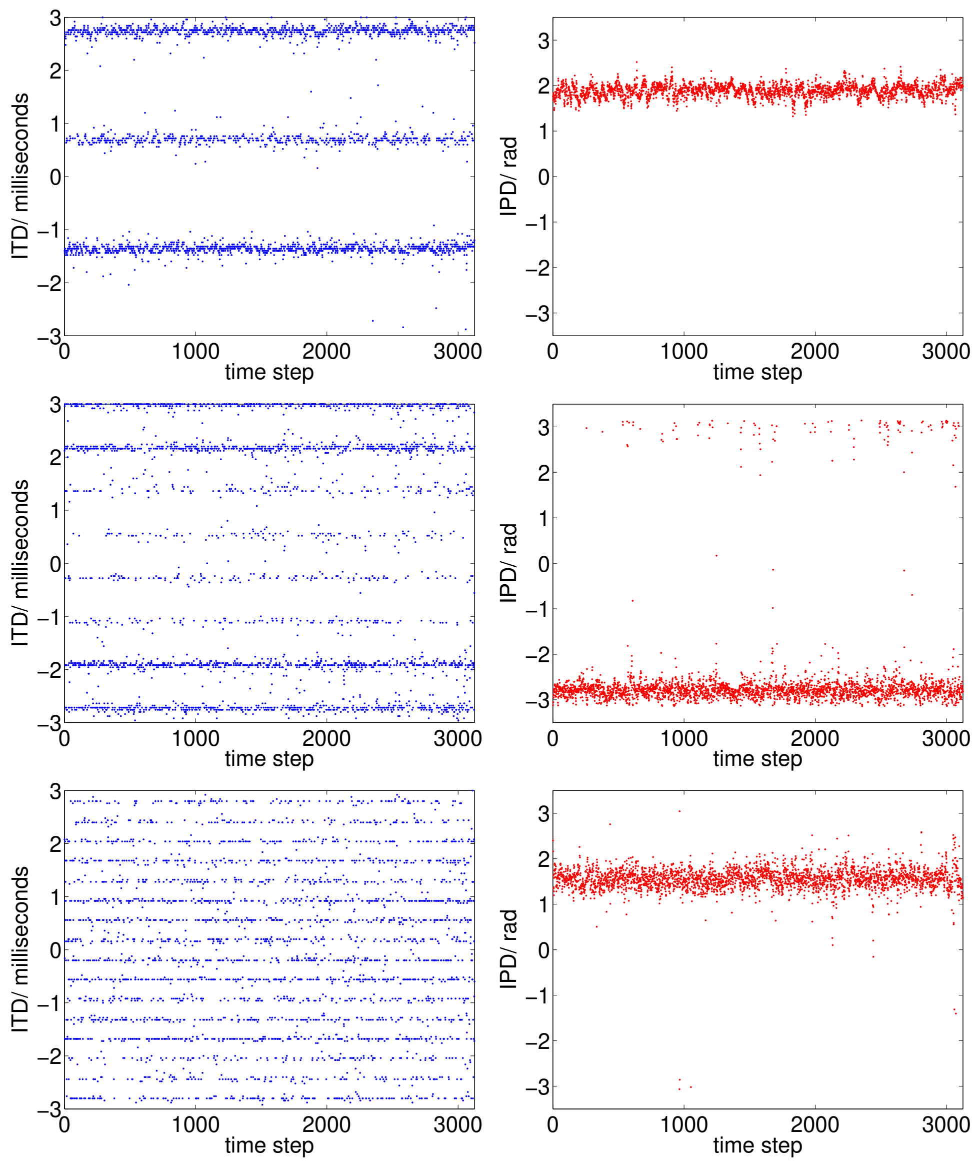


Figure D.1.: Time series of the argument of the maximum of the ICCF (left) and of the IPD (right) for different frequencies, without noise. Top row: band 11 (630 Hz); Mid row: band 20 (1400 Hz); lower row: band 30 (3100 Hz); Azimuth and elevation are always 60° and 0°, respectively.

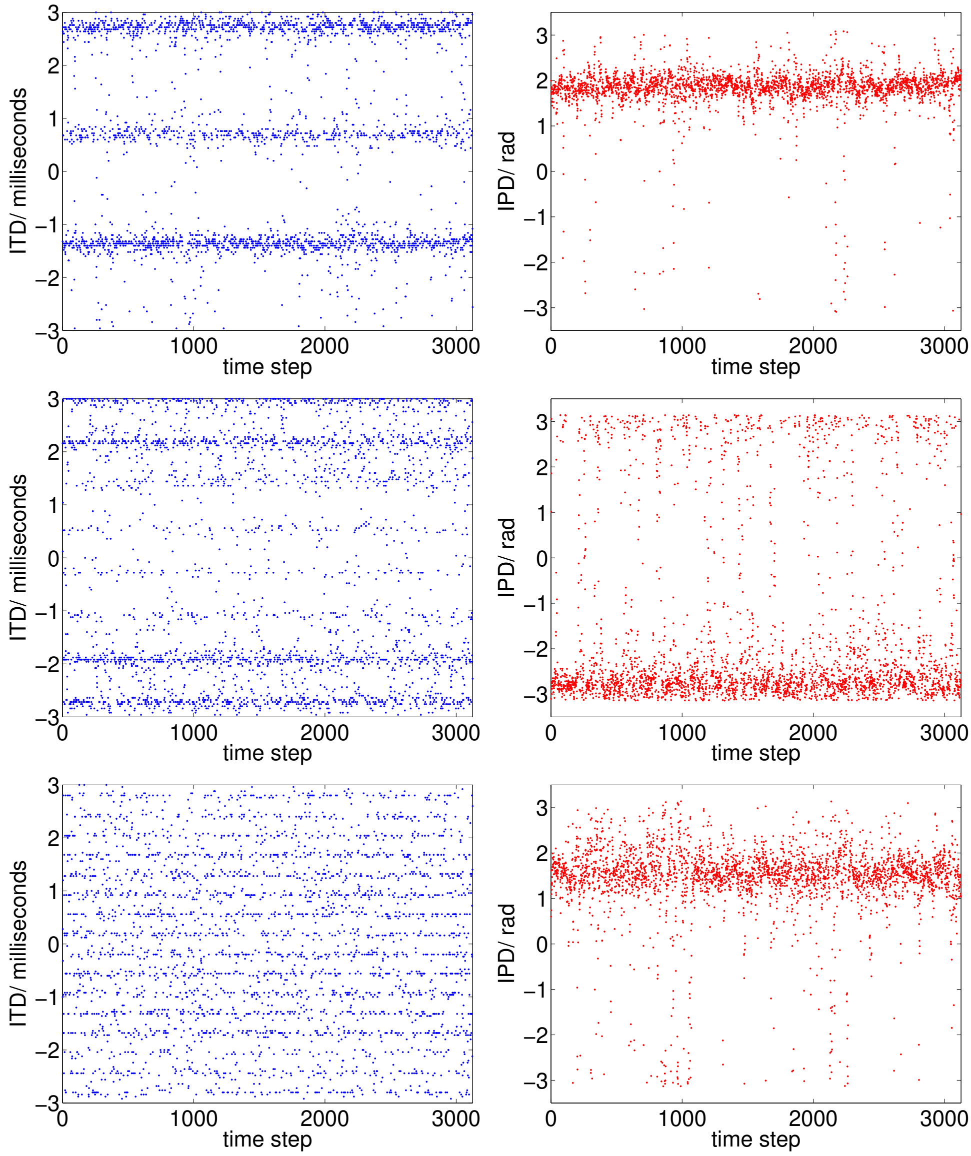


Figure D.2.: Time series of the argument of the maximum of the ICCF (left) and the IPD (right) for different frequencies, and an SNR of 20 dB. Top row: band 11 (630 Hz); mid row: band 20 (1400 Hz); lower row: band 30 (3100 Hz). Azimuth and elevation are always 60° and 0°, respectively.

E. Distributions of Interaural Level and Phase Differences as Functions of Signal-To-Noise Ratio and Frequency

Figure E.1 and Fig.E.2 show the mean value and the standard deviation of the ILD and the IPD as a function of frequency for two sound sources at different azimuth pairs (0° and 5°, 0° and 15°, and 60° and 70°). Both sources are mixed with cafeteria noise at an SNR of 5 dB. The difference of the mean values of the ILD caused by the different azimuths is much smaller than the standard deviation, even at high frequencies. For the IPD, at least at higher frequencies, the distributions shown have a smaller overlap, but because of the cyclical nature of the phase variable, the mean values become ambiguous for frequencies higher than 1200 Hz (17.0 ERB).

In Figs. E.3 and Figs. E.4, histograms of ILD and IPD for a signal from one direction mixed with cafeteria noise at several SNRs are shown. With decreasing SNR, the distribution of the ILD values becomes much wider and changes its shape first to a skewed PDF (observe the contour lines), and then to a broad parable. Also, the mean value moves towards zero. Corresponding changes occur with the distribution of the IPD; at low SNR, it becomes similar an uniform circular distribution.

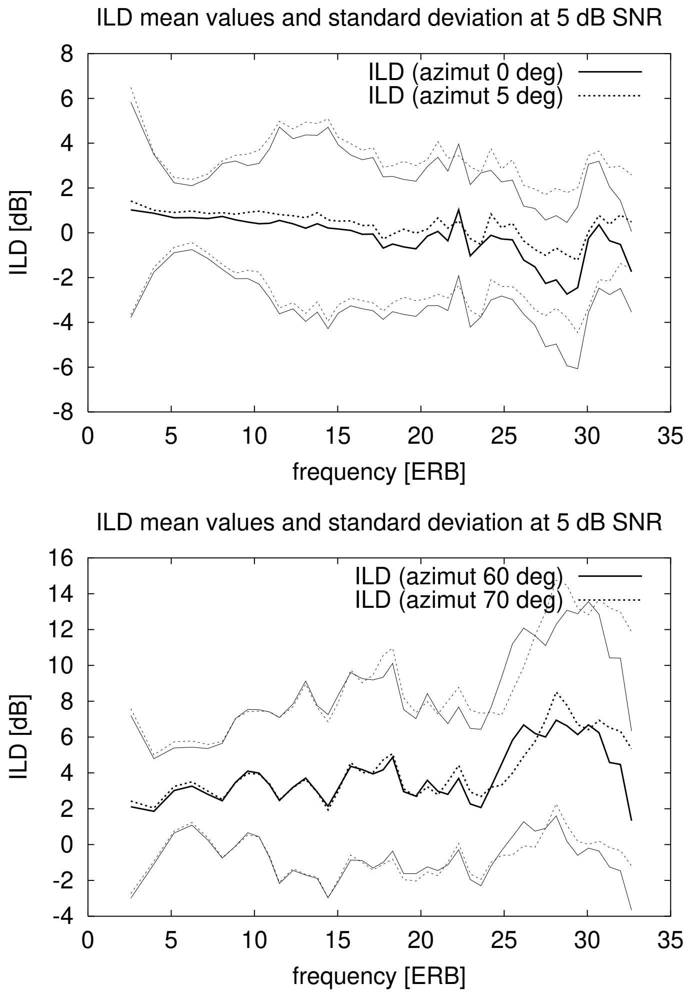


Figure E.1.: Upper panel: Mean value (thick lines) $\pm$ standard deviation (thin lines) of ILD for 0° azimuth (solid lines) and 5° azimuth (dashed). Abscissa: Frequency in ERB, ordinate: ILD in dB. The signal is speech in cafeteria noise at 5 dB SNR. Lower panel: Same as upper panel, with azimuth 60° (solid lines) and 70° (dashed). The elevation is 0° in both cases.

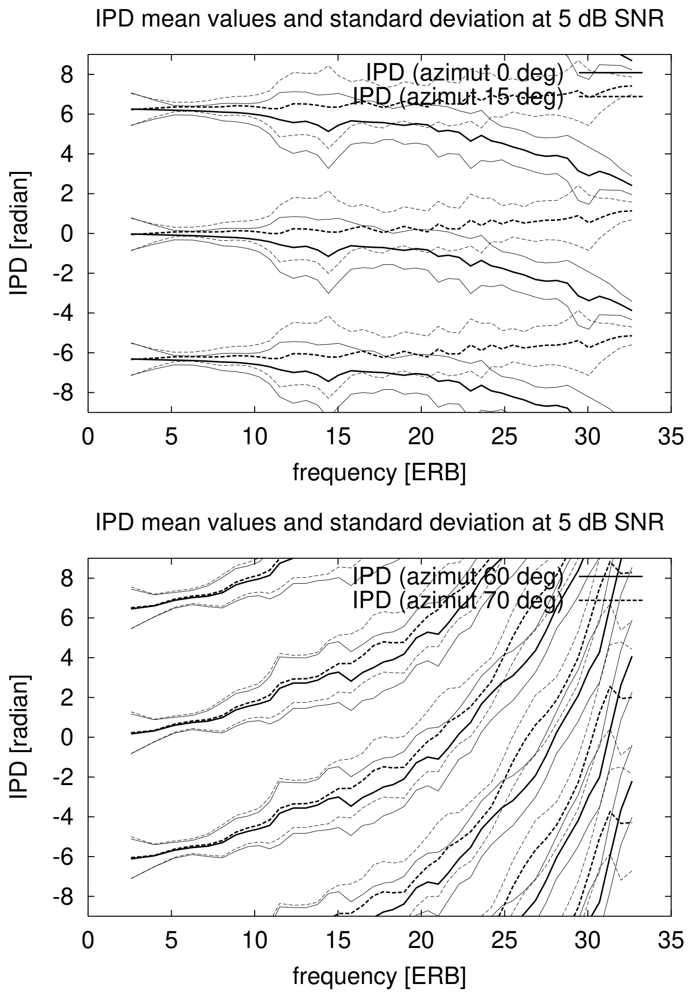


Figure E.2.: Upper panel: Mean value (thick lines) $\pm$ standard deviation (thin lines) of IPD for 0° azimuth (solid lines) and 15° azimuth (dashed). Abscissa: Frequency in ERB, ordinate: IPD in radian. To account for the cyclical nature of the IPD, the curves have been unwrapped, shifted and repeated vertically at multiples of 2π, and an interval of $[-3\pi \ldots 3\pi]$ is displayed. In the lower panel, the azimuths are 60° (solid lines) and 70° (dashed). The signal is speech in cafeteria noise at 5 dB SNR, the elevation is 0° in both cases.

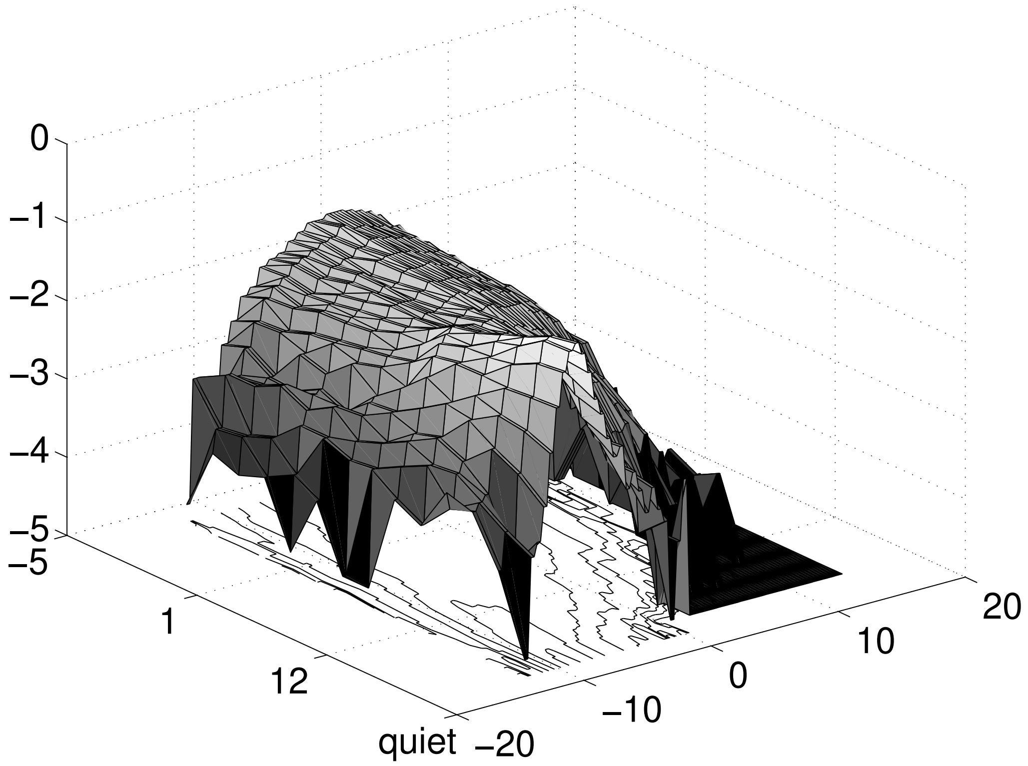


Figure E.3.: Histograms of ILD in cafeteria noise for a signal from -45° azimuth at SNRs from -5 dB to silence, and a frequency of 3100 Hz (approx. 24.8 ERB) The abscissa give the ILD in decibel on the x-axis, and the SNR of the signal on the y-axis. The ordinate is the frequency of occurrence of each ILD value, logarithmically scaled with base 10. The figure shows that at lower SNRs, the histograms become broader, change their shape, and that the maxima of the histograms shift toward zero. Note that in this log-scaled representation, histograms matching a Gaussian PDF have parabolic shape.

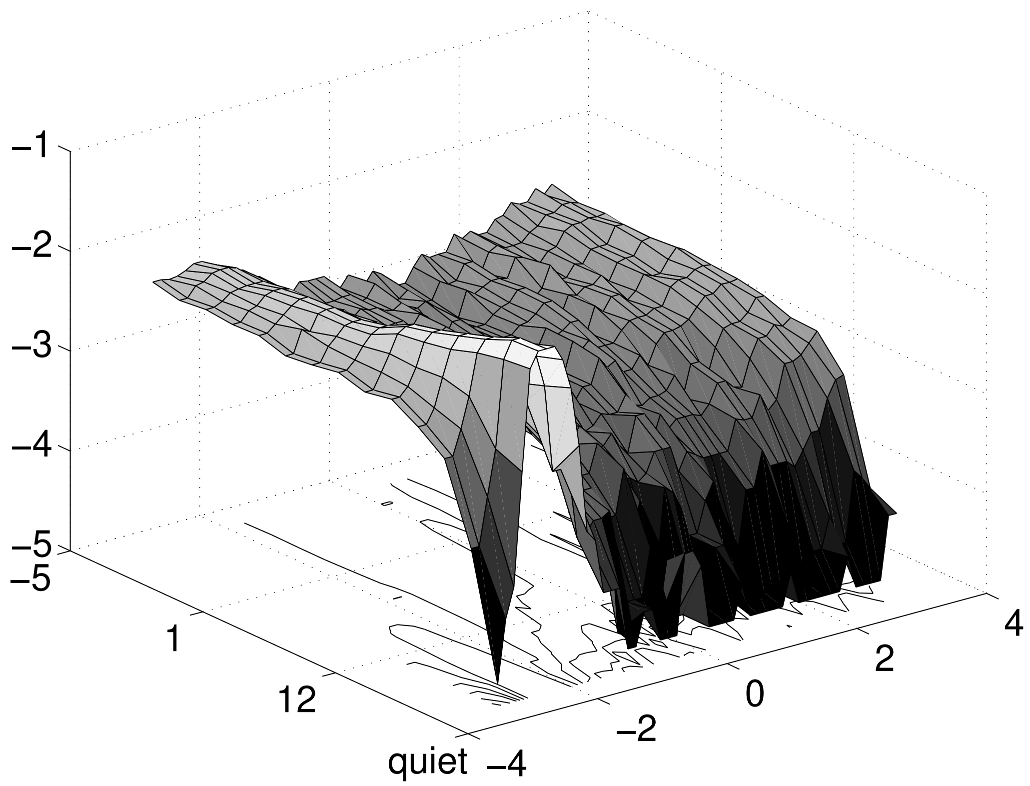


Figure E.4.: Histograms of IPD in cafeteria noise, same signals as in Fig. E.3, but for a frequency of 634 HZ (approx. 12.3 ERB). The abscissa are the value of the IPD in radian (x-axis), and the SNR of the signal (z-axis). The ordinate is the frequency of occurrence of each IPD value, logarithmically scaled with base 10. At lower SNRs, the histograms become broader and similar to a uniform distribution for the interval $[-\pi \ldots \pi]$.

Bibliography

Acero, A., Altschuler, S., and Wu, L. (**2000**), “Speech/noise separation using two microphones and a VQ model of speech signals,” in “Proc. Int. Conf. on Spoken Lang. Proc. ICLSP 2000,” (Beijing), volume 4, pp. 532–535, [Online] `http://www.ee.columbia.edu/~dpwe/papers/AceroAW00-vqica.pdf`.

Albani, S., Peissig, J., and Kollmeier, B. (**1996**), “Model of binaural localization resolving multiple sources and spatial ambiguities,” in Kollmeier (1996), pp. 227–232.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (**2001**), “The CIPIC HRTF database,” in “Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics,” (Mohonk Mountain House, New Paltz, NY), pp. 99–102, [Online] `http://interface.cipic.ucdavis.edu/data/doc/CIPIC_HRTF_Database.pdf`.

Allen, J. B. (**1977**), “Short term spectral analysis, synthesis and modification by discrete Fourier transform,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-25**(3):235 – 238.

Allen, J. B., Berkley, D. A., and Blauert, J. (**1977**), “Multimicrophone signal-processing technique to remove room reverberation form speech signals,” *J. Acoust. Soc. Am.*, **62**(4):912–915.

Allen, J. B., and Rabiner, L. R. (**1977**), “A unified approach to short-time Fourier analysis and synthesis,” in “Proceedings of the IEEE,” (IEEE), volume 65, pp. 1558–1564.

Anemüller, J. (**2001**), *Across-frequency processing in convolutive blind source separation*, Ph.D. thesis, University of Oldenburg, Germany.

Anemüller, J., and Kollmeier, B. (**2000**), “Amplitude modulation decorrelation for convolutive blind source separation,” in P. Pajunen, and J. Karhunen (editors), “Proceedings of the second international workshop on independent component analysis and blind signal separation,” pp. 215–220, http://www.physik.uni-oldenburg.de/Docs/medi/members/ane/pub.

Anemüller, J., and Kollmeier, B. (**2003**), "Adaptive separation of acoustic sources for anechoic conditions: A constrained frequency domain approach," *Speech Communication*, **39**(1-2):79–95, [Online] http://sccn.ucsd.edu/~jorn/webpage/bib/pdf/AnemullerK_SpeCom_2003.pdf.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (**2002**), "A tutorial on particle filters for online nonlinear/non-Gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, **50**(2):174–188, [Online] http://www.robots.ox.ac.uk/~parg/mlrg/papers/pfilt.ps.

Barker, J., Cooke, M., and Ellis, D. (**2005**), "Decoding speech in the presence of other sources," *Speech Communication*, **45**(1).

Belin, P., and Zatorre, R. J. (**2000**), "'What', 'where', and 'how' in auditory cortex," *Nature Neuroscience*, **3**(10):965–966, [Online] http://www.nature.com/neuro/journal/v3/n10/full/nn1000_965.html.

Bell, A. J., and Sejnowski, T. J. (**1995**), "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, **7**(6):1129–1159, [Online] ftp.salk.edu/pub/tony/bell.blind.ps.Z.

Bendat, J. S., and Piersol, A. G. (**1986**), *Random data – Analysis and measurement procedures* (John Wiley & Sons, New York, Chichester, Brisbane), second edition.

Berzuini, C., and Gilks, W. (**2001**), "RESAMPLE-MOVE filtering with cross-model jumps," in A. Doucet, N. de Freitas, and N. Gordon (editors), "Sequential Monte Carlo Methods in Practice," (Springer), Statistics for Engineering and Information Science, chapter 6, pp. 117–138.

Beutelmann, R., and Brand, T. (**2005**), "Modelling binaural speech intelligibility for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, (submitted).

Blake, A., and Isard, M. (**1998**), *Active Contours: The Application of Techniques from Graphics,Vision,Control Theory and Statistics to Visual Tracking of Shapes in Motion* (Springer, New York).

Blauert, J. (**1983**), *Spatial Hearing - The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, Massachusetts).

Bodden, M. (**1992**), *Binaurale Signalverarbeitung: Modellierung der Richtungserkennung und des Cocktail-Party-Effekts*, Ph.D. thesis, Ruhr-Universität Bochum, Düsseldorf.

Bodden, M. (**1993**), "Modeling human sound source localization and the cocktail-party-effect," *Acta Acustica*, **1**(1):43–55.

Bodden, M. (**1996**a), "Auditory demonstrations of a cocktail-party-processor," *Acta Acustica*, **82**(2):356 – 357.

Bodden, M. (**1996**b), "Binaural models and cocktail party processors," *Acta Acustica*, **82**(Suppl. 1):86.

Boll, S. F. (**1979**), "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-27**(2):113 – 120.

Boll, S. F. (**1992**), "Speech enhancement in the 1980s: Noise supression with pattern matching," in S. Furui, and M. M. Sondhi (editors), "Advances in Speech Signal Processing," (Marcel Dekker, Inc., New York), pp. 309–325.

Borisyuk, A., Semple, M. N., and Rinzel, J. (**2002**), "Adaptation and inhibition underlie responses to time-varying interaural phase cues in a model of inferior colliculus neurons," *Journal of Neurophysiology*, **88**(4):2134–2146.

Braasch, J. (**2002**a), *Auditory Localization and Detection in Multiple-Sound Source Scenarios*, Ph.D. thesis, Ruhr-Universität Bochum, Düsseldorf, VDI-Verlag. Fortschritts-Berichte VDI Reihe 10 Nr.707.

Braasch, J. (**2002**b), "Localization in presence of a distracter and reverberation in the frontal horizontal plane. II. Model algorithms," *ACUSTICA/acta acustica*, **88**:956–969.

Braasch, J., and Hartung, K. (**2002**), "Localization in presence of a distracter and reverberation in the frontal horizontal plane. I. Psychoacoustical data," *ACUSTICA/acta acustica*, **88**:942–955.

Brainard, M. S., Knudsen, E. I., and Esterly, S. D. (**1992**), "Neural derivation of sound source location: Resolution of spatial ambiguities on binaural cues," *J. Acoust. Soc. Am.*, **91**(2):1015–1027.

Brandstein, M., and Ward, D. (editors) (**2001**), *Microphone Arrays. Signal Processing Techniques and Applications.* (Springer, Berlin).

Breebaart, J., van de Par, S., and Kohlrausch, A. (**1999**), "The contribution of static and dynamically varying ITDs and IIDs to binaural detection," *J. Acoust. Soc. Am.*, **106(2)**:979–992.

Bregman, A. S. (**1990**), *Auditory Scene Analysis: The perceptual Organization of Sound* (MIT Press, Cambridge, Massachusetts).

Bregman, A. S. (**1993**), "Auditory scene analysis: Hearing in complex environments," in McAdams and Bigand (1993), chapter 2, pp. 10–36.

Bronkhorst, A. W. (**2000**), "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acustica*, **86**(1):117–128.

Bronkhorst, A. W., and Plomp, R. (**1989**), "Binaural speech intelligibility in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, **86**(4):1374–1383.

Brown, G. J. (**1992**), *Computational auditory scene analysis: A representational approach*, Ph.d. thesis, University of Sheffield, Department of Computer Science, Sheffield, England, UK.

Brown, G. J., and Cooke, M. P. (**1994**), "Computational auditory scene analysis," *Computer Speech and Language*, **8**(4):297–336.

Brugge, J. F. (**1992**), "An overview of central auditory processing," in Popper and Fay (1992), chapter 1, pp. 1–33.

Burns, G., Daoud, R., and Vaigl, J. (**1994**), "LAM: An Open Cluster Environment for MPI," in "Proceedings of Supercomputing Symposium," pp. 379–386, [Online] `http://www.lam-mpi.org/download/files/lam-papers.tar.gz`.

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., and Hallett, M. (**2003**), "Neural correlates of cross-modal binding," *Nature Neurosci.*, **6**(2):190–195.

Butler, R. A. (**1986**), "The bandwidth effect on monaural and binaural localization," *Hear Res.*, **21(1)**:67–73.

Caird, D., and Klinke, R. (**1987**), "Processing of interaural time and intensity differences in the cat inferior colliculus," *Experimental Brain Research*, **68**(2):379 – 392.

Cappé, O. (**1994**), "Elimination of the musical noise phenomenon with the Epharim and Malah noise suppressor," *IEEE Trans. Speech and Audio Processing*, **2**(2):345–349.

Carlyon, R. P., and McAdams, R. S. (**1992**), "The psychophysics of current sound segregation," *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences*, **336**(1278):347–355.

Cherry, E. C. (**1953**), "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, **25**(5):975 – 979.

Cherry, E. C. (**1959**), "Two ears – but one world," in W. A. Rosenblith (editor), "Sensory communication : Contribtions to the Symposium on Principles of Sensory Communication," MIT (The MIT Press, Cambridge, Mass.), pp. 99–117.

Cherry, E. C., and Wiley, R. (**1967**), "Speech communication in very noisy environments." *Nature*, **214**(93):1164.

Cheveigné, d., Alain (**1993**), "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *J. Acoust. Soc. Am.*, **93**(6):3271–3290.

Choo, K., and Fleet, D. (**2001**), "People tracking using hybrid Monte Carlo filtering," in "Proc. IEEE International Conference on Computer Vision, Vancouver," volume 2, pp. 321–328, [Online] `http://www2.parc.com/spl/members/fleet/research/Papers/iccv2001.pdf`.

Chung, W., Carlile, S., and Leong, P. (**2000**), "A performance adequate computational model for auditory localization," *J. Acoust. Soc. Am.*, **107**(1):432–445.

Clarey, J. C., Barone, P., and Imig, T. J. (**1992**), "Physiology of thalamus and cortex," in Popper and Fay (1992), chapter 5, pp. 232–334.

Cohen, I., and Berdugo, B. (**2001**), "Speech enhancement for non-stationary noise environments," *Signal Processing*, **81**:2403–2418.

Colburn, H. S. (**1996**), "Computational models of binaural processing," in Hawkins *et al.* (1996), chapter 8, pp. 332–400.

Colburn, H. S., and Hawley, M. L. (**1996**), "Models of impaired binaural hearing," *Acta Acustica*, **82**(Suppl. 1):S86.

Colburn, H. S., Zurek, P. M., and Durlach, N. I. (**1987**), "Binaural directional hearing - impairments and aids," in W. A. Yost, and G. Gourevitch (editors), "Directional Hearing," (Springer, New York), chapter 11, pp. 261 – 278.

Compernolle, D. v. (**2001**), "Future directions in microphone array processing," in Brandstein and Ward (2001), chapter 18, pp. 389–394.

Cooke, M. (**1993**), *Modeling Auditory Processing and Organization*, Distinguished Dissertations in Computer Science (Cambridge University Press), ph.D. Thesis.

Cooke, M., Brown, G. J., and Green, P. (**1993**), "Computational auditory scene analysis: Listening to several things at once," *Endeavour*, **17**(4):186–190.

Cooke, M., and Ellis, D. P. W. (**2001**), "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, **35**(3-4):141–177.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**a), "Robust automatic speech recognition with missing and uncertain acoustic data," *Speech Communication*, **34**(3):267–285.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**b), "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, **34**(3):267–286.

Culling, J. F., and Summerfield, Q. (**1995**), "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," *J. Acoust. Soc. Am.*, **98**(2):785–797.

Daly, M., Reilly, J. P., and Manton, J. (**2004**), "A bayesian approach to blind source recovery," in "Asilomar Conference in Signals, Sensors and System," [Online] `http://www.ece.mcmaster.ca/~reilly/html/papers/mike/Asilomar2004.pdf`.

Damaske, P., and Wagner, B. (**1969**), "Richtungshörversuche über einen nachgebildeten Kopf," *Acta Acustica*, **21**:30–35.

Datum, M. S., Palmieri, F., and Moiseff, A. (**1996**), "An artificial neural-network for sound localization using binaural cues," *J. Acoust. Soc. Am.*, **100**(1):372–383.

Domnitz, R. H., and Colburn, H. S. (**1976**), "Analysis of binaural detection models for dependence on interaural target parameters," *J. Acoust. Soc. Am.*, **59**(3):598–601.

Doucet, A., de Freitas, N., and Gordon, N. (**2001**), "An introduction to sequential Monte Carlo methods," in A. Doucet, N. de Freitas, and N. Gordon (editors), "Sequential Monte Carlo Methods in Practice," (Springer), Statistics for Engineering and Information Science, chapter 1, pp. 3–14.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**), "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing aid assessment," *Audiology*, **40**:148–157.

Duda, R. O. (**1997**), "Elevation dependence of the interaural transfer function," in Gilkey and Anderson (1997), chapter 3, pp. 49–75.

Durlach, N. I., and Colburn, H. S. (**1978**), "Binaural phenomena," in E. C. Carterette, and M. P. Friedman (editors), "Handbook of Perception - Hearing," (Academic Press, New York), volume 4, chapter 10, pp. 365 – 466.

Durlach, N. I., and Pang, X. D. (**1986**), "Interaural magnification," *J. Acoust. Soc. Am.*, **80**(6):1849–1850.

Edwards, B. (**2004**), "Hearing aids and hearing impairment," in S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (editors), "Speech Processing in the Auditory System," (Springer, New York), Springer Handbook of Auditory Research, pp. 339–421.

Ellis, D. P. W. (**1996**), *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, Massachusetts Institute of Technology, MA, [Online] `http://sound.media.mit.edu/~dpwe/pdcasa/doc.html`.

Ephraim, Y. (**1992**), "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Transactions on Signal Processing*, **40**(4):725.

Ephraim, Y., and Malah, D. (**1984**), "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-32**(6):1109–1121.

Ephraim, Y., and Malah, D. (**1985**), "Spech enhancement using a minimum mean-square error log spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-33**(2):443–445.

Ephraim, Y., Malah, D., and Juang, B. H. (**1989**a), "On the application of hidden Markov-models for enhancing noisy speech," *IEEE Transactions on Acoustics, Speech and Signal Processing ASSP*, **37**(12):1846–1856.

Ephraim, Y., Malah, D., and Juang, B.-H. (**1989**b), "Speech enhancement based upon hidden markov modeling," in "International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89.", (Glasgow, UK), volume 1, pp. 353–356.

Evans, E. F. (**1992**), "Auditory processing of complex sounds: An overview," *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences*, **336**(1278):295–306.

Fisher, N. I. (**1993**), *Statistical analysis of circular data* (Cambridge University Press, Cambridge, New York).

Fitzpatrick, D., Batra, R., and Stanford, T. (**1997**), "A neural population code for sound localization," *Nature*, **388**:871–874.

Fong, W., Godsill, S. J., Doucet, A., and West, M. (**2002**), "Monte Carlo smoothing with application to audio signal enhancement," *IEEE Transactions on Signal Processing*, **50**(2):438–449, [Online] `http://www-sigproc.eng.cam.ac.uk/~sjg/`.

Gales, M. J., and Young, S. J. (**1993**), "Parallel model combination for speech recognition in noise," Technical Report CUED/F-INFENG/TR135, Cambridge University Engineering Department, [Online] `http://mi.eng.cam.ac.uk/reports/abstracts/gales_tr135.html`.

Gallmeister, B. O. (**1995**), *POSIX.4: Programming for the real world* (O'Reilly, Sebastopol, CA).

Gandhi, M. A., and Hasegawa-Johnson, M. A. (**2004**), "Source separation using particle filters," in "Proceedings of the ICSLP 2004," (ISCA), pp. 2673–2676, [Online] `http://www.ifp.uiuc.edu/speech/pubs/2004/gandhi_icslp2004.pdf`.

Garofolo, J. (**1998**), *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database*, National Institute of Standards and technology (NIST), Gaithersburgh, Maryland.

Gelb, A. C. (editor) (**1994**), *Applied Optimal Estimation* (The MIT Press, Cambridge, Ma.), 12th edition.

Gilkey, R. H., and Anderson, T. R. (editors) (**1997**), *Binaural and Spatial Hearing in Real and Virtual Environments* (Lawrence Erlbaum Assoc., Mahwah, New Jersey).

Godsill, S. J., and Rayner, P. J. (**1998**), *Digital Audio Restoration* (Springer, Berlin).

Godsmark, D., and Brown, G. J. (**1999**), "A blackboard architecture for computational auditory scene analysis," *Speech Communication*, **27**(3-4):351–366, [Online] `http://www.dcs.shef.ac.uk/~guy/pdf/spcom99.pdf`.

Good, M. D., and Gilkey, R. H. (**1996**), "Sound localization in noise: The effect of signal-to-noise ratio," *J. Acoust. Soc. Am.*, **99**(2):1108–1117.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (**1993**), "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *Radar and Signal Processing, IEE Proceedings F*, **140**(2):107–113.

Greenberg, J. E., Desloge, J. G., and Zurek, P. M. (**2003**), "Evaluation of array-processing algorithms for a headband hearing aid," *J. Acoust. Soc. Am.*, **113**(3):1646–1657.

Greenberg, J. E., and Zurek, P. M. (**2001**), "Microphone-array hearing aids," in Brandstein and Ward (2001), chapter 11, pp. 229–253.

Griffiths, L. J., and Jim, C. W. (**1982**), "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagat.*, **30**(1):27–34.

Hancock, K. E., and Delgutte, B. (**2004**), "A physiologically based model of interaural time difference discrimination," *Journal of Neuroscience*, **24**(32):7110–7117.

Harding, S., and Meyer, G. (**2001**), "A case for multi-resolution auditory scene analysis," in "Proceedings of Eurospeech, Aalborg," (ISCA), pp. 159–162.

Harper, N. S., and McAlpine, D. (**2004**), "Optimal neural population coding of an auditory spatial cue," *Nature*, **430**:682–686.

Hawkins, H. L., McMullen, T. A., Popper, A. N., and Fay, R. R. (editors) (**1996**), *Auditory Computation*, volume 6 of *Springer Handbook of Auditory Research* (Springer, New York).

Hawley, M., Litovsky, R., and Colburn, S. (**1999**), "Speech intelligibility and localization in a multi-source environment," *J. Acoust. Soc. Am.*, **105**(6):3436–3448.

Hawley, M., Litowsky, R. Y., and Colburn, S. (**1998**), "Speech intelligibility in a complex environment by listeners with hearing impairment," in "Proc. ICASSP," pp. 2609 – 2610.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**), "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *J. Acoust. Soc. Am.*, **115**(2):833–843.

Henning, G. B. (**1973**), "Effect of interaural phase on frequency and amplitude discrimination," *J. Acoust. Soc. Am.*, **54**(5):1160–1178.

Hoffman, M. W., and Buckley, K. M. (**1995**), "Robust time-domain processing of broadband microphone array data," *IEEE Transactions on Speech and Audio Processing*, **3**(3):193–203.

Hofman, P. M., van Riswick, J. G. A., and van Opstal, A. J. (**1998**), "Relearning sound localization," *Nature Neuroscience*, **1**:417–421.

Hohmann, V., Albani, S., and Nix, J. (**1999**), "Application of localization models to noise suppression in hearing aids," *ACUSTICA - acta acustica*, **85**:S225, talk presented at the 137th meeting of the Acoustical Society of America joint with Forum Acusticum 1999 in Berlin. See also JASA **105**, 1151.

Hohmann, V., Nix, J., Grimm, G., and Wittkop, T. (**2002**a), "Binaural noise reduction for hearing aid," in "Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)," (IEEE), volume 4, pp. 4000–4003.

Hohmann, V., Nix, J., Grimm, G., and Wittkop, T. (**2002**b), "Cocktail party processing based on interaural parameters," *Acta acustica united with acustica*, **88**(Suppl. 1):1–6.

Häusler, R., Colburn, H. S., and Marr, E. (**1983**), "Sound localization in subjects with impaired hearing," *Acta Oto-Laryngologica*, **S400**:1–62.

Isabelle, S. K., Janko, J. A., and Gilkey, R. H. (**1998**), "A model of auditory localization using neural networks," *J. Acoust. Soc. Am.*, **103**(5):2845.

Isard, M. (**1998**), "The condensation algorithm," [Online] `http://www.robots.ox.ac.uk/~misard/condensation.html`.

Ito, Y., Colburn, H. S., and Thompson, C. L. (**1982**), "Masked discrimination of interaural time delays with narrow-band signal," *J. Acoust. Soc. Am.*, **72**(6):1821–1826.

Jacobsen, T. (**1976**), "Localization in noise," Technical Report 10, The Acoustics Laboratory, Technical University of Denmark.

Janko, J. A., Anderson, T. R., and Gilkey, R. H. (**1997**), "Using neural networks to evaluate the viability of monaural and interaural cues for sound localization," in Gilkey and Anderson (1997), chapter 26, pp. 557–570.

Jeffress, L. A. (**1948**), "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, **41**:35–39.

Jenison, R. L. (**2000**), "Correlated cortical populations can enhance sound localization performance," *J. Acoust. Soc. Am.*, **107**(1):414–421.

Joris, P. X., and Yin, T. C. T. (**1996**), "Envelope coding in the lateral superior olive. I. Sensitivity to interaural time difference," *Journal of Neurophysiology*, **73**(3):1043–1062.

Jourjine, A., Rickard, S., and Yilmaz, O. (**2000**), "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in "Proc. ICASSP2000, June 5-9, 2000, Istanbul, Turkey," volume 5, pp. 2985–2988, [Online] `http://www.math.princeton.edu/~srickard/bss/DUET_ICASSP2000.ps`.

Kailath, T. (editor) (**1977**), *Linear Least-Squares Estimation*, number 17 in Benchmark Papers in Electrical Engineering and Computer Science (Dowden, Hutchinson & Ross, Inc., Stroudsburg, Pennsylvania).

Kailath, T. (**1981**), *Lectures on Wiener and Kalman Filtering*, volume 140 of *CISM Courses and Lectures* (Springer).

Kalman, R. E. (**1960**), "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, **82**(Series D):35–45, reproduced in (Kailath, 1977, pp. 254 – 264).

Kidd, G., Jr., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (**1998**), "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns," *J. Acoust. Soc. Am.*, **104**(1):422–431.

Kitagawa, G. (**1996**), "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models," *Journal of computational and graphical statistics*, **5**(1):1–25.

Knudsen, E. I. (**1982**), "Auditory and visual maps of space in the optic tectum of the owl," *J. Neurosci.*, **2**:1177 – 1194.

Knudsen, E. I., and Konishi, M. (**1978**), "A neural map of auditory space in the owl," *Science*, **200**(4343):795 – 797.

Koch, R., Püschel, D., and Kollmeier, B. (**1991**), "Simulation des Cocktail-Party-Effektes, Störgeräuschreduktion in räumlichen Hörsituationen mit Hilfe binauraler Modulationsspektren," in "Fortschritte der Akustik - DAGA 1991," (DPG), pp. 797–800.

Kollmeier, B. (editor) (**1996**), *Psychoacoustics, Speech and Hearing Aids* (World Scientific Publishing, Singapore).

Kollmeier, B., and Koch, R. (**1994**), "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Am.*, **95**(3):1593–1602.

Kollmeier, B., Peissig, J., and Hohmann, V. (**1993**), "Real-time multiband dynamic compression and noise reduction for binaural hearing aids," *J. Rehabil. Res. Dev.*, **30**(1):82–94.

Kompis, M., and Dillier, N. (**1994**), "Noise reduction for hearing aids: Combining directional microphones with an adaptive beamformer," *J. Acoust. Soc. Am.*, **96**(3):1910–1913.

Kopp, B. (**1978**), "Hierarchical classification III: Average-linkage, median, centroid, WARD, flexible strategy," *Biometrical J.*, **20**(7/8):703–711.

Korompis, D., Wang, A., and Yao, K. (**1995**), "Comparison of microphone array designs for hearing aid," in "Proceedings of the ICASSP," (IEEE), pp. 2739–2742.

Kuwada, S., and Yin, T. C. T. (**1983**), "Binaural interaction in low-frequency neurons in inferior colliculus of the cat. I. Effects of long interaural delays, intensity, and repetition rate on interaural delay function," *Journal of Neurophysiology*, **50**(4):981–999.

Kuwada, S., and Yin, T. C. T. (**1987**), "Physiological studies of directional hearing," in W. A. Yost, and G. Gourevitch (editors), "Directional Hearing," (Springer, New York), chapter 6, pp. 146–176.

Kysela, J., and many others (**2000-2005**), "Linux advanced sound architecture (ALSA)," [Online] `http://www.alsa-project.org`.

Langtangen, H. P. (**2004**), *Python Scripting for Computational Science*, volume 3 of *Texts in Computational Science and Engineering* (Springer).

Larocque, J. R., Reilly, J. P., and Ng, W. (**2002**), "Particle filters for joint detection, estimation and tracking of an unknown number of directions of arrival," *IEEE Transactions on Signal Processing*, **50**(12):2926–2937, [Online] `http://www-sigproc.eng.cam.ac.uk/~kfn20/papers/particle_tracking.pdf`.

Lee, K. Y., and Jung, S. (**2000**), "Time-domain approach using multiple Kalman filters and EM algorithm to speech enhancement with nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, **8**(3):282–291.

Levitt, H. (**2001**), "Noise reduction in hearing aids: A review," *Journal of Rehabilitation Reserach and Development*, [Online] `http://www.vard.org/jour/01/38/1/pdf/levitt.pdf`.

Lewicki, M. S. (**2002**), "Efficient coding of natural sounds," *Nat. Neuroscience*, **5**(4):356–363, [Online] `http://www.cnbc.cmu.edu/~tai/readings/circuit/lewicki_natural-sounds.pdf`.

Li, K. P., Hughes, G. W., and House, A. S. (**1969**), "Correlation characteristics and dimensionality of speech spectra," *J. Acoust. Soc. Am.*, **46**(4):1019–1025.

Lim, J. S., Oppenheim, A. V., and Braida, L. D. (**1978**), "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal Processing*, **26**(4):354 – 358.

Lim, S., Jae (**1978**), "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, **26**(5):471–472.

Lindemann, W. (**1986**a), "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," *J. Acoust. Soc. Am.*, **80**(6):1608 – 1622.

Lindemann, W. (**1986**b), "Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front," *J. Acoust. Soc. Am.*, **80**(6):1623–1630.

Liu, C., Rosenhouse, J., and Sideman, S. (**1997**), "A targeting-and-extracting technique to enhance hearing in the presence of competing speech," *J. Acoust. Soc. Am.*, **101**(5):2877–2891.

Liu, C., Wheeler, B. C., O'Brien, W. D., Jr., and Bilger, R. C. (**2000**), "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Am.*, **108**(4):1888–1905.

Liu, C., Wheeler, B. C., O'Brien, W. D., Jr., Lansing, C. R., Bilger, R. C., Jones, D. L., and Feng, A. S. (**2001**), "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," *J. Acoust. Soc. Am.*, **110**(6):3218–3231.

Lutz, M. (**2001**), *Programming Python* (O'Reilly, Sebastopol, CA).

Lutz, M., and Ascher, D. (**2003**), *Learning Python* (O'Reilly, Sebastopol, CA).

Lyon, R. F. (**1983**), "A computational model of binaural localization and separation," in "Proceedings of the International Conference on Acoustics, Speech and Signal Processing ICASSP' 83," (IEEE), volume 3, pp. 1148–1151.

MacCormick, J., and Blake, A. (**2000**), "A probabilistic exclusion principle for tracking multiple objects," *International Journal of Computer Vision*, **39**(1):57–71, [Online] `http://www.robots.ox.ac.uk/~vdg/`.

MacCormick, J., and Isard, M. (**2000**), "Partitioned sampling, articulated objects, and interface-quality hand tracking," in "Proceedings of the 6th European Conference on Computer Vision," (Springer, London), volume 2, pp. 3–19, [Online] `http://www.robots.ox.ac.uk/~vdg/abstracts/eccv2000-handtracking.html`.

MacKay, D. J. C. (**1999**), "Introduction to Monte Carlo methods," in M. I. Jordan (editor), "Learning in Graphical Models," (MIT Press, Cambridge, MA), chapter 7, pp. 175 – 204, [Online] `ftp://wol.ra.phy.cam.ac.uk/pub/mackay/erice.ps.gz`.

Macpherson, E. A., and Middlebrooks, J. C. (**2002**), "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Am.*, **111**(5):2219–2236.

Marin, C. M. H., and McAdams, S. (**1991**), "Segregation of concurrent sounds. II Effects of spectral envelope tracing, frequency-modulation coherence, and frequency-modulation width," *J. Acoust. Soc. Am.*, **89**(1):341–351.

Marr, D. (**1980**), *Vision* (W. H. Freedman Press, San Francisco).

Marzinzik, M. (**2000**), *Noise reduction schemes for digital hearing aids and their use for the hearing impaired*, Ph.D. thesis, Universität Oldenburg, Medical Physics, [Online] `http://www.bis.uni-oldenburg.de/dissertation/2001/marnoi00/marnoi00.html`.

McAdams, S. (**1993**), "Recognition of sound sources and events," in McAdams and Bigand (1993), chapter 6, pp. 147–197.

McAdams, S., and Bigand, E. (editors) (**1993**), *Thinking in Sound: The Cognitive Psychology of Human Audition* (Clarendon Press).

McAlpine, D., and Grothe, B. (**2003**), "Sound localization and delay lines - do mammals fit the model ?" *Trends in Neurosciences*, **26**(7):347–350.

Mehrgardt, S., and Mellert, V. (**1977**), "Transformation characteristics of the external human ear," *J. Acoust. Soc. Am.*, **61**(6):1567–1576.

Mellinger, D. K., and Mont-Reynaud, B. (**1996**), "Scene analysis," in Hawkins *et al.* (1996), chapter 7, pp. 271–329.

Meyer, R., Nix, J., and Hohmann, V. (**2005**), "Sequentielle Monte-Carlo-Verfahren zur Grundfrequenzerkennung überlagerter realer und synthetischer Vokale," in "Fortschritte der Akustik- DAGA'2005," (DEGA Deutsche Gesellschaft für Akustik, Oldenburg).

Middlebrooks, J. C., Makous, J. C., and Green, D. M. (**1989**), "Directional sensitivity of sound-pressure levels in the human ear," *J. Acoust. Soc. Am.*, **86**(1):89–108.

Moore, B. C. J. (**1989**a), *An Introduction to the Psychology of Hearing*, chapter 7, volume 1 of Moore (1989c), third edition, pp. 229–253.

Moore, B. C. J. (**1989**b), *An Introduction to the Psychology of Hearing*, chapter 3, volume 1 of Moore (1989c), third edition, pp. 100–101.

Moore, B. C. J. (**1989**c), *An Introduction to the Psychology of Hearing* (Academic Press).

Morton, A. (**2001**), "Kernel patches for low-latency scheduling," [Online] `http://www.zip.com.au/~akpm/linux/schedlat.html`.

Nakashima, H., Chisaki, Y., Usagawa, T., and Ebata, M. (**2003**), "Frequency domain binaural model based on interaural phase and level differences," *Acoust. Sci. & Tech.*, **24**(4):172–178.

Neti, C., Young, E. D., and Schneider, M. H. (**1992**), "Neural network models of sound localization based on directional filtering by the pinna," *J. Acoust. Soc. Am.*, **92**(6):3140–3156.

Nichols, B., Buttlar, D., and Proulx Farrel, J. (**1996**), *Pthreads Programming* (O'Reilly, Sebastopol, CA).

Nix, J. (**2005**), "Audio examples for sound source demixing controlled by Bayesian sound localization," [Online] `http://medi.uni-oldenburg.de/demo/demo_voice_unmixing.html`.

Nix, J., and Hohmann, V. (**1999**), "Statistics of binaural parameters and localization in noise," in T. Dau, V. Hohmann, and B. Kollmeier (editors), "Psychophysics, Physiology and Models of Hearing," (World Scientific Publishing Co., Singapore), pp. 263–266.

Nix, J., and Hohmann, V. (**2000**), "Robuste Lokalisation im Störgeräusch auf der Basis statistischer Referenzen," in A. Sill (editor), "Fortschritte der Akustik – DAGA'2000," (DEGA (Deutsche Gesellschaft für Akustik e. V.), Oldenburg), pp. 384–385.

Nix, J., and Hohmann, V. (**2001**), "Enhancing sound sources by use of spatial cues," in P. Dalsgaard (editor), "Proceedings of the Eurospeech 2001 Workshop on Consistent and Reliable Acoustical Cues (CRAC)," (International Speech Communication Association), pp. 1–4, [Online] `http://www.ee.columbia.edu/crac/program.html`.

Nix, J., Kleinschmidt, M., and Hohmann, V. (**2003**a), "Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction," in "Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)," (ICSA International Speech Communication Association, Bonn, Germany), pp. 1441–1444.

Nix, J., Kleinschmidt, M., and Hohmann, V. (**2003**b), "Computational scene analysis of cocktail-party situations based on sequential Monte Carlo methods," in "Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems, & Computers," (IEEE Signal Processing Society), volume 1, pp. 735 –739.

Nokia Corporation, ., (**2005**), "Python for Nokia Series 60 Cell Phones," [Online] `http://www.forum.nokia.com/main/0,,034-821,00.html`.

Olshausen, B. A., and Field, D. J. (**2004**), "Sparse coding of sensory inputs," *Current Opinion in Neurobiology*, **14**:481–487.

Otten, J. (**2001**), *Factors influencing acoustical localization*, Ph.D. Thesis, Universität Oldenburg, Oldenburg, Germany, [Online] `http://docserver.bis.uni-oldenburg.de/publikationen/dissertation/2001/ottfac01/ottfac01.html`.

Papoulis, A. (**1965**), *Probability, Random Variables, and Stochastic Processes*, Systems Sciences (McGraw-Hill Kogakusha, Tokyo), 9th edition.

Parra, L., and Spence, C. (**2000**), "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, **8**(3):320–327, [Online] `http://newton.bme.columbia.edu/~lparra/publish/ieee-tsap.pdf`.

Peissig, J. (**1992**), *Binaurale Hörgerätestrategien in komplexen Störschallsituationen*, Ph.D. thesis, Universität Göttingen.

Perret, S., and Noble, W. (**1997**), "The contribution of head motion cues to localization of low-pass noise," *Percept. Psychophys.*, **59**:1018–1026.

Peña, J. L., and Konishi, M. (**2001**), "Auditory spatial receptive fields created by multiplication," *Science*, **292**(5515):249–252.

Popper, A. N., and Fay, R. R. (editors) (**1992**), *The Mammalian Auditory Pathway: Neurophysiology*, volume 2 of *Springer Handbook on Auditory Research* (Springer Verlag, New York).

Rayleigh, L. S. J. W. (**1907**), "On our perception of sound direction," *Philos. Mag.*, **13**:214–232.

Reyes-Gomez, M., Raj, B., and Ellis, D. P. W. (**2003**), "Multi-channel source separation by beamforming trained with factorial HMMs," in "ICASSP Hong Kong 2003," volume I, pp. 664–667, [Online] `http://www.ee.columbia.edu/~mjr59/speakerseparation3.pdf`.

Reyes-Gomez, M. J., Jojic, N., and Ellis, D. P. W. (**2004**), "Towards single-channel unsupervised source separation of speech mixtures: The layered harmonics/formants separation/tracking model," in "Workshop on Statistical and Perceptual Audio Processing (SAPA), Korea 2004," [Online] http://www.ee.columbia.edu/~dpwe/pubs/sapa04-transform.pdf.

Roman, N., Wang, D., and Brown, G. J. (**2003**), "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, **114**(4):2236–2252.

Roweis, S. (**2000**), "One microphone source separation," in "Proc. NIPS (Neural Information Processing Systems) 2000," pp. 793–799.

Sachs, L. (**1992**), *Angewandte Statistik (Applied Statistics)* (Springer, Berlin Heidelberg), 7th edition.

Sameti, H., and Deng, L. (**2002**), "Nonstationary-state hidden Markov model representation of speech signals for speech enhancement," *Signal Processing*, **82**:205–227.

Sameti, H., Sheikhzadeh, H., Deng, L., and Brennan, R. L. (**1998**), "HMM-based strategies for enhancement of speech embedded in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, **6**(5):445–455.

Schauer, C., Zahn, T., Paschke, P., , and Gross, H. M. (**2000**), "Binaural sound localization in an artificial neural network," in "Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2000," (IEEE), volume 2, pp. 865 – 868.

Schiel, F. (**2003**), "BAS (Bayrisches Archiv für Sprachsignale, Universität München): PhonDat 2 speech corpus," .

Schiel, F., Draxler, C., and Tillmann, H. (**1997**), "The Bavarian Archive for Speech Signals: Resources for the speech community," in "Proceedings of the Eurospeech, Rhodos, Greece," pp. 1687–1690, [Online] http://www.phonetik.uni-muenchen.de/Publications/Granada-98-Corpora.ps.

Schneider, W. (**1986**), *Anwendung der Korrelationstheorie der Hirnfunktion auf das akustische Figur-Hintergrund Problem (Cocktailparty Effekt)*, Dissertation, Universität Göttingen.

Schwander, T., and Levitt, H. (**1987**), "Effect of two-microphone noise reduction on speech recognition by normal-hearing listeners." *Journal of Rehabilitation Research and Development*, **24**(4):87–92.

Scott, S., and Johnsrude, I. S. (**2003**), "The neuroanatomical and functional organization of speech perception," *Trends in Neurosciences*, **26**(2):100–107.

Searle, C. L., Braida, L. D., Cuddy, D. R., and Davis, M. F. (**1975**), "Binaural pinna disparity: Another localization cue," *J. Acoust. Soc. Am.*, **57**(2):448–455.

Shackleton, T., Meddis, R., and Hewitt, M. J. (**1992**), "Across frequency integration in a model of lateralization," *J. Acoust. Soc. Am.*, **91**(4):2276–2279.

Shaw, E. A. (**1997**), "Acoustical features of the human external ear," in Gilkey and Anderson (1997), chapter 2, pp. 25–47.

Sheikhzadeh, H., Brennan, R. L., and Sameti, H. (**1995**), "Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications," in "Proc. ICASSP '95," (Detroit, MI), volume 1, pp. 808–811.

Shields, P. W., and Campbell, D. R. (**1997**), "Multi-microphone sub-band adaptive signal processing for improvement of hearing aid performance: Preliminary results using normal-hearing volunteers," in "Proc. ICASSP," (IEEE), volume 1, pp. 415–418.

Shinn-Cunningham, B. G., Santarelli, S., and Kopco, N. (**2000**), "Tori of confusion: Binaural localization cues for sources within reach of a listener," *J. Acoust. Soc. Am.*, **107**(3):1627–1636.

Slatky, H. (**1993**), *Algorithmen zur richtungsselektiven Verarbeitung von Schallsignalen - die Realisierung eines binauralen Cocktail-Party-Prozessor-Systems (Algorithms for direction-selective processing of sound signals – realization of a binaural cocktail-party-processor system)*, Dissertation, Ruhr-Universität Bochum.

Soede, W. (**1990**), *Improvement of Speech intelligibility in noise (Development and evaluation of a new directional hearing instrument based on array technology)*, Ph.D. thesis, Universität Delft, Niederlande.

Soede, W., Berkhout, A. J., and Bilsen, F. A. (**1993**), "Development of a directional hearing instrument based on array technology," *J. Acoust. Soc. Am.*, **94**(2):785–798.

Spitzer, M. W., and Semple, M. N. (**1991**), "Interaural phase coding in auditory midbrain: Influence of dynamic stimulus features," *Science*, **254**(5032):721–724.

Squyres, J. M., and Lumsdaine, A. (**2003**), "A Component Architecture for LAM/MPI," in "Recent Advances in Parallel Virtual Machine and Message Passing Interface: Proceedings, 10th European PVM/MPI Users' Group Meeting," (Springer-Verlag, Venice, Italy), number 2840 in Lecture Notes in Computer Science, pp. 379 – 387.

Stern, R. M., and Bachorski, R. J. (**1983**), "Dynamic cues in binaural perception," in R. Klinke, and R. Hartmann (editors), "Hearing - Physiological Bases and Psychophysics," (Springer, Heidelberg), pp. 209–215.

Stern, R. M., and Trahiotis, C. (**1997**), "Models of binaural perception," in Gilkey and Anderson (1997), chapter 24, pp. 499–531.

Stern, R. M., Zeiberg, A. S., and Trahiotis, C. (**1988**), "Lateralization of complex binaural stimuli: A weigthed-image model," *J. Acoust. Soc. Am.*, **84**(1):156–165.

Stern, R. M., Jr., Slocum, J. E., and Phillips, M. S. (**1983**), "Interaural time and amplitude discrimination in noise," *J. Acoust. Soc. Am.*, **73**(5):1714–1722.

Strube, H. W. (**1981**), "Separation of several speakers recorded by two microphones (cocktail party processing)," *Signal Processing*, **3**:355–364.

Strube, H. W., and Wilmers, H. (**1999**), "Noise reduction for speech signals by operations on the modulation freqeuncy spectrum," *J. Acoust. Soc. Am.*, **105**(2):1092.

Summerfield, Q., and Stubbs, R. (**1990**), "Strengths and weaknesses of procedures for separating simultaneous voices," *Acta Oto-Laryngologica*, **S469**:91–100.

Tonning, F. (**1973**), "Directional audiometry. VII. The influence of azimuth on the perception of speech in aided and unaided patients with binaural hearing loss," *Acta Oto-Laryngologica*, **75**:425–431.

Tribolet, J. M. (**1977**), "A new phase unwrapping algorithm," *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-25**(2):170 – 177.

Unoki, M., and Akagi, M. (**1999**), "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, **27**(3-4):261–279.

Van Compernolle, D., Ma, W., Xie, F., and Diest, M. V. (**1990**), "Speech recognition in noisy environments with the aid of microphone arrays," *Speech Communication*, **9**(5–6):433–442.

van der Kouwe, A. J. W., Wang, D. L., and Brown, G. J. (**2001**), "A comparison of auditory and blind separation techniques for speech segregation," *IEEE Trans. Speech and Audio Processing*, **9**(3):189–195, [Online] `http://www.dcs.shef.ac.uk/~guy/pdf/tsap2001.pdf`.

van Rossum, G., *et al.* (**1998-2005**), "Python Website," [Online] `http://www.python.org`.

Varga, A. P., and Moore, R. K. (**1990**), "Hidden Markov model decomposition of speech and noise," in "Proc. ICASSP 1990," (Albuquerque, New Mexico), volume 2, pp. 845–848.

Vermaak, J., Andrieu, C., Doucet, A., and Godsill, S. J. (**2002**), "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Transactions Speech and Audio Processing*, **10**(3):173–185, [Online] `http://www.ee.mu.oz.au/staff/doucet/vermaak_drieu_doucet_tvar_seq.ps.gz`.

von der Malsburg, C., and Schneider, W. (**1986**), "A neural cocktail-party processor." *Biol Cybern*, **54**(1):29–40.

Wagner, H. (**1991**), "A temporal window for lateralization of interaural time difference by barn owls," *Journal of comparative Physiology A*, **169**:281–289.

Wallach, H. (**1940**), "The role of head movements and vestibular and visual cues in sound localization," *J. Exp. Psychol.*, **27**:339–368.

Wang, A., Yao, K., Hudson, R. E., Korompis, D., Lorenzelli, F., Soli, S. F., and Gao, S. (**1996**), "A high performance microphone array system for hearing aid applications," in "Proc. ICASSP," (IEEE), volume 6, pp. 3197–3200.

Wang, D. L. (**2000**), "On connectedness: A solution based on oscillatory correlation," *Neural Computation*, **12**:131–139.

Wang, D. L., and Brown, G. J. (**1999**), "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, **10**(3):684–697, [Online] `citeseer.ist.psu.edu/wang99separation.html`.

Ward, D. B., Lehmann, E. A., and Williamson, R. C. (**2003**), "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, **11**(6):826–836.

Ward, J. H., Jr. (**1963**), "Hierarchical grouping to optimize an objectice function," *Journal of the American Statistical Association*, **58**(301):236–244.

Widrow, B. (**2000**), "A microphone array for hearing aids," in "Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC," IEEE (IEEE), pp. 7–11.

Wightman, F. L., and Kistler, D. J. (**1989**a), "Headphone simulation of free-field listening. I: Stimulus synthesis," *J. Acoust. Soc. Am.*, **85**(2):858–867.

Wightman, F. L., and Kistler, D. J. (**1989**b), "Headphone simulation of free-field listening. II: Psychophysical validation," *J. Acoust. Soc. Am.*, **85**(2):868–878.

Wightman, F. L., and Kistler, D. J. (**1999**), "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, **105**(5):2841–2853.

Wightman, F. R., and Kistler, D. J. (**1989**c), "Headphone simulation of free-field listening. I. Stimulus synthesis," *J. Acoust. Soc. Am.*, **85**(2):858–867.

Wittkop, T. (**2001**), *Two-channel noise reduction algorithms motivated by models of binaural interaction.*, Ph.D. thesis, Universität Oldenburg.

Wittkop, T., Albani, S., Hohmann, V., Peissig, J., Woods, W. S., and Kollmeier, B. (**1997**), "Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction," *Acustica united with Acta Acustica*, **83**(4):684–699.

Wittkop, T., and Hohmann, V. (**2003**), "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Communication*, **39**:111–138.

Xie, F., and van Compernolle, D. (**1996**), "Speech enhancement by spectral magnitude estimation — a unifying approach," *Speech Communication*, **19**(2):89–104.

Yost, W. A. (**1991**), "Auditory image perception and analysis: The basis for hearing," *Hearing Research*, **56**:8–18.

Yost, W. A. (**1997**), "The cocktail party problem: Forty years later," in Gilkey and Anderson (1997), pp. 329–347.

Zahorian, S. A., and Rothenberg, M. (**1981**), "Principal-components analysis for low redundancy encoding of speech spectra," *J. Acoust. Soc. Am.*, **69**(3):832–845.

Zheng, Y., and Hasegawa-Johnson, M. (**2004**), "Formant tracking by mixture state particle filter," in "ICASSP 2004," [Online] `http://www.ifp.uiuc.edu/speech/pubs/conferences.html`.

Zurek, P. M. (**1991**), "Probability distributions of interaural phase and level differences in binaural detection stimuli," *J. Acoust. Soc. Am.*, **90**(4):1927–1932.

Zurek, P. M., Greenberg, J. E., and Rabinowitz, W. M. (**1996**), "Prospects and limitations of microphone-array hearing aids," in Kollmeier (1996), pp. 233–244.

Index of Citations

Acero *et al.* (2000), 4, 12, 116
Albani *et al.* (1996), 9, 17, 20, 60, 61
Algazi *et al.* (2001), 90, 96
Allen *et al.* (1977), 8
Allen and Rabiner (1977), 29
Allen (1977), 29, 62, 96
Anemüller and Kollmeier (2000), 113
Anemüller and Kollmeier (2003), 3, 81, 88
Anemüller (2001), 3, 123
Arulampalam *et al.* (2002), 15, 89, 91, 93, 94, 100
Barker *et al.* (2005), 87
Belin and Zatorre (2000), 11
Bell and Sejnowski (1995), 3, 88
Berzuini and Gilks (2001), 114
Beutelmann and Brand (2005), 1
Blake and Isard (1998), 15, 17, 89, 115
Blauert (1983), 20, 60, 61
Bodden (1992), 8
Bodden (1993), 8, 20
Bodden (1996a), 8, 21, 60, 82
Bodden (1996b), 60, 80
Boll (1979), 4, 8, 88
Boll (1992), 5, 17, 88, 114
Borisyuk *et al.* (2002), 20
Braasch and Hartung (2002), 54
Braasch (2002a), 54
Braasch (2002b), 54
Brainard *et al.* (1992), 9, 11, 51, 54, 61
Brandstein and Ward (2001), 82
Breebaart *et al.* (1999), 20
Bregman (1990), 7, 8, 86
Bregman (1993), 7
Bronkhorst and Plomp (1989), 1, 8
Bronkhorst (2000), 1, 8, 87, 88
Brown and Cooke (1994), 86
Brugge (1992), 20
Burns *et al.* (1994), 101
Bushara *et al.* (2003), 11, 110
Butler (1986), 52
Caird and Klinke (1987), 20
Cappé (1994), 4
Carlyon and McAdams (1992), 7
Cherry and Wiley (1967), 1, 8, 86, 87
Cherry (1953), 1, 11, 85, 86, 125
Cherry (1959), 1, 86
Cheveigné (1993), 124
Choo and Fleet (2001), 117
Chung *et al.* (2000), 20, 56
Clarey *et al.* (1992), 20
Cohen and Berdugo (2001), 4
Colburn *et al.* (1987), 8
Colburn and Hawley (1996), 8
Colburn (1996), 20
Compernolle (2001), 2, 5, 82
Cooke *et al.* (1993), 87
Cooke *et al.* (2001a), 112
Cooke *et al.* (2001b), 87, 112
Cooke and Ellis (2001), 7, 8, 10, 86, 87, 110, 112
Cooke (1993), 12, 87
Culling and Summerfield (1995), 1
Daly *et al.* (2004), 116

Damaske and Wagner (1969), 61
Datum *et al.* (1996), 20, 56, 60, 81
Domnitz and Colburn (1976), 21
Doucet *et al.* (2001), 15, 91, 94, 115
Dreschler *et al.* (2001), 139
Duda (1997), 20, 21, 24, 54, 60, 61, 80
Durlach and Colburn (1978), 8, 20
Durlach and Pang (1986), 8
Edwards (2004), 3–5, 119
Ellis (1996), 12, 87, 110
Ephraim *et al.* (1989a), 5, 88, 114, 116, 123, 124
Ephraim *et al.* (1989b), 5
Ephraim and Malah (1984), 4
Ephraim and Malah (1985), 4
Ephraim (1992), 5
Evans (1992), 86
Fisher (1993), 32, 55, 138
Fitzpatrick *et al.* (1997), 55
Fong *et al.* (2002), 17, 112, 116, 122
Gales and Young (1993), 5, 87, 88
Gallmeister (1995), 132
Gandhi and Hasegawa-Johnson (2004), 89, 112, 123
Garofolo (1998), 96
Gelb (1994), 13
Godsill and Rayner (1998), 5, 112
Godsmark and Brown (1999), 86, 87
Good and Gilkey (1996), 52
Gordon *et al.* (1993), 15, 89
Greenberg *et al.* (2003), 3, 60, 82, 119
Greenberg and Zurek (2001), 2, 3, 60, 82, 119
Griffiths and Jim (1982), 2, 88
Häusler *et al.* (1983), 8
Hancock and Delgutte (2004), 55
Harding and Meyer (2001), 12
Harper and McAlpine (2004), 55, 120
Hawley *et al.* (1998), 1
Hawley *et al.* (1999), 119
Hawley *et al.* (2004), 1, 3
Henning (1973), 21
Hoffman and Buckley (1995), 3, 82
Hofman *et al.* (1998), 82
Hohmann *et al.* (1999), 19
Hohmann *et al.* (2002b), 8
Isabelle *et al.* (1998), 60, 81
Isard (1998), 15, 17
Ito *et al.* (1982), 51
Janko *et al.* (1997), 20, 56
Jeffress (1948), 20, 60, 81
Jenison (2000), 53
Joris and Yin (1996), 20
Jourjine *et al.* (2000), 5
Kailath (1981), 4, 13
Kalman (1960), 13
Kidd *et al.* (1998), 8
Kitagawa (1996), 15, 89, 94
Knudsen and Konishi (1978), 9, 11, 51
Knudsen (1982), 20
Kollmeier *et al.* (1993), 20
Kollmeier and Koch (1994), 5, 8, 9, 20, 88, 124
Kompis and Dillier (1994), 2, 82
Kopp (1978), 33, 67, 97
Korompis *et al.* (1995), 82
Kuwada and Yin (1983), 20
Kuwada and Yin (1987), 20
Kysela and many others (2000-2005), 131
Langtangen (2004), 101, 132
Larocque *et al.* (2002), 114
Lee and Jung (2000), 5
Levitt (2001), 1, 3–5
Lewicki (2002), 12, 124
Li *et al.* (1969), 88, 113, 123, 124
Lim *et al.* (1978), 5
Lim (1978), 4, 88
Lindemann (1986a), 60, 80
Lindemann (1986b), 60
Liu *et al.* (1997), 5, 8
Liu *et al.* (2000), 20, 55, 60, 61, 81
Liu *et al.* (2001), 3, 60, 82, 119
Lutz and Ascher (2003), 132
Lutz (2001), 132, 134

Lyon (1983), 54, 60, 61, 80, 81
MacCormick and Blake (2000), 116
MacCormick and Isard (2000), 116
MacKay (1999), 15
Macpherson and Middlebrooks (2002), 53
Marin and McAdams (1991), 7
Marr (1980), 110
Marzinzik (2000), 2, 4, 88, 119
McAdams and Bigand (1993), 7, 86
McAdams (1993), 7
McAlpine and Grothe (2003), 20, 120
Mehrgardt and Mellert (1977), 20, 60, 90
Mellinger and Mont-Reynaud (1996), 7
Meyer *et al.* (2005), 125
Middlebrooks *et al.* (1989), 60, 90
Moore (1989a), 7
Moore (1989b), 29, 97
Morton (2001), 131
Nakashima *et al.* (2003), 54, 55, 80
Neti *et al.* (1992), 20, 56, 60
Nichols *et al.* (1996), 132
Nix *et al.* (2003a), 85
Nix *et al.* (2003b), 85
Nix and Hohmann (1999), 19
Nix and Hohmann (2000), 19
Nix and Hohmann (2001), 59
Nix (2005), 79
Nokia Corporation (2005), 134
Olshausen and Field (2004), 12, 14, 124
Otten (2001), 27, 66
Papoulis (1965), 26
Parra and Spence (2000), 3, 81, 88, 114
Peña and Konishi (2001), 11, 51
Peissig (1992), 1, 8, 9, 17, 60, 82
Perret and Noble (1997), 82
Rayleigh (1907), 60
Reyes-Gomez *et al.* (2003), 4, 12, 88
Reyes-Gomez *et al.* (2004), 12
Roman *et al.* (2003), 3, 60, 80, 82
Roweis (2000), 5
Sachs (1992), 31, 137
Sameti *et al.* (1998), 5, 88
Sameti and Deng (2002), 5, 12
Schauer *et al.* (2000), 54, 80, 81
Schiel *et al.* (1997), 96
Schiel (2003), 96
Schneider (1986), 12
Schwander and Levitt (1987), 2
Scott and Johnsrude (2003), 11, 110
Searle *et al.* (1975), 51, 61
Shackleton *et al.* (1992), 54
Shaw (1997), 20, 52
Sheikhzadeh *et al.* (1995), 17
Shields and Campbell (1997), 82
Shinn-Cunningham *et al.* (2000), 61
Slatky (1993), 26, 60
Soede *et al.* (1993), 2, 60, 82
Soede (1990), 2
Spitzer and Semple (1991), 20
Squyres and Lumsdaine (2003), 100, 132
Stern *et al.* (1983), 51
Stern *et al.* (1988), 54
Stern and Bachorski (1983), 51
Stern and Trahiotis (1997), 54
Strube and Wilmers (1999), 5, 88, 124
Strube (1981), 12
Summerfield and Stubbs (1990), 5, 7, 112
Tonning (1973), 8
Tribolet (1977), 55
Unoki and Akagi (1999), 12, 87
Van Compernolle *et al.* (1990), 2
Varga and Moore (1990), 87
Vermaak *et al.* (2002), 17, 89, 115, 116, 122, 124
Wagner (1991), 9, 22
Wallach (1940), 82
Wang *et al.* (1996), 82
Wang and Brown (1999), 87, 110
Wang (2000), 87
Ward *et al.* (2003), 15, 81, 89, 115
Ward (1963), 33, 67, 97
Widrow (2000), 3, 82
Wightman and Kistler (1989a), 41

Wightman and Kistler (1989b), 20, 49
Wightman and Kistler (1989c), 90
Wightman and Kistler (1999), 82
Wittkop *et al.* (1997), 20, 21, 31, 54, 82
Wittkop and Hohmann (2003), 8, 9, 60
Wittkop (2001), 3, 8, 9, 82
Xie and van Compernolle (1996), 4, 88
Yost (1991), 1, 7, 86
Yost (1997), 1, 8, 87
Zahorian and Rothenberg (1981), 124
Zheng and Hasegawa-Johnson (2004), 15, 89, 112
Zurek *et al.* (1996), 2
Zurek (1991), 21
van Rossum *et al.* (1998-2005), 132
van der Kouwe *et al.* (2001), 3, 87, 88, 110
von der Malsburg and Schneider (1986), 12, 87

Index

a posteriori probabilities, 32, 65, 75
a priori knowledge, 94, 121
  of interaural parameters, 21, 32, 65, 67
  of spectro-temporal features, 96, 98, 113
a priori probabilities, 94, 98
abstraction of features, 110
accuracy of direction estimation, 49, 67
acquisition of *a priori* data, 26, 67, 98
across-frequency processing, 139
  discrimination by, 51
  in neurophysiology, 54
  of common onsets, 86, 92
  of correlation values, 54
  of envelope correlations, 113
  of interaural parameters, 56, 81
  of probabilities, 22, 52, 57, 62, 70, 83
activity
  neural, interpretation of PDF, 55
  of ITD-tuned neurons, 56
adaptation time
  of BSS methods, 3
  of localization algorithm, 75, 79, 81, 83
  of SMC algorithm, 103, 105, 106, 114, 117
adaptive beamforming, 2–4, 60, 68, 82, 121, 123
additional observables, 125
advantages of SMC methods, 15, 17, 113
Alpha 21264 CPU, 130
ALSA (Advanced Linux Sound Architecture), 131, 132
ambiguity, 83
  caused by noise, 4, 61
  of direction, 52, 60, 70, 81
  of IPD, 53, 61, 81, 145
  of ITD, 54, 81
  of observations, 12, 87, 112
  of state, *see* state uncertainty
  resolution, 11, 52, 54, 61, 81
AMD Opteron CPU, 83
amplitude modulation, 27, 86, 98
analysis window, 24, 29, 62, 96, 98
AND operation, 11, 54
anechoic conditions
  assumption, 90
  measurement, 66
  results, 50
anechoic room, 26, 66
angle to the median plane, 45, 48
angular velocity, 90
applications
  of localization algorithm, 55, 59, 79, 123
  of SMC methods, 15, 115
approaches for sound localization, 54, 60, 80, 81, 89
AR coefficients, *see* autoregressive coefficients
array processing, *see* microphone arrays
artifacts, 79, 108, 115
ASR (Automatic Speech Recognition), 5,

82, 88, 112, 126
assembler code, 128
assumption
of constant time delay, 81
of free-field condition, 90, 115
of independence, 3, 53, 65, 97
of linear combination, 3
of linear superposition, 68, 81
of number of voices, 90, 114
of signal type, 90, 114
of similar level, 90, 115
of stationarity, 4, 41, 88, 116, 117
assumptions
for envelope tracking, 89–90, 114
for localization in noise, 22, 64
of algorithms, 2, 4, 88
of SMC methods, 89, 100
relaxation, 114
asymmetry of the HRTF, *see* pinna disparity cues
audio drivers, 131
audio hardware, 131
auditory
approach, 6, 22, 55, 96, 113, 121
filters, 96
grouping, 86, 112, *see also* binding problem
induction effects, 86
maps, 20
models, 125
processing, 11, 89, 123
scene analysis, 7–8, 119
simulation, *see* CASA
system, 86
accomplishments, 1, 3, 87
Automatic Speech Recognition, *see* ASR
autoregressive coefficients, 116, *see* LPC
averaging across frequency bands, 29, 62, 97
averaging of IPD, *see* cyclical variable
azimuth, 97, 99
definition, 24
dynamics, 90, 97
expected value, 100, 105
influence on PDF, 41
movement, 90, 102, 103
sampled directions, 66
tracking, 103

bandwidth, *see also* frequency bands
and localization accuracy, 52
of frequency analysis, 22, 29, 64, 97
of ICCF, 56, 61, 141
barn owl (*tyto alba*), 9, 20, 61
Bayes's formula, 32, 65, 91
Bayesian
analysis of information content, 32, 51–54
classification, 82
and neural networks, 56
and width of tuning curves, 55
criterion for direction discrimination, 31
estimation, 21, 32, 55
estimation, multidimensional, 10, 32, 64–65, 91
framework for CASA, 89, 91, 113, 117
localization algorithm, 82, *see* localization algorithm
a priori data, 65
applications, 121
integration with SMC methods, 125
summary, 64, 120–121
beamforming, 113, 121
adaptive, 68, 82
algorithms for hearing aids, 60, 82
approaches, 2–4
based on direction estimation, 3
controlled by localization algorithm, 68
drawbacks, 11
for speech enhancement, 88
binary masking, 5
binaural

cues for sound localization, 22, 120
features, 8–10, 90
recordings, 9, 26–28, 62, 65
signal, 94, 96, 97
binding problem, 11, 86, 119, *see also* auditory grouping
blackboard system, 87
block diagram
of bootstrap algorithm, 95
of localization algorithm, 62
of real-time environment, 129
bootstrap algorithm, 124
convergence properties, 105, 106, 115, 125
general properties, *see* SMC methods
implementation, 100
performance, 101–110, 121–122
steps, 93, 94
structure, 95
Bregman's Bs, 7, 9
BSS (Blind Source Separation), 2, 3, 11, 81, 88, 113

C/C++ programming language, 100, 128
cafeteria noise, 27, 34, 52, 66
cancellation of desired signal, 3
car noise, 27, 34, 42, 43, 66
CASA (Computational Auditory Scene Analysis), 7–8, 86, 119, 125
competing hypotheses, 110
earlier approaches, 86
for noise reduction, 117
integration of principles, 110
key challenges, 88
summary of results, 121
used cues, 7, 8, 10, 86
channel, *see* frequency subbands
Chapman-Kolmogorov equation, 91
CIPIC HRTF database, 96, 98
circular
kurtosis, 32, 42, 138
mean, 32, 138, 145
skew, 32, 42, 138
standard deviation, 32, 37, 138
circular kurtosis, 41
circular skew, 31, 41
clock frequency, 130
cluster analysis, 62, 67, 97
cluster, workstations, 129, 133
Cocktail Party Effect, 1, 8, 87
and hearing impairment, 1
Cocktail Party Processor, 1, 20
Cocktail Party situation, *see* multi-talker situation and nonstationary noise environment
codebook, 94, 98
entry, 97, 139
generation, 97, 99
lookup error, 100
retrieval from, 99
size, 97, 101, 116
codewords, *see* codebook entries
coefficients, *see also* components
autoregressive, *see* LPC
coincidence detection, 20, 54
combination of features, *see* auditory grouping
combination of probabilities
and physiological models, 11
in localization algorithm, 22, 26, 54, 62, 70, 82
of source features, 97, 113
common
amplitude modulation, 86, 92, 98
fate, 86
frequency modulation, 86
onset, 7, 8, 10, 86, 92, 98
common onset, 124
communication in groups, 1, 87
comparison of state and observation, 96, 100
competing hypotheses, 11, 110
complex variables
statistics of, 30, 55

compound random variable, 25, 65, 98, 113
compression of histograms, 30, 33, 62, 67
computation time, 100, 102
    of algorithm, *see* computational complexity
    of hardware, *see* speed of hardware
computational complexity, 6
    of Bayesian algorithm, 67
    of bootstrap algorithm, 124
    of SMC methods, 17, 99, 100
        causes, 115–117
        comparison, 122–123
        reduction, 117
concourse, *see* train station concourse
concurrent talkers, localization performance, 78
concurrent voices
    auditory discrimination, 87
    localization, 61, 69, 81
    separation of envelopes, 62, 69, 90, 106, 121
Condensation Algorithm, *see* SMC methods
conditional probability, 32, 65, 75, 91, 93
cones of confusion, 45, 51–53, 61, 64
conflicting explanations, 87, 112
confusion
    of directions, 52
    of elevations, 45
    of front-back directions, 45, 52, 61, 67, 70, 75
continuity of source directions, 81
continuity principle, 8, 81, 86, 112
contralateral, 37
convergence
    of particle filter, 106, 115
    time, 105, 117
        improvement, 125
        of BSS approaches, 81
        of localization, 81
        of SMC algorithm, 114
convolution, 96
correlations between frequency bands, 123, 139
    use, 53, 89, 113, 121
correlations, tracked envelopes, 109
Cosmea M hearing aid, 26, 66
coupled oscillators, 87
CPU (Central Processing Unit)
    of DSP system, 31
    of real-time environment, 62
cross correlation, 88
    coefficient, 109
    interaural, 54
    theorem, 25, 55, 141
cross spectrum, 82
cross-spectrum, 30
crosstalk between spectral estimates, 108
cues
    for scene analysis, *see* CASA, used cues
    for sound localization, 24, 51, 52
cyclic variable, statistics of, 30, 55, 61, 138, 145
cyclical moment coefficients, 69, 138

d' analysis, 31, 42
DAT (Digital Audio Tape) recordings, 65, 66
data entries for algorithm, 94
decay of distributions, 37
decibel scale, 97
decision histogram, 33, 46, 70, 72
degeneration, 115
delays, *see* latencies
demixing of concurrent sources, 67–69, 79–80
design choices for bootstrap algorithm, 94
detectability, 31, 42, 50, 51
detection theory, *see* signal detection theory
deterministic approaches, 88

deviation of estimated azimuths, 46
deviation of estimated levels, 114
dimensionality of observation, 92
dimensionality of state space, 12, 14, 92, 115, 122, 124
direction
  coordinates, 24, 64, 65, 97
  index, 26
direction estimate, *see* MAP estimate
direction index, 64
directional
  cues, 9, 86
  filtering, 20, 82
  information, extraction, 82
  microphones, 2
  tracking, 103
discrete PDF, 98, 99
discretization, *see* temporal discretization
discrimination of directions, 42
disjoint orthogonality, 5
distance measure, 54, 80, 100
distinction of front and back directions, *see* front-back confusions
distinction of objects by multiple dimensions, 13
distortion component, 69
distributed noise, 26
distribution of interaural parameters, 33–42
distribution percentiles, 35
DOA (Direction Of Arrival), *see* localization
driver layer of real-time environment, 131
DSP (Digital Signal Processor), 31, 127, 128
duration of evaluated signals, 109
duration of recordings, 27, 66

ear-independent transfer function, 68
efficiency, 17, 62, 99, 100, 115
elevation, 97, 99
  definition, 24
  dynamics, 97
  errors, 70, 75
  expected value, 100
  influence on PDF, 41
  sampled directions, 66
enhancement of speech, *see* noise reduction
envelope delays, 53, 54
envelope estimation, 2, 4, 113
Equivalent Rectangular Bandwidth, *see* ERB
ERB (Equivalent Rectangular Bandwidth), 29, 97
  definition, 29
  subbands, 29
ergodicity, 26
error
  measures, 67
  of direction estimate, 49, 61
  of DSP libraries, 128
  of elevation estimate, 75
  of envelope tracking, 108
  of positioning, 27
  of spectral estimation, 103, 106
  patterns of direction estimate, 47
  signal, 3, 69, 108
estimated
  azimuth, 45, 75, 78, 89, 103, 105
  directions, 45, 46, 75
  envelopes, 106
  envelopes, evaluation, 109
  level, 103
estimation of directions, *see* localization algorithm
estimation of PDF, 26
Euclidian distance, 100
excess, *see* kurtosis
exchange of references, 74
excitation, 11, 55, 110
expectation-based processing, 11, 87
expected values, 31, 69
  computation, 93, 96, 100, 110, 137
exploratory programming, 129

extraction of features, 110

Fast Fourier Transform, *see* FFT
feature
  abstraction, 110
  extraction, 110
  integration, 110, 112, 117, 119
    in auditory system, 11
  vector, 25, 65, 88, 91
features for CASA, 86
FFT (Fast Fourier Transform) length, 29, 62, 96
filtered PDF, 94, 100
filtered spectrum, expected value, 100
filtering operation of SMC algorithm, 99
fixed rules, 8, 12, 88
fluctuations, 9, 21, 83
  of interaural parameters, 9, 34, 41, 50, 57, 64, 120
    causes, 22
    consequences, 50–51
fMRI experiments, 11, 110
formants, 89, 108, 112
forward transformation, 114, 117
forward-backward algorithm, 122
frame rate, 29, 62, 96
free-field apparatus, *see* TASP
free-field spectrum, 64, 67
frequency
  bands, 22, 29, 54, 67, 97, 141
  integration, 51, 56, 62, 92, 139
  integration of probabilities, 22, 54, 57, 62, 70, 71, 83, 120
  range, 27, 53
  resolution, 29, 62, 97, 123
frequency-domain representation, 20, 54, 61, 80, 108, 120
front-back confusions, 33, 52, 61, 70, 75
  definition, 67
  percentage, 48, 49
full duplex processing, 129, 133
fundamental frequency, 5, 13, 89, 124, 125
fuzzy computation, 11, 56, 62

Gaussian PDFs (Probability Density Functions), 31, 99, 100
  moments, 137
generality of reference data, 70
generator functions, 132
Gestalt Laws, 7, 8, 86
global inhibition, 110
goose tracks, 15
Graphical User Interface, *see* GUI
grid-based Bayesian approach, 122
group delay, 53
grouping rules, 86, 88, 112
grouping, auditory, 8, 86, 112
GUI (Graphical User Interface), 133
gurgling, 4

Hann window, 29, 62, 96
Hanning window, *see* Hann window
hardware
  for real-time implementation, 31, 83
  of real-time environment, 130
harmonic structure, 108
harmonicity, 86
Head Related Transfer Function, *see* HRTF
head rotations, 81
headrest, 27
hearing aids, 9, 60, 82, 133
  requirements, 1, 88
hearing impairment
  binaural hearing, 8
  consequences, 1
hiding of desired signal, 5, 112
hierarchical cluster analysis, 97
hierarchy of cues, 7, 87, 112
high-dimensional integration, 91
high-dimensional state space, 92
higher-order moments of linear and cyclical variables, 31, 37–42, 51, 137–138
histogram
  data reduction, 62, 67

dependence on SNR, 37
measurement, 26, 30, 62
of direction estimates, 33, 46, 70, 72
of ILD, 36
of interaural parameters, 21, 26, 67
of IPD, 37
of spectral transitions, 98
HMM (Hidden Markov Models), 3, 5, 13, 87, 88, 91, 114, 122, 126
host computer, 128
how-stream, 11
HRTF (Head Related Transfer Functions), 10, 20, 88, 123
database, 94
filtering, 24, 64, 89, 90, 96, 99
patterns, 60
properties, 37, 41, 81
quotient, 24, 68
symmetry, 52, 60
human performance, 1, 3, 5, 53
hypotheses, 94
competition of, 11, 87, 110
number, *see* number of particles
hypothetical
azimuth, 99
elevation, 99
spectra, 96, 97, 99–100, 114, 117
state, computation, 99–100

ICA (Independent Component Analysis), 4, 88
ICCF (Interaural Cross Correlation Function), 60, 61, 141
ambiguity, 141
ICCF (Interaural Cross-Correlation Function), 25, 54, 56
iceberg strategy, 134
IID (Interaural Intensity Difference), *see* ILD
ILD (Interaural Level Difference), 20, 60
computation, 24, 30–31, 62, 64, 67
definition, 25
statistics, 64, 145
ill-posedness, *see* underdetermined equation
implementation
of localization algorithm, 62
of SMC algorithm, 100
improbable state transitions, 122
improvement of SNR, *see* SNR improvement
impulse responses, 96
In-The-Ear hearing aids, *see* ITE
Independent Component Analysis, *see* ICA
independent components, assumption of, 3, 26, 45, 65, 97
information retrieval, 50
inhibition, 11, 110
initialization of bootstrap algorithm, 93, 98, 103
integration
of auditory approach and source modeling, 113, 125
of Bayesian localization and envelope-tracking, 125
of different auditory cues, 10, 117
of features, 11, 22, 51–52, 92, 110, 112, 117
temporal, *see* temporal integration
intelligibility, 119
of demixed voices, 79
other noise reduction methods, 5
spectral estimation methods, 4
inter-process communication, 131
interaural
axis, 61
Cross Correlation Function, *see* ICCF
cross-power spectrum, 25
group delay, 24
Intensity Difference, *see* ILD
Level Difference, *see* ILD
parameters, 9
computation, 29–31, 64
dependence on direction, 41, 51

dependence on frequency, 34
dependence on noise condition, 42, 50, 61
dependence on SNR, 34, 50
statistics, 9, 33
phase delay, 24
Phase Difference, *see* IPD
Time Difference, *see* ITD
timing cues, representations, 24, 54–56
Transfer Function, *see* ITF
interfering talkers, 60, 74, 87, 89
tests with, 69
interpolation, 96
inverse filtering, 69
inversion
of mixing matrix, 68, 79, 83
of the observable, 114
IPC (Inter-Process Communication), 131
IPD (Interaural Phase Difference), 60, 61
comparison with time domain representation, 54, 81, 141
computation, 30–31, 62, 64, 67
definition, 25
expected value, 37, 41, 69, 138
standard deviation, 41, 138
statistics, 64, 138
ipsilateral, 37
ITD (Interaural Time Difference), 20, 60, 61
time series, 141
ITE (In-The-Ear) hearing aids, 26, 65
ITF (Interaural Transfer Function)
definition, 24
elevation dependence, 45
for demixing signals, 68, 79

Jeffress Model, 20, 60, 80, 120
JND (Just Noticeable Difference), 53

Kalman Filter, 5, 13, 92
kernel modifications, real-time processing, 131
kurtosis, 31
definition, 31
of cyclical variables, 138
of linear variables, 137
properties of ILD and IPD, 37, 41

LAM (Local Area Multicomputer), 100, 132
latencies for real-time processing, 133
latencies of algorithms, 6
layers of real-time environment, 129
learning stage, *see* training stage
left-right coordinate, 48, 49
LENS Algorithm, 60, 82
level differences
between voices, 78, 90
interaural , *see* ILD
library modules, 132
limitations of algorithms, 4, 6, 81, 115
linear distortion, 69
linear superposition, 64, 68, 81, 96, 99
Linux system, 131
Local Area Multicomputer, *see* LAM
local excitation - global inhibition, 11, 87, 89, 110
localization
algorithm, 3, 17, 119
a priori data, 65
applications, 82, 121
comparison, 54
decision histogram, 52
drawbacks, 81
evaluation, 33
generality, 70
integration with SMC methods, 125
motivation, 61
overview, 20
results, 42, 45–50
robustness, 57
summary, 57, 64, 120–121
at low SNRs, 21, 52, 64, 81
existing approaches, 80
in noise, 70

of concurrent talkers, 74–78, 83
performance, 66, 67, 69, 121
with neural networks, 56, 65, 81
logarithmic scale
interpretation as activity, 55
localization decisions, 46
polar plots, 37, 38
spectra, 97, 100
lookup error, 100
lookup in codebook, 98, 99
loosing track, 108
loudspeakers, 27
low-latency patches, 131
low-latency processing, 129
low-pass filter, 30, 62, 79
LPC (Linear Predictive Coding), 124
LSF (Line Spectrum Frequencies), 124

magnitude spectrum, 96, 103
mammals, 20
manifold, 12
MAP (Maximum *A Posteriori*) estimate, 32, 53, 65, 75, 78
map of auditory space, 20
marginal distributions, 26, 32, 53, 65, 112
market noise, 27
Markov model, *see also* Hidden Markov Model (HMM)
extensions, 114
formalism, 90
implementation, 113
Markov series of spectra, 97, 98, 139
masking, 8
masking of sounds, 86, 89, 112
matching conditions, 49, 61, 70
matrix inversion, 79
matrix inversion of mixing matrix, 68
Maximum A Posteriori, *see* MAP
mean, 99
circular, 32, 69, 138
value of ILD, 37, 145
value of IPD, 41, 42
measurement
errors, 53
noise, 91
of PDF, 20, 62, 98
details, 30
generality, 53
indoor, 55
precision, 53
signals, 66
summary, 57
theory, 26
median plane, 24, 45, 51, 120
median plane, confusions, *see* cones of confusion
memory locking, 131
Message Passing Interface Library, *see* MPI
metal workshop, 27, 66
methods for speech enhancement, 88
microphone arrays, 2, 82, 123
applications, 5, 82
non-adaptive, 2
microphone distance, 5, 61
microphone noise, 92
Mises, *see* von Mises
mismatch of test and training data, 70
missing features approach, 112
mixture of talkers, 66, 78, 90, 103
mobile telephony, 82, 134
models of auditory processing, 125
models of binaural processing, 20, 22, 55–57, 82
modular programming, 128
modulation frequencies, 5, 98, 124
moment coefficients
for ILD, 120, 137
for IPD, 120, 138
of probability density functions, 137
monaural cues, 53
moving average, 22, 30
moving sound sources, 80, 81, 90, 103, 114
MPI (Message Passing Interface), 100, 132
multi-channel methods, 113

multi-microphone methods for noise reduction, 82
multi-talker situation, 1, 75, 114
multicomputer, 129
multidimensional
  Bayesian estimation, 10
  statistical filtering, 6, 12–17, 85, 88, 119
multimodal neural processing, 86, 110
multimodal PDFs, 92, 100
multiplication
  of neural activities, 11, 56
  of probabilities, 26, 56, 62, 70, 112
multiprocessor hardware, *see* SMP workstations
multivariate Gaussian PDF, 100
musical recordings, restoration, 5, 112
musical tones, 4

narrow-band filtering, 60
natural sound signals, 12
network connection, 133
network of coupled oscillators, 110
neural
  binding, 11, 110, 117, 119
  maps, 11, 20, 120
  networks, 56, 65, 87
    for sound localization, 81
  processing, parallel, 110
  processing, probabilistic interpretation, 11, 125
  responses to ITD, 20
  tuning curves, 55–57, 120
noise
  bursts, 79
  condition, 42, 49, 52
  environments, 27, 28, 43, 49, 66, 70
  from train station concourse, 66
  recordings, 27, 66
  reduction, 60
    algorithms, overview, 1–5
    by spectral estimation, 4, 5
    by spectral subtraction, 88
    for communication technologies, 5, 82
    speech quality, 4
    with HMM, 88
non-deterministic properties, speech, 92
non-Gaussian
  PDF, 15, 26, 113
    of ILD , 37
    of IPD, 41
    of spectral transitions, 98
  state space model, 12, 85, 88
non-hierarchical model, 112, 117
non-matching conditions, 70
nonlinear
  estimation, 88
  non-Gaussian state space methods, 12, 88, 92
  tracking, 15, 85
  transformation, 56, 92
nonstationary
  noises
    and CASA, 86
    and speech enhancement, 88
    assumption, 89
    enhancement of music recordings, 112
    localization, 60
    objectives, 2
    recordings, 28
    speech enhancement, 114
  signals
    and interaural parameters, 21, 22, 64
    and ITD, 54, 80
    and speech enhancement, 88
    generation, 27
    information content, 57
    separation, 62
normal distribution, *see* Gaussian PDF
normalization step, 93, 110
normalized weights of particles, 112
number

of particles, 101
of voices, 101
number of
codebook entries, 97, 101, 116
coefficients, 97, 115
directions estimates, 33, 67
discrete directions, 27, 33, 64, 70
entries in transition matrix, 98
frequency bands, 97
interferers, 3, 74
particles, 100, 108, 115, 122
per coordinate, 116
particles in the universe, 122
possible state sequences, 122
voices, 90, 97, 100
numerical libraries, 128
Numpy vector library, 132

object-oriented programming, 132
objectives of thesis, 1–2, 11–12, 88, 119
observable, 94
auditory approach, 113, 124
computation, 96
directional, 125
perceptual evaluation, 113
statistics, 91, 112
PDF, 100
observations, uncertainty, 12, 13, 70, 91
old-plus-new principle, 8, 10, 86
on-line algorithm, 62, 83, 85
advantage, 114
requirements, 13
structure, 94, 99
why investigated, 6
onset cues, *see* common onset
operation stage, 62
Opteron CPU, 83
original signals, 103
oscillatory correlation, 87, 110
oscillatory movement, 12, 102
outliers, 30
overview on chapters, 17

parallel
processing, 110, 129
processing, neural system, 110
programming, 100
parameters
of bootstrap algorithm, 101–102
of short-term frequency analysis, 29, 62, 96
particle filter, 89, 92
algorithm, *see* SMC methods
name, 94
particles, 94
definition, 94
weights, 94, 100, 112
partitioned sampling, 116
pattern-matching, 5, 78, 88, 114
patterns, 60
PCA (Principal Component Analysis), 124
PDA (Personal Digital Assistant), 134
PDF (Probability Density Function)
multimodality, 92, 100
of ILDs, 34
of interaural parameters, 21, 25
of IPDs, 37
of observation statistics, 100
of source features, 89
of spectral transitions, 98, 113, 125, 139
of system dynamics, 91, 97, 112
of system state, 91, 96
percentage of front-back confusions, 45, 48, 49, 74
perceptual representation, 89
perceptual restoration, 5, 112
perceptually relevant features, 113
performance
of envelope tracking, 106
of humans, 51, 52
of localization algorithm, 66, 67, 69, 121
peripheral auditory processing, 96, 123, 124

Personal Digital Assistant, *see* PDA
phase
 effects, 99
 unwrapping, 25, 32, 55
phase delay, *see* interaural phase delay
PHONDAT database, 96, 98
physiological models, 60
pinna asymmetries, 51, 52, 120
pinna disparity cues, 45, 52
place theory, 20, 80, 120
platforms for real-time environment, 134
population code, 55
portability, 134
POSIX IV threads, 131
prediction step, 93, 96, 99
prediction-driven CASA, 87
preprocessing, 62, 97
probabilistic
 approach, 6, 11, 59, 80, 81, 120
 interpretation, 11, 85, 125
 metric, 80
Probability Density Function, *see* PDF
processing streams, 11
programming environments, real time, 127
programming languages
 C, 100, 128
 C++, 100, 133
 for DSP systems, 128
 Python, 62, 83, 100, 132, 134
properties of speech, 113
protection of run-time environment, 128
Psychoacoustics of sound localization, 20, 51–54
pulsation threshold, 8
Python
 for real-time processing, 62
 script language, 62, 83, 100, 121, 128, 132
 speed, 100

quality
 of codebook, 101
 of reconstruction, 115
 of speech, 101, 102, 116
quantization
 effects, 108
 noise, 92, 115
 of short-term spectra, 97–99
quantized PDF, 98, 99
quasi-continuity of state space, 92, 114
quasi-stationarity of speech, 98, 112
questions of research, 11
quotient transfer function, 25

railway station, *see* train station concourse
random
 process, 26
 variable, 64
  ILD and IPD, 25
  observation, 91, 100
  system state, 91
re-use of source code, 132
real-time
 processing, 31, 114, 121, 129
 processing, development of, 127
 scheduling, 131, 133
real-world signals, 86, 123
 parameter fluctuations, 21, 55, 64, 80
 recordings, 26, 27, 65, 66
 test of localization, 22, 61, 65
receptive fields, 56
reconstruction of desired signal, 5, 88, 120
recorded positions, 27
recordings
 of noise environments, 27, 66
 of reference signals, 26–29
 of talkers, 66, 79
 of test and reference signals, 65–66
redundancy of speech signals, 1, 88, 112, 124
reference statistics, 62, 69, 70, 80
relaxation of assumptions, 114–115
reliability of features, changing, 87

reliability of localization, *see* robustness
representation of
 interaural timing, 20
 observable, 124
 observation statistics, 100
 source states, 113
 state, 115
representational streams, 11, 110
requirements for multidimensional statistical methods, 13, 92
resampling step, 93, 96, 100, 110, 112
research questions, 119
restoration of signals, 5, 88, 112, 120
restriction of coefficients, 79
result of particle filtering, 100
retrieval of spectra, 99
reverberation, 3, 90, 115
reweighting step, 93, 96, 100, 112
RMS (Root Mean Square)
 of azimuth error, 49
 of level, 29
 of level error, 103
robustness
 of feature integration, 11–17
 of sound localization, 49, 52, 56, 57, 78, 81, 123
 to environmental noise, 123
Root Mean Square, *see* RMS
run-time environment, 128

sampled directions, 27, 64, 66
sampling frequency, 27, 28, 62
sampling of PDF, 15, 94, 100, 115
scene analysis in the auditory system, *see* auditory scene analysis
scope of work, 6
script language, 62, 121, 127, 128
script language, real-time environment, 132
segmentation, 110
separation of envelopes, 106
separation of voices, 1, 83, 121
Sequential Monte Carlo Method, *see* SMC Methods
shape of PDFs, 41
shared memory, 131
shift of mean values, 37, 145
short-term
 frequency analysis, 29, 62, 96, 133, *see also* FFT
  consequences, 22
 spectra, 97, 98, 139
 time series, 141
signal detection theory, 21
Signal-To-Noise Ratio, *see* SNR
SIMD (Single Instruction, Multiple Data), 132
simultaneous talkers, *see* concurrent talkers
single instruction, multiple data, 132
skew, 31, 37, 56
 of cyclical variables, 138
 of linear variables, 137, 145
SMC (Sequential Monte Carlo)
 Algorithm, *see also* bootstrap algorithm
SMC (Sequential Monte Carlo) Methods, 15–17, 89, 91, 121
 advantages, 113, 123
 applications, 15, 89
 combination of observations, 125
 computation time, 17, 102, 115, 122
 convergence, 105, 106, 114, 115, 125
 evaluation, 101–103, 121–122
 further literature, 94
 goal, 91, 119
 implementation, 100
 optimizing, 115, 117, 124
 output, 100
 principle, 15, 94
 properties, 92, 112
SMP (Symmetric Multiprocessing) workstation, 127, 129, 133
SNR (Signal-to-Noise Ratio), 87

definition, 29
dependence, 34–41, 49
definition, 52
dependence, 49, 52
impact, 52, 53, 64
improvement, 79
by envelope tracking, 109
for steered beamforming algorithm, 69, 82, 121
by envelope tracking, 108
of reference signals, 49
of test signals, 49
software environment for algorithm development, 100
sound
field, 25
localization algorithm, *see* localization algorithm
localization in median plane, 120
source direction, advantage for CASA, 86
source coding, 113, 116, 125
source movements, 81, 90, 102, 103, 125
sparse representations, 124
spatial
distributed noise, 28
filtering, 2–4, 60, 113, 119
superposition, *see* linear superposition
specification of task, 6
spectral
estimation, 4, 106
power densities, 97
subtraction, 4, 88
transitions, 86, 98
spectro-temporal
cues, 86
dynamics, 11, 89, 90, 97–98, 110, 117, 121
features, 7, 12, 17, 86, 90, 113, 119
of natural sounds, 86
of speech, 139
spectrogram, 107
speech
database, 96–98
enhancement, 60, 113, *see* noise reduction
pause detection, *see* VAD
quality
of algorithms, 4, 79, 115
of codebook, 101
Reception Threshold, *see* SRT
recognition, *see* ASR
signals, 26–27, 65–66, 79, 96
speed (complexity) of processing, *see* computational complexity
speed of hardware, 62, 114, 128, 130
SRT (Speech reception Threshold), 2, 82
stages in neural processing, 110
stages of localization algorithm, 62, 65, 74
standard deviation, 99
computation, 31, 137
of azimuth dynamics, 99, 102, 103, 105, 106
of azimuth estimate, 67
of estimates, *see* RMS
of ILD, 34, 37, 41, 145
of level estimate, *see* RMS
start values, 98
state
coding, 124
dynamics, 124
of sound source, 113
transitions, improbable, 122
uncertainty, 91, 103
vector, 96
definition, 97
state prediction, *see* prediction step
state-space
approach, 12, 88, 90, 112
definition, 90
dimensionality, 92, 116
methods, requirements, 92

representations, advantages, 12, 113
station concourse, *see* train station concourse
stationarity of noise, assumption, 4
statistical independence, 26, 45
statistics of spectro-temporal dynamics, 89, 90
steered beamforming, 3, 82, 121, 123
stencil filter method, 81
stereo signals, *see* binaural signals
stochastic process, 25
strategies for noise reduction, 6–17
stream analysis, *see* auditory scene analysis
streams, representational, 110
street market, 27
structure
of bootstrap algorithm, 95
of processing for direction estimation, 22
subbands, *see* frequency bands
subspace of probable states, 12
substrates for feature integration, 10
suggestions for future research, 123–125
summary of goals, 119
Symmetric Multiprocessing, *see* SMP
symmetry of HRTFs, 60, 61, 120
system dynamics, 97, 100, 113
system libraries, 131

target signal, 27
TASP (Two Arc Source Positioning) apparatus, 27, 66
telephony codes, 116
temporal discretization, 108
temporal integration, 22, 33, 51, 62
temporal variability, *see* non-stationarity
test signals, 66, 96
threshold criterion, 30
time constants, 22, 30, 33, 79
time delay, interaural, *see* ITD
time series, 26
timing cues, representations, 22
TIMIT database, 96, 98
tip-of-the-iceberg strategy, 134
TMS320C40, 31
top-down processing, 11, 87
tori of confusion, 61
tracking, 15
error, 108
in high-dimensional spaces, 12, 122
of envelopes, 106, 121
of parameters, 13
of short-term level, 103
of sound source direction, 103
of two voices, 105–106
principle, 12
speed, 99
uncertainty, 105, 106
traffic noise, 27, 66
train station concourse, 27
recording, 66
trained references, 49, 62, 67
training stage, 62, 65
trajectory, 12, 13
transfer function, *see* HRTFs
transition matrix, 94, 98, 113, 139
transition probabilities, 125
measurement, 98
properties, 98
tree-based vector quantization, 124
trials for bootstrap algorithm, 103
Tribolet's algorithm, 55
trigonometric moment coefficients, 138
Tukey's rule, 31
tuning curves, 55–57
sharpening, 56
TVAR (time variant autoregressive) model, 116
Two Arc Source Positioning System, *see* TASP
two interfering talkers, 75
tyto alba, 11
tyto alba, 9, 61

uncertainty, *see also* ambiguity

of observation, 91
of state, 91, 105
underdetermined equation, 3, 88, 114
uniform circular distribution, 37, 138, 145
untrained references, 49, 53
unwrapping of phase variable, 25, 32, 55
update equation, 94, 96, 97, 99

VAD (Voice Activity Detection), 4, 113
varying reliability, 87
vector library, 100, 132
Vector Quantization, 97, 99, 108, 116, 121, 124
vector strength, 32, 41, 138
vocal apparatus, 124
von Mises Distribution, 55
VQ, *see* vector quantization

Ward's method, 62, 67, 97
weight of particles, 94, 96, 100, 112
where-stream, 11
width of distributions, 34
Wiener Filter, 4, 60
Wiener-Khintchine theorem, 25, 55, 141
Wiener-Lee relation, 55
window shift, 29, 62, 96
workshop, 27
workstation cluster, 129
worst-case latencies, 133

Acknowledgments

I want to say thanks. First to Prof. Dr. Dr. Birger Kollmeier, for his constant support and important encouragement, and the excellent working conditions in the Medical Physics Group which are to a great part due to his effort. My thanks also to Dr. Volker Hohmann, who was decisive in initiating this project and always took time for discussion of new ideas, approaches and results, for his support to develop new possible solutions to unsolved problems.

Thanks to Michael Kleinschmidt for important discussions, his helpful sharing of knowledge and ideas, and also for the enjoyment of working with him.

I thank Steve Greenberg and Bastiaan Kleijn for their constructive critics to my writing. Also, I thank Jörn Anemüller, Helmut Riedel, Jesko Verhey, and Jennifer Brown for proofreading parts of the manuscript, and many helpful suggestions. Giso Grimm, Oliver Fobel, and Michael Granzow helped in dealing with the computers and their often mysterious interactions. Also, I am indebted to Jürgen Weiss and Benno Hubert, who did a magnificent job in keeping the Linux cluster going. Thanks also to the technical staff of the Medical Physics Group, especially Anita Goerges, and Frank Grunau, for their friendly and patient support. I also want to thank the other members of the group – it has been enjoyable to work with you.

This work has benefitted greatly from many persons which provided excellent software and precious personal help for free. My special thanks go to Travis Oliphant, Andrew Morton, and Paul Barton-Davies. Also, thanks to Peter Kootsookos for his important reference.

I want to thank my parents which kindled my curiosity from childhood, and supported my wish to work in science. Also, thanks to my siblings, and to my close friends, Gudrun, Shu-Chiu, Holger, and Feng for their sharing, their listening, and their affectionate support in often challenging times.

Biographical Note

Johannes Nix was born at May 20, 1967 in Köln (Cologne), Germany. When he was five years old, it was recognized first that he had acquired a severe cochlear hearing loss. Thanks to the rapid prescription of hearing aids and the continuous support of his parents, he was able to visit a regular school.

In 1987, he enrolled in physics at the Rheinisch Westfälische Technische Hochschule in Aachen, where he studied Spanish, philosophy, and political science as well. After the intermediary exam and an one-year stay at the Phillips Universität Marburg, he continued his study in 1991 with courses in applied physics at the Carl-von-Ossietzky Universität Oldenburg. In 1993, he spent one year in Colombia, South America, where he attended a master course on mechanical engineering and measurement devices in the Group of Alvaro Pinilla, at the Universidad de Los Andes, Bogotá; During this time, he contributed in simulations and development of measurements devices for small-scale wind power plants. Back in Oldenburg, he worked in several employments in software development for commercial and research purposes. He joined the Medical Physics group directed by Prof. Birger Kollmeier in 1998, finishing his diploma thesis in 1999.

He then stayed in Oldenburg as a PhD student and joined the European Graduate School in 'Neurosensory Science, Systems and Applications' and later the 'Center of Excellence in Hearing Aid Technology"; with this, he began an in-depth research in algorithms and principles which have potential to improve future hearing aids. Additionally, he organized part of the systems and network administration of the group, taught lab courses in digital signal processing, and worked jointly with the Hörzentrum Oldenburg GmbH in procedures for testing and development of hearing aid algorithms.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig verfasst habe und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

Oldenburg, den 21. April 2005

(Johannes Nix)