Binaural Scene Analysis: Localization, Detection and Recognition of Speakers in Complex Acoustic Scenes

May, Tobias; Van de Par, S.L.J.D.E.; Kohlrausch, A.G.

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Binaural Scene Analysis

Localization, Detection and Recognition of Speakers in Complex Acoustic Scenes

The work described in this thesis was performed jointly at the Eindhoven University of Technology, The Netherlands, and the University of Oldenburg, Germany. The work was financially supported by Philips Research, Eindhoven, The Netherlands.

Printing: Universiteitsdrukkerij Technische Universiteit Eindhoven

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

May, Tobias

Binaural Scene Analysis: Localization, Detection and Recognition of Speakers in Complex Acoustic Scenes / by Tobias May. - Eindhoven: Technische Universiteit Eindhoven, 2012. - Proefschrift. -

A catalogue record is available from the Eindhoven University of Technology Library

ISBN 978-90-386-3230-8

Keywords: computational auditory scene analysis / binaural processing / speaker identification / missing data / binary mask estimation

Binaural Scene Analysis

Localization, Detection and Recognition of Speakers in Complex Acoustic Scenes

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op maandag 8 oktober 2012 om 16.00 uur

door

Tobias May

geboren te Oldenburg, Duitsland

Dit proefschrift is goedgekeurd door de promotoren:

Prof. Dr. ir. S.L.J.D.E. van de Par
en
Prof. Dr. A.G. Kohlrausch

Glossary

List of acronyms

ASA	Auditory Scene Analysis
ASR	Automatic Speaker Recognition
BRIR	Binaural Room Impulse Response
CASA	Computational Auditory Scene Analysis
CMN	Cepstral Mean Normalization
CMVN	Cepstral Mean and Variance Normalization
DCT	Discrete Cosine Transform
DRR	Direct-to-Reverberation Ratio
EM	Expectation-Maximization
ERB	Equivalent Rectangular Bandwidth
FFT	Fast Fourier Transform
FP	False Positive
GCC	Generalized Cross-Correlation
GSC	Generalized Sidelobe Canceller
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HRTF	Head Related Transfer Function

HTK Hidden Markov Model ToolKit

IBM Ideal Binary Mask
ILD Interaural Level Difference
IPD Interaural Phase Difference
ITD Interaural Time Difference

KEMAR Knowles Electronic Manikin for Acoustic Research
KLD Kullback-Leibler Divergence

LC Local Criterion

MC Monte-Carlo
MAP Maximum A Posteriori
MD Missing Data
MDL Minimum Description Length
MFCC Mel Frequency Cepstral Coefficient
MIT Massachusetts Institute of Technology
ML Maximum Likelihood
MMSE Minimum Mean Square Error
MVDR Minimum Variance Distortionless Response

NR Noise Reduction

PDF Probability Density Function
PHAT PHAse Transform

RASTA RelAtive SpecTrAl
RIR Room Impulse Response
ROC Receiver Operating Characteristic

SCOT Smoothed COherence Transform
SCOTM Smoothed COherence Transform Modified
SFD Speech Fragment Decoder
SID Speaker IDentification
SNR Signal-to-Noise Ratio

SRT	Speech Reception Threshold
SSC	Speech Separation Challenge database
SSR	Signal-to-Signal Ratio
STFT	Short-Time Fourier Transform
STSA	Short-Time Spectral Amplitude
TIMIT	Acoustic-phonetic continuous speech corpus, recorded by Texas Instruments (TI), and transcribed at the Massachusetts Institute of Technology (MIT)
T-F	Time-Frequency
TP	True Positive
UBM	Universal Background Model
VAD	Voice Activity Detector
WARA	Wearable Augmented Reality Audio

List of symbols and variables

$A_f(t,\tau)$	Normalized auto-correlation function of channel f, time frame t and lag τ
$\mathcal{A}$	True number of sound sources
$\hat{\mathcal{A}}$	Estimated number of sound sources
α	Over-estimation factor, smoothing constant
α_{Frag}	Smoothing constant for fragment-based integration
b	Sample index
B	Frame length in samples
β	Spectral floor
$C_f(t,\tau)$	Normalized cross-correlation function of channel f, time frame t and lag τ
d	Dimension index

D	Dimension of the feature space
$\hat{\delta}_f(t)$	Interpolated peak position relative to the integer peak position $\hat{\tau}_f(t)$ in channel f and time frame t
e_f, $\bar{e}_f$	Smoothed envelope of channel f and its short-term average
η	Smoothing constant
f	Gammatone channel index, auditory filter index
f_c	Set of center frequencies of F auditory filters
f_s	Sampling frequency in Hz
f_ω	Frequency corresponding to frequency bin index ω
F	Number of gammatone channels, number of auditory filters
$\mathcal{F}_{\mathrm{AD}}(t,f)$	Mean average deviation feature of time frame t and channel f
$\mathcal{F}_{\mathcal{R}}(t,f)$	Ratemap feature of time frame t and channel f
$\mathcal{F}_{\mathcal{R},n}(t,f)$	*Better ear* ratemap feature of the nth speech source
$\mathcal{F}^{\mathrm{L}}_{\mathcal{R}}(t,f)$	Ratemap feature of the left ear signal
$\mathcal{F}^{\mathrm{R}}_{\mathcal{R}}(t,f)$	Ratemap feature of the right ear signal
$\mathcal{F}_{\mathrm{S}}(t,f)$	Synchrony feature of time frame t and channel f
h_f	Hair cell response of channel f
h^{L}_f, $\bar{h}^{\mathrm{L}}_f$	Left ear hair cell response of channel f and its short-term average
h^{R}_f, $\bar{h}^{\mathrm{R}}_f$	Right ear hair cell response of channel f and its short-term average
$h_{\mathrm{FB}}(\omega,f)$	Matrix containing the triangular auditory filter weights for frequency bin ω and auditory filter f
$H[k]$	Azimuth histogram of sound source direction k
$\hat{\mathrm{ild}}_f(t)$	Estimated ILD at channel f and time frame t
$\hat{\mathrm{itd}}_f(t)$	Estimated ITD at channel f and time frame t
i	Discrete time index
j	Gaussian component index
k	Sound source direction index

K	Number of discrete sound source directions
L	Left ear, left side
L	Set of azimuth histogram bin indices of all detected sources
L^{Speech}	Set of azimuth histogram bin indices of all detected speech sources
ℓ_m	Azimuth histogram bin index of the mth candidate
ℓ_n^{Speech}	Azimuth histogram bin index of the nth speech source
$\mathcal{L}(t, f, k)$	Log-likelihood of sound source direction k at time frame t and channel f
$\tilde{\mathcal{L}}(t, f, k)$	Log-likelihood subsequent recursive smoothing
λ_φ	GMM of sound source direction φ
λ_{f,φ_k}	GMM of sound source direction φ_k in frequency channel f
λ_{Speech}	GMM-based speech model
λ_{Noise}	GMM-based noise model
m	Speech source candidate index
M	Number of speech source candidates
$\mathcal{M}(t, f)$	Binary mask at time frame t and frequency channel f
$\mathcal{M}_m(t, f)$	Binary mask of the mth candidate at time frame t and frequency channel f
$\vec{\mu}_j$	Mean vector of the jth Gaussian component
n	Speech source index
$\left\lvert\hat{N}(t, \omega)\right\rvert^2$	Estimated noise power spectrum at time frame t and frequency bin ω
$\hat{N}^2_{\text{FB}}(t, f)$	Estimated noise power at time frame t at the output of auditory filter f
O	Frame shift in samples
ω	Frequency bin index
p_m	Log-likelihood ratio of the mth candidate
p_n^{Speech}	Log-likelihood ratio of the nth speech source
$p_n^{\text{Speech,W}}$	Weighted log-likelihood ratio of the nth speech source
$\Phi^{\text{LL}}(\omega)$	Auto power spectrum of the left ear signal at frequency bin ω

$\Phi^{RR}(\omega)$	Auto power spectrum of the right ear signal at frequency bin ω
$\Phi^{LR}(\omega)$	Cross power spectrum between the left and the right ear signals at frequency bin ω
$\Psi_{PHAT}(\omega)$	PHAT weighting function at frequency bin ω
$\Psi_{SCOTM}(\omega)$	SCOTM weighting function at frequency bin ω
r	Reliable feature component index
R	Right ear, right side
$\mathcal{R}$	Sub-vector containing reliable feature components
r_φ	Set of $\mathcal{A}$ sound source directions
$\hat{r}_\varphi$	Set of $\hat{\mathcal{A}}$ estimated sound source directions
$\hat{S}(t,\omega)$	Estimated speech spectrum at time frame t and frequency bin ω
$\hat{S}^2_{FB}(t,f)$	Estimated speech power at time frame t at the output of auditory filter f
$\mathcal{S}$	True number of speech sources
$\hat{\mathcal{S}}$	Estimated number of speech sources
$\sigma_{j,d}$	Variance of the jth Gaussian component at feature dimension d
Σ_j	Covariance matrix of the jth Gaussian component
t	Time frame index
T	Number of time frames, number of observations
T_{60}	Reverberation time in seconds
τ	Time lag
τ_{max}	Maximum delay which is evaluated for the synchrony feature
$\hat{\tau}_f(t)$	Time lag corresponding to the maximum peak in the normalized cross-correlation function at channel f and time frame t
θ_c	Threshold on the height of the primary peak of the normalized cross-correlation function
θ_e	Entropy threshold
θ_h	Histogram threshold
θ_φ	Absolute error threshold in degrees
u	Unreliable feature component index
$\mathcal{U}$	Sub-vector containing unreliable feature components

φ	Sound source direction
$\hat{\varphi}^{\mathsf{T}}(t)$	Estimated sound source direction at time frame t
$\hat{\varphi}^{\mathsf{TF}}(t,f)$	Estimated sound source direction at time frame t and frequency channel f
$\mathcal{V}$	Number of Gaussian components
$\mathcal{V}_{\mathsf{min}}$	Minimum number of Gaussian components
$\mathcal{V}_{\mathsf{max}}$	Maximum number of Gaussian components
w_j	Weight of the jth Gaussian component
w_{H}	Hamming window
$x(i)$	Noisy discrete time domain signal at time index i
$\vec{x}$	D-dimensional feature vector
$\vec{x}_{t,f}$	D-dimensional feature vector at time frame t and frequency channel f
x^{T}	Transpose of x
$X(t,\omega)$	Noisy signal spectrum at time frame t and frequency bin ω

Contents

List of Figures

List of Tables

1 General Introduction

1.1 Auditory scene analysis: Humans versus machines

In everyday life, we are exposed to an overwhelming variety of acoustic signals reaching our ears. The human auditory system has the fascinating capability to process this mixture of individual sound sources and to intentionally focus on a desired target source. The ability to extract individual sound sources from a complex acoustic mixture is termed auditory scene analysis (ASA). According to Bregman (1990), this ability of the human auditory system is facilitated by two processes, namely segmentation and grouping. First, the acoustic input is decomposed into spectro-temporal units, where each individual unit is assumed to be primarily dominated by one single sound source. Secondly, individual units that are likely to belong to the same acoustic object are subsequently grouped together, forming a coherent *stream*. In this way, the human auditory system is able to segregate a desired target source from interfering signals in challenging acoustic environments. One of the most frequently used examples to illustrate this ability of the human auditory system to analyze complex acoustic scenes is the so-called *cocktail party effect*, which describes the phenomenon of attending to a conversation with one talker while other talkers are speaking at the same time (Cherry, 1953, Bronkhorst, 2000, Haykin and Chen, 2005). Bregman postulated a set of primitive grouping principles that are presumably utilized by the human auditory system to perceptually organize and group individual acoustic

components that originate from the same acoustic object, among them proximity in time and frequency, periodicity, common onsets and offsets, amplitude and frequency modulation and common spatial location.

Although humans seem to be able to effortlessly analyze even rather complex acoustic scenes consisting of multiple competing sound sources, the development of machine listening algorithms that are able to automatically retrieve information about complex acoustic scenes (e.g., estimating the number of active sound sources and the corresponding azimuthal positions or recognizing the identity of a speaker) has been proven to be extremely difficult. The field of research dealing with computational approaches to perform auditory scene analysis is called computational auditory scene analysis (CASA), where the goal is to *produce a computational description of the objects and their spatial locations in a physical scene from sensory input* (Wang, 2007).

One particular challenging aspect of performing ASA with machines is when the analysis is restricted to binaural signals (Wang and Brown, 2006). It is generally known that the performance of microphone arrays usually improves with increasing number of microphones, and a common requirement for beamforming techniques, such as the adaptive Generalized Sidelobe Canceller (GSC) proposed by Griffiths and Jim (1982), is that the number of microphones must be larger than the number of interfering sound sources to successfully attenuate the interfering signals (overdetermined case). In contrast, the human auditory system is able to simultaneously deal with a large number of competing sources by analyzing only the binaural input that is supplied by the two ears. Consequently, the human auditory system is clearly able to deal with the underdetermined case. For this reason, binaural approaches to CASA are of special interest and may help to provide new insights into potential mechanisms that are employed by the human auditory system.

Although machine listening has made enormous progress over the past decades, many studies verify that the human ability to analyze acoustic scenes is superior to the performance that is achieved by machines. In the context of speech recognition, Lippmann (1997) compared the recognition performance of human listeners with the performance of machines for a wide range of different speech stimuli. He reported that the word error rates of human listeners are often more than an order of magnitude below the error rates obtained by machines. Furthermore, he found that human performance was significantly more robust against the degradation of the speech material. A similar trend was observed for the task of speaker verification. Whereas machines were pretty competitive in clean

acoustic conditions, which might be attributed to the *effectively larger memory resources*, and sometimes outperformed human listeners, humans were significantly more robust against signal degradation (Schmidt-Nielsen and Crystal, 2000) and session variability (Lu *et al.*, 1997). More recent studies determined the gap between human and machine performance in terms of the signal-to-noise ratio (SNR). Considering the task of recognizing consonants in the presence of speech-shaped noise, the human advantage was estimated to be in the range of 10 dB (Sroka and Braida, 2005). Similarly, a 15 dB advantage of human listeners was observed in the context of phoneme recognition (Meyer *et al.*, 2011).

1.2 Common approaches to machine listening

The automatic analysis of acoustic scenes, including the extraction of the position of speech sources and the recognition of the speaker identities would provide valuable information for a wide range of applications, such as teleconference systems, hands-free communication systems and voice-controlled systems. Also, the estimated position of the speech source could be used to control self-steering hearing aids (Rohdenburg *et al.*, 2008) in the presence of interfering noise sources, where information about the spatial position of the target source is used to determine the steering vector for a Minimum Variance Distortionless Response (MVDR) beamformer. In addition, more general knowledge about the acoustic scene can be used to automatically select the signal processing strategies of hearing aids which have been optimized for a particular acoustic situation (Wittkop and Hohmann, 2003).

In state-of-the art speech signal processing systems for automatic speech recognition or automatic speaker recognition, the speech signal is commonly analyzed and characterized by its spectral content. One of the most frequently-used feature representations of speech signals are the Mel Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980), which are short-term cepstral features based on the mel-frequency scale. Cepstral coefficients are obtained by applying an orthogonal transformation, such as the discrete cosine transform (DCT), to the short-term logarithmic mel-frequency scaled spectrum of the speech signal, which results in a largely decorrelated set of MFCC coefficients. Because each individual cepstral coefficient reflects properties of the global spectral shape of the speech signal, it is sufficient to retain the lowest 13 coefficients, which effectively reduces the dimensionality of the feature vector, therefore allowing for a compact representation.

Based on a sequence of MFCC features, a classifier based on hidden Markov models (HMMs) or Gaussian mixture models (GMMs) can be trained with the aim to discriminate the MFCC patterns corresponding to different words or different speaker identities. Given a set of MFCC features, e.g., which correspond to an unknown speaker identity, the classifier is essentially trying to match the observed MFCC pattern with one of the patterns previously learned in the training stage. These classifiers are typically trained to function in a particular acoustic environment, which does often not agree with the application environment however. Hence, any mismatch between the MFCC features obtained in the training and in the testing condition, which might be induced by environmental noise, room reverberation or the presence of interfering talkers, will substantially reduce the classification performance.

One approach that has been suggested to alleviate the mismatch between the features extracted from the training and the testing material is to apply feature normalization techniques which aim at compensating for time-invariant, convolutional distortions and channel effects. More specifically, cepstral mean normalization (CMN) is commonly applied to increase the robustness of automatic speech recognition and speaker identification systems where the long-term mean of each cepstral coefficient is subtracted from the MFCC feature vector (Atal, 1974, Rosenberg *et al.*, 1994). This concept of CMN has been extended to also consider the long-term standard deviation of the feature vector, resulting in cepstral mean and variance normalization (CMVN) (Openshaw and Mason, 1994, Tibrewala and Hermansky, 1997). In order to capture environmental changes, a segmental normalization of the mean and variance can be performed where the statistics of the cepstral coefficients are estimated over a segment length of interest (Viikki and Laurila, 1998). An alternative approach is the RelAtive SpecTrAl (RASTA) technique where the temporal trajectory of the feature vector is filtered in such a way that spectral changes that occur faster than the modulation range of speech components are reduced (Hermansky and Morgan, 1994). More sophisticated approaches try to enhance the feature representation by applying speech enhancement algorithms (Berouti *et al.*, 1979, Boll, 1979, Ephraim and Malah, 1984, 1985, Ephraim, 1992, Srinivasan *et al.*, 2007, Loizou, 2007) in order to eliminate the impact of environmental noise. However, these techniques require a robust estimation of the noise power which is very difficult to obtain in non-stationary noise conditions. Whereas all the aforementioned approaches can - to some extent - improve the robustness of speech signal processing systems in adverse acoustic environments, they do not really address the fundamental problem that the computation of the feature vector during testing is based on a noisy speech signal where no distinction is made between the target signal and interfering sound sources. As a consequence, the resulting feature vector

will reflect the properties of all active sound sources, which in turn limits the detectability of the target source.

1.3 Monaural sound source segregation

A major step towards noise robustness is to segregate the target source from the noisy background, which is achieved by introducing the concept of missing and unreliable acoustic information (Cooke *et al.*, 1994, 2001). The missing data (MD) concept separates the T-F representation of a noisy speech signal into reliable and unreliable T-F units. This separation is accomplished by an ideal binary mask (IBM), which indicates whether a spectrographic T-F unit is dominated by the target signal (reliable T-F unit) or if it is contaminated by noise or interfering sources (unreliable T-F unit). In general, there are two different approaches on how to deal with unreliable information. The first method, termed *imputation*, attempts to recover the true values of the unreliable spectral components. The theoretical advantage of this method is that no modification of the classifier is required and that conventional features such as MFCCs can be extracted based on the reconstructed spectral representation. Nevertheless, this approach imposes the substantial challenge of precisely recovering the missing information. For an overview and comparison of different data imputation methods, the reader is referred to Raj (2000). The second alternative, called *marginalization*, modifies the structure of the classifier in such a way that the classification is based solely on the reliable components, effectively ignoring the unreliable components. Further knowledge about the bounds of unreliable T-F units can be incorporated in both methods, imposing additional constraints on the possible range of values that the unreliable components might have had. In the context of speech recognition, it has been shown that the usage of the IBM allows for excellent speech recognition performance in the presence of additive noise for a wide range of SNRs and that the marginalization technique is superior to the imputation approach (Cooke *et al.*, 2001). Theoretically, an arbitrary number of sound sources can be processed by the MD approach, assuming that the appropriate set of target-dominated T-F units can be robustly identified for each individual sound source.

Whereas the concept of missing data may be first and foremost seen as a mathematical tool to deal with uncertain, missing information, there are some analogies with human auditory processing. The motivation to treat noise-dominated T-F units as unreliable acoustic information is closely related to the phenomenon of masking where *one sound*

may be obscured, or rendered inaudible, in the presence of other sounds (Moore, 2003). Furthermore, the observation that speech-dominated T-F units tend to be sparse in the presence of background noise has led Cooke to propose a glimpsing model of human speech perception (Cooke, 2005, 2006). It is based on the idea that the detection of glimpses, which are defined as contiguous groups of adjacent T-F units that contain reliable information about the target source, may be employed by human listeners when recognizing speech in the presence of background noise. He observed that the percentage of detected glimpses in noisy speech is a good indicator for human intelligibility. Furthermore, a computational model based on MD recognition and *a priori* information about the position of glimpses in the T-F plane was shown to successfully predict the consonant identification scores of human listeners as a function of the number of competing talkers that contributed to the babble-modulated noise.

Besides the success of incorporating the concept of MD in the field of speech recognition, it has also been shown that applying the IBM directly to a noisy speech signal in order to gate noise-dominant T-F units can significantly increase speech intelligibility for both normal hearing and hearing-impaired human listeners (Anzalone *et al.*, 2006, Brungart *et al.*, 2006, Li and Loizou, 2008a, Kjems *et al.*, 2009, Brungart *et al.*, 2009). Furthermore, the influence of the local SNR criterion (LC) to separate reliable and unreliable T-F units on speech intelligibility has been investigated. When varying the LC threshold while keeping the SNR of the noisy speech mixture fixed, a plateau was found where almost 100% speech understanding was possible for human listeners for a wide range of LC values (Brungart *et al.*, 2006, Li and Loizou, 2008a). It has been suggested that this plateau, ranging from about -20 dB to 5 dB, may indicate that the pattern of the IBM itself is most important for human listeners, not the local LC value (Li and Loizou, 2008a). At a local SNR criterion, e.g. -10 dB, a considerable amount of T-F units in the IBM is dominated by the masker, yet almost 100% of the speech was correctly recognized. Because the IBM seems to direct the attention of the human listener to the target-dominated T-F units, the additional energy of the masker at LC values as low as -20 dB apparently does not degrade the ability of human listeners to recognize speech. In terms of frequency resolution, a number of 16 frequency channels was found to be already sufficient to produce highly intelligible speech signals (Wang *et al.*, 2008, Li and Loizou, 2008b). Furthermore, the formulation of the IBM has been extended to reverberant conditions (Roman and Woodruff, 2011). A reflection boundary was introduced in order to separate the direct path and early reflections of the target signal from the late reflections. Improved speech reception thresholds (SRTs) are obtained when considering the direct path and early reflections of up to 50 ms of the target signal

as part of the desired signal and when treating the late reflections as noise (Roman and Woodruff, 2011).

Unfortunately, the usage of the IBM implies that *a priori* knowledge about the spectro-temporal position of reliable and unreliable T-F units is available, which is rarely the case in practical applications. In practice, the IBM needs to be estimated by analyzing the noisy speech signal. Based on the encouraging experimental results that were obtained with the IBM, Wang suggested that the main goal of CASA is to estimate the ideal binary mask (Wang, 2005). To achieve this, a wide variety of different mask estimation techniques have been proposed to determine the set of reliable T-F units that are associated with the target source.

In order to estimate the ideal binary mask, several monaural cues have been exploited. The reliability of an individual T-F unit can be interpreted as the ratio of the target energy to the energy of interfering noise. Therefore, a common approach to determine target-dominant T-F units is to estimate the local signal-to-noise ratio (SNR) in individual T-F units by tracking the level of the background noise. A local SNR threshold is defined to separate T-F units according to target-dominated and noise-dominated sets of T-F units (Drygajlo and El-Maliki, 1998a, Vizinho *et al.*, 1999, Cooke *et al.*, 2001, Renevey and Drygajlo, 2001, Ris and Dupont, 2001, May *et al.*, 2012a). It has been shown that a local SNR threshold of LC = 0 dB is optimal in terms of the theoretical SNR gain (Li and Wang, 2008, 2009); therefore, this criterion is commonly chosen for the SNR-based mask estimation. For the particular problem of detecting reliable T-F units in the presence of reverberation, a modulation-based mask estimation technique has been proposed, which focuses on the direct sound by detecting speech onsets (Palomäki *et al.*, 2004a, 2006). However, it should be noted that both the SNR-based and the modulation-based estimation of the ideal binary mask may only be valid for acoustic mixtures with one target source, since these approaches can not distinguish between multiple competing target sources. Thus, other cues are required to organize the T-F representation of multi-source mixtures. Drawing inspiration from the auditory domain, the concept of periodicity and the temporal continuity of pitch tracks has been exploited for the estimation of the IBM of multiple speech sources (Wu *et al.*, 2003, Roman and Wang, 2006, Ma *et al.*, 2007). Furthermore, these pitch-related features have been jointly analyzed with modulation-based features (Hu and Wang, 2001, 2004). However, these approaches can only be applied to voiced speech. In order to extend the working range of the pitch-based approaches, information about onsets and offsets (Hu and Wang, 2007) can be incorporated as recently reported by Han and Wang (2011).

1.4 Binaural approaches to sound source segregation

The input from two ears is a fundamental advantage that is utilized by the human auditory system. Although it is argued that common spatial location might not be the cue that is predominantly involved in auditory grouping (Darwin, 1997, 2008), which is supported by the fact that human listeners can segregate a target source from monaural signals, many researchers have emphasized that the ability of the human auditory system to exploit spatial information plays a major role in analyzing and understanding speech in complex multi-source scenarios (Cherry, 1953, Kidd *et al.*, 1998, Hawley *et al.*, 1999, Bronkhorst, 2000, Hawley and Litovsky, 2004). In addition, the usage of binaural cues is very attractive, because the unique combination of interaural time differences (ITDs) and interaural level differences (ILDs) is inherently connected to the physical azimuthal position of sound sources in the acoustic space, regardless of the type of source. Consequently, many computational models attempt to replicate mechanisms of binaural processing to enhance speech in challenging acoustic scenarios. One of the first cocktail party processors that was based on binaural cues was the model proposed by Lyon (1983). The model was capable of separating two competing sounds emerging from different spatial locations. Bodden proposed a cocktail party processor which combined a binaural model of human sound source localization (Lindemann, 1986a, Gaik, 1993) with a time-varying Wiener filter to suppress interfering sound sources from undesired spatial directions (Bodden, 1993). The system was able to improve speech intelligibility for hearing-impaired listeners in anechoic multi-source scenarios. Kollmeier and colleagues presented a binaural noise reduction scheme for enhancing a target source emerging from the frontal direction, while suppressing lateral noise sources and reverberation (Kollmeier *et al.*, 1993). The dereverberation stage aimed at distinguishing between the direct sound and reflections and was based on the interaural coherence. Further extensions have been proposed to combine monaural and binaural cues by incorporating cues related to pitch (Denbigh and Zhao, 1992, Shamsoddini and Denbigh, 2001, Christensen *et al.*, 2007, Wohlmayr and Képsi, 2007, Woodruff and Wang, 2010a,b) and amplitude modulation (Kollmeier and Koch, 1994).

One of the first approaches that utilized binaural cues for the estimation of the ideal binary mask was the model proposed by Palomäki *et al.* (2001, 2004b). The model implemented some aspects of the precedence effect by incorporating an inhibitory mechanism that emphasizes acoustic onsets (Palomäki *et al.*, 2004b). The resulting model was proposed to be able to cope with small room reverberation. Around the same time, machine

learning techniques were employed to create a probabilistic model that incorporated knowledge about the distribution of binaural cues in anechoic conditions (Roman *et al.*, 2002, 2003). A classifier based on a maximum a posteriori (MAP) decision rule was trained to jointly analyze the patterns of ITDs and ILDs in different frequency channels. It was shown that binaural cues can be used to approximate the performance of the IBM in anechoic conditions (Roman *et al.*, 2002, 2003). A substantial SNR gain of up to 12 dB was reported for anechoic three-source scenarios when the estimated binary mask was used to segregate the target source (Roman and Wang, 2003). Furthermore, the probabilistic approach consistently outperformed the model proposed by Bodden (1993) in all experimental conditions. Nevertheless, handling the impact of reverberation and the presence of multiple competing sound sources has remained a challenging problem which is addressed in this thesis.

1.5 Objectives

The primary objective of this dissertation is to create a binaural scene analyzer that can automatically retrieve information about sound sources that are acoustically active in a complex acoustic scene. More specifically, the goal is to develop an algorithm which itself is able to simultaneously localize, detect and recognize multiple target speakers in the presence of reverberation and interfering noise sources based on the analysis of binaural signals. Whereas the term *auditory* in CASA may suggest that the computational approaches are exclusively based on auditory principles, we want to clarify that it is not the intention of this work to build a physiologically plausible computer algorithm. Rather, we use principles of human auditory processing and combine them with machine learning techniques in order to reduce the performance gap between humans and machines and to obtain robust performance similar to human auditory processing in adverse conditions.

1.6 Outline of this thesis

In **Chapter 2**, a probabilistic model for robust sound source localization is presented that is based on supervised learning of azimuth-dependent binaural cues, namely interaural time and level differences, in individual frequency channels. The distribution of these binaural cues depends on a variety of factors, among them room reverberation, the presence

of multiple competing sound sources, and changes in the source-receiver configuration. Because of this, multi-conditional training is performed to account for the uncertainty of binaural cues resulting from complex acoustic scenarios. The performance of the proposed localization model is systematically evaluated in multi-source scenarios with up to four competing speech sources in reverberant environments with a reverberation time of up to $T_{60} = 0.69\,\text{s}$. Experimental results show that the proposed localization model achieves significantly higher localization accuracy compared to binaural state-of-the art localization techniques. Furthermore, the ability of the proposed model to predict the number of active sound sources in an acoustic mixture is explored. Relevant publications related to this part of the thesis are May *et al.* (2009, 2011a).

The localization model presented in Chapter 2 is used in **Chapter 3** to create a 2D localization map by grouping individual T-F units according to the spatial location of the most dominant sound source. In order to increase the reliability of the estimated sound source direction of individual T-F units, a spectro-temporal integration stage is proposed, which integrates spatial evidence across a set of T-F units that are believed to belong to the same acoustic source. We find that this integration stage reduces the overall localization error and increases the amount of available information. The relevant publication related to this part of the thesis is May *et al.* (2010).

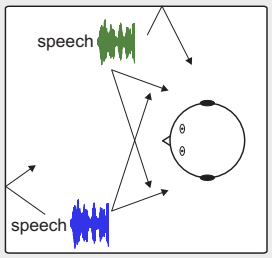
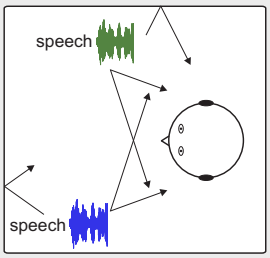
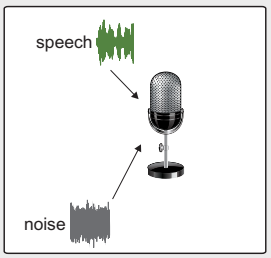
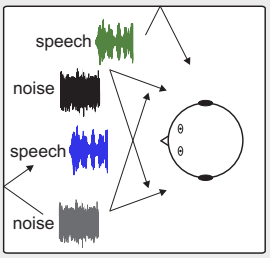
Although missing data techniques provide a powerful framework to deal with multiple overlapping sound sources and interfering noise, the approach requires that the feature space can be separated into reliable and unreliable T-F units. Therefore, it is inherently limited to the usage of spectral features that have a limited performance as compared to the more compact MFCC features. In **Chapter 4**, we incorporate the adaptation of speaker models using universal background models (UBMs) into the framework of MD-based speaker identification. This integration aims at improving the robustness of speaker models that are based on spectral features in the context of MD recognition. Our experimental results show that this combination substantially decreases the sensitivity of the resulting speaker models to errors in the binary mask estimation, which is especially advantageous in acoustic scenarios with highly fluctuating noise. Furthermore, the problem of estimating the ideal binary mask (IBM) in the presence of background noise is addressed. Therefore, a variety of different techniques to obtain an estimation of the local signal-to-noise ratio in individual T-F units is compared, and the influence on speaker recognition performance is discussed. The relevant publication related to this part of the thesis is May *et al.* (2012a).

In **Chapter 5**, a binaural scene analyzer is presented which combines the binaural front-end developed in Chapter 2 with the robust MD-based speaker recognition framework presented in Chapter 4. The estimation of the ideal binary mask is based on spatial evidence, which allows a simultaneous recognition of multiple target speakers. In order to link the estimated sound source activity supplied by the binaural front-end with the automatic speaker recognition stage, a speech detection module is proposed that is able to detect a predefined number of speech sources in the presence of interfering noise sources. In this way, the binaural scene analyzer is able to selectively focus on processing speech sources and does not require *a priori* information about the spatial position of the speakers. The ability of the proposed binaural scene analyzer to simultaneously localize and recognize a predefined number of competing target speakers in the presence of interfering noise sources and reverberation is systematically evaluated. Relevant publications related to this part of the thesis are May *et al.* (2011b,c, 2012b).

Finally, **Chapter 6** summarizes the main findings of this dissertation and suggests future research.

An overview about the different acoustic conditions that are used in the individual chapters is given in Tab. 1.1. Chapter 2 and Chapter 3 focus on the development of the binaural

Table 1.1: Overview of the acoustic scenarios for each individual chapter.

Chapter 2	**Chapter 3**	**Chapter 4**	**Chapter 5**
Localization of multiple sound sources	Time-frequency-based localization map	Noise-robust speaker identification	Localization and recognition of multiple speakers
			
Receiver			
binaural	binaural	microphone	binaural
Room characteristic			
reverberant	reverberant	anechoic	reverberant
Target sources			
1-4 speech sources	2 speech sources	1 speech source	1-2 speech sources
Interfering sources			
-	-	1 noise source	1-3 noise sources

front-end for robust sound source localization. Therefore, binaural signals consisting of multiple competing speech sources in simulated reverberant conditions are considered. To address the impact of noise, Chapter 4 presents a noise-robust MD-based speaker recognition system which is evaluated with monaural noisy speech signals. Finally, the individual contributions are combined in Chapter 5 and the resulting binaural scene analyzer is systematically evaluated with binaural mixtures consisting of several target sources and interfering noise source in simulated reverberant conditions.

This chapter is based on:

- May *et al.* (2009): "A probabilistic model for robust acoustic localization based on an auditory front-end," in *Proceedings of the NAG/DAGA*, Rotterdam, The Netherlands, p. 254 (A).
- May *et al.* (2011a): "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing* **19**(1), pp. 1–13.

2

A Probabilistic Model for Robust Localization based on a Binaural Auditory Front-end

Although extensive research has been done in the field of machine-based localization, the degrading effect of reverberation and the presence of multiple sources on localization performance has remained a major problem. Motivated by the ability of the human auditory system to robustly analyze complex acoustic scenes, the associated peripheral stage is used in this chapter as a front-end to estimate the azimuth of sound sources based on binaural signals. One classical approach to localize an acoustic source in the horizontal plane is to estimate the interaural time difference (ITD) between both ears by searching for the maximum in the cross-correlation function. Apart from ITDs, the interaural level difference (ILD) can contribute to localization, especially at higher frequencies where the wavelength becomes smaller than the diameter of the head, leading to ambiguous ITD information. The interdependency of ITD and ILD on azimuth is a complex pattern that depends also on the room acoustics, and is therefore learned by azimuth-dependent Gaussian mixture models (GMM). Multi-conditional training is performed to take into account the variability of the binaural features which results from multiple sources and the effect of reverberation. The proposed localization model outperforms state-of-the-art localization techniques in simulated adverse acoustic conditions.

2.1 Introduction

The ability to localize sound sources in adverse acoustic environments is necessary for a wide range of applications, e.g. communication devices and hearing aids. Although extensive research has been done in the field of localization, the localization of multiple sources in adverse acoustic conditions has remained a challenging task. The performance of microphone array localization depends on the array configuration and generally increases with the number of microphones (DiBiase *et al.*, 2001). In contrast to microphone-array based techniques, the performance of the human auditory system is very robust against the presence of multiple competing sources (cocktail party scenario) for tasks related to localization (Hawley *et al.*, 1999) and speech recognition (Bronkhorst and Plomp, 1992), despite only exploring the acoustic mixture arriving at both ears. These remarkable capabilities of the human auditory system imply that in principle it is possible to analyze complex acoustic scenes and to independently localize and identify a desired source within an acoustic mixture based on binaural signals. This human ability to analyze complex acoustic mixtures is referred to as auditory scene analysis (ASA) (Bregman, 1990). One influential view on ASA is that the underlying mechanisms to retrieve information about a specific sound source can be divided into two processes: first, the human auditory system parses the acoustic scene into fragments (regions in the time-frequency plane), and second, fragments which may belong to the same acoustic object are grouped together. Because of this ability of the human auditory system to segment, group and integrate information of multiple sources in adverse acoustic conditions, research dealing with models of computational auditory scene analysis (CASA) is a growing field. A comprehensive overview of CASA and its relevance for applications in the field of automatic speaker recognition (ASR) and speech segregation can be found in Wang and Brown (2006). The reliable estimation of the source position based on binaural signals is relevant not only for CASA systems, e.g. as applied in the area of robust speech recognition (Palomäki *et al.*, 2004b, Harding *et al.*, 2006), but furthermore is required for hearing aids and systems related to wearable augmented reality audio (WARA) (Härmä *et al.*, 2004). Inspired by the robustness of the human auditory system, several studies have incorporated stages of human auditory processing to improve sound source localization in adverse acoustic conditions (Bodden, 1993, Faller and Merimaa, 2004, Wilson and Darrell, 2005).

The two major cues which are exploited by the human auditory system to localize acoustic sources are interaural time and level differences. One of the classical approaches to

measure the interaural time difference is to search for the main peak in the generalized cross-correlation (GCC) function (Knapp and Carter, 1976), which estimates the interaural time difference (ITD) between the left and the right ears. The mathematical operation of the uniformly-weighted GCC between binaural signals is equivalent to the coincidence model suggested first in 1948 for describing human sound source localization (Jeffress, 1948). Since then, several modifications and extensions have been proposed to explain the results obtained from psychoacoustic experiments. Another important cue exploited by the human auditory system is the interaural level difference (ILD), which is attributed to the head shadowing effects. The ILD was taken into account by incorporating the mechanisms of contralateral and temporal inhibition into the cross-correlation model (Lindemann, 1986a). In this way, several psychoacoustic phenomena related to the precedence effect could be successfully predicted (Lindemann, 1986b). A comprehensive review of the recent development of binaural models can be found in Braasch (2005).

As indicated, the ILD cue plays an important role in localization, especially at higher frequencies where the wavelength becomes smaller than the diameter of the head, leading to ambiguous ITD information. Nevertheless, there have been only a few attempts to combine both cues in a model for binaural sound localization. For example, the peak selection in the cross-correlation analysis was steered by the ILD cue in order to select the correct peak at higher frequencies where the cross-correlation function becomes ambiguous (Viste and Evangelista, 2004). Nonetheless, the presence of reverberation has a stronger impact on the ILD cue than on the ITD (Shinn-Cunningham *et al.*, 2005), thus making such a peak selection procedure less reliable in adverse acoustic conditions. In addition, the dependence of ITD and ILD on azimuth is a complex, multimodal pattern that also depends on the reverberation and the presence of competing sources, and accordingly it can be best exploited by using a probabilistic model.

A combined evaluation of binaural cues has been successfully applied as a front-end for anechoic sound source segregation (Roman *et al.*, 2003), where a joint feature space consisting of ITDs and ILDs was trained in ideal acoustic conditions in order to segregate a target source from interfering sources at different azimuth positions. The modeling was performed by an adaptive kernel density method, because the use of Gaussian mixture models (GMM) had been reported to lead to issues related to the initialization process and to the problem of selecting the number of Gaussian components (Roman *et al.*, 2003).

In Brown *et al.* (2006) and Harding *et al.* (2006), the effect of reverberation was included

in a probabilistic model to predict a missing data mask for speaker recognition. A histogram technique was utilized to model the probability density function (PDF) of the binaural cues associated with the target source located at 0° azimuth. Although the performance in reverberation was reported to be comparable to the system established under anechoic acoustic conditions (Roman *et al.*, 2003), the performance was sensitive to the source/receiver configuration that was used to train the model for the recognition of the binaural cues (Brown *et al.*, 2006). In Nix and Hohmann (2006), PDFs of interaural cues based on the fast Fourier transform (FFT), namely interaural phase differences (IPDs) and ILDs were measured by histograms in order to perform localization in non-stationary noise conditions. Those cues were integrated by combining their probabilities across frequency.

In this chapter, a sound source localization model is presented that is robust against the presence of multiple sources, changes in the source/receiver configuration and the impact of reverberation. Based on an auditory front-end, the complex interaction of ITDs and ILDs is learned by a probabilistic model that is trained under various acoustic conditions to obtain robustness. For single sound source localization, long analysis windows between 100-200 ms length are commonly applied to increase the robustness of localization in reverberation (Champagne *et al.*, 1996, Chen *et al.*, 2005). To be able to resolve the time-dependent azimuth position of the most salient source in complex multi-source mixtures, which is required for many relevant applications (e.g. beamforming, tracking and CASA systems), a relatively short analysis window of 20 ms is used in this chapter. Gaussian mixture models (GMM) are used to learn the azimuth-dependent distribution of the binaural feature space consisting of both ITDs and ILDs. This feature space, based on time-domain analysis, is then compared to an FFT-based feature space consisting of IPD and ILD as described in Nix and Hohmann (2006). Furthermore, the selection procedure for the number of Gaussian components will also be addressed. The straightforward approach is a manual selection by visual inspection, which will be compared to an unsupervised learning of Gaussian mixtures (Figueiredo and Jain, 2002), where the model complexity is automatically determined. The performance of the GMM-based localization method is evaluated and compared with some state-of-the-art binaural localization techniques in multi-source reverberant conditions. Finally, the ability of the model to generalize to unknown source/receiver configurations is discussed.

This chapter is organized as follows. The next section describes the extraction of the binaural cues. Section 2.3 explains the details of the probabilistic model for sound source localization. In Section 2.4, the evaluation procedure is described and the performance in

simulated reverberant multi-source scenarios is presented in Section 2.5. A summary of the main findings and concluding remarks will be given in Section 2.6.

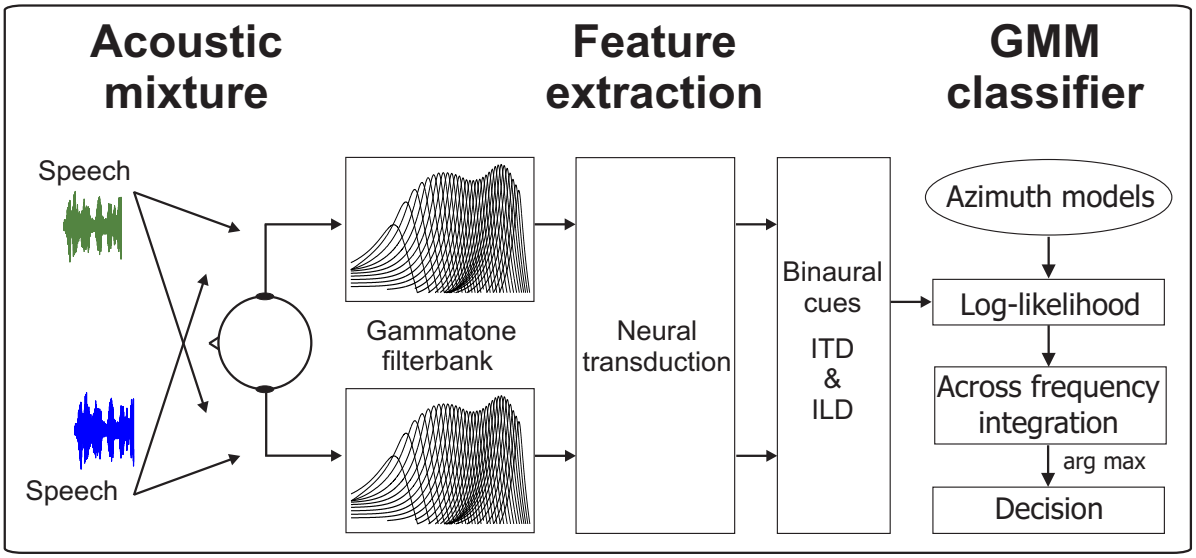


Figure 2.1: Schematic diagram of the binaural sound source localization model based on a GMM classifier. See Section 2.2 and Section 2.3 for details.

2.2 Binaural cue extraction

The two main cues enabling human sound source localization in the horizontal plane are interaural time and level differences, ITDs and ILDs respectively. The ITD cue is most robust at low frequencies, whereas the ILD cue is predominantly used at higher frequencies (Blauert, 1997). The direction-dependent spectral modifications largely associated with the complex pinna shape are especially important for the elevation detection and to resolve front/back ambiguities. Because the localization task in this work is restricted to the frontal horizontal plane (zero elevation), the ability to discriminate sources in the vertical domain will not be investigated.

2.2.1 Auditory front-end

A schematic overview of the proposed binaural sound source localization model is shown in Fig. 2.1. The peripheral processing of the human auditory system is simulated by an auditory front-end consisting of a gammatone filterbank followed by inner hair cell-processing. This front-end is adopted from Roman *et al.* (2003). In order to resemble the

frequency selectivity of the human cochlea, the signals arriving at the left and the right ear are decomposed into $F = 32$ auditory channels using a fourth-order gammatone filterbank. More specifically, phase-compensated gammatone filters are used to synchronize the binaural cues across auditory channels at a common time instance (Brown and Cooke, 1994). The gammatone channel responses are aligned by compensating for the group delays of the gammatone filters at their nominal center frequencies. According to Glasberg and Moore (1990), the channel center frequencies are equally distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5 kHz. Similar to the model described in Roman *et al.* (2003), channel-dependent gains are applied to simulate the middle-ear transfer function, as determined by Moore *et al.* (1997). The neural transduction process in the inner hair cells is approximated by halfwave-rectification followed by a square-root compression. The resulting binaural auditory signals of the left and the right ear are represented by h_f^{L} and h_f^{R} respectively, where the subscript f indicates the gammatone channel index. Binaural cues are estimated using a rectangular window of 20 ms at a sampling frequency of $f_{\mathrm{s}} = 44.1$ kHz (corresponding to a frame length of $B = 882$ samples). An overlap between successive frames of 50% was applied, corresponding to a frame shift of 10 ms. This high temporal resolution was chosen to capture rapid changes in multi-source scenarios.

2.2.2 ITD

The time difference between the binaural auditory signals in the fth channel is estimated using the normalized cross-correlation function, which is defined as a function of time lag τ and the frame number t

$$C_f(t,\tau) = \frac{\sum_{b=0}^{B-1} \left(h_f^{\mathrm{L}}(t\cdot B/2-b)-\bar{h}_f^{\mathrm{L}}\right)\left(h_f^{\mathrm{R}}(t\cdot B/2-b-\tau)-\bar{h}_f^{\mathrm{R}}\right)}{\sqrt{\sum_{b=0}^{B-1}\left(h_f^{\mathrm{L}}(t\cdot B/2-b)-\bar{h}_f^{\mathrm{L}}\right)^2}\sqrt{\sum_{b=0}^{B-1}\left(h_f^{\mathrm{R}}(t\cdot B/2-b-\tau)-\bar{h}_f^{\mathrm{R}}\right)^2}}. \quad (2.1)$$

$\bar{h}_f^{\mathrm{L}}$ and $\bar{h}_f^{\mathrm{R}}$ denote the mean values of the left and right auditory signals and these are estimated over the frame number t. The normalized cross-correlation function is evaluated for time lags within the range of $[-1, 1]$ ms, and its maximum corresponds to the estimated ITD (in samples)

$$\hat{\tau}_f(t) = \arg\max_{\tau} C_f(t,\tau). \quad (2.2)$$

The resolution of the ITD cue is limited by the sampling interval, whereas the actual time delay can lie between two successive samples. In order to increase the ITD accuracy while keeping the computational complexity moderate, exponential interpolation can be applied around the estimated maximum $\hat{\tau}_f(t)$ of the normalized cross-correlation function (Zhang and Wu, 2005)

$$\hat{\delta}_f(t) = \frac{\log C_f(t, \hat{\tau}_f(t)+1) - \log C_f(t, \hat{\tau}_f(t)-1)}{4\log C_f(t, \hat{\tau}_f(t)) - 2\log C_f(t, \hat{\tau}_f(t)-1) - 2\log C_f(t, \hat{\tau}_f(t)+1)}. \quad (2.3)$$

The fractional part $\hat{\delta}_f(t)$ can be considered to describe the interpolated peak position relative to the estimated integer peak position $\hat{\tau}_f(t)$, and the overall ITD estimate $\hat{\text{itd}}_f(t)$ is then given in seconds by the combination of the integer and the fractional estimate

$$\hat{\text{itd}}_f(t) = \left(\hat{\tau}_f(t) + \hat{\delta}_f(t)\right)/f_s. \quad (2.4)$$

In addition to the exponential interpolation, another classical approach was also tested, which describes the peak of the band-limited cross-correlation function by a parabola (Jacovitti and Scarano, 1993). The performance of both interpolation methods were almost identical in low gammatone channels up to 1.5 kHz, whereas the exponential interpolation gave better results at higher frequencies and was therefore selected in the current chapter.

2.2.3 ILD

The interaural level difference is estimated by comparing the energy integrated across the time interval B between the left and right ears. The ILD cue in the fth gammatone channel expressed in dB is given by

$$\hat{\text{ild}}_f(t) = 20\log_{10}\left(\frac{\sum_{b=0}^{B-1} h_f^{\text{R}}(t\cdot B/2-b)^2}{\sum_{b=0}^{B-1} h_f^{\text{L}}(t\cdot B/2-b)^2}\right). \quad (2.5)$$

Note that in Eq. (2.5), 20 instead of 10 is used to compensate for the square-root compression of the neural transduction process. A sound source positioned at the left-hand side will result in a negative ILD whereas a positive ILD will be caused by a source lateralized to the right-hand side.

2.3 Model architecture

A probabilistic model is used to estimate the position of a sound source from the set of binaural cues described in Section 2.2. Therefore, GMMs are trained to recognize the azimuth-dependent pattern of the binaural cues. The training and the architecture of the model is described in the following sections.

2.3.1 Multi-conditional training

To achieve a robust localization performance, the model is trained in various simulated acoustic conditions to account for the variability of the binaural features caused by multiple sources and the effect of additional reverberation. As analyzed by Roman *et al.* (2003), the distribution of binaural cues is dependent on the presence of an interfering source and its strength relative to the target source. To incorporate this effect into the model, the training sequences consist of a target source within the azimuth range of $[-50°, 50°]$ with an interfering source positioned at $\pm 5°$,$\pm 10°$,$\pm 20°$,$\pm 30°$ and $\pm 40°$ relative to the target azimuth (see Fig. 2.3). All target-/interfering-source combinations were presented at three different global signal-to-noise ratios (SNRs) of 20, 10 and 0 dB, as defined prior to spatialization. Moreover, the uncertainty of the binaural features attributed to the room reverberation is taken into account by using simulated Binaural Room Impulse Responses (BRIRs). This approach is similar to the training procedure described in Brown *et al.* (2006), Harding *et al.* (2006), where mixtures of multiple sources (target plus interfering source) were used in reverberation to obtain a more reliable model for identifying time-frequency elements (binary mask), which are associated with the target source only. This mask was used in the context of missing data speech recognition, where the recognition stage is based on the reliable components, while excluding time-frequency elements which are dominated by the interfering source. The authors showed that their model is robust against changes in the simulated room absorption which were not considered in the training stage. However, their system was sensitive to the relative placement of the source and the receiver within the room (source/receiver configuration) (Brown *et al.*, 2006). To improve the model performance in this respect, the training data in the current study were created using multiple source/receiver configurations, where the BRIRs were synthesized by using the room simulation package developed by Campbell *et al.* (2005). This software package combines a database of Head-Related Transfer Functions (HRTFs) (Gardner and Martin, 1994) measured in anechoic conditions, with room reflections simulated according

to the image source model (Allen and Berkley, 1979). To create the target and the interfering source signals, utterances of male speakers that were randomly selected from the TIMIT database (Garofolo *et al.*, 1993) were convolved with the simulated BRIRs. Throughout the multi-conditional training phase, the frequency-dependent absorption coefficients of the room were chosen to yield a constant reverberation time $T_{60} = 0.5\,\text{s}$, in order to introduce the same amount of uncertainty for all gammatone channels. Note that only one level of reverberation was used to train the localization model. To obtain a reliable estimate of the target position, it is essential that the probabilistic model is trained only with binaural features that are associated with the target source. Thus, the following four criteria were employed to select the frames where the binaural features are dominated by the target source:

1. An energy-based voice activity detector (VAD) was used to monitor the activity of the target source, and, a frame was considered to be silent and excluded if the energy level drops by more than 40 dB below the global maximum.

2. Frames were considered for training only if the target source was stronger than the interfering source. This analysis compared the energy of the target source to the energy of the interfering source after spatialization. The signals of the left and the right ear were added prior to energy computation.

3. Frames were removed when the height of the primary peak in the normalized cross-correlation function was less then a threshold θ_c, assuming that the associated binaural cues are dominated by the room reflections. This third criterion was motivated by the fact that the amplitude of the normalized cross-correlation reveals information about the ratio between the direct sound and the room reflections, which becomes low when the signals at the left and the right ear are dominated by reflections. The threshold was set to $\theta_c = 0.3$ by inspection, which still considers frames with low correlation between the binaural signals to incorporate the uncertainty of binaural cues resulting from adverse acoustic conditions into the training procedure.

4. The fourth criterion removed frames from the training set, if the maximum of the normalized cross-correlation function corresponded to one of the most lateral time lags ($\tau = \pm 44$)[1]. For those time lags, it is assumed that the corresponding ITD of $[-1, 1]$ ms is outside the plausible range for the human head.

1 Valid for a sampling frequency of 44.1 kHz.

Note that the last three criteria are monitored in all gammatone channels independently, whereas the first criterion (VAD) is based on the signal prior to gammatone analysis. Based on the requirement that all four criteria have to be satisfied, about 50% of the frames were removed.

2.3.2 Binaural feature space

As already pointed out, the ITD and the ILD cues contain complementary information about the source position, and can therefore be combined in a two-dimensional binaural feature space

$$\begin{aligned} X_f &= \{\vec{x}_{1,f}, \ldots, \vec{x}_{T,f}\} \\ &= \left\{ \left(\hat{\text{itd}}_f(1), \hat{\text{ild}}_f(1)\right), \ldots, \left(\hat{\text{itd}}_f(T), \hat{\text{ild}}_f(T)\right) \right\}, \end{aligned} \tag{2.6}$$

where T represents the number of observations for gammatone channel f. This joint feature space of ITDs and ILDs is shown in Fig. 2.2 for a speech source at 0° azimuth using the multi-conditional training. Each dot represents an observation of the binaural feature space for a single frame within a specific gammatone channel. The receiver (KEMAR head) was placed in the middle of the room, whereas the target source and the interfering source were positioned at a radial distance of 1.5 m with respect to the receiver. The binaural cues were simulated in a room measuring 5.1 x 7.1 x 3 m with a reverberation time of $T_{60} = 0.5$ s.

It can be observed in Fig. 2.2 that the interdependency of ITDs and ILDs results in complex patterns. At higher frequencies, where the wavelength is smaller than the diameter of the head, the ITD information becomes ambiguous. This effect is reflected by the number of distinct clusters in the binaural feature space, which systematically increases with the gammatone center frequency. The spread of the clusters can be related to the reverberation and the presence of an interfering source. Considering a target source at 0° azimuth in anechoic conditions without an interfering source, the distribution of ITDs and ILDs would be very narrow and hardly any side peaks would be observed.

To estimate the position of a sound source from a set of binaural cues, the complex pattern of the binaural feature space is learned by a probabilistic model. In Brown *et al.* (2006), the probability density function (PDF) of the binaural feature space depending on the sound source azimuth was modeled by a histogram technique. In that study, two

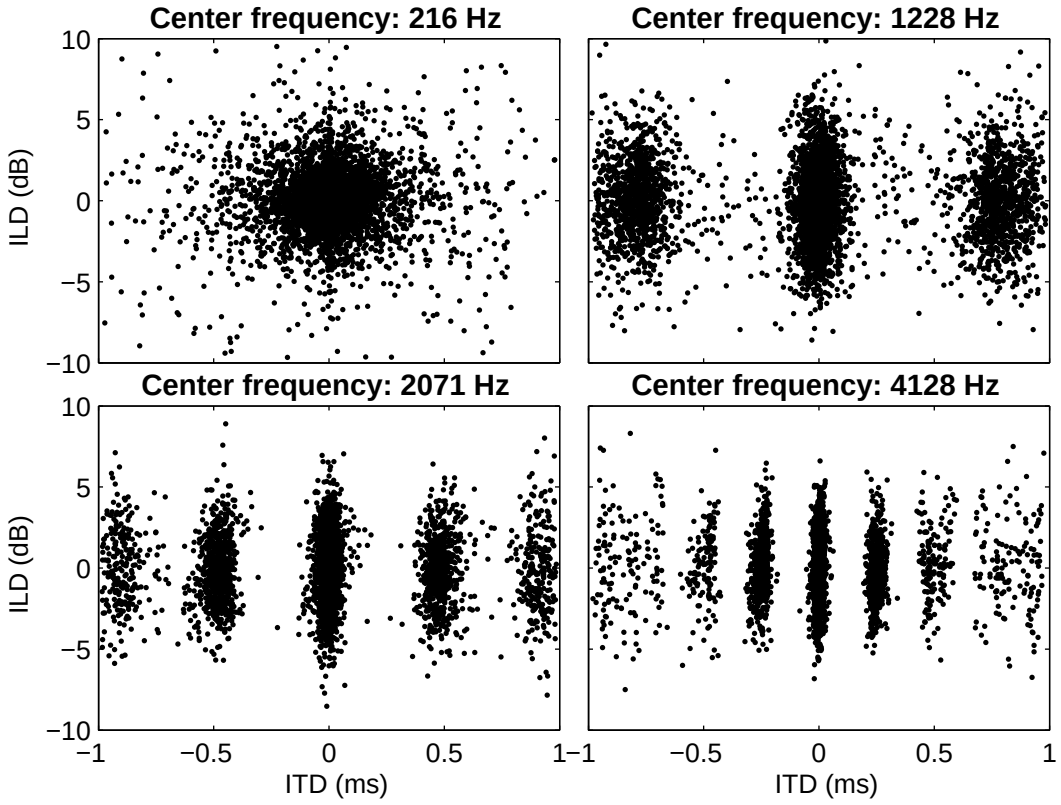


Figure 2.2: Binaural feature space computed for a target source at 0° azimuth at different gammatone filter center frequencies under reverberation condition ($T_{60} = 0.5\,\mathrm{s}$). The number of clusters increases along the ITD feature dimension due to the ambiguous nature of the cross-correlation function at high frequencies.

histograms were computed: one analyzing the binaural feature space for both target and interfering sources, and the second for the observations related to the target source only. The relation between these two histograms was used to derive the probability of a region which is dominated by the target source. The bin size of the histogram is the result of a trade-off between the PDF resolution and the amount of data required for a sufficient training of the model. Furthermore, a threshold needs to be set for the histogram in order to control the potential effect of insufficient training on the PDF, which may occur for certain binaural feature combinations. Hence, ITD-ILD combinations were removed from the histogram if the number of counts was below a certain threshold, producing better estimates. It was also reported in Brown *et al.* (2006) that the performance was sensitive to the histogram threshold and, more importantly, to the source/receiver configuration used for the simulation of the training data.

In order to overcome these limitations, Gaussian mixtures are chosen in the current study to model the probability density of the binaural cues for all azimuth positions. Because the binaural features tend to cluster in the feature space, the azimuth-dependent PDF can be

modeled by the sum of superimposed Gaussian components. Also, the use of GMMs results in a smoother decision area than the histogram technique, which is expected to reduce the sensitivity of the model to unknown source/receiver configurations.

2.3.3 Gaussian mixture modeling

Gaussian mixture models are used to describe the direction-dependent distribution of the binaural feature space. Considering one specific sound source direction φ, denoted by λ_φ, a Gaussian mixture density for a D-dimensional feature vector $\vec{x}$ is the weighted sum of $\mathcal{V}$ Gaussian components (Reynolds and Rose, 1995):

$$p(\vec{x}|\lambda_\varphi) = \sum_{j=1}^{\mathcal{V}} w_j \, p_j(\vec{x}), \tag{2.7}$$

where $\vec{x}$ corresponds to the output of one specific gammatone channel. Each mixture component j is characterized by the component weight w_j, its mean vector $\vec{\mu_j}$ and the covariance matrix Σ_j. The variable λ_φ represents the sound source properties. Furthermore, the mixture weights w_j satisfy $\sum_{j=1}^{\mathcal{V}} w_j = 1$. Each of the $\mathcal{V}$ components is a D-variate Gaussian function given by

$$p_j(\vec{x}) = \frac{1}{(2\pi)^{D/2}\,|\Sigma_j|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu_j})^{\mathsf{T}}\Sigma_j^{-1}(\vec{x}-\vec{\mu_j})\right]. \tag{2.8}$$

The parameters required to uniquely describe the GMM can be summarized by the following notation:

$$\lambda_\varphi = (w_j, \vec{\mu}_j, \Sigma_j) \quad \forall \quad j = 1, \ldots, \mathcal{V}. \tag{2.9}$$

Diagonal covariance matrices are used to describe the dependency between the ITD and the ILD features, since the clusters are orientated perpendicularly with respect to the ILD dimension (see Fig. 2.2). Moreover, the correlation between feature vector elements can be modeled, in principle, by a larger number of diagonal covariance matrices than the full covariance matrices, which are computationally more expensive (Reynolds and Rose, 1995).

Let $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$ be a set of T observations of the D-dimensional feature vector $\vec{x}$. The log-likelihood of the sound source direction φ for a single time frame t can be

computed as follows:

$$\log p(\vec{x}_t|\lambda_\varphi) = \log \sum_{j=1}^{\mathcal{V}} w_j \, p_j(\vec{x}_t). \tag{2.10}$$

One GMM was chosen to model the pattern of the binaural feature space within each gammatone channel f for each sound source direction φ independently. Extending the log-likelihood computation to F gammatone channels indicated by the index f and to K equally likely sound source directions represented by $\{\lambda_{f,\varphi_1}, \lambda_{f,\varphi_2}, ..., \lambda_{f,\varphi_K}\}$, the estimated sound source location is found by maximizing the log-likelihood of the current observation $\vec{x}_{t,f}$

$$\hat{\varphi}^{\top}(t) = \arg\max_{1\leq k\leq K} \underbrace{\sum_{f=1}^{F} \underbrace{\log p(\vec{x}_{t,f}|\lambda_{f,\varphi_k})}_{\text{log-likelihood}}}_{\text{across frequency integration}} . \tag{2.11}$$

This azimuth decision is made on a frame-by-frame basis in order to capture the time-dependent characteristics of multiple acoustic sources. In contrast to summing the binaural cues across frequency (Shackleton *et al.*, 1992), the evidence about a sound source location is accumulated by combining the log-likelihood function across all F gammatone channels. In this way, the uncertainty associated with the azimuth estimate of a particular gammatone channel is taken into account, making an optimal use of the available information. This probabilistic integration of cues was also proposed by Nix and Hohmann (2006). Gaussian mixtures are trained to recognize the binaural feature space in steps of 5°. To increase the localization accuracy, an exponential interpolation (Zhang and Wu, 2005) is applied around the maximum of the log-likelihood function accumulated across frequency in Eq. (2.11).

2.3.4 GMM parameter estimation

The most common approach to estimate the set of GMM parameters λ_{f,φ_k} is to use the iterative Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). After initializing the GMM parameters, each EM iteration consists of two steps, namely the E-step and the M-step, respectively. First, the E-step determines the membership of each training sample by assigning it to the Gaussian cluster which is most likely to have generated it. Based on the new membership estimation, the GMM parameters are recalculated in the M-step. This iterative procedure continues until the difference in likelihood between two successive iterations is less than a predefined threshold ϵ. Although Gaussian mixture

models can, in theory, be established to approximate arbitrarily complex probability density functions, the quality and the robustness of the estimated GMM parameters depend on the number of Gaussian components and the way they are initialized prior to using the EM algorithm.

It is difficult to select the optimal number of Gaussian components $\mathcal{V}$, because the "true" number is usually unknown. An extensive number of Gaussian components reduces the ability of the GMM model to generalize to observations which were not included in the training data. By choosing too few components, however, the essential characteristics of the feature space can not be properly represented. One straightforward approach to choosing the number of GMM components is to visually identify the number of clusters, assuming that these clusters would also be recognized by the initialization procedure. Recently, an algorithm for unsupervised learning of Gaussian mixtures was presented, which automatically selects the optimal number of Gaussian components by minimizing a cost function based on the minimum description length (MDL) criterion (Figueiredo and Jain, 2002). Furthermore, the algorithm was reported to reduce the sensitivity of the EM algorithm to the initialization procedure by starting with significantly more components than required, and successively removing the unnecessary components using the MDL selection criterion. In this study, both manual and automatic approaches were used to approximate the binaural feature space, of which the effect on localization performance will be compared in Section 2.5.

The set of Gaussian parameters listed in Eq. (2.9) needs to be initialized prior to running the EM algorithm. A random initialization or the use of clustering algorithms is one of the common approaches (McLachlan and Peel, 2000). In this study, the k-means clustering algorithm (Lloyd, 1982) is used to find the initial parameters of the $\mathcal{V}$ Gaussian components. As the ITD and the ILD features have different scales, a variance normalization is performed prior to equalize both dimensions.

2.4 Evaluation setup

2.4.1 Baseline systems

The proposed localization model is compared with three baseline systems by using the performance evaluation described in Section 2.4.4. All baseline algorithms were imple-

mented to perform localization by using the same framing parameters as the proposed model: analysis window of 20 ms with a frame shift of 10 ms. The ITD estimates of all baseline systems were also refined by exponential interpolation (Zhang and Wu, 2005).

GCC Gammatone

The first baseline system is based solely on the ITD analysis (cf. Eq. (2.1)) with the same auditory front-end as the GMM-based localization model. Each local peak in the normalized cross-correlation function is replaced by an impulse of the same height and convolved with a Gaussian kernel (Roman *et al.*, 2003, Palomäki *et al.*, 2004b). The same parameters were used as suggested in Palomäki *et al.* (2004b). This process sharpens the peaks in the normalized cross-correlation function, and therefore is beneficial especially when there are multiple sources spatially close to one another. The final azimuth estimate is given by transforming the ITD according to a frequency-dependent mapping function, which takes into account the diffraction effects of the head and shoulders (Bodden, 1993, Roman *et al.*, 2003, Palomäki *et al.*, 2004b), and integrating the information across gammatone channels.

FFT PHAT and FFT SCOTM

The FFT-based generalized cross-correlation (GCC) technique is also used for comparison (Knapp and Carter, 1976). More specifically, two commonly used weighting functions for the GCC are explored: the phase transform (PHAT) (Knapp and Carter, 1976) and a modified[2] version of the smoothed coherence transform (SCOTM) (Carter *et al.*, 1973). Let $\Phi^{\mathrm{LL}}(\omega)$ and $\Phi^{\mathrm{RR}}(\omega)$ be the auto power spectrum of a 20 ms time segment of the left and the right ear signal, respectively. Prior to spectral analysis, the time segments were multiplied by a Hamming window and padded with zeros to reach a window length corresponding to the next highest power of two. Furthermore, let $\Phi^{\mathrm{LR}}(\omega)$ denote the cross power spectrum between the two signals. The frequency-specific weighting functions are given by $\Psi_{\mathrm{PHAT}}(\omega) = 1/\left|\Phi^{\mathrm{LR}}(\omega)\right|$ and $\Psi_{\mathrm{SCOTM}}(\omega) = 1/\left[\Phi^{\mathrm{LL}}(\omega)\Phi^{\mathrm{RR}}(\omega)\right]^{0.3}$. Channel-dependent pre-whitening is applied to the binaural signal in the spectral domain prior to

2 The square-root operator in the denominator of the conventional SCOT weighting was changed to cube-root compression. This modification was found to improve localization performance in reverberation.

the GCC analysis, to reduce the dependence of localization on the structure of the source signal (Chen *et al.*, 2005). Similarly to the *GCC Gammatone*, a mapping function was employed to relate the broadband ITD estimate to the corresponding azimuth. The corresponding mapping functions were derived by learning the azimuth-dependent responses of the two broadband FFT-based localization models to a speech source that was presented systematically at locations in the azimuth range of $[-50°, 50°]$.

2.4.2 GMM settings

As discussed before, the GMMs were trained with the number of Gaussian components selected either manually (Nabney and Bishop, 2001-2004) or automatically (Figueiredo and Jain, 2002). The automatic selection was constrained between $\mathcal{V}_{\text{min}} = 5$ and $\mathcal{V}_{\text{max}} = 25$ Gaussian components. In addition, the stopping criterion of the EM algorithm was set to $\epsilon = 1e^{-5}$ for both training methods with a maximum of 300 iteration steps.

2.4.3 Acoustic conditions

Binaural cues were simulated at various locations in a room of dimensions 5.1 x 7.1 x 3 m, as depicted in Fig. 2.3. The circles correspond to all possible receiver positions (KEMAR head) in the training phase. The diamonds represent the positions of the receiver at which the localization model was evaluated for various reverberation times. Note that only the receiver position 5 was used for both training and evaluation in order to study the influence of known/unknown receiver positions. The receiver was always oriented towards $-90°$ and placed at 1.75 m above the ground, where the source azimuth was varied, at a radial distance of 1.5 m, with respect to receiver position. The positioning of sound sources (filled triangles) is sketched for one training and one evaluation scenario. The black triangle shows the placement of a target source at $-50°$ with respect to receiver position 15, and the positions of an interfering source (crosses) placed at $\pm5°$,$\pm10°$,$\pm20°$,$\pm30°$ and $\pm40°$ relative to the target source, which were systematically processed in the multi-conditional training. As can be seen from Fig. 2.3, the maximum lateral sound source position had to be limited between $\pm90°$ since some receiver positions were too close to the room boundary (e.g. receiver position 11). On the other hand, the placement of the interfering source required an spatial offset of $\pm40°$ with respect to the target source position. Therefore, the GMM localization models were trained and evaluated for a narrower range of target

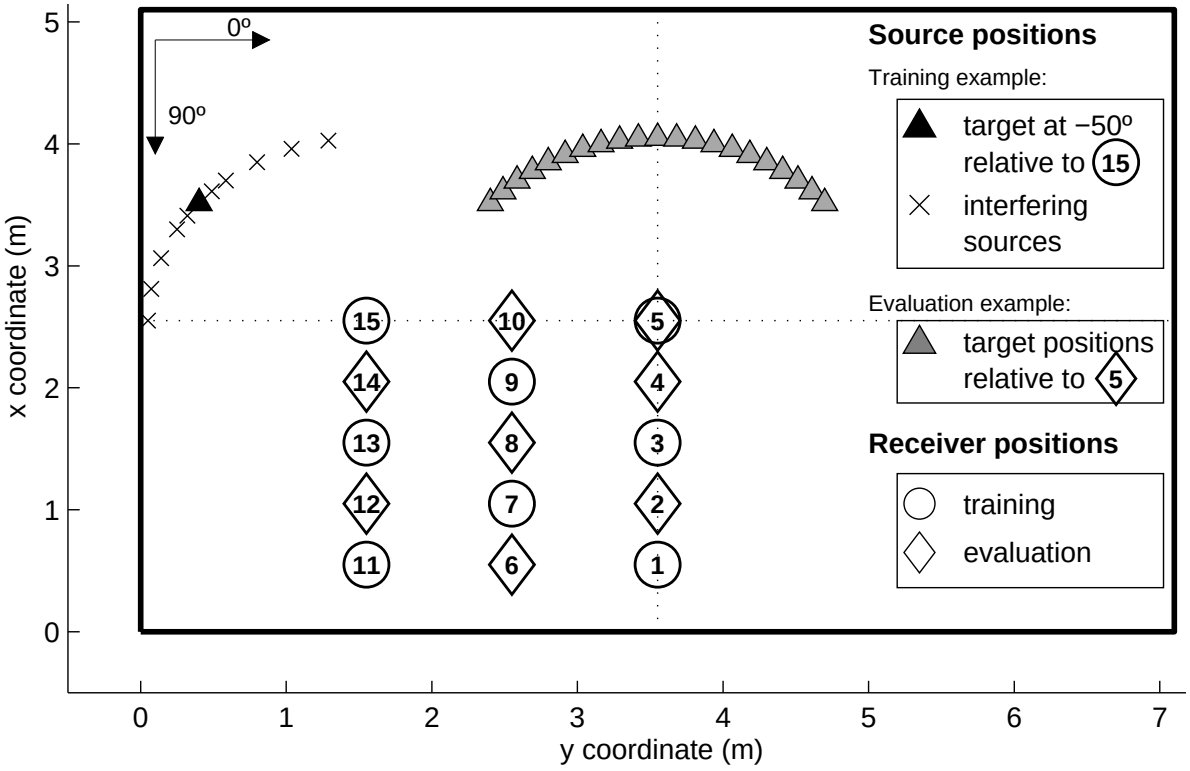


Figure 2.3: Diagram showing the room dimension with all receiver positions used for training (circles) and for evaluation (diamonds). The triangles show exemplarily the positioning of sound sources for one training and one evaluation scenario. See Section 2.4.3 for details.

azimuths between $\pm 50°$ at every $5°$, which resulted in 21 possible sound source locations. The gray triangles represent all 21 target positions with respect to receiver position 5 which were used for evaluation. If not stated otherwise, the localization model was trained with binaural cues extracted for all 8 training positions (denoted by circles in Fig. 2.3) and tested at 7 evaluation positions (denoted by diamonds).

For evaluation, the surface *Acoustic plaster* was selected to characterize the reverberation of the room within the room simulation software (Campbell *et al.*, 2005), where the frequency-dependent absorption coefficients were used for all room boundaries. In order to take into account mild-to-strong reverberation, several different sets of absorption coefficients were used for the room simulation and the frequency-dependent and average reverberation time T_{60} for all experimental conditions are listed in Tab. 2.1. Each row represents one experimental condition used for evaluation. Note that this reverberation with a low-pass characteristic is different from the frequency-independent reverberation time $T_{60} = 0.5\,\mathrm{s}$ which was used to perform the multi-conditional training. This mismatch between training and testing conditions is incorporated to analyze to what extend the proposed approach is able to generalize to *unknown* acoustic conditions.

Table 2.1: Reverberation times T_{60} in seconds for all experimental conditions.

Surface	Frequency in Hz						mean
	125	250	500	1000	2000	4000	
Acoustic plaster T_{60} in s	0.48	0.33	0.13	0.08	0.06	0.06	0.19
	0.67	0.48	0.24	0.15	0.11	0.10	0.29
	0.81	0.61	0.36	0.23	0.18	0.15	0.39
	0.93	0.75	0.46	0.30	0.24	0.19	0.48
	1.09	0.89	0.56	0.39	0.30	0.24	0.58
	1.26	1.03	0.69	0.48	0.37	0.29	0.69

2.4.4 Performance evaluation

The localization performance was evaluated on a frame-by-frame basis where an absolute error threshold $\theta_\varphi = 5\,^\circ$ was considered to classify the estimated source azimuth as either correct or anomalous (Ianniello, 1982, Champagne *et al.*, 1996). Correct estimates were further analyzed by means of the bias and the standard deviation. The performance of all localization techniques was analyzed given the results of a series of Monte-Carlo (MC) simulation experiments which were carried out for four sets of acoustic scenarios with 1, 2, 3 and 4 simultaneously active sound sources. Acoustic sources were represented by male speech selected from the TIMIT database (Garofolo *et al.*, 1993), which were different from those used in the training stage. Reference azimuth labels were obtained from the anechoic speech files by using an energy-based VAD. The average length of the acoustic mixtures was 2.86 s. Regardless of the number of active sources in the acoustic mixture, each localization method produced one azimuth estimate per frame associated with the primary peak, while secondary peaks were ignored for a robust localization. In the case of a multi-source mixture, the localization estimate was considered to be correct, only if the error was within the threshold θ_φ relative to one of the reference azimuth positions. Localization performance was evaluated at all 21 sound source positions within the azimuth range of $\pm 50^\circ$. The sound sources in a multi-source mixture were positioned randomly, but the distance between nearby sources was constrained to at least 10°. The energy of each source was adjusted prior to spatialization to maintain a global SNR of 0 dB. All four scenarios, each consisting of 21 mixtures, were presented three times using different sentences at all 8 receiver positions selected for evaluation (see Fig. 2.3). Therefore, a total of 2016 (4 x 21 x 3 x 8) acoustic mixtures were tested for each reverberation time.

2.5 Localization experiments

2.5.1 Experiment 1: Influence of GMM model complexity

The first experiment investigated the influence of the GMM model complexity on localization performance. As described before, two different methods were used to determine the number of Gaussian components $\mathcal{V}$. First, the binaural feature space was approximated by a manually determined number of components, which was fixed across all azimuth angles and gammatone channels. The second method automatically selected the optimal model complexity for each azimuth angle and each gammatone channel independently, resulting in a variable model complexity.

The percentage of anomalies per frame is listed in Tab. 2.2 as a function of the GMM model complexity, where the performance was averaged across all 4 acoustic scenarios. With increasing number of Gaussian components, the model performed better in all reverberation conditions. In particular, the improvement was significant from 5 to 11 Gaussian components, but the performance gain started to saturate as the model complexity was further increased. For example, the average percentage of anomalies changed only by 0.09% between the order 15 and 21. In addition, the models with an extensive amount of Gaussian components (e.g. 25 or 31) performed slightly worse, which may indicate that the model was overtrained with the limited set of training data.

The last row in Tab. 2.2 shows the performance of the GMM model with variable model complexity. The average number of automatically determined Gaussian components was 17 across azimuth angles and gammatone channels[3] and the performance was similar to the fixed GMM model with 21 components. However, the training procedure for this method takes significantly longer because the model is fitted for the whole complexity range between $\mathcal{V}_{\text{min}}$ and $\mathcal{V}_{\text{max}}$ Gaussian components. Furthermore, the similar performance of both procedures suggests that the requirement for learning the binaural feature space does not change across azimuth directions or gammatone channels. Thus there seems to be no advantage in individualizing the training of the model for each azimuth direction and gammatone channel. Considering the performance saturation and the computational costs, the number of the GMM components was set to 15, constant across gammatone

3 Note that this average number of automatically determined Gaussian components did not vary substantially across frequency channels.

Table 2.2: Experiment 1: Percentage of anomalous localization estimates depending on the number of Gaussian components $\mathcal{V}$.

GMM complexity $\mathcal{V}$	Reverberation time in seconds 0.19	0.29	0.39	0.48	0.58	0.69	mean
Fixed 5	6.65	10.37	15.58	20.52	25.48	30.16	18.13
Fixed 11	2.20	4.30	8.05	12.28	17.16	22.03	11.00
Fixed 15	2.16	4.15	7.76	11.87	16.72	21.50	10.70
Fixed 21	2.11	4.06	7.71	11.77	16.62	21.40	10.61
Fixed 25	2.14	4.09	7.73	11.79	16.63	21.40	10.63
Fixed 31	2.12	4.09	7.76	11.85	16.67	21.52	10.67
Variable	2.14	4.05	7.70	11.76	16.61	21.39	10.61

channels for the simulations presented in this study.

2.5.2 Experiment 2: Selection of binaural cues

The second experiment analyzed the impact of either using ITD or ILD only, or performing localization based on a joint two-dimensional binaural feature space. In Fig. 2.4, the percentage of anomalies averaged across 4 acoustic scenarios is shown as a function of the reverberation time. Also, the performance of the gammatone-based feature extraction (solid lines) defined in Eq. (2.6) is compared with a feature space obtained by an FFT-based auditory front-end (dashed lines), consisting of IPD and ILD, according to Nix and Hohmann (2006). The same parameters were used to simulate the auditory periphery for the two implementations (see Section 2.2.1), where no temporal smoothing was performed across frames.

The results show that the exclusive use of the interaural level cue (*GMM ILD*) is not sufficient to reliably determine the location of acoustic sources. Because the GMM models were trained for the reverberant environment, the localization performance slightly improved as the reverberation time increased up to $T_{60} = 0.3\,\mathrm{s}$, but overall, the error rate of *GMM ILD* was, in general above 50%. In contrast to *GMM ILD*, a reasonable localization performance could be achieved with the ITD model (*GMM ITD*). For example, even for a relatively long reverberation time of $T_{60} = 0.69\,\mathrm{s}$, the average percentage of anomalies was only about 28.39%.

It is apparent in Fig. 2.4 that the joint evaluation of both ITD and ILD produced the best result (*GMM ITD & ILD*). Although the isolated ILD cue does not allow for robust localization, it can significantly improve the localization performance by disam-

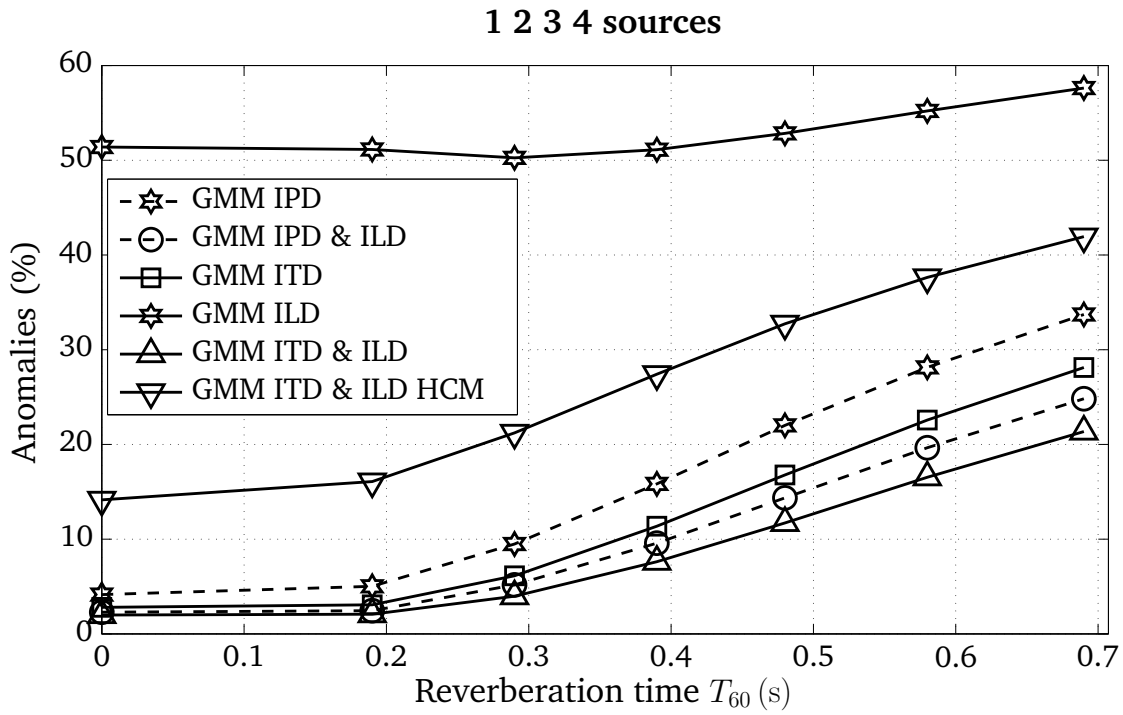


Figure 2.4: Experiment 2: Effect of binaural cue selection on localization performance. The percentage of anomalies is averaged across all 4 acoustic scenarios (1, 2, 3 and 4 sources). Cues were extracted using either the gammatone-based front-end (solid lines) or the FFT-based representation (dashed lines). See Section 2.5.2 for details.

biguating the ITD information especially in acoustic conditions with strong reverberation.

Comparing the gammatone-based (solid lines) and the FFT-based auditory front-end (dashed lines), the GMM model using ITD performed noticeably better compared to the GMM using IPD only. The performance gap between the *GMM ITD* and the *GMM IPD* increased with the reverberation time, which may indicate that the azimuth-dependent IPD pattern is not so systematically modified by the reverberation time as the ITD pattern, and therefore, the GMM classifier cannot utilize it effectively. Using the joint feature space, the performance gap between the gammatone-based (*GMM ITD & ILD*) and FFT-based front-end (*GMM IPD & ILD*) was reduced, but the gammatone-based front-end performed consistently better than its counterpart.

So far, a basic neural transduction model for the inner hair cells was used for the simulation where the signals are half-wave rectified and square-root compressed. However, more detailed models typically employ a high-order low-pass filter to simulate the loss of phase-locking in the auditory nerve at higher frequencies (Bernstein and Trahiotis, 1996, Bernstein *et al.*, 1999), as was used in this study for the model de-

noted *GMM ITD & ILD HCM*. Compared to the performance of the basic hair cell model (*GMM ITD & ILD*), localization accuracy is significantly worse, which clearly reflects the importance of the ITD fine structure at higher frequencies for a better localization performance. Therefore, the basic hair cell model excluding the low pass filter is used for the following experiments.

2.5.3 Experiment 3: Dependency on source/receiver configuration

In the third experiment, the effect of using multiple receiver positions for the multi-conditional training of the localization model was evaluated. Two GMM models, denoted as *GMM Pos 5* and *GMM Pos 7*, were trained with binaural cues simulated at one specific receiver position (position 5 and 7, respectively), and compared to a model, *GMM All*, which was trained with binaural cues extracted for all 8 training positions (1, 3, 5, 7, 9, 11, 13 and 15). The average percentage of anomalies for the receiver position considered for the evaluation is shown in Fig. 2.5, where the performance was averaged across all four acoustic scenarios (from one to four sources). To maximize the influence of receiver position on the binaural cues, long reverberation times (on average $T_{60} = 0.69\,\mathrm{s}$) were used in this experiment, but the performance was found to be similar regardless of the reverberation condition.

The model *GMM All* performed best among all three models, as shown in Fig. 2.5. Only at receiver position 5, the *GMM Pos 5* achieved a lower percentage of anomalies but the difference was rather small (4.19%) considering that the model had been trained for this very position. Indeed, the average performance was best with the *GMM All*, implying that the model is robust and can localize acoustic sources even from untrained receiver positions.

Overall, the *GMM Pos 5* produced the highest percentage of anomalies. The receiver position 5 is located in the center of the room, farthest from any room boundary. On the other hand, all the evaluation positions are relatively close to the room boundaries, where the pattern of the binaural cues can easily be modified or shifted by the acoustic reflection. For example, the performance is particularly low at receiver positions 6 and 12, which are close to the room boundary (see Fig. 2.3). Indeed, the model is quite sensitive to the placement of the receiver (the performance difference between position 5 and 12 is 17.58%), which obviously resulted in large variances shown by the error bars.

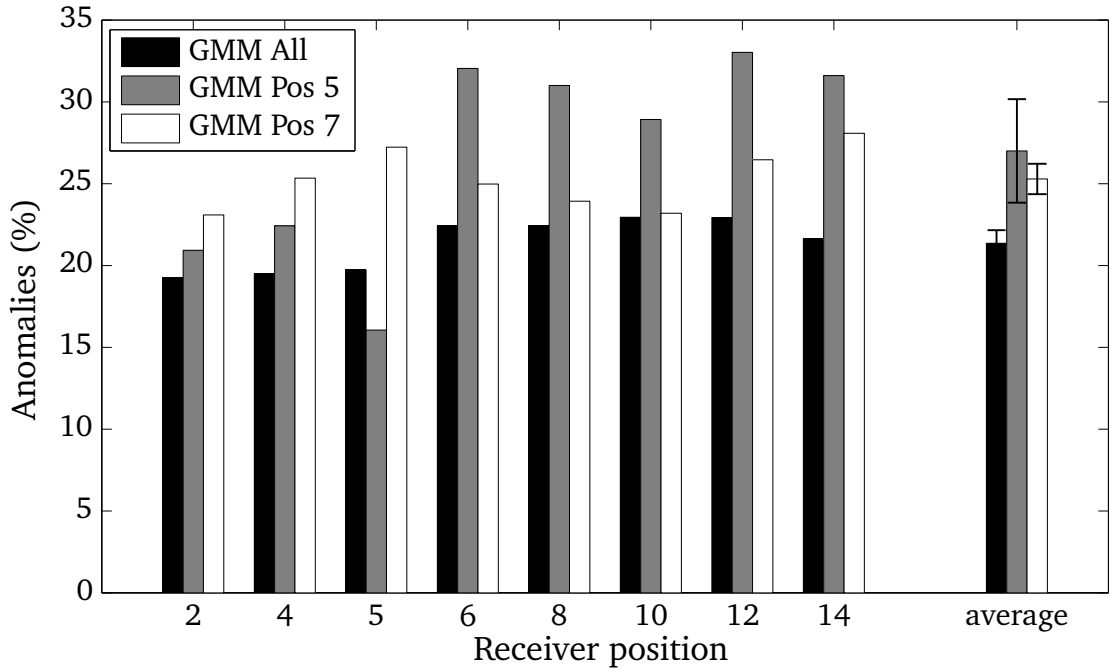


Figure 2.5: Experiment 3: Percentage of anomalies evaluated at various receiver positions under reverberation condition ($T_{60} = 0.69\,\mathrm{s}$). The GMM localization model was trained using either one specific training position (Pos 5, Pos 7) or all 8 training positions.

Compared to the receiver position 5 (see Fig. 2.3), position 7 is closer to the evaluation positions, and therefore the *GMM Pos 7* performed better than the *GMM Pos 5*. Especially at receiver position 6, 8 and 10, which are close to the model training position 7, the overall percentage of anomalies is almost as low as that for the *GMM All*. Nevertheless, the performance is, in general, better with the GMM model trained with multiple receiver positions than with one specific position.

So far, the radial distance between the acoustic sources and the receiver was kept constant at 1.5 m for both the training and the evaluation conditions. To analyze the effect of the radial distance on localization performance, the proposed localization model trained with binaural cues at a radial distance of 1.5 m was evaluated at receiver position 2 for five different radial distances. Whereas ITDs can be considered to be fairly independent of the distance between the source and the receiver, ILDs can change considerably with distance in the proximal region, in particular for nearby sources at distances below 0.5 m (Brungart and Rabinowitz, 1999). However, for distances larger than 1 m, the distance-dependent changes of binaural cues are assumed to be negligible (Brungart *et al.*, 1999). The effect of distance is incorporated in the room simulation software (Campbell *et al.*, 2005) by simulating the distance-dependent circular wave attenuation and by modeling the air absorption with a low-pass filter.

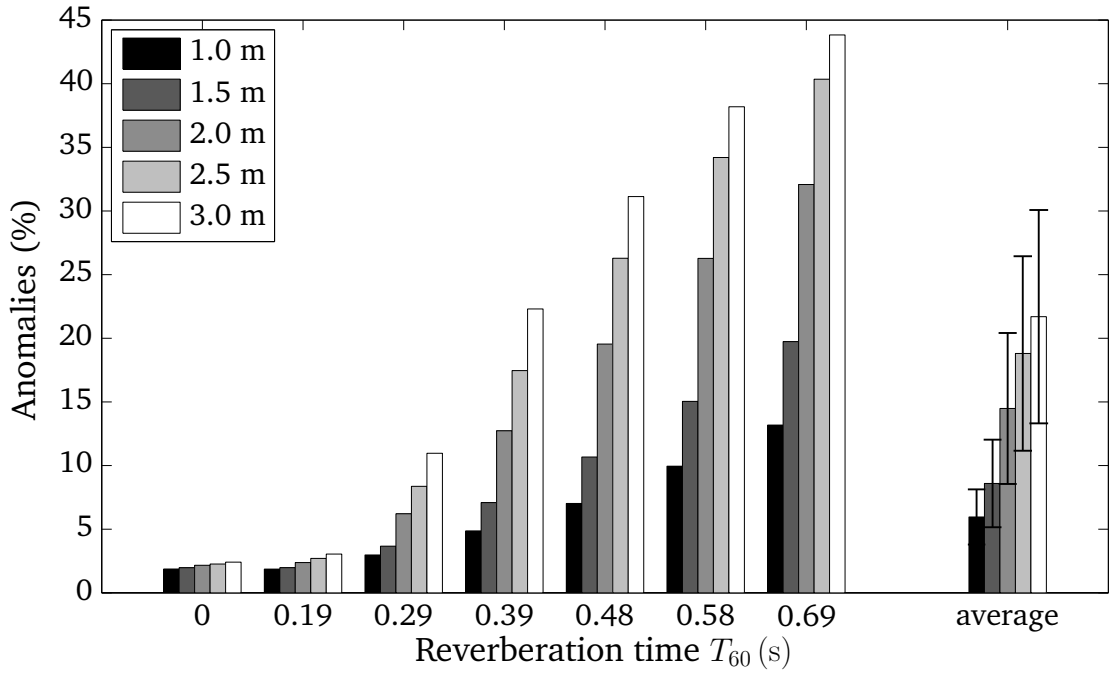


Figure 2.6: Experiment 3: Percentage of anomalies depending on reverberation time evaluated at various distances between the source and receiver position 2. The GMM localization model was trained with binaural cues at a radial distance of 1.5 m using all 8 training positions.

The direct-to-reverberation ratio (DRR) decreases with increasing radial distance, which may reduce the reliability of the measured binaural cues. The percentage of anomalies is presented in Fig. 2.6 averaged over all 4 acoustic scenarios. As shown in Fig. 2.6, the localization prediction of the model becomes less reliable with increasing source-receiver distance. With longer reverberation time, the localization error is more frequent, and the difference in performance is quite noticeable when going from 1.5 m to 2.0 m. This result shows that the GMM model trained at a certain fixed distance may not successfully be applied for the localization prediction of more distant sources. However, the fact that the model performance improves at a shorter distance indicates that the model is capable to generalize to distances which have not been included in the training phase. The poorer performance at larger distances seems to be mainly determined by the reduced reliability of the binaural cues, resulting from the larger distance between the source and the receiver. This is in line with the expectation that the direct-to-reverberation ratio decreases with increasing radial distance and consequently affects the reliability of the binaural cues.

2.5.4 Experiment 4: Effect of the number of active sources

Experiment 4 compared the performance of the GMM-based localization model to the baseline systems described in Section 2.4.1, where the effect of the number of active sources was also analyzed in terms of the localization performance. The localization results are presented in Fig. 2.7 for all evaluated systems as a function of the reverberation time T_{60}. Panels (A)-(D) show the individual results for all four acoustic scenarios. Furthermore, the percentage of anomalies and the standard deviation of the correct estimates averaged over all four acoustic scenarios are shown in Fig. 2.8, respectively.

As expected, the average percentage of anomalies increased with reverberation time for all localization methods. For the single source scenario shown in panel (A) of Fig. 2.7, both broadband FFT-based models gave more accurate predictions than the gammatone-based GCC method, where the SCOTM weighting performed slightly better than the PHAT weighting. However, the performance of both broadband FFT-based methods, SCOTM and PHAT, significantly deteriorated, when the number of sources increased.

The cross-correlation analysis based on the auditory front-end, *GCC Gammatone*, outperformed the broadband FFT-based methods in all multi-source conditions, which implies that the frequency selective cue extraction effectively resolved the dominant localization information of multiple sources. Nevertheless, the percentage of anomalies increased quite rapidly with the reverberation time.

The localization error was significantly reduced, when the ITD cue was employed in combination with Gaussian mixtures. Although the same information is available as in the case of *GCC Gammatone*, the presence of reverberation affected the performance less due to the multi-conditional training and the probabilistic integration of source evidence across channels. In addition, localization performance was more robust in multi-source scenarios because the binaural cues modified by competing sources were also considered in the multi-conditional training phase.

As Fig. 2.7 clearly shows, the joint evaluation of ITD and ILD by Gaussian mixtures performed best in all acoustic scenarios, where the additional ILD information was especially important to cope with strong reverberation. Indeed the performance of the broadband FFT-based methods was strongly affected by the number of active sources, for which the GMM model using both ITD and ILD cues was almost independent.

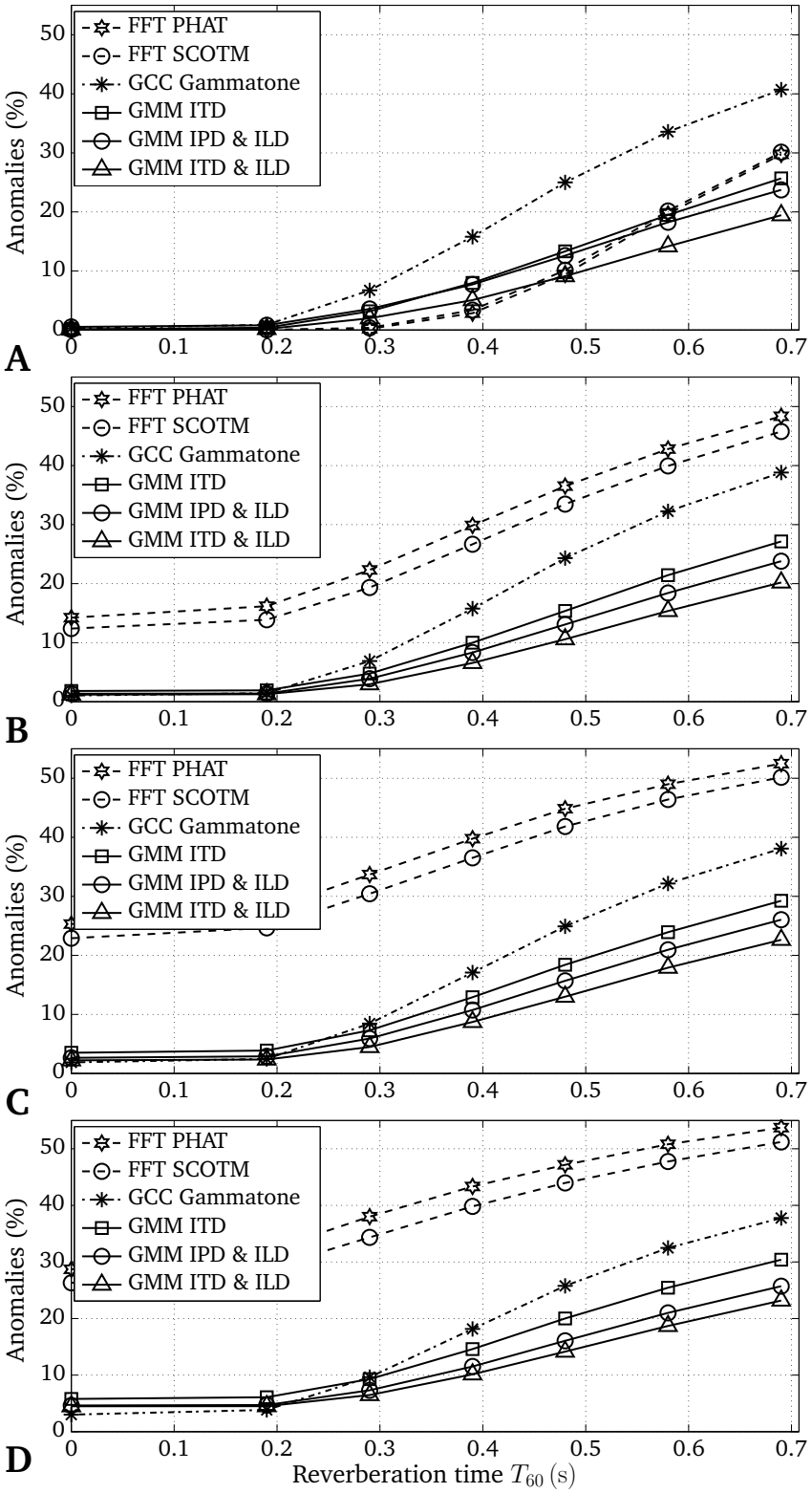


Figure 2.7: Experiment 4: Percentage of anomalies depending on reverberation time T_{60} of all baseline methods and GMM-based localization algorithms evaluated in four acoustic scenarios (A)-(D), consisting of 1, 2, 3 and 4 sources. Results are shown for all three categories of localization methods, namely the broadband FFT-based methods (dashed lines), the gammatone-based GCC (dash-dotted lines) and the GMM-based models (solid lines).

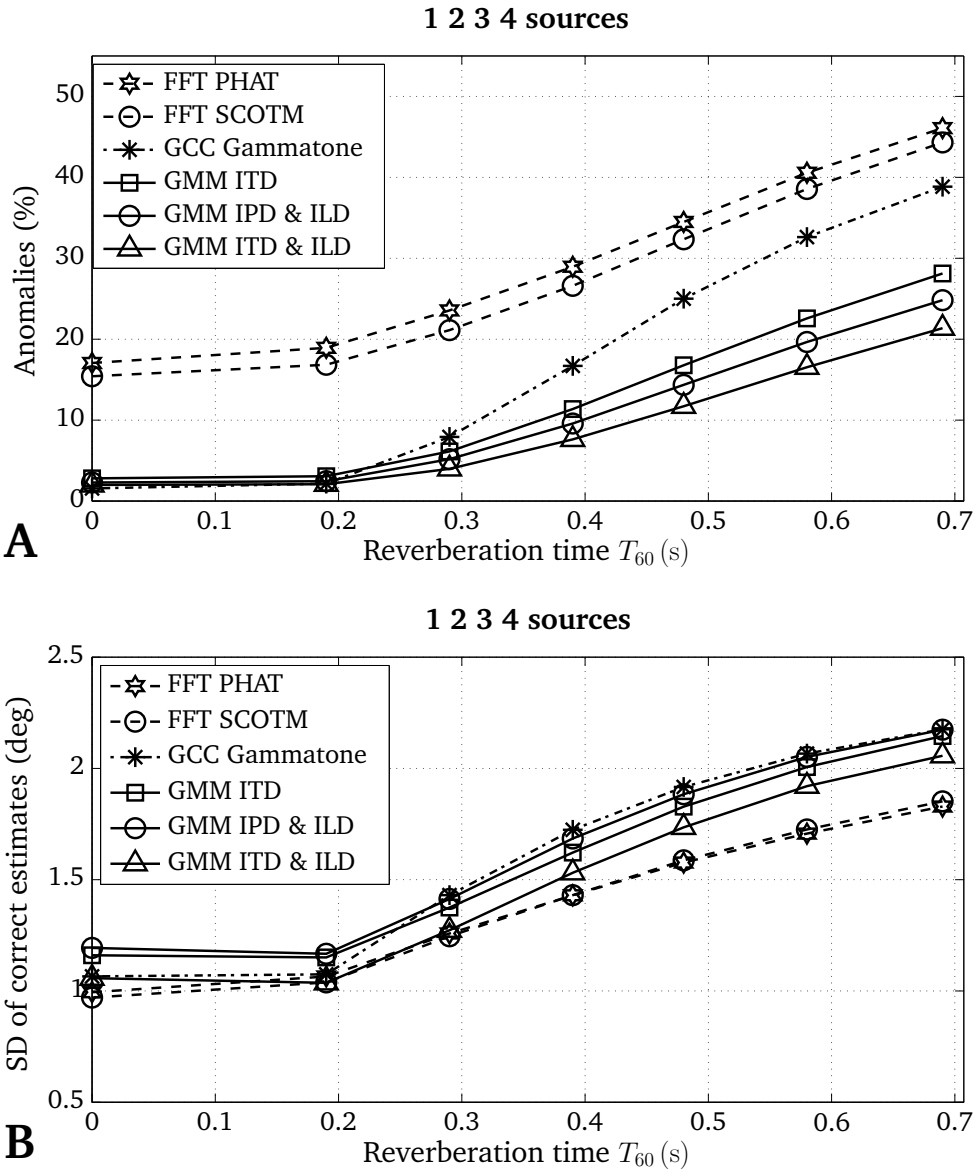


Figure 2.8: Experiment 4: (A) Percentage of anomalies and (B) the standard deviation of the correct localization estimates summarized across all 4 acoustic scenarios (1, 2, 3 and 4 sources). Results are shown as a function of the reverberation time T_{60} for all three categories of localization methods, namely the broadband FFT-based methods (dashed lines), the gammatone-based GCC (dash-dotted lines) and the GMM-based models (solid lines).

2.5.5 Experiment 5: Blind estimation of the number of acoustic sources

The number of active sources in an acoustic mixture can be a valuable information, which might be used for blind source separation algorithms, to control and steer the beams of a microphone array, or to post-process the frame-by-frame localization estimates. In Experiment 5, the capability of the proposed localization method to predict the number of active sources in an acoustic mixture was explored. Therefore, a histogram based on the frame-by-frame azimuth estimates is computed with a resolution of 5° by pooling the azimuth estimates of all frames over the entire acoustic mixture. After normalization, all local peaks in the histogram that fall below a predefined threshold θ_h are discarded. The remaining histogram peaks are assumed to be caused by active sound sources and are therefore selected as source candidates.

The histogram-based procedure is illustrated in Fig. 2.9 for an acoustic mixture consisting of three competing speech sources that are located at 35°, 5° and -30° in a reverberant room ($T_{60} = 0.69\,\mathrm{s}$). The left panel shows the three active speech sources and the frame-based azimuth estimates of the proposed localization model. The right panel presents the azimuth histogram. Whereas five local peaks are detected in the azimuth histogram, only the three sound source positions above the histogram threshold are considered as active sound sources.

The following performance measure is used to take into account errors which are caused by either selecting more or fewer source candidates than the true number of active sources. Let $r_\varphi = \{r_{\varphi_1}, ..., r_{\varphi_\mathcal{A}}\}$ be a set of $\mathcal{A}$ reference source positions which were present simultaneously in the acoustic mixture, and let $\hat{r}_\varphi = \left\{\hat{r}_{\varphi_1}, ..., \hat{r}_{\varphi_{\hat{\mathcal{A}}}}\right\}$ be a set of $\hat{\mathcal{A}}$ estimated source candidates. The percentage of correctly identified number of sources is given by

$$p_\mathrm{c} = 100 \cdot \frac{|r_\varphi \cap \hat{r}_\varphi|}{|r_\varphi \cup \hat{r}_\varphi|}. \tag{2.12}$$

The intersection of the reference source positions r_φ and the estimated source candidates $\hat{r}_\varphi$ is related to the union of the two. The operator $|.|$ represents the cardinality of a set, which is a measure of the number of elements of a set. In this way, performance decreases if more than the true number of sources are selected, although all reference positions might have been correctly identified.

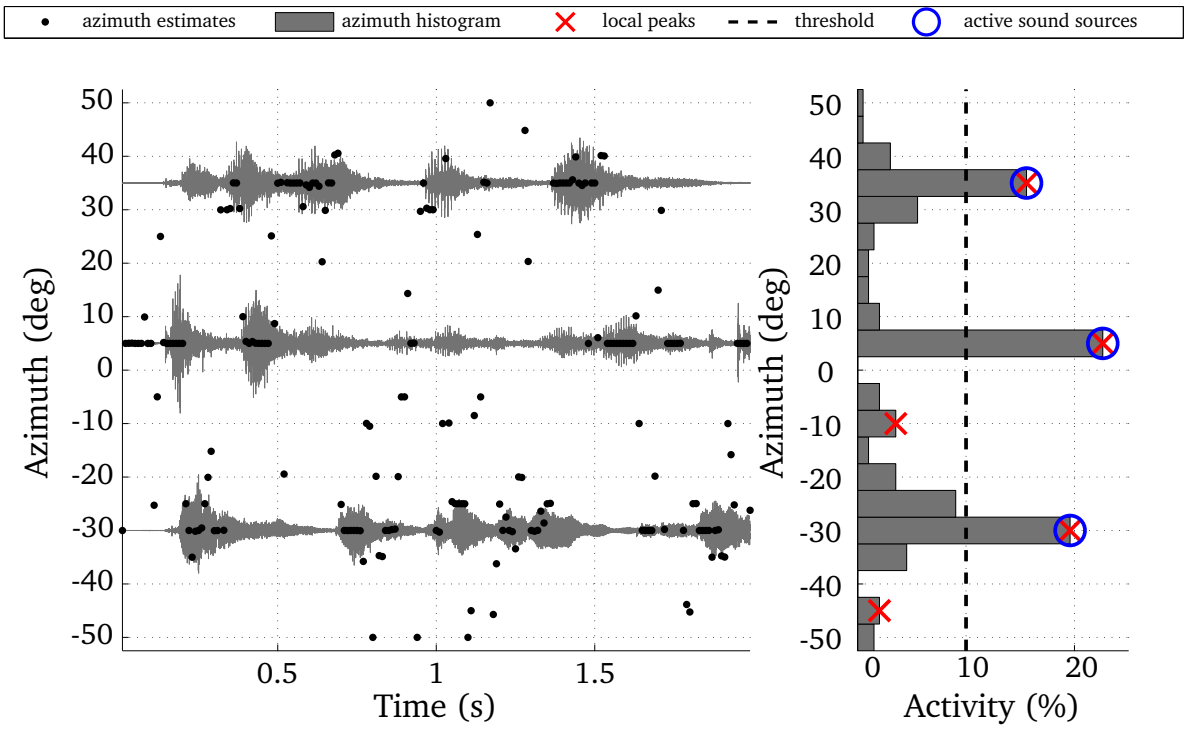


Figure 2.9: Histogram-based detection of three competing speech sources (35 °, 5 ° and −30 °) in reverberant conditions (T_{60} = 0.69 s) based on the frame-based azimuth estimates of the proposed localization model.

The performance was evaluated for various histogram thresholds θ_h, and results are reported for the threshold which gave the overall best result for each localization method independently. The percentage of correctly identified number of sources is shown in panel (A) of Fig. 2.10. The performance was averaged over four acoustic scenarios, containing between 1 and 4 sources and over all 8 evaluation positions. As already shown in experiment 4, the presence of multiple sources has a severe effect on the performance of the FFT-based localization methods. Even under anechoic conditions, the performance of estimating the number of active sources was below 85%. With increasing reverberation time, the performance decreased to about 55%. Again, the SCOTM filtering performed consistently better than the PHAT weighting. Up to a reverberation time of T_{60} = 0.2 s, the gammatone-based GCC method is slightly superior to the GMM-based models. But with stronger reverberation, the performance of the GCC-based method rapidly decreased to 53% at a reverberation time of T_{60} = 0.69 s. Due to the multi-conditional training of the Gaussian mixtures, localization performance is robust against the impact of reverberation. As a consequence, the GMM-based models led to a cleaner histogram of source positions, which allowed for a more reliable identification of the number of active sources. Even at a reverberation time of T_{60} = 0.69 s, about 86% of the multi-source mixtures were correctly classified using the *GMM ITD & ILD*. In panel (B) of Fig. 2.10, the performance

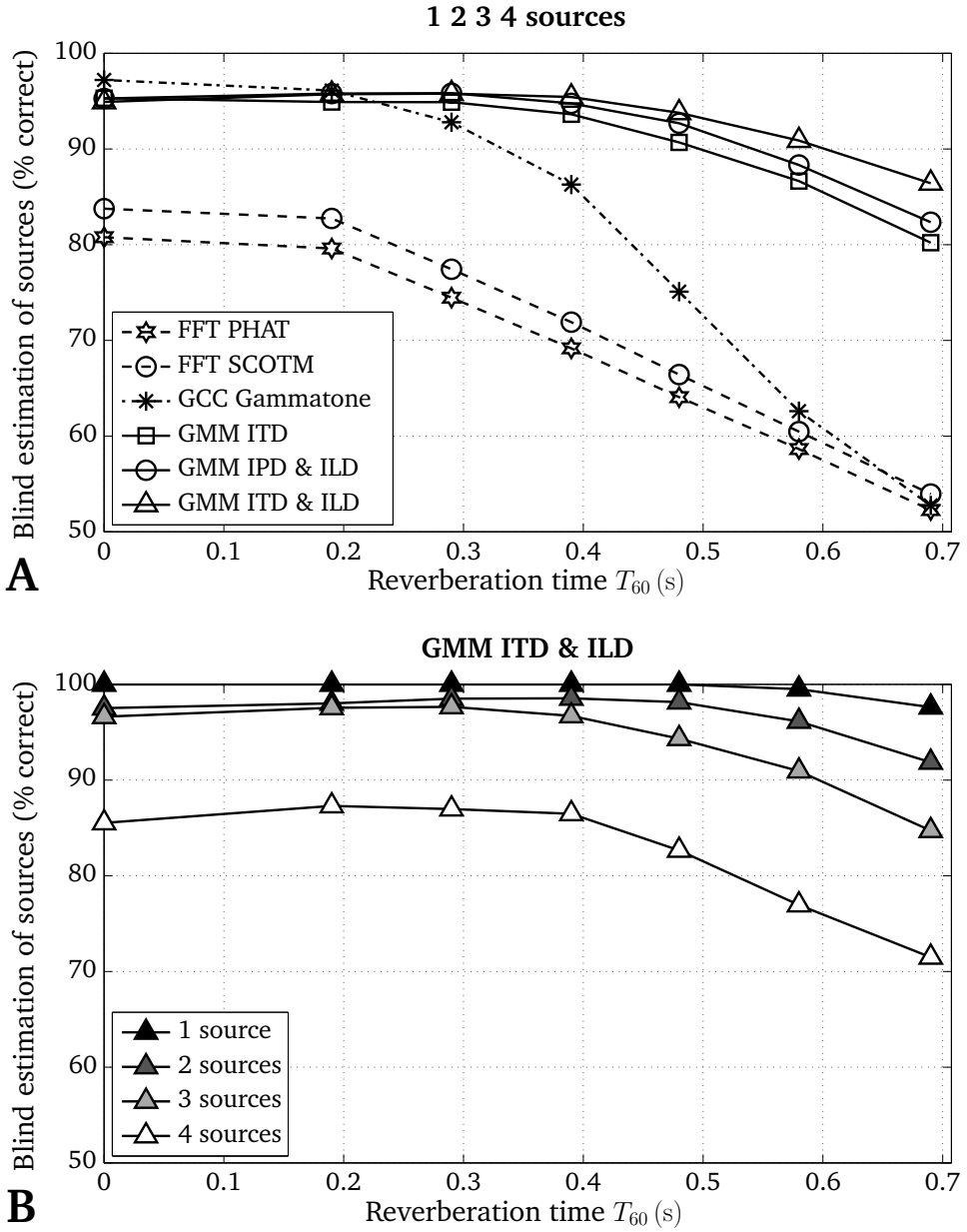


Figure 2.10: Experiment 5: Performance of estimating the number of active sources in mixtures consisting of 1, 2, 3 and 4 sources as a function of reverberation time T_{60}. In (A), the average performance is presented for all baseline methods and GMM-based localization algorithms. In (B), the dependence of performance on the number of sources is shown for the GMM-based evaluation of both ITD and ILD.

of *GMM ITD & ILD* is presented depending on the number of acoustic sources. Whereas classification was quite robust for 1, 2 and 3 sources, performance significanlty decreased for mixtures consisting of four sources.

2.6 Discussion and conclusions

A robust acoustic binaural localization model was presented, which is based on the supervised learning of azimuth-dependent binaural feature maps consisting of ITD and ILD. The model was evaluated in simulated adverse acoustic scenarios and outperformed binaural state-of-the-art localization techniques, especially in multi-source scenarios. Furthermore, the model was capable of generalizing to unknown source/receiver configurations which were not included in the training stage. Based on the frame-by-frame localization estimates, an efficient histogram technique allowed to robustly estimate the number of active sources in acoustic multi-source mixtures.

The robustness of the model against reverberation and the presence of multiple sources is attributed to three factors: First, due to the auditory front-end, the frequency-dependent, dominant localization information of multiple sources can be spectrally resolved, allowing for a robust estimation of the binaural cues. Second, GMMs are used to evaluate the joint binaural feature space and to accumulate evidence of possible source locations across frequency in a probabilistic way, taking into account the available information in an optimum way. Third, the multi-conditional training incorporates the uncertainty of the binaural cues caused by room effects, reverberation, the presence of multiple sources and changes in the source/receiver configurations.

It was shown, that integrating the probabilities of possible sound source locations across frequency is superior to accumulating the localization cues directly. Moreover, the joint evaluation of ITD and ILD disambiguates the information derived from the ITD cue, especially in strong reverberation, increasing the robustness of the localization model.

Although the concept of using either IPD or ITD might provide similar information, the localization performance using the ITD cue was superior to the IPD cue. This comparison is based on the assumption that the distribution of both ITD and IPD can be modeled equally well using a sum of Gaussian distributions. One possible explanation might be that the multiple clusters in the ITD feature space, which reflect the ambiguous ITD information, are warped to the interval between $\pm\pi$ in the IPD representation. Since the centered ITD clusters are more sharp and the more lateral clusters are more broad, this differentiated analysis is lost in the IPD representation, presumably decreasing the localization performance. Furthermore, it is worthwhile to note that the GMM localization model using a basic hair cell model outperformed the more detailed hair cell model including

higher order low-pass filtering. Thus, the probabilistic model is capable of exploiting the ITD in the fine structure at higher frequencies, which is generally accepted not to be accessible by the human auditory system (Klumpp and Eady, 1956, Zwislocki and Feldman, 1956). Thus, the model is not strictly limited by the processing which is believed to be performed by the human auditory system.

The broadband FFT localization (PHAT and SCOTM) is prone to errors in multi-source scenarios, because it actually averages the directional cues of all sources over a short time segment, which can lead to a phantom source that does not necessarily reflect the source position of the most dominant source. This is especially likely to happen if the energetic contribution of sources is equally strong and the sources are located symmetrically but in opposite directions with respect to the receiver (e.g. $-40\,^\circ$ and $40\,^\circ$). Furthermore, a noticeably longer analysis window than 20 ms is commonly used to increase the robustness in adverse conditions (Chen *et al.*, 2005). However, this is only reasonable for single-source localization.

The current analysis of the localization model was restricted to the frontal horizontal plane, whereas front-back discrimination and localization in the elevation domain are tasks for further investigations. Using the framework of Gaussian mixture models, the binaural feature space could be readily extended by additional descriptive features which are depending on the position of sound sources. For example, in order to extend the model to the elevation domain, the use of spectral cues might be beneficial (Zakarauskas and Cynader, 1993).

The radial distance between the source and the receiver, which determines the relation between the direct and the reverberated sound, was a sensitive parameter. Similar to the reverberation time, the radial distance is a source of uncertainties which modifies the distribution of the binaural cues. Localization performance significantly decreased at larger radial distances, which is in line with behavioral data observed for humans (Devore *et al.*, 2007). In order to improve the working range of the model, it might be beneficial to train the model either with binaural cues simulated at a larger radial distance or with binaural cues corresponding to various radial distances.

The reported localization performance was achieved by applying the probabilistic model on a frame-by-frame basis, accumulating evidence over frequency channels. However, accumulating evidence of sound source locations across a larger time span could further increase localization performance. Integrating the localization cues over patches across time and frequency, which are believed to belong to a single source was reported

to significantly improve ITD-based localization performance (Christensen *et al.*, 2007, 2008, 2009). Instead of integrating the localization cues directly, the proposed probabilistic localization model could combine likelihoods of sound source locations across patches.

Due to its robustness and high temporal resolution, the localization model presented in this study might be very suitable as a front-end for CASA algorithms that segregate and recognize sound sources in complex acoustic mixtures.

This chapter is based on:

- May *et al.* (2010): "The effect of spectro-temporal integration in a probabilistic model for robust acoustic localization," in *2nd International Symposium on Auditory and Audiological Research (ISAAR 2009)*, Helsingør, Denmark, pp. 125–134.

3

The Effect of Spectro-Temporal Integration in a Probabilistic Model for Robust Acoustic Localization

A robust acoustic localization model will be presented, which is based on the supervised learning of azimuth-dependent binaural feature maps consisting of interaural time differences (ITD) and interaural level differences (ILD). Motivated by the robust localization performance of the human auditory system, the associated peripheral stage is used in this study as a front-end for binaural cue extraction. Multi-conditional training is performed to take into account the variability of the binaural features which results from the combination of multiple sources, the effect of reverberation and changes in the source/receiver configuration. One way of accumulating evidence of possible sound source locations is to combine information across auditory channels. Alternatively, integrating evidence across groups of time-frequency (T-F) units, so-called fragments, which are believed to belong to a single source, was reported to significantly improve ITD-based localization performance (Christensen et al., 2007). Instead of accumulating the localization cue directly, the proposed model combines likelihoods, taking into account the uncertainty which is associated with the azimuth estimate of a particular T-F unit. Various procedures of controlling the spectro-temporal integration will be discussed and the influence on sound source localization will be presented.

3.1 Introduction

The human auditory system is able to identify and localize acoustic objects in adverse acoustic conditions. Regarding speech perception and localization tasks, the robustness of the human auditory system is superior to computer algorithms that have access to the information also available to the human auditory system (binaural signal). Bregman's auditory scene analysis (ASA) is an approach to describe how the human auditory system derives a description of complex acoustic scenes (Bregman, 1990). First, the auditory input is segregated into fragments, representing groups of T-F units which are dominated by a single source. In the second step, fragments which correspond to the same acoustic source are grouped together to form an acoustic description of the sources present in the scene. In order to perform this higher level analysis of complex acoustic scenes, a front-end is required to partition the T-F plane into groups of coherent T-F units, which are believed to be dominated by one source.

The segregation of multiple sources based on localization cues was shown to lead to high performance in anechoic conditions (Roman *et al.*, 2003). However, the accuracy of localization information based on individual T-F units rapidly decreases with increasing reverberation time. In order to increase the reliability of the localization estimate, information could be grouped and integrated across T-F units that are dominated by the same source. Recently, a two-stage model for exploiting spatial cues across groups of T-F units was proposed to improve frame-based localization in reverberation (Christensen *et al.*, 2007, 2008, 2009). First, a pitch tracking algorithm was employed to group T-F units together to so-called *fragments*, according to common pitch information. Those fragments of consistent pitch information were used in the second stage to integrate the ITD information across the corresponding spectro-temporal regions.

This chapter presents a model to create T-F-based localization maps of multi-source scenarios in reverberation, by integrating evidence of sound source direction across groups of T-F units. There is strong evidence that the formation of auditory objects is not primarily driven by spatial cues (Darwin, 2008). Nevertheless, it is interesting to explore whether in realistic acoustic settings with reverberation, there is enough localization information to derive fragments based on common spatial information. Compared to the pitch-based fragment generation, a method based on spatial cues would work for arbitrary acoustic signals (voiced/unvoiced/noise). In contrast to accumulating spatial cues as proposed by Christensen and colleagues, in this study the evidence will be accumulated by integrating likelihoods of sound source locations across fragments. In this way, the

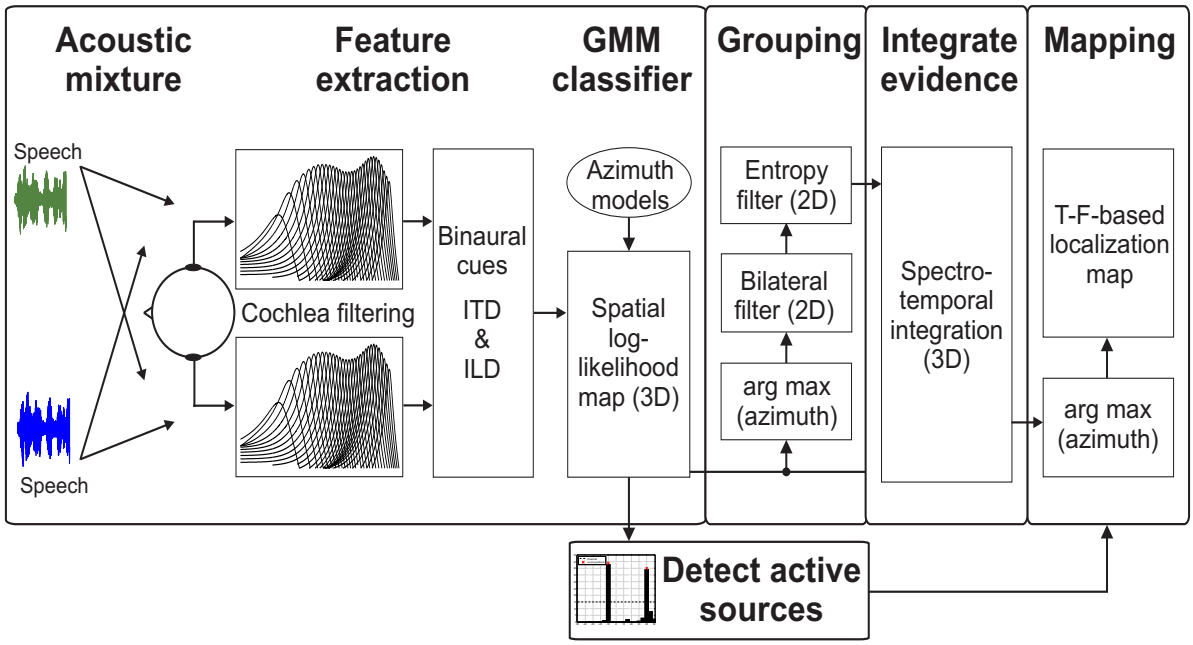


Figure 3.1: Model architecture for creating a T-F-based localization map. See Section 3.2 for details.

uncertainty of localization estimates of individual T-F units is taken into account. With respect to frame-based localization performance, it was shown in Chapter 2 that the integration of log-likelihoods is superior to integrating the normalized cross-correlation function across frequency.

The remainder of this chapter is structured as follows. Section 3.2 will explain the model architecture. The evaluation setup and the experimental results will be given in Section 3.3 and 3.4. Section 3.5 will conclude the chapter.

3.2 Model architecture

The model that we propose to derive the localization map consists of four major building blocks, which are shown in Fig. 3.1. The assumed input is a binaural signal consisting of several acoustic sources at fixed spatial locations. Given this binaural input, the proposed model first computes the log-likelihood of azimuth-dependent sound source activity for individual T-F units. Secondly, T-F units are grouped together to fragments based on consistency in localization information. Those fragments are used in the third block to accumulate log-likelihoods across the corresponding spectro-temporal regions of each fragment. Fourthly, the modified log-likelihood map is transformed into the final

localization map. Each of the four building blocks is explained in detail in the following sections. The description is supported by visualizing outputs of various processing steps involved in estimating the localization map of a two-source mixture in reverberation ($T_{60} = 0.39\,\mathrm{s}$).

3.2.1 Spatial log-likelihood map

In Chapter 2 we have presented a probabilistic model for sound source localization that can be used for the computation of a spatial log-likelihood map. In this model, the acoustic signal is decomposed into $F = 64$ frequency channels using a fourth-order gammatone filterbank. Spatial ITD and ILD cues are extracted independently for each frequency channel using a time window of 20 ms with a 10 ms frame shift. The joint distribution of both spatial cues resulting from multiple sources and reverberation is learned by a Gaussian mixture model (GMM) classifier for a range of $K = 21$ azimuth positions between $\pm 50\,^\circ$ in steps of $5\,^\circ$. The frequency- and azimuth-dependent GMMs are denoted by $\{\lambda_{f,\varphi_1}, \ldots, \lambda_{f,\varphi_K}\}$. Given the binaural feature vector $\vec{x}_{t,f}$ which consists of the estimated ITDs and ILDs, the model computes a three-dimensional spatial log-likelihood map $\mathcal{L}$, which represents the log-likelihood that the kth sound source direction is active at time frame t and frequency channel f:

$$\mathcal{L}(t, f, k) = \log p(\vec{x}_{t,f} | \lambda_{f,\varphi_k}), \tag{3.1}$$

The aim of the final localization map is to group and label the T-F representation of multi-source mixtures according to azimuth information of the active sources. In order to achieve this, the number and the azimuth of active sources is required. Figure 3.2 shows the estimated azimuth map before grouping and spectro-temporal integration for an acoustic mixture consisting of two male speakers in reverberation ($T_{60} = 0.39\,\mathrm{s}$). In panel (A), the most likely source position for individual T-F units φ^{TF} is presented

$$\hat{\varphi}^{\mathrm{TF}}(t, f) = \underset{1 \le k \le K}{\arg\max}\, \mathcal{L}(t, f, k). \tag{3.2}$$

Although dominant azimuth locations seem to be detectable by visual inspection, an automatic detection of active sources using the T-F-based localization information is a nontrivial task.

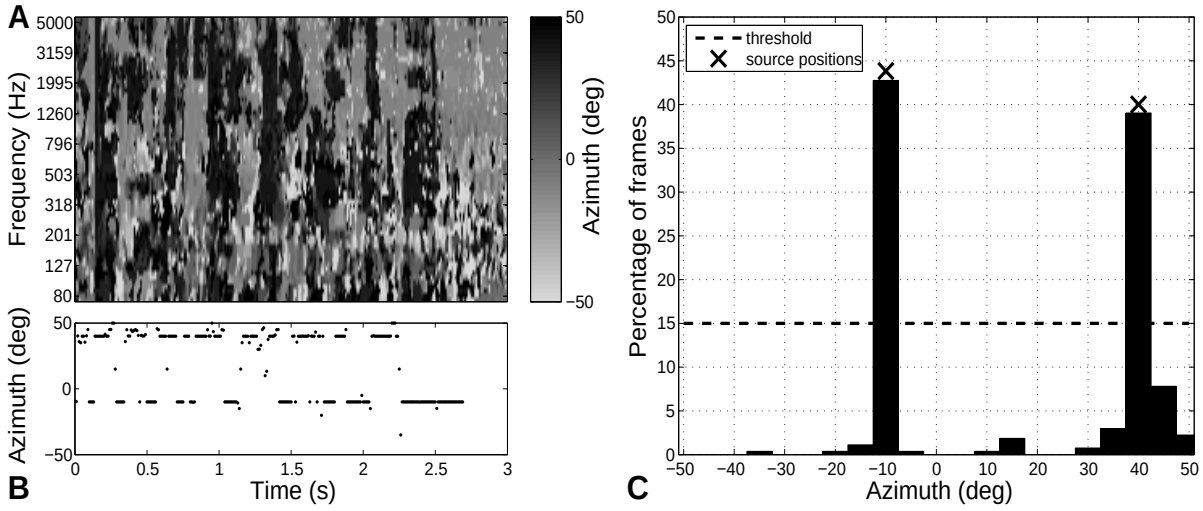


Figure 3.2: Estimated localization information for an acoustic mixture consisting of 2 male speakers (at 40° and −10° azimuth) in reverberant conditions ($T_{60} = 0.39\,\mathrm{s}$) based on: (A) individual T-F units and (B) time frames after integrating evidence across frequency channels. As illustrated in (C), a histogram of the frame-based localization estimates can be used to detect the number of active sources in the mixture and the corresponding azimuth positions.

Integrating the log-likelihood across frequency and estimating the most dominant sound source direction per frame $\hat{\varphi}^{\mathrm{T}}$ according to

$$\hat{\varphi}^{\mathrm{T}}(t) = \underset{1\leq k\leq K}{\arg\max} \sum_{f=1}^{F} \mathcal{L}(t, f, k) \tag{3.3}$$

leads to a good frame-by-frame indicator of the azimuth of the most dominant source (see panel (B)). Using these frame-based localization estimates, an efficient histogram technique can be utilized to detect active sound sources and the corresponding azimuth positions by applying a threshold criterion θ_h (see Section 2.5.5), as shown in panel (C). Note that the histogram analysis is performed on a sentence basis for each acoustic mixture.

3.2.2 Grouping of consistent localization information across T-F units

In this study, consistent localization information is grouped across T-F units leading to fragments. For this purpose, the 3D log-likelihood map is transformed into a preliminary 2D localization map by recursively smoothing the log-likelihood map $\mathcal{L}$ across frames and selecting the most likely source position for each T-F unit according to Eq.(3.2).

The recursive smoothing is intended to reduce fluctuations of the azimuth information across time

$$\tilde{\mathcal{L}}(t, f, k) = \alpha_{\text{Frag}}\, \tilde{\mathcal{L}}(t-1, f, k) + (1 - \alpha_{\text{Frag}})\, \mathcal{L}(t, f, k)\,, \tag{3.4}$$

where the smoothing constant is set to $\alpha_{\text{Frag}} = 0.7165$. This choice is adopted from Christensen *et al.* (2007) to match the average length of speech fragments in multi-source scenarios, which was reported to be in the range of 30 ms. The preliminary localization map is shown in panel (A) of Fig. 3.3.

In the next step, the estimated localization information is averaged across neighboring T-F units by applying a bilateral filter (Tomasi and Manduchi, 1998). The output of the bilateral filter is a weighted sum of the azimuth of neighboring T-F units, where the weighting function is data-dependent. The weighting function is a multiplication of two Gaussian functions. The first one decreases with the distance of neighboring T-F units to the center T-F units, putting more emphasis on the T-F units that are in the proximity of the center unit. The second weighting function decreases with increasing azimuth difference to the azimuth of the center T-F unit, using predominantly T-F units with similar azimuth values for the averaging. The latter weight provides the edge preserving property of the bilateral filter. The width of the Gaussian functions are specified by two parameters, σ_{TF} and σ_{φ}, controlling the decrease of the Gaussian weight in the T-F domain and in the azimuth domain, respectively. To ensure that the edge-preserving property of the bilateral filter is maintained for arbitrary source configurations, the decrease of the Gaussian weight along the azimuth domain σ_{φ} is adjusted by using the estimated source positions described in the previous section. The minimum spatial distance between all detected sources is taken as the change in azimuth which should be preserved by the bilateral filter. Therefore, σ_{φ} is changing linearly as a function of the minimum source distance. In this way, the smoothing of the bilateral filter is stronger if acoustic sources are spatially further apart, whereas less smoothing is applied if sources are spatially close to one another. The effect of the bilateral filter can be seen in Fig. 3.3, by comparing the preliminary localization map in panel (A) with the output of the bilateral filter in panel (B). Although groups of connected T-F units which correspond to either of the two source positions can be visually identified in panel (A), the azimuth within those areas is quite noisy and fluctuates around the true azimuth position. The localization information after bilateral filtering is consistent across large areas of T-F units and small azimuth deviations are smoothed, whereas boundaries between two different source positions are maintained.

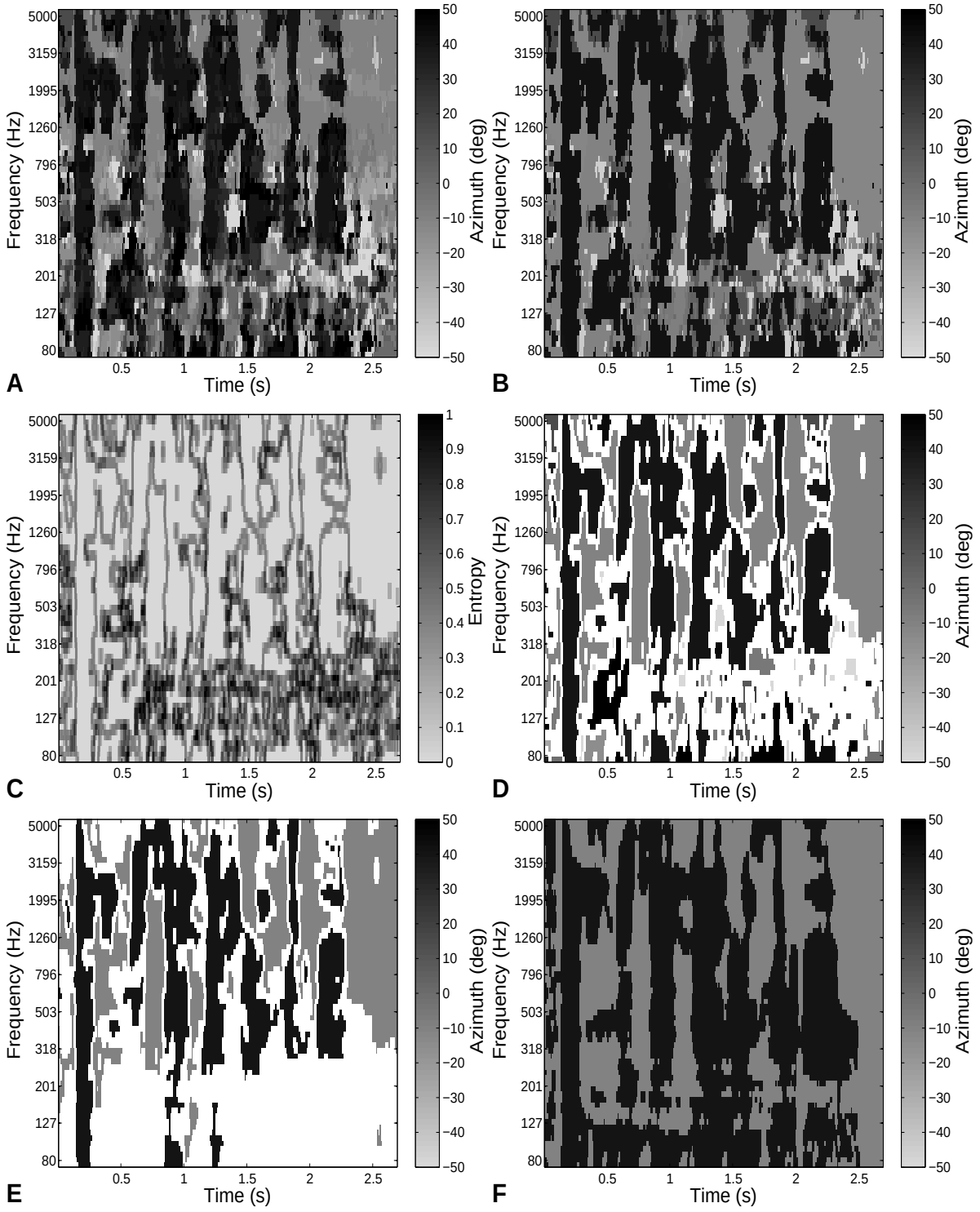


Figure 3.3: Process of creating a localization map of an acoustic mixture consisting of 2 male speakers (at 40° and −10° azimuth) in reverberant conditions ($T_{60} = 0.39\,\mathrm{s}$): (A) Preliminary localization map, (B) Localization map after bilateral filter, (C) Entropy map of filtered localization map, (D) Localized groups of T-F units, (E) Final localization map and (F) Ideal grouping based on the *a priori* SNR. White color indicates the background.

After bilateral filtering, fragments of consistent localization information are extracted to complete the grouping procedure. To achieve this, the variability of azimuth information is computed over a window of adjacent T-F units. More specifically, the variability is measured by calculating the entropy of the azimuth values across a window of 3×3 neighboring T-F units. Panel (C) of Fig. 3.3 shows the entropy map of the localization information presented in panel (B). The entropy map is close to zero if the azimuth is consistent across the analysis window (low variability) and approaches 1 if the azimuth is completely random. Hence, the entropy filter can be used to effectively group T-F units and to extract the boundaries between areas of consistent localization information. An experimentally determined threshold of $\theta_e = 0.3$ is used to transform the entropy map into a binary image, labeling consistent and inconsistent T-F units with 1 and 0, respectively. T-F units labeled with 0 are considered as background, and are therefore not considered for the fragment creation. T-F units labeled with 1 are grouped across time and frequency to obtain the final fragments. A pair of T-F units is considered to belong to the same fragment, if the two spectro-temporal units are connected either horizontally or vertically (4-neighbors connectivity).

3.2.3 Spectro-temporal integration

Based on the extracted fragments, the spectro-temporal integration accumulates evidence of source locations across all corresponding T-F units. For each fragment, the log-likelihoods corresponding to a specific azimuth are integrated to obtain an estimate of the likelihood that the source is at this specific azimuth. This 2D integration is done for all 21 azimuth positions of the log-likelihood map independently. Various weighting schemes are discussed in Christensen *et al.* (2009) to control the contribution of each T-F unit to the overall average of the fragment. But because the log-likelihood map already reflects the uncertainty which is associated with local T-F units, a uniform combination of log-likelihoods seems to be optimal.

3.2.4 Mapping

After performing spectro-temporal integration, a 2D localization map is formed by estimating the most likely source location for each fragment. The azimuth labeling is done only for those T-F units which do correspond to a fragment. The remaining T-F units are labeled as background. This process can be seen in panel (D) of Fig. 3.3. While the

estimated azimuth of most fragments correspond to either of the real sound sources at 40 ° and −10 °, some fragments in lower frequency channels point to different azimuth positions.

After assigning a localization label for each fragment in the T-F plane, a post-processing is performed to remove T-F units from the localization map which do not correspond to one of the active sound source positions. The detection of source activity is based on the histogram technique described in Section 3.2.1. The final localization map after post-processing is depicted in panel (E). For comparison, the ideal localization map based on the *a priori* signal-to-noise ratio (SNR) between both sources is presented in panel (F).

3.3 Evaluation setup

3.3.1 Acoustic conditions

The performance of the proposed system was evaluated in simulated two-source scenarios. Binaural mixtures were generated by convolving anechoic speech files with simulated binaural room impulse responses (BRIR). The speech files were randomly selected from the TIMIT database (Garofolo *et al.*, 1993). The BRIRs were synthesized for a room of dimensions 5.1 x 7.1 x 3 meter by using the room simulation package (Campbell *et al.*, 2005). The preset *Acoustic plaster* was selected in the software package to model a frequency-dependent absorption characteristic of the room. Different sets of absorption coefficients were used to realize mild-to-strong reverberation. Each binaural mixture consisted of two, fully overlapping speech files of different speakers which were mixed at an SNR of 0 dB, as defined after spatialization. The probabilistic model was trained to recognize locations between ±50 ° in steps of 5 °, resulting in 21 possible source positions. 21 mixtures consisting of two sources were created by randomly selecting all 21 source positions, but the minimum sources distance was constrained to be at least 10 °. All 21 mixtures were created and presented five times using different sentences, leading to a total of 105 mixtures for each reverberation condition.

3.3.2 Performance evaluation

The accuracy of the estimated localization map is evaluated by comparing it to the ideal binary mask (IBM) (Wang, 2005), which represents the ideal grouping of T-F units based on the a priori SNR of the target and interfering source. Because this work focuses on competing sources rather then separating a target from interfering sources, the IBM was modified to take the azimuth of the dominant source per T-F unit. A T-F unit is considered to be correct, if the estimated azimuth corresponds to the azimuth of the IBM. Thus, the percentage of localization errors is computed by dividing the number of correctly identified T-F units by the number of T-F units in the estimated localization map.

The estimated localization map is not fully occupied, but only contains information which is believed to belong to one of the detected sources. The remaining T-F units are labeled as background. Hence, an additional metric is used to measure the amount of available information, expressed in percentage of T-F units. Ideally, the percentage of errors should be minimized while maximizing the percentage of available information. But there is an interdependency between both measures. At a certain amount of information, a further increase is only possible at the cost of increasing the localization errors.

3.3.3 Algorithms

The performance of the proposed method is compared against two baseline systems. The first system *T-F units* produced localization maps on the basis of individual T-F units without any grouping and integration according to Eq.(3.2). The second system *Leaky* is based on the preliminary localization map which is depicted in panel (A) of Fig. 3.3. Compared to *T-F units*, it explores the time-dependent development of sound source likelihoods by applying a leaky integrator across frames. The third system *Fragments* used the proposed spectro-temporal integration by accumulating evidence across fragments (see panel (D) of Fig. 3.3). Note that the post-processing described in the mapping section is applied to the localization maps of all methods.

3.4 Experimental results

Panel (A) of Fig. 3.4 shows the error rate of the estimated localization maps for acoustic mixtures consisting of two sources in reverberation. The first method *T-F units*, which computed localization maps based on individual T-F units, achieved an error rate of 5.7% in anechoic conditions. However, performance rapidly deteriorated with reverberation time, showing that the estimated localization information based on individual T-F units is not robust in adverse acoustic conditions. Although significantly lower error rates were accomplished by the *Leaky* system, the general dependency of localization accuracy on reverberation time remained. The proposed method *Fragments*, which performed spectro-temporal integration based on fragments, outperformed both baseline systems in all reverberation conditions. Furthermore, the localization error per T-F unit was almost independent of the reverberation time, ranging from 4.5% in anechoic conditions to 5.5% in strong reverberation ($T_{60} = 0.69\,\mathrm{s}$). This demonstrates the benefit of grouping and integrating localization information across groups of T-F units.

The percentage of available localization information is presented in panel (B) of Fig. 3.4. With increasing reverberation time, the estimated localization information becomes more noisy and therefore, the percentage of T-F units which correspond to one of the detected source positions decreases. However, especially for reverberation times below $T_{60} = 0.4\,\mathrm{s}$, the localization maps produced by the *Fragments* system contained significantly more information compared to the two baseline systems.

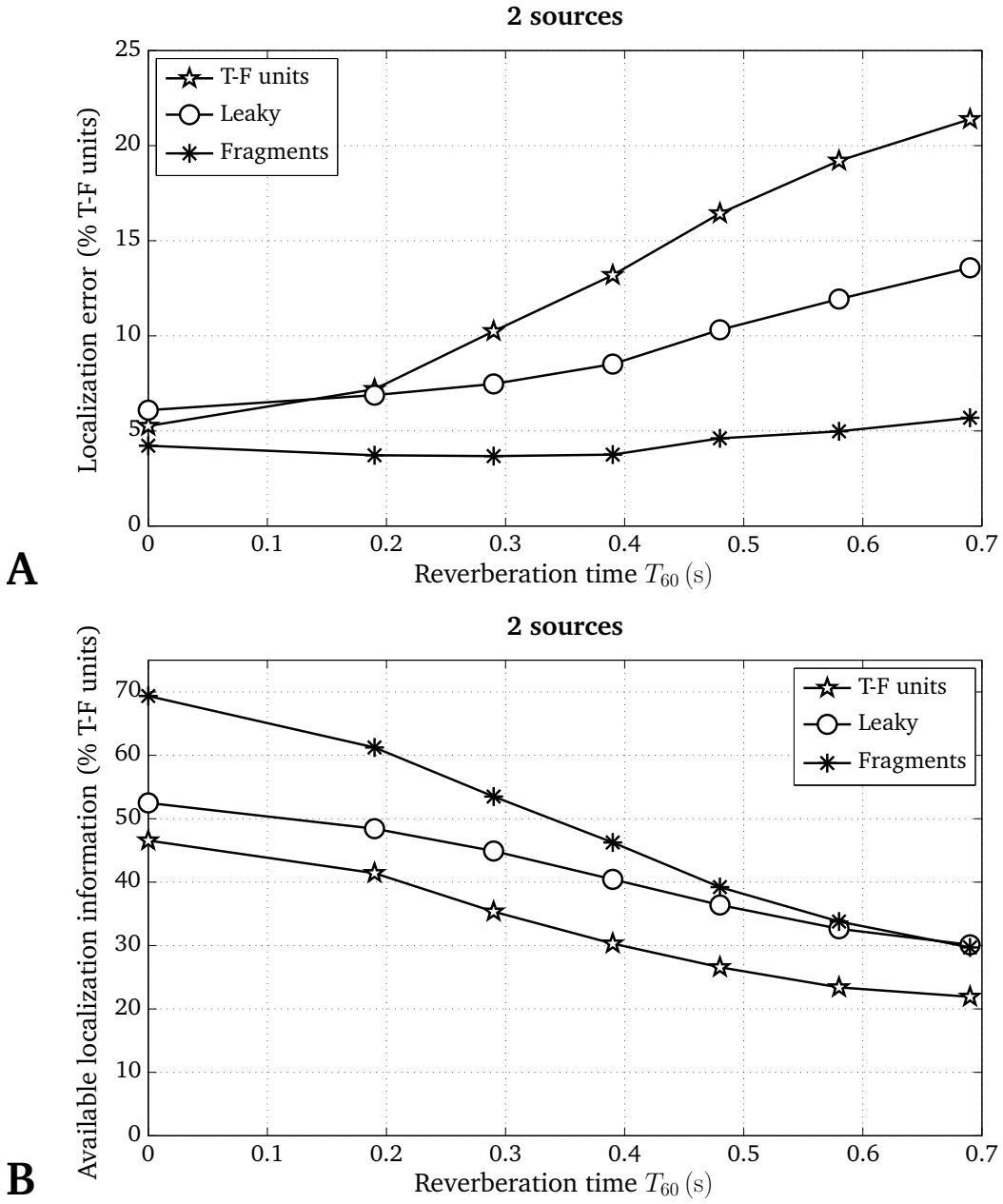


Figure 3.4: The accuracy of the estimated localization maps for two-source mixtures, expressed in (A) percentage of localization errors and (B) percentage of available localization information per T-F unit.

3.5 Conclusions

This chapter presented a method to create T-F-based localization maps for multi-source mixtures. By accumulating sound source evidence across groups of T-F units which are believed to be dominated by one source, the estimated localization information was shown to be robust in simulated, adverse acoustic conditions. The new method for extracting fragments which was proposed, is based on common spatial information. Compared to

pitch-based grouping, no explicit assumption about the sound source is required, which makes the method generally applicable to arbitrary acoustic scenarios. Future work will investigate the use of primitive grouping principles, e.g. onset and offset analysis (Hu and Wang, 2007), to further enhance the fragment generation process. In addition, the estimated localization map will be used as a front-end for higher order analysis of complex acoustic scenes.

This chapter is based on:

- May *et al.* (2012a): "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *IEEE Transactions on Audio, Speech, and Language Processing* **20**(1), pp. 108–121.

4

Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling

Although the field of automatic speaker recognition (ASR) has been the subject of extensive research over the past decades, the lack of robustness against background noise has remained a major challenge. This chapter describes a noise-robust speaker recognition system that combines missing data (MD) recognition with the adaptation of speaker models using a universal background model (UBM). For MD recognition, the identification of reliable and unreliable feature components is required. For this purpose, the signal-to-noise ratio (SNR) based mask estimation performance of various state-of-the art noise estimation techniques and noise reduction schemes is compared. Speaker recognition experiments show that the usage of a UBM in combination with missing data recognition yields substantial improvements in recognition performance, especially in the presence of highly non-stationary background noise at low SNRs.

4.1 Introduction

Whereas single speaker recognition can be performed quite robustly in clean acoustic conditions, the recognition performance severely degrades in the presence of background noise (Lippmann, 1997). The reduced performance is caused by the mismatch between the features which have been learned by the classifier under clean acoustic conditions and the features which are observed in adverse acoustic scenarios. Common approaches to reduce this mismatch are feature compensation methods such as cepstral mean normalization (CMN) (Atal, 1974) and relative spectral (RASTA) processing (Hermansky and Morgan, 1994).

In contrast to compensating for the effect of environmental noise, a major step towards noise-robust speaker recognition is to modify the structure of the recognizer such that it only considers feature components which are believed to contain reliable information about the target signal. Considering a two-dimensional time-frequency (T-F) representation of a noisy speech signal, some T-F components will be dominated by the target signal, whereas other regions will be contaminated by background noise. In missing data (MD) recognition, the classification is performed only on that part of the observed spectro-temporal feature space that is believed to be reliable (Cooke *et al.*, 2001). Two different approaches exist to deal with missing features. *Imputation* refers to the technique of replacing missing features with an estimate of the feature value, whereas *marginalization* basically ignores missing features. Marginalization has been shown to be superior to imputation in the context of speech recognition (Cooke *et al.*, 2001). In order to perform marginalization, a mask is constructed which classifies the feature space into reliable and unreliable components. Like the spectral feature space, the mask is defined as a function of time and frequency. In the second step, the classification is solely based on the reliable feature components whereas the unreliable components are assumed to be masked by the background noise.

One major drawback of the MD framework is that it needs to be based on spectral features. This limitation is caused by the required correspondence between the mask and individual feature components. In contrast to spectral features, mel frequency cepstral coefficients (MFCCs) are orthogonalized by the discrete cosine transform (DCT) and therefore can be used for a more compact feature representation (Davis and Mermelstein, 1980), where each cepstral feature is representing properties of the global spectral shape. As a result, cepstral features are more accurately modeled by recognizers which commonly assume independence of feature components (e.g., diagonal Gaussian mixture models).

This may account for the better accuracy of cepstral-based recognition systems under clean acoustic conditions compared to MD-based recognition systems (Cooke *et al.*, 2001, Palomäki *et al.*, 2004a). It is also interesting to note that in the context of automatic speech recognition, the performance of an MD-based speech recognizer significantly decreases as the vocabulary size increases (Srinivasan *et al.*, 2006), which is generally less problematic for cepstral-based recognition systems. Possibly a similar effect may occur for speaker recognition, where the number of speakers which can be discriminated using the spectral feature representation may be limited due to the covariance between the feature components, which is not effectively modeled by a diagonal covariance matrix. Thus, whereas MD systems provide considerable advantages over cepstral-based techniques in terms of noise robustness, MD systems are limited by their inherent dependence on spectral features and therefore, a proper modeling of the speaker-dependent characteristics becomes especially important for MD-based recognition systems in order to provide a substantial benefit over conventional MFCC-based recognizers.

Many state-of-the art speaker recognition systems approximate the speaker-dependent distribution of features by Gaussian mixture models (GMMs) (Reynolds and Rose, 1995). GMM-based speaker models are predominantly used for MD-based speaker recognition (Drygajlo and El-Maliki, 1998a,b, Shao and Wang, 2003, 2006, Kühne *et al.*, 2008, Pullella *et al.*, 2008). In cepstral-based speaker recognition systems, the usage of a universal background model (UBM) in combination with GMMs is well established and was shown to outperform GMM-based speaker recognition (Reynolds *et al.*, 2000). A UBM represents the speaker-independent distribution of features and speaker models are obtained by adapting the well-trained UBM parameters to the speaker-dependent speech material. Despite its superior performance, the possible benefit of using a UBM in combination with MD-based speaker recognition has not been investigated.

In this chapter, we combine the UBM-based adaptation of speaker models with missing data recognition. It is expected that the representation of spectral features can be substantially improved by using a UBM model, especially for recognizing speakers for which there is a limited set of training material. Because the UBM is trained on the pooled speech material of many speakers, it is possible to significantly increase the number of Gaussian mixture components and therefore, develop a more precise model of the feature distribution without the risk of over-training. In order to show the potential benefit of combining missing data with the UBM-based adaptation of speaker models, a missing data mask is required that indicates whether a feature component is reliable or missing. Because the estimated mask is the most critical component in missing data systems that limits the

overall recognition performance, an extensive comparison of methods for estimating the missing data mask based on the local signal-to-noise ratio (SNR) is performed. Therefore, various strategies for obtaining an estimation of the noise and the clean speech spectrum are systematically compared and evaluated in non-stationary noise conditions. In this way, the best performing method for the GMM-based missing data recognizer is found and will serve as a baseline for the newly proposed approach.

The remainder of the chapter is organized as follows. Section 4.2 gives an overview about missing data classification, the adaptation of UBM-based speaker models and discusses various ways to derive an estimate of reliable feature components based on a local SNR criterion. Section 4.3 outlines the evaluation procedure and the baseline system. In Section 4.4, speaker recognition experiments are conducted to analyze the benefit of using a UBM in combination with MD recognition and to evaluate various mask estimation procedures. Section 4.5 summarizes the main findings and concludes the chapter.

4.2 Automatic speaker recognition system

In this section, the missing data-based speaker recognition system and the adaptation of speaker models from a universal background model (UBM) are described. Furthermore, various methods will be presented for deriving the required missing data mask based on the spectral estimation of the noise and speech components.

4.2.1 Missing data recognition using adapted Gaussian mixture models

Gaussian mixture models are used to approximate the probability distribution of the D-dimensional feature vector $\vec{x}$ for the task of speaker recognition. Assuming $\mathcal{V}$ diagonal Gaussian mixture components, the probability density function (PDF) of a GMM is given by (Reynolds and Rose, 1995)

$$p(\vec{x}|\lambda) = \sum_{j=1}^{\mathcal{V}} w_j \prod_{d=1}^{D} \mathcal{N}(x_d, \mu_{j,d}, \sigma_{j,d}^2), \tag{4.1}$$

where w_j is the component weight and $\mathcal{N}(x_d, \mu_{j,d}, \sigma^2_{j,d})$ is a uni-variate Gaussian distribution with mean $\mu_{j,d}$ and variance $\sigma^2_{j,d}$

$$\mathcal{N}(x_d, \mu_{j,d}, \sigma^2_{j,d}) = \frac{1}{\sqrt{2\pi\sigma^2_{j,d}}} \exp\left(-\frac{(x_d - \mu_{j,d})^2}{2\sigma^2_{j,d}}\right). \tag{4.2}$$

The model for each specific speaker can be summarized by the following set of parameters

$$\lambda = \left(w_j, \vec{\mu}_j, \vec{\sigma}^2_j\right) \quad \forall \quad j = 1, \ldots, \mathcal{V}. \tag{4.3}$$

In missing data recognition, the feature vector $\vec{x}$ is split into two sub-vectors, according to reliable $\mathcal{R}$ and unreliable $\mathcal{U}$ components, and both are treated differently during the classification process. The evidence of the reliable feature components is directly used to estimate the likelihood of the speaker identity λ. Although the unreliable components are assumed to be dominated by additive noise, they do contain information about the maximum energy of the target speech component. The assumption that the unreliable feature components are bounded between zero and the observed spectral energy values x_u is exploited by bounded marginalization (Cooke *et al.*, 2001), where the average likelihood is computed across the range of all possible levels that the unreliable components might have had (also called *counter-evidence*)

$$p(\vec{x}|\lambda) = \sum_{j=1}^{\mathcal{V}} w_j \prod_{r\in\mathcal{R}} \mathcal{N}(x_r, \mu_{j,r}, \sigma^2_{j,r}) \times \underbrace{\prod_{u\in\mathcal{U}} \frac{1}{x_{\text{high},u} - x_{\text{low},u}} \int_{x_{\text{low},u}}^{x_{\text{high},u}} \mathcal{N}(x_u, \mu_{j,u}, \sigma^2_{j,u}) dx_u}_{\text{counter-evidence}}. \tag{4.4}$$

The integral in Eq. (4.4) can be evaluated as the vector difference of error functions (Cooke *et al.*, 2001), and Eq. (4.4) can be rewritten as

$$p(\vec{x}|\lambda) = \sum_{j=1}^{\mathcal{V}} w_j \prod_{r\in\mathcal{R}} \mathcal{N}(x_r, \mu_{j,r}, \sigma^2_{j,r}) \times \prod_{u\in\mathcal{U}} \frac{1}{x_{\text{high},u} - x_{\text{low},u}} \frac{1}{2} \left[\text{erf}\left(\frac{x_{\text{high},u} - \mu_{j,u}}{\sqrt{2\sigma^2_{j,u}}}\right) - \text{erf}\left(\frac{x_{\text{low},u} - \mu_{j,u}}{\sqrt{2\sigma^2_{j,u}}}\right)\right]. \tag{4.5}$$

The bounds were set to $[x_{\text{low},u}, x_{\text{high},u}] = [0, x_u]$.

The speaker-dependent set of GMM parameters λ listed in Eq. (4.3) is commonly initialized by k-means clustering (Lloyd, 1982) and further refined using the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). The objective of selecting the number of Gaussian components $\mathcal{V}$ is to find the minimum model complexity which is required to accurately model the characteristics of all speakers (Reynolds and Rose, 1995). As discussed in the introduction, instead of estimating the GMM parameters for each speaker independently, a speaker-independent universal background model (UBM) is used, which is trained on the pooled speech material of many speakers using k-means clustering and the EM algorithm (Reynolds *et al.*, 2000). A speaker-dependent model is derived by adapting the well-trained UBM parameters to the speech material of the corresponding speaker using maximum a posteriori (MAP) estimation. During the adaptation process, only those Gaussian components of the UBM are adapted, which show sufficient probabilistic alignment with the speaker-dependent speech material. In this way, the parameters of Gaussian components which are potentially under-represented are not updated to the new data, making the model adaptation robust even to a small amount of training data (Reynolds *et al.*, 2000). The MAP adaptation was shown to outperform the estimation of GMM parameters using the maximum likelihood (ML) approach (Reynolds *et al.*, 2000).

4.2.2 Spectral features

Spectral features are computed using the short-time Fourier transform (STFT) on a frame-by-frame basis. First, the input signal $x(i)$ is processed with a first-order pre-emphasis filter using a coefficient of 0.97 in order to enhance the spectral representation of high frequencies. This is a standard pre-processing technique for computing mel frequency cepstral coefficients, which is typically not applied if spectral features are extracted. However, it was found to be beneficial for missing data recognition in pilot experiments. Then, the signal is transformed into overlapping segments and the B-point STFT is computed as

$$X(t,\omega) = \sum_{b=0}^{B-1} w_{\text{H}}(b)\, x(t\,O + b) \exp\left(\frac{-j2\pi\omega b}{B}\right), \tag{4.6}$$

where t indexes the frame number, ω represents the frequency bin index corresponding to the frequency $f_\omega = \omega f_{\text{s}}/B$, f_{s} specifies the sampling frequency, w_{H} is a Hamming window function and O determines the frame shift in samples. To reduce the number of spectral

components, the spectrum $X(t,\omega)$ is passed through an auditory filterbank that resembles the frequency resolution of the human auditory system, resulting in an auditory power spectrum

$$X_{\mathrm{FB}}^2(t,f) = \sum_{\omega=0}^{B-1} |h_{\mathrm{FB}}(\omega,f) \cdot X(t,\omega)|^2 \tag{4.7}$$

for $f = 1, 2, ..., F$, where $F = 32$ is the number of auditory filters and $h_{\mathrm{FB}}(\omega,f)$ is a matrix containing the frequency-dependent auditory filter weights. The set of center frequencies $f_c = \{f_c(1), \ldots, f_c(F)\}$ of the auditory filterbank are equally distributed on the equivalent rectangular bandwidth (ERB) scale (Glasberg and Moore, 1990) using a spacing of 1 ERB between 80 Hz and 5000 Hz. The set of triangular auditory filter weights is computed as

$$h_{\mathrm{FB}}(\omega,f) = \begin{cases} 0 & \text{for } f_\omega < f_c(f-1) \\ \frac{f_\omega - f_c(f-1)}{f_c(f) - f_c(f-1)} & \text{for } f_c(f-1) \le f_\omega < f_c(f) \\ \frac{f_\omega - f_c(f+1)}{f_c(f) - f_c(f+1)} & \text{for } f_c(f) \le f_\omega < f_c(f+1) \\ 0 & \text{for } f_\omega \ge f_c(f+1). \end{cases} \tag{4.8}$$

In the following, the two-dimensional, time- and frequency-dependent auditory power spectrum will be referred to as T-F representation. Finally, the auditory power spectrum is loudness compressed by raising it to the power of 0.33 to obtain the spectral features which are used for recognition.

4.2.3 Mask estimation

In order to perform missing data classification, a mask is required which classifies the T-F representation into reliable and unreliable components. The underlying concept of the mask is that T-F units are assumed to be reliable if they are dominated by the target source, whereas the unreliable T-F units are considered to be dominated by interfering noise. This formulation implies that the local signal-to-noise ratio (SNR) is known for individual T-F components. To establish an upper performance limit, it is common to employ an ideal binary mask (IBM), which assumes *a priori* knowledge about the local SNR (Cooke *et al.*, 2001). It was shown that such an ideal binary mask yields excellent speech recognition performance (Cooke *et al.*, 2001) and can significantly increase speech intelligibility in multi-talker scenarios (Brungart *et al.*, 2006). Therefore, the estimation of the IBM was suggested to be the main goal of computational auditory scene analysis (CASA) (Wang, 2005). Various strategies have been proposed to estimate the IBM

based on auditory grouping principles (Hu and Wang, 2004, 2007, Ma *et al.*, 2007), binaural interaction (Roman *et al.*, 2003, Palomäki *et al.*, 2004b, Harding *et al.*, 2006) and assessing the local SNR (Drygajlo and El-Maliki, 1998a, Vizinho *et al.*, 1999, Cooke *et al.*, 2001). It is, however, outside the scope of this chapter to analyze and compare all existing approaches. Because the focus of the present study is to robustly identify speakers in the presence of noise, the mask $\mathcal{M}(t,f)$ is determined by estimating the local SNR in individual T-F units. A local SNR criterion (Cooke *et al.*, 2001) of LC = 0 dB is applied to decide whether a T-F unit is reliable

$$\mathcal{M}(t,f) = \begin{cases} 1 & \text{if } 10\log_{10}\frac{\hat{S}^2_{\text{FB}}(t,f)}{\hat{N}^2_{\text{FB}}(t,f)} > \text{LC} \\ 0 & \text{otherwise.} \end{cases} \tag{4.9}$$

The local SNR is obtained by comparing the estimated auditory power spectrum of speech $\hat{S}^2_{\text{FB}}(t,f)$ to the estimated auditory power spectrum of noise $\hat{N}^2_{\text{FB}}(t,f)$ in individual T-F units. The estimation of speech and noise components is carried out in the spectral domain before applying the auditory filterbank. This was shown to be superior to performing the spectral estimation within auditory bands (Flores and Young, 1994). After estimating the spectral magnitude of speech and noise, both spectra are transformed to the auditory domain in analogy to Eq. (4.7) :

$$\hat{S}^2_{\text{FB}}(t,f) = \sum_{\omega=0}^{B-1} \left| h_{\text{FB}}(\omega,f) \cdot \hat{S}(t,\omega) \right|^2 \tag{4.10}$$

$$\hat{N}^2_{\text{FB}}(t,f) = \sum_{\omega=0}^{B-1} \left| h_{\text{FB}}(\omega,f) \cdot \hat{N}(t,\omega) \right|^2 . \tag{4.11}$$

Because neither the noise nor the speech spectrum is generally known *a priori*, they need to be estimated based on the noisy signal spectrum $X(t,\omega)$. A plethora of methods exist to perform this task and the choice of both the noise estimation technique and the method to obtain an estimate of the clean speech spectrum can potentially influence the quality of the estimated binary mask. However, only a few studies have investigated the effect of some basic noise estimation techniques on MD recognition performance (Vizinho *et al.*, 1999, Ris and Dupont, 2001). Since the estimated binary mask is the most critical component in missing data recognition systems, we will describe the most important

1 The notation of parameters corresponds to the notation in the corresponding reference. Therefore, the same parameter may have a different meaning across references.

Table 4.1: Evaluated noise estimation techniques.

Method	Parameters[1]	Reference
Initial50ms	average spectrum	Vizinho *et al.* (1999)
Hirsch95	$\alpha = 0.9$, $\beta = 2.5$	Hirsch and Ehrlicher (1995)
Doblinger95	$\alpha = 0.8$, $\beta = 0.96$, $\gamma = 0.998$	Doblinger (1995)
Cohen02	same as in reference	Cohen and Berdugo (2002)
Lin03	$\alpha_{\text{final}} = 0.9$, $Q = 2$	Lin *et al.* (2003)
Lin03Mod	$\alpha_{\text{final}} = 0.9$, $Q = 2$, $\eta = 0.6$	Lin *et al.* (2003)
Martin06	same as in reference	Martin (2006)
Rangachari06	same as in reference	Rangachari and Loizou (2006)

methods to derive an estimate of the noise and the speech spectrum and their influence on the estimated binary mask will be assessed and systematically evaluated in terms of speaker recognition performance in Sections 4.4.2 and 4.4.3. A brief explanation of the compared algorithms will be given in the following. For a detailed description, the reader is referred to the corresponding references.

Noise estimation

The estimate of the noise power spectrum is derived from the noisy signal spectrum $X(t, \omega)$. Various noise estimation techniques have been proposed to deal with stationary and fluctuating noise types. A comprehensive overview and implementational details can be found in Loizou (2007). In this study, the most relevant developments are compared in the context of speaker recognition. The most simple method *Initial50ms* is estimating the noise by averaging the power spectrum of the initial frames (Vizinho *et al.*, 1999), assuming that no speech is present. The weighted average method *Hirsch95* introduced by Hirsch and Ehrlicher is using a first order recursion and employs an adaptive threshold to stop the recursion when speech activity is detected (Hirsch and Ehrlicher, 1995). An alternative approach *Lin03* is adjusting the first order recursion based on the estimated *a posteriori* SNR (Lin *et al.*, 2003). More elaborated methods, such as *Doblinger95*, *Cohen02* and *Martin06*, recursively average the noise power by tracking minima in the noisy spectrum (Doblinger, 1995, Martin, 2001, Cohen and Berdugo, 2002, Martin, 2006). A modification *Rangachari06* that aims at reducing the adaptation time of the noise estimate especially for highly non-stationary conditions is using a smoothing factor based on speech presence probability (Rangachari *et al.*, 2004, Rangachari and Loizou, 2006). Finally, a modified version *Lin03Mod* of the SNR-dependent recursive averaging (Lin *et al.*, 2003) was implemented by smoothing the noisy power spectrum with a first order

Table 4.2: Evaluated gain curves for estimating the clean speech spectrum.

Gain function	Parameters[1]	Reference
SpecSubPowα_1	$\alpha = 1, \beta = 0.01, \gamma_1 = 1, \gamma_2 = 1$	Virag (1999)
SpecSubPowα_3	$\alpha = 3, \beta = 0.01, \gamma_1 = 1, \gamma_2 = 1$	Virag (1999)
SpecSubMagα_1	$\alpha = 1, \beta = 0.01, \gamma_1 = 2, \gamma_2 = 0.5$	Virag (1999)
SpecSubMagα_3	$\alpha = 3, \beta = 0.01, \gamma_1 = 2, \gamma_2 = 0.5$	Virag (1999)
MMSE STSA	$\alpha = 0.98$	Ephraim and Malah (1984)
MMSE log-STSA	$\alpha = 0.98$	Ephraim and Malah (1985)

recursion prior to estimating the SNR-dependent smoothing parameter. The frame-based smoothing was performed with a filter coefficient of $\eta = 0.6$ and aimed at reducing the variance of the resulting noise estimate. In Tab. 4.1, the parameter settings of all evaluated noise estimation techniques are listed. As far as possible, parameters were chosen according to the recommendations of the authors.

Speech estimation

In addition to the estimated noise power spectrum $\left|\hat{N}(t,\omega)\right|^2$, an estimate of the clean speech spectrum $\hat{S}(t,\omega)$ is required to construct the missing data mask according to Eq. (4.9). In the context of noise reduction, much effort has been directed at improving the perceived quality of the estimated speech signal by attenuating the amount of residual noise while keeping speech and background noise artifacts at a minimum. However, these perceptual constraints are not necessarily relevant for the estimation of the ideal binary mask and it is a question, if perceptual improvements can be quantified in terms of recognition performance. To answer this question, the most common approaches are briefly reported and the influence on the estimated binary mask will be assessed in Section 4.4.3.

An estimate of the speech spectrum can be derived by subtracting the estimated noise spectrum from the corrupted signal spectrum. The most frequently used technique to accomplish this is to perform spectral subtraction (Boll, 1979) by applying an SNR-dependent gain function in the frequency domain. The residual noise that exhibits strong temporal fluctuations after processing is often referred to as *musical noise* (Cappe, 1994). To reduce this problem of musical noise in spectral subtraction-based noise reduction schemes, an over-estimation of the noise in combination with introducing a spectral floor was found to be beneficial (Berouti *et al.*, 1979). In the context of detecting reliable feature components based on spectral subtraction, the optimal over-estimation factor

was reported to be close to $\alpha = 3$ (Drygajlo and El-Maliki, 1998b, Renevey and Drygajlo, 2000). An alternative approach is to estimate the optimal minimum mean square error (MMSE) short-time spectral amplitude (STSA), which was reported to significantly reduce the problem of musical noise by recursively smoothing the *a priori SNR* (Ephraim and Malah, 1984, 1985). More recently, model-based approaches (Ephraim, 1992, Srinivasan *et al.*, 2007, Zhao *et al.*, 2008) have been reported to further improve the performance of speech enhancement systems especially in the presence of non-stationary noise at the expense of increasing computational complexity. But bearing in mind that the mask estimation is part of the front-end for missing data classification, we limited the estimation of the clean speech spectrum to SNR-based gain functions, which can be efficiently applied in the spectral domain.

To study the effect of the above described approaches on the estimation of the clean speech spectrum, and consequently on the estimation of the ideal binary mask, the following four gain functions are evaluated: magnitude spectral subtraction, power spectral subtraction, MMSE STSA and MMSE log-STSA. The gain functions were implemented using the speech processing toolbox *VOICEBOX* (Brookes, 2009). Parameters of all tested gain functions are listed in Tab. 4.2. Note that due to the convention in the corresponding references, the parameter α has a different meaning within the spectral subtraction and the MMSE framework. Spectral subtraction-based gain functions are characterized by the over-estimation factor α, the spectral floor β and the two exponents γ_1 and γ_2 in the generalized spectral subtraction scheme, which define the suppression rule to be either magnitude or power spectral subtraction. The MMSE-based gain curves are computed using a smoothing constant α to recursively estimate the *a priori SNR*.

4.3 Evaluation setup

4.3.1 Acoustic mixtures

Speaker recognition performance was evaluated on a closed set of 34 speakers (18 males and 16 females) using the SSC database (Cooke and Lee, 2006). The database consists of 17.000 clean utterances, 500 utterances per speaker. The audio signals were down-sampled to a sampling frequency of $f_s = 16\,\text{kHz}$. To ensure that the speech material used to train the UBMs is different from the material used for the speaker recognition experiments, speech files of all 34 speakers were randomly partitioned into two equal sized

sets, each consisting of 250 sentences per speaker. The first half was used to train the speaker-independent but gender-dependent UBMs for all recognizers. From the second half of the SSC database (again comprising 250 sentences per speaker), a certain amount of speech files was randomly selected for the speaker recognition experiments presented in Section 4.4. This limitation of speech files is motivated by the fact that the amount of available speech material is often a constraint for practical applications. The amount of speech material involved in the training of speaker-dependent models (using either a conventional GMM recognizer or the UBM-based adaptation, see Section 4.3.2) and the evaluation of speaker recognition accuracy is reported for each experiment, individually. To investigate the influence on speaker recognition performance, the amount of available speech files is systematically varied in Section 4.4.4. In general, for each speaker 70% of the available speech material was randomly selected to train the corresponding speaker model and the remaining 30% was used for evaluation.

Because the assessed recognition accuracy will depend to some extent on the random selection of the training and the testing material, results are reported as the mean speaker identification accuracy over a series of 20 simulations, each containing a new randomly selected set of speech files for training and testing. Speaker recognition experiments were performed on either the full set of 34 speakers or with a subset of 10 randomly selected speakers. For each of the 20 simulations, a new subset of 10 speakers was randomly selected from the set of 34 speakers. In the testing phase, utterances were digitally mixed at various SNRs with noise signals drawn from the NOISEX database (Varga *et al.*, 1992). Five different noise types were used for evaluation: factory noise, cockpit noise, speech babble, destroyer operation engine noise and car noise. The SNR was computed by comparing the A-weighted energy of the speech signal to the A-weighted energy of the noise signal. The A-weighting filter was applied to ensure that the SNR is adjusted predominantly within the frequency range that is relevant for speech. The design of the A-weighting filter was implemented according to American National Standards Institute (1983). To prevent that the energy of speech is underestimated due to silent parts, an energy-based voice activity detector (VAD) was used to only consider signal segments with relevant speech activity. A frame was considered to contain relevant speech activity, if its energy level was within 40 dB of the global maximum.

4.3.2 GMM and UBM parameters

In this study, two different types of recognizers were used to model the speaker-dependent distribution of features. The first recognizer, denoted as *GMM*, was a conventional GMM[2] recognizer. Based on features extracted from the speaker-dependent speech material, speaker models were first initialized by 20 iterations of the k-means algorithm and further trained using the EM algorithm. The EM algorithm iteratively refines the model parameters λ by maximizing the likelihood of the resulting GMM. The stopping criterion of the EM algorithm was set to $1e^{-5}$ with a maximum of 300 iterations. The second recognizer, referred to as *GMM-UBM*, utilized two gender-dependent UBMs, reflecting the distribution of speaker-independent features for male and female speech, respectively. The training of both UBMs involved one half of the SSC database (see Section 4.3.1) and was accomplished by 20 initial k-means iterations followed by the EM algorithm using a stopping criterion of $1e^{-5}$ with a maximum of 300 iterations. Speaker-dependent models were obtained by adapting the trained UBM parameters to the speaker-dependent speech material. Therefore, first the gender selection was performed by selecting the UBM which showed the higher probabilistic alignment with the speaker-dependent speech material. Secondly, as suggested by Reynolds *et al.* (2000), only the mean vectors of the UBM were adapted using a relevance factor of 16. The adaptation was performed by 10 iterations of the EM algorithm. Note that the two recognizers were used both within the MD framework and also to represent the MFCC-based feature vectors (see Section 4.4.5). Only speech material with relevant speech activity was included in the training stage of both systems by using the previously described VAD.

4.3.3 Baseline system

A conventional robust speaker recognition system was trained using a 26-dimensional feature vector consisting of 13 static MFCC coefficients, including the 0th order coefficient, and first order temporal derivatives, so-called *delta coefficients*. The static MFCC coefficients were computed using the RASTAMAT toolbox (Ellis, 2005). Parameters[3] were chosen to reproduce MFCC coefficients according to the hidden Markov models toolkit (HTK), which is a commonly used front-end for speaker recognition experiments.

2 The GMM modeling was performed using the NETLAB package (Nabney and Bishop, 2001-2004).

3 `melfcc(in, fs, 'lifterexp', -22, 'nbands', 20, 'dcttype', 3, 'maxfreq', 8000, 'fbtype', 'htkmel', 'sumpower', 0, 'wintime', 20e-3, 'hoptime', 10e-3, 'numcep', 13)`. For details see `http://labrosa.ee.columbia.edu/matlab/rastamat/mfccs.html`

The delta coefficients were computed using the trend derived from linear regression over a window of 5 frames (Soong and Rosenberg, 1988). In a pilot experiment, it was observed that the additional use of second order temporal derivatives, so-called *acceleration* coefficients reduced speaker recognition performance at low SNRs and therefore, acceleration coefficients were not appended. Cepstral mean and variance normalization (CMVN) was applied for improved robustness, where the feature statistics are measured over the duration of one utterance (Openshaw and Mason, 1994, Tibrewala and Hermansky, 1997). Compared to cepstral mean normalization (CMN) (Atal, 1974), CMVN substantially improved MFCC-based recognition performance.

4.4 Experiments

A series of speaker recognition experiments was conducted. The first experiment investigated the benefit of combining MD recognition with the UBM-based adaptation of speaker models under idealized conditions, assuming that the required ideal binary mask is known *a priori*. In reality, the IBM is not known and needs to be estimated. Therefore, experiments two and three aimed at exploring various techniques to estimate the IBM in a variety of different background noise conditions. In addition, a detailed analysis of errors made by the mask estimation process is presented. Based on these findings, the best performing method for estimating the missing data mask was selected in the fourth experiment and the performance improvement with the UBM-based missing data recognizer as a function of available speech material is investigated. Finally, the fifth experiment compared speaker recognition performance of the UBM-based missing data recognizer with a GMM-based MD recognizer and with state-of-the art MFCC-based recognizers (with and without UBM adaptation).

4.4.1 Experiment 1: Effect of UBM using an IBM

The first experiment studied the effect of using a UBM within the MD framework. To isolate the effect of the UBM on speaker recognition performance, this initial experiment used the ideal binary mask for MD recognition. In order to further investigate the effect of erroneously classified T-F components on missing data recognition, the ideal binary mask was modified by randomly labeling unreliable T-F components as reliable. A spectral feature component corresponding to an unreliable T-F unit is dominated by background

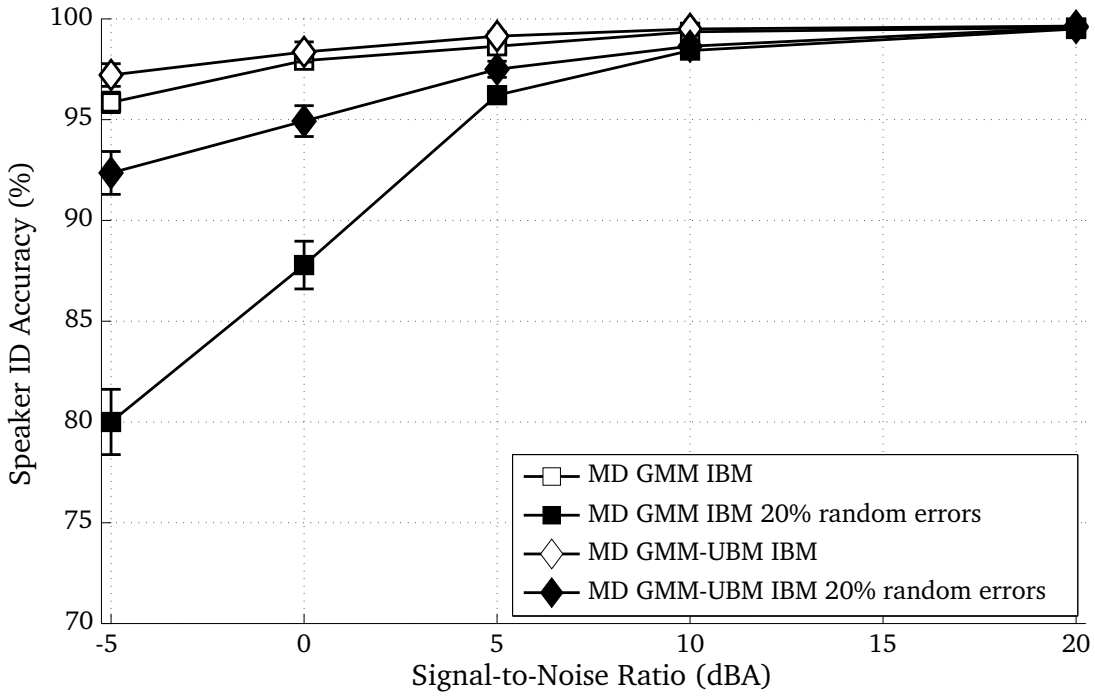


Figure 4.1: Experiment 1: SNR-dependent speaker recognition performance for 10 speakers in the presence of factory noise using the ideal binary mask (IBM). The average recognition performance over a series of 20 simulations is presented for both recognizers, the GMM-based missing data system *MD GMM IBM* (squares) and the system including a universal background model *MD GMM-UBM IBM* (diamonds). The error bars represent the standard error of recognition performance across all 20 simulations.

noise and thus is likely to cause a mismatch between the training and the testing situation. The number of randomly modified components was chosen to be 20% of the number of reliable T-F components. Speaker recognition accuracy was evaluated on a subset of 10 speakers. The amount of available speech material per speaker was limited to 25 sentences, using 18 for training and 7 for testing. The optimal model complexity of both recognizers, *MD GMM IBM* and *MD GMM-UBM IBM*, was individually selected based on pilot experiments. For both systems, a model complexity of 64 Gaussian components was chosen.

The average SNR-dependent speaker recognition accuracy in the presence of factory noise is presented in Fig. 4.1. Note that the standard error of recognition performance across all 20 simulations was below 2% for all experimental conditions. Open symbols represent the IBM and black symbols indicate that the IBM has been modified by randomly labeling unreliable T-F units as being reliable. When using the IBM, the recognition performance of both systems *MD GMM IBM* and *MD GMM-UBM IBM* is almost identical. However, it can be seen that the MD recognizer in combination with a UBM is significantly more

robust when unreliable T-F components are randomly labeled as reliable. Especially at low SNRs, the advantage of the UBM-based system in terms of speaker recognition accuracy is larger than 12%. When using the *MD GMM* system, all speaker-dependent model parameters are trained using the training data of one speaker only, and as a result, observations which have not been seen during the training stage can cause erroneous likelihood values which can potentially bias the class decision in very different ways across different speakers. Regarding the UBM-based recognition system, all speaker-dependent models share the same initial model parameters that have been learned based on the pooled speech material of the SSC database. Furthermore, only the mean values of the Gaussian components are adapted to the speaker-dependent training data that show sufficient probabilistic alignment. As a result, the UBM-based system is significantly less sensitive to observations which were not included in the training stage.

4.4.2 Experiment 2: Influence of noise estimation algorithms

The second experiment investigated the influence of various noise estimation algorithms on the estimation of the IBM in terms of speaker recognition performance. Independent of the noise estimation technique, the speech spectrum was obtained by applying the MMSE log-STSA gain function to the noisy input spectrum. Speaker models were trained using a GMM-based missing data recognizer *MD GMM* with 16 Gaussian components. Speaker recognition accuracy was evaluated on a subset of 10 speakers, involving a total of 25 sentences per speaker (18 sentences for training and 7 sentences for testing). The SNR-dependent recognition accuracy for all evaluated noise estimation techniques is presented in Tab. 4.3.

The upper five panels of Tab. 4.3 show recognition performance for different noise types, whereas the last table depicts the average performance over all noise conditions. The corresponding standard error of recognition performance across all 20 simulations was below 2% for all experimental conditions. Results are ranked for different noise types according to recognition performance, starting with the lowest performance for babble noise and ending with the highest performance in the presence of car noise. This ranking coincides with the stationarity of the background noise, ranging from very non-stationary (babble and factory noise) to more stationary conditions (car noise).

Table 4.3: Experiment 2: Average missing data speaker recognition accuracy over a series of 20 simulations for a subset of 10 speakers in the presence of different types of background noise. The binary mask was estimated using the *MMSE log-STSA* noise reduction scheme and various noise estimation techniques listed in Tab. 4.1.

babble noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	29.86	58.57	84.50	96.79	98.86
Hirsch95	20.43	45.00	76.50	94.93	98.86
Doblinger95	20.50	46.50	79.43	95.14	98.79
Cohen02	**34.36**	**69.71**	90.86	**97.86**	98.93
Lin03	16.29	36.50	66.29	90.64	98.86
Martin06	22.29	51.50	79.50	95.79	**99.07**
Rangachari06	23.07	50.86	82.57	96.07	98.93
Lin03Mod	33.07	68.29	**91.50**	97.71	99.00
factory noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	35.64	61.93	84.93	95.64	98.86
Hirsch95	23.00	45.86	75.43	93.57	**99.07**
Doblinger95	24.43	48.86	79.36	94.93	98.79
Cohen02	**39.86**	69.50	**89.43**	**96.57**	98.86
Lin03	16.43	32.64	60.57	86.57	98.86
Martin06	26.86	54.00	81.07	94.86	98.86
Rangachari06	30.29	61.29	87.00	**96.57**	98.64
Lin03Mod	39.29	**70.93**	88.86	96.07	98.79
destroyer noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	45.79	71.93	89.43	96.21	99.00
Hirsch95	24.57	49.00	79.00	94.29	99.00
Doblinger95	31.14	62.00	86.86	96.29	98.71
Cohen02	45.57	77.64	**92.50**	97.36	**99.07**
Lin03	13.93	29.07	56.86	84.57	98.71
Martin06	35.86	64.36	86.50	96.07	99.00
Rangachari06	**55.57**	**82.50**	94.64	**97.50**	98.93
Lin03Mod	50.79	77.14	91.36	96.79	99.00
cockpit noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	79.50	93.93	97.36	**98.79**	99.00
Hirsch95	77.57	93.29	97.36	98.64	**99.07**
Doblinger95	81.57	94.21	**97.57**	98.21	98.71

Continued on next page. . .

Table 4.3 (continued)

Cohen02	73.86	92.64	97.07	98.29	99.00
Lin03	26.00	57.43	85.14	94.79	98.79
Martin06	73.71	91.64	97.50	98.64	99.00
Rangachari06	81.71	**94.93**	97.14	98.07	98.86
Lin03Mod	**83.50**	93.64	96.71	97.93	98.93
car noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	72.07	85.00	91.43	95.57	98.14
Hirsch95	60.21	75.57	87.50	94.71	**98.50**
Doblinger95	60.64	77.07	87.64	95.50	97.64
Cohen02	80.50	92.93	96.29	97.57	**98.50**
Lin03	22.07	28.36	45.93	68.43	96.43
Martin06	67.29	80.29	87.86	93.29	98.29
Rangachari06	72.43	84.43	91.29	96.21	98.21
Lin03Mod	**86.14**	**93.43**	**96.71**	**98.07**	**98.50**

average	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	52.57	74.27	89.53	96.60	98.77
Hirsch95	41.16	61.74	83.16	95.23	**98.90**
Doblinger95	43.66	65.73	86.17	96.01	98.53
Cohen02	54.83	80.49	**93.23**	**97.53**	98.87
Lin03	18.94	36.80	62.96	85.00	98.33
Martin06	45.20	68.36	86.49	95.73	98.84
Rangachari06	52.61	74.80	90.53	96.89	98.71
Lin03Mod	**58.56**	**80.69**	93.03	97.31	98.84

At an SNR of 20 dBA, no substantial difference was observed between the evaluated noise estimation techniques. Considering lower SNRs, several methods performed well for one particular noise type but recognition performance was considerably lower compared to other methods if the speech material was corrupted by other types of background noise. *Rangachari06* performed well under the influence of both cockpit and destroyer noise, but recognition performance was significantly lower than other methods in the presence of factory, babble and car noise. *Doblinger95* and *Hirsch95* performed best in the condition of cockpit noise but performance was way below *Cohen02* and *Lin03Mod* in all other

background noise scenarios. This sensitivity to the type of background noise suggests that the corresponding noise estimation methods are most suitable for certain types of background noise or that the corresponding parameters may have been optimized for a particular noise condition. On average, the recognition performance using *Martin06* was about 10% below *Lin03Mod* at low SNRs, which might have been caused by the rather conservative adaptation of the noise spectrum and the comparably long initialization phase.

The overall lowest recognition performance was obtained by *Lin03*. This method was initially designed to estimate the noise spectrum in auditory bands. Clearly, the spectral variance of frequency bins is much larger than the variation within auditory bands and therefore, this method failed to successfully predict the noise power spectrum. In contrast to the above mentioned techniques, *Cohen02* and *Lin03Mod* were consistently among the best methods across all evaluated noise types. Compared to *Lin03*, the additional recursion in *Lin03Mod* significantly increased recognition performance.

In order to gain more understanding of the underlying factors that influence the recognition performance, receiver operating characteristic (ROC) curves (Fawcett, 2006) were computed by comparing the estimated binary mask to the ideal binary mask which was computed using *a priori* knowledge of the speech and the noise spectra. The ROC graph visualizes the trade-off between correctly identified T-F components which are dominated by the target signal (true positive rate) and misclassified T-F elements which are dominated by background noise (false positive rate). The higher the true positive rate and the smaller the false positive rate, the higher the quality of the estimated binary mask. Based on the experimental results reported in Tab. 4.3, the corresponding SNR-dependent ROC curves are shown for all evaluated noise estimation techniques in Fig. 4.2. The ROC curves are averaged across all 5 background noise types. With decreasing SNR, the true positive rate decreased for all noise estimation techniques. Although *Lin03* achieved the highest true positive rates, the false positive rates are above 20% for all SNR conditions. The additional first order recursion employed in *Lin03Mod* substantially reduced the false positive rate. It can be seen that the two best performing noise estimation methods *Lin03Mod* and *Cohen02* are most conservative in labeling reliable T-F components, keeping the false positive rate down to about 2%. The small advantage of *Lin03Mod* over *Cohen02* at low SNRs is manifested in a slightly higher true positive rate. Thus it seems that false positive errors are most problematic for MD recognition using a GMM classifier.

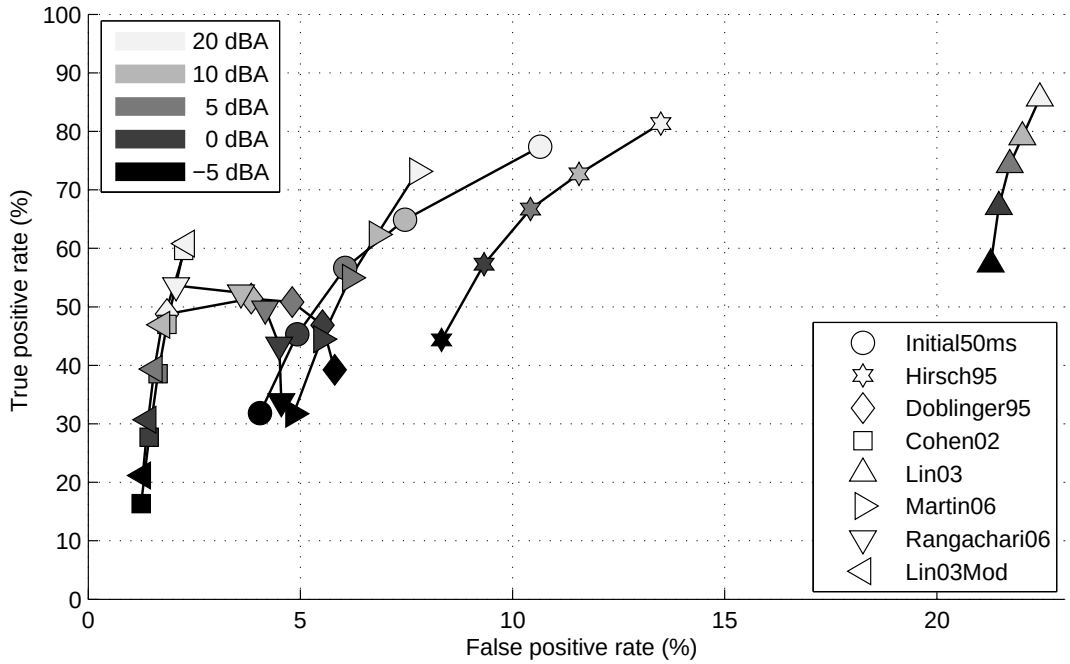


Figure 4.2: Experiment 2: SNR-dependent ROC curves for all evaluated noise estimation algorithms. The ROC curves are averaged over all five noise conditions.

4.4.3 Experiment 3: Influence of speech estimation algorithms

In the third experiment, the impact of deriving an estimate of the clean speech spectrum $\hat{S}(t,\omega)$ on the estimated binary mask was analyzed. The clean speech spectrum is estimated by subtracting the estimated noise spectrum from the noisy input spectrum using various SNR-based gain curves (see Tab. 4.2). The objective of those gain curves is to enable maximum noise suppression while simultaneously minimizing the amount of speech distortion. The strength of the noise suppression determines the amount of noise energy which might leak into the estimated speech power spectrum, and consequently, will affect the amount of T-F elements in the estimated binary mask which are erroneously identified as reliable components due to the over-estimation of the speech power spectrum.

The MD-based speaker recognition performance for the estimated binary mask based on all tested methods to derive an estimate of the clean speech spectrum is presented in Tab. 4.4 depending on the SNR and the type of background noise. The standard error of recognition performance across all 20 simulations was below 2% for all experimental conditions. The noise power spectrum was estimated by *Lin03Mod*, which was the most consistent method among all evaluated noise estimation techniques in the second experiment. The remaining experimental conditions were identical to the second experiment (see Section 4.4.2).

Similar to the second experiment, the SNR-dependent ROC curves corresponding to the experimental results in Tab. 4.4 are presented in Fig. 4.3.

Table 4.4: Experiment 3: Average missing data speaker recognition accuracy over a series of 20 simulations for a subset of 10 speakers in the presence of different types of background noise. The binary mask was estimated using the *Lin03Mod* noise estimation technique and various noise reduction schemes listed in Tab. 4.2.

babble noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	14.64	31.36	61.14	90.71	98.57
SpecSubPowα_3	19.21	42.50	75.86	95.29	**99.14**
SpecSubMagα_1	21.93	49.57	83.79	96.79	99.00
SpecSubMagα_3	32.07	67.29	90.71	97.57	99.00
MMSE STSA	28.57	60.36	88.43	97.21	99.00
MMSE log-STSA	**33.07**	**68.29**	**91.50**	**97.71**	99.00
factory noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	14.86	28.86	57.71	88.21	98.86
SpecSubPowα_3	18.93	42.86	77.93	94.79	**99.21**
SpecSubMagα_1	24.93	54.43	84.14	95.93	98.93
SpecSubMagα_3	35.71	69.57	**89.21**	**96.29**	98.71
MMSE STSA	33.43	64.71	86.36	96.07	98.93
MMSE log-STSA	**39.29**	**70.93**	88.86	96.07	98.79
destroyer noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	15.29	25.50	51.86	85.07	98.93
SpecSubPowα_3	18.50	37.93	71.79	93.29	**99.14**
SpecSubMagα_1	21.50	46.86	79.93	94.50	98.93
SpecSubMagα_3	33.79	65.79	86.71	95.64	98.50
MMSE STSA	45.36	74.29	90.29	**96.86**	**99.14**
MMSE log-STSA	**50.79**	**77.14**	**91.36**	96.79	99.00
cockpit noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	19.21	41.93	76.57	93.50	98.93
SpecSubPowα_3	47.50	82.21	95.64	98.14	99.00
SpecSubMagα_1	67.57	91.79	96.57	97.93	99.00
SpecSubMagα_3	76.64	91.00	95.43	97.29	98.71
MMSE STSA	**86.71**	**94.71**	**97.14**	**98.29**	**99.07**
MMSE log-STSA	83.50	93.64	96.71	97.93	98.93

Continued on next page...

Table 4.4 (continued)

car noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	16.57	18.50	30.50	52.14	93.71
SpecSubPowα_3	25.50	39.93	55.71	80.50	98.43
SpecSubMagα_1	32.36	54.86	72.50	91.86	**98.71**
SpecSubMagα_3	71.07	88.14	94.86	97.93	98.36
MMSE STSA	82.93	91.71	96.43	**98.36**	98.64
MMSE log-STSA	**86.14**	**93.43**	**96.71**	98.07	98.50

average	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	16.11	29.23	55.56	81.93	97.80
SpecSubPowα_3	25.93	49.09	75.39	92.40	**98.99**
SpecSubMagα_1	33.66	59.50	83.39	95.40	98.91
SpecSubMagα_3	49.86	76.36	91.39	96.94	98.66
MMSE STSA	55.40	77.16	91.73	**97.36**	98.96
MMSE log-STSA	**58.56**	**80.69**	**93.03**	97.31	98.84

Regarding spectral subtraction-based gain functions, magnitude-based spectral subtraction with an over-estimation factor of 3 *SpecSubMag*α_3 enabled the strongest attenuation of the estimated noise and consequently, lowered the level of the estimated speech spectrum. Thus, *SpecSubMag*α_3 conservatively selected reliable T-F components and as a result, achieved the highest recognition performance among all spectral subtraction-based algorithms. This observation is consistent across all evaluated background noise scenarios and can be confirmed by comparing the corresponding ROC curves, which show systematically decreasing false positive rates for noise reduction schemes with stronger noise attenuation. Although the MMSE-based gain functions *MMSE STSA* and *MMSE log-STSA* produced slightly higher false positive rates, *MMSE log-STSA* outperformed *SpecSubMag*α_3 in terms of recognition accuracy, especially in conditions with low SNRs. Whereas moderate improvements are observed in the presence of highly non-stationary noise (babble and factory noise), more substantial benefits of *MMSE log-STSA* over the spectral subtraction-based methods are found for more stationary scenarios, namely destroyer and car noise. This advantage is presumably caused by the higher true positive rates of *MMSE log-STSA*, which effectively produced an estimated binary mask which is less sparse than the mask obtained by *SpecSubMag*α_3.

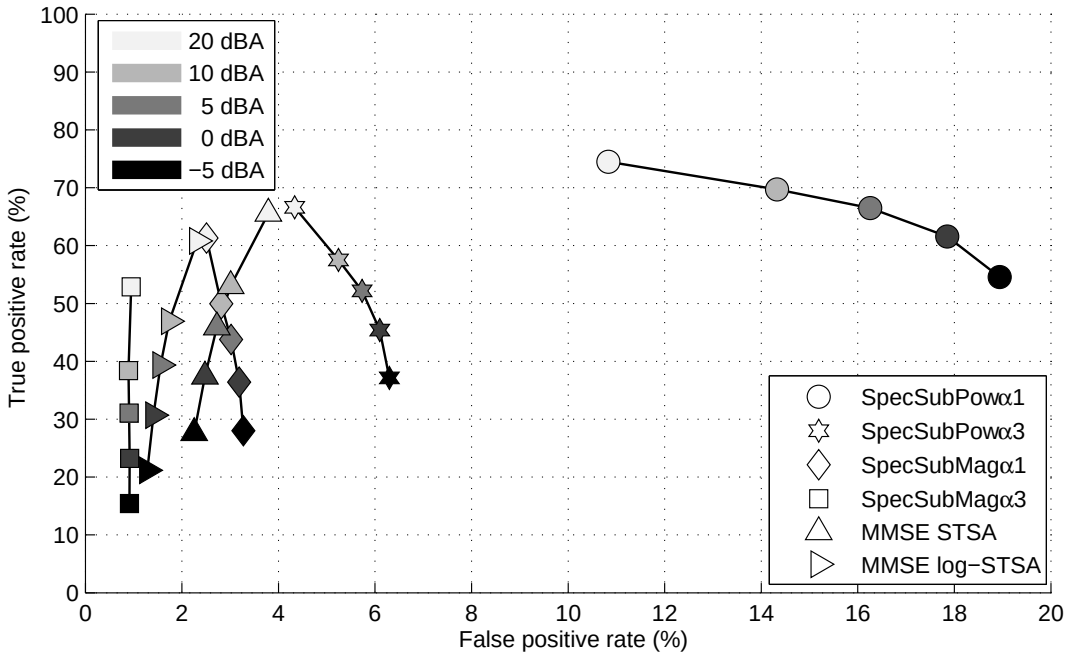


Figure 4.3: Experiment 3: SNR-dependent ROC curves for all evaluated noise suppression rules. ROC curves are averaged over all five noise conditions.

4.4.4 Experiment 4: Effect of UBM using an estimated binary mask

The first experiment showed a significant benefit when using a UBM within the MD framework under ideal conditions (i.e. when the ideal binary mask is known *a priori*). Consequently, the fourth experiment aimed at verifying this benefit for scenarios in which the ideal binary mask is not known *a priori* and needs to be estimated. Based on findings of the second and third experiment, the IBM was estimated by combining the *Lin03Mod* noise estimation with the *MMSE log-STSA* gain function. In addition, the dependency of the reported performance gain on the amount of available speech material is investigated. The speech material that can be used for training and testing the speaker models consist of 250 sentences per speaker, which might not be available for practical applications. Furthermore, the amount of available speaker-dependent speech material and the complexity of the recognizer (number of Gaussian components used by the recognizer) are interdependent and both parameters are expected to influence the comparison between both speaker recognition systems *MD GMM* and *MD GMM-UBM*. Therefore, the influence of the following parameters is analyzed:

- Number of speaker-dependent sentences: 13, 25, 50, 125 and 250 of the SSC database (70% for training and 30% for testing),
- Number of GMM components: 16, 32, 64, 128 and 256,

- Type of recognizer: *MD GMM* and *MD GMM-UBM*.

The relative difference in speaker recognition accuracy between *MD GMM-UBM* and *MD GMM* is presented in the left panels of Fig. 4.4 as a function of the SNR and the number of Gaussian components. In addition, the right panels show the corresponding absolute speaker recognition accuracy using the *MD GMM-UBM* recognizer. Speaker recognition was performed on a subset of 10 speakers in the presence of factory noise. The amount of available speech material is systematically increased, ranging from 13 sentences in the top panels to 250 sentences in the bottom panels. It can be observed that the benefit of the *MD UBM-GMM* is greatest when a limited amount of speech material is available for training. In general, rather moderate improvements are observed for conditions with a high SNR down to 10 dBA, because the corresponding recognition performance is close to 100%, but considerable improvements are achieved for low SNRs. If only 13 sentences of speaker-dependent material are involved in the speaker recognition experiments as depicted in panel 4.4(A), the improvement can be as high as 23% for conditions with low SNRs. With increasing availability of speech material, the difficulty of properly training speaker-dependent GMMs decreases and as a consequence, the relative benefit of the UBM-based missing data system decreases. When all 250 sentences per speaker are used (see panel 4.4(E)), the benefit of applying a UBM is only about 10% at low SNRs. Overall, the usage of a UBM in conjunction with MD recognition shows a consistent and substantial benefit, especially in challenging noise conditions with low SNR.

The optimal number of Gaussian components moderately depends on the amount of available training material. If the amount of speech material is limited (to up to 25 sentences), the *MD GMM-UBM* recognizer with 64 Gaussian components performed best. This is consistent with the observation that the UBM-based recognizer with 64 components using the ideal binary mask performed best in the first experiment (see Section 4.4.1). The use of Gaussian mixture models with a higher complexity led to a decrease in recognition performance (see panels 4.4(F) and 4.4(G)), which may indicate that the resulting speaker models were overtrained given the limited amount of speech material. Having access to at least 50 sentences of the SSC database, the recognizer with 128 Gaussian mixtures slightly outperformed systems with lower complexity at low SNRs. Speaker recognition performance saturated at a model complexity of 128 Gaussian components and a further increase did not lead to significant improvements.

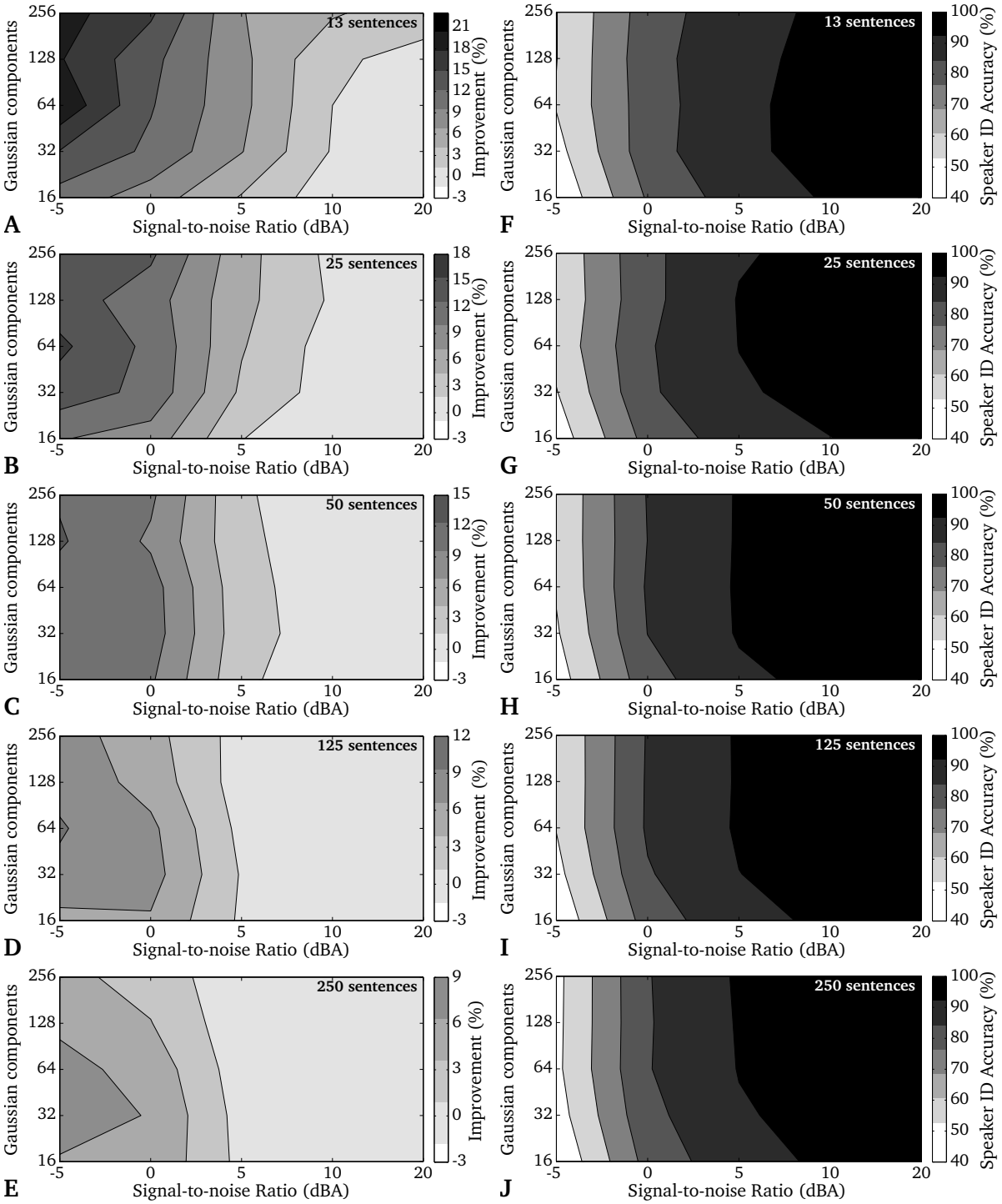


Figure 4.4: Experiment 4: Speaker identification improvement of the *MD GMM-UBM* method compared to the *MD GMM* system on a closed set of 10 speakers in the presence of factory noise. The left panels show the speaker identification improvement as a function of the SNR and the number of Gaussian components for experiments involving (A) 13, (B) 25, (C) 50, (D) 125, and (E) 250 sentences of speaker-dependent speech material. The right panels (F)-(J) show the corresponding speaker identification accuracy of the *MD GMM-UBM* recognizer. Both speaker identification improvement and speaker identification accuracy are reported as the mean over a series of 20 simulations. The standard error of both measures was below 3% for all conditions.

4.4.5 Experiment 5: MD recognition versus MFCCs

The last experiment compared the proposed UBM-based missing data recognizer with a conventional MFCC-based recognition system (see Section 4.3.3). The distribution of both spectral and MFCC features was learned by the two classifiers *GMM* and *GMM-UBM* (see Section 4.3.2). For MD recognition, the required mask was estimated using the *Lin03Mod* noise estimation technique and the *MMSE log-STSA* gain function. Furthermore, MD recognition was performed using the ideal binary mask in order to show the theoretical upper performance limit. In addition to the conventional MFCC-based recognizer, another baseline system is used that employed a noise reduction (NR) stage prior to computing the MFCC feature vector. Among all evaluated methods, the *Hirsch95* noise estimation technique in combination with the *MMSE log-STSA* gain function performed best during initial tests. This pre-processing is indicated by *NR*. Furthermore, analogously to the ideal binary mask, an ideal noise reduction front-end *NR IDEAL* was applied that used *a priori* knowledge about the noise power spectrum. Similar to the first experiment, the optimal model complexity of each recognizer was selected based on pilot experiments. Whereas the two MFCC-based recognizers *MFCC GMM* and *MFCC GMM NR* performed best with 32 Gaussian components, all other recognizers utilized 128 Gaussian components.

The average speaker recognition performance is shown as a function of the SNR in Fig. 4.5 and Fig. 4.6, respectively. Figure 4.5 presents speaker recognition performance on a subset of 10 speakers, whereas the results for the full set of 34 speakers are shown in Fig. 4.6. For each speaker, a total of 25 sentences were used (18 sentences for training and 7 sentences for testing). The standard error across all 20 simulations was below 2% for all experimental conditions. Black symbols signify recognizers that are based on UBM adaptation. According to the results obtained in the second and third experiment (see Section 4.4.2 and Section 4.4.3), the five background noise types are grouped into two categories, namely highly non-stationary (babble and factory noise) and more stationary scenarios (destroyer, car and cockpit noise). Speaker recognition performance is separately shown as the average of highly non-stationary (top panels) and more stationary noise conditions (bottom panels).

Compared to the GMM-based cepstral recognizer *MFCC GMM* (△), the additional use of a UBM (▲) significantly improved recognition performance in all conditions. But already at moderate SNRs, the speaker recognition performance of both MFCC-based recognizers *MFCC GMM* and *MFCC GMM-UBM* rapidly decreased. Considering more stationary

noise conditions, a consistent and significant performance gain was obtained when noise reduction was performed prior to computing the MFCC features (▼). However, this benefit is noticeably reduced in highly non-stationary conditions, probably because of the inability of the noise estimation technique to quickly adapt to sudden changes in noise level. In turn, MD-based speaker recognition was superior to all MFCC-based recognition systems. In the presence of highly non-stationary noise (babble and factory noise), the usage of a UBM in combination with MD recognition (●) substantially improved recognition performance, especially at low SNRs. The relative performance improvement of the *MD GMM-UBM* recognizer over the *MD GMM* system (○) was in the range of 20% at very low SNRs. This improvement was found for both sets consisting of 10 and 34 speakers. Whereas the benefit of the UBM for the MFCC-based recognizer decreased at lower SNRs, the improvement for the MD recognizer in highly non-stationary noise conditions increased with decreasing SNR. Regarding the more stationary noise types, the benefit of the UBM was not significant. One possible explanation might be that for more stationary noise types, the false positive rate of the estimated binary mask is substantially below the error rates obtained in fluctuating noise scenarios. As a result, a larger amount of T-F units is erroneously labeled as speech for fluctuating noise, increasing the mismatch between training and testing. Compared to the GMM-based MD recognizer *MD GMM*, the UBM-based MD recognizer *MD GMM-UBM* is not as sensitive to this mismatch, providing a benefit especially for non-stationary conditions.

Considering the *MD GMM-UBM* system in the presence of both highly non-stationary and more stationary noise scenarios, the speaker recognition performance was close to that of the recognizer using the ideal binary mask *MD GMM-UBM IBM* for SNRs as low as 5 dBA, which suggests that the *Lin03Mod* noise estimation combined with the *MMSE log-STSA* gain function produces high quality missing data masks. Comparing the overall recognition performance between the two sets consisting of 10 and 34 speakers, it can be noticed that the speaker recognition accuracy is lower for the larger set of speakers, especially at lower SNRs. A similar dependency was observed in the context of speech recognition, where the recognition accuracy decreased when the vocabulary size was increased (Srinivasan *et al.*, 2006).

The fifth experiment also showed the fundamental limitation of noise reduction schemes to improve speaker recognition performance of MFCC-based recognizers in the presence of noise. As long as *a priori* information about the noise power spectrum is available, the MFCC recognizer including the ideal noise reduction front-end *MFCC GMM-UBM NR IDEAL* shows excellent recognition performance and is comparable to the MD

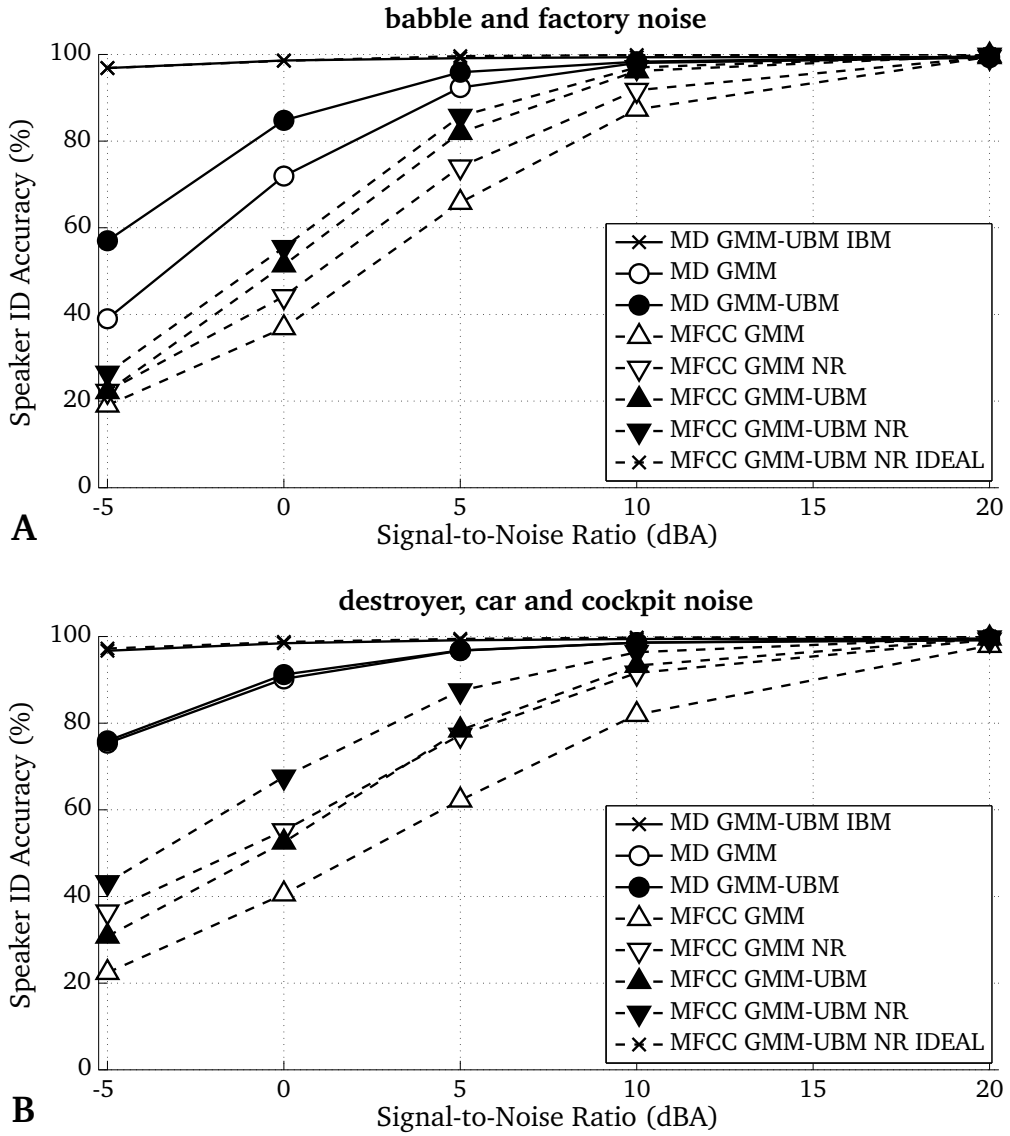


Figure 4.5: Experiment 5: SNR-dependent speaker recognition performance on a set consisting of 10 speakers. Results are shown for two groups of background noise scenarios; (A) highly non-stationary (babble and factory noise) and (B) more stationary noise conditions (destroyer, car and cockpit noise). Recognition performance is presented as the average recognition performance over a series of 20 simulations. The standard error of the reported recognition performance across all 20 simulations was below 2% for all experimental conditions. Results are presented for two types of recognizers, the MFCC-based recognizers (dashed lines) and the MD recognizers (solid lines). Black symbols indicate that the corresponding recognizers are based on UBM adapation. The two recognizers marked by crosses utilize *a priori* information about the noise spectrum.

system using the ideal binary mask *MD GMM-UBM IBM*. However, for MFCC-based recognition, there is a tremendous difference between the system that uses the ideal noise power and the one that employs the best operating front-end for noise estimation *MFCC GMM-UBM NR*. This performance difference between the idealized and the realistic scenario is significantly smaller for the MD recognizer, although both systems employ

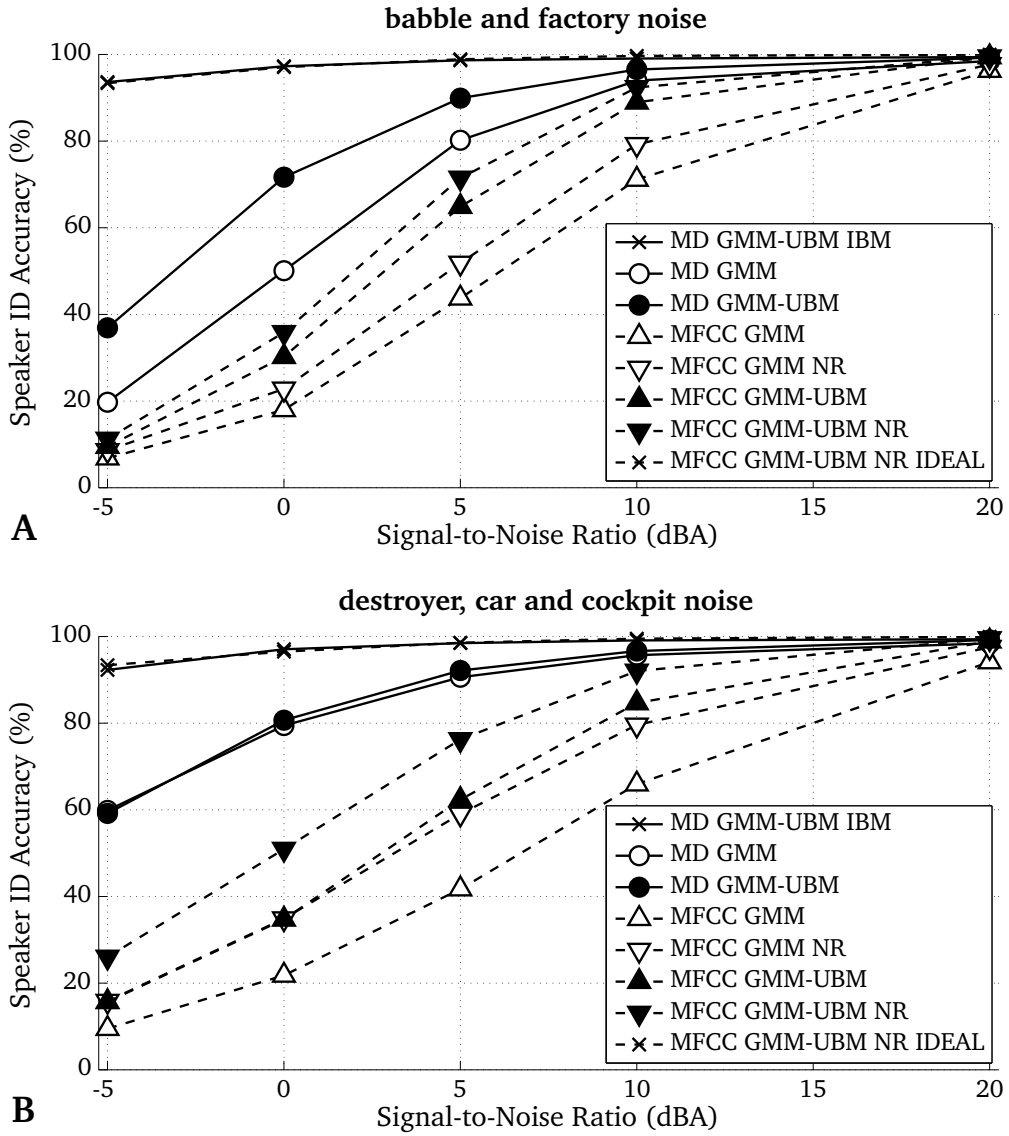


Figure 4.6: Experiment 5: SNR-dependent speaker recognition performance on a set consisting of 34 speakers. Results are shown for two groups of background noise scenarios; (A) highly non-stationary (babble and factory noise) and (B) more stationary noise conditions (destroyer, car and cockpit noise). Recognition performance is presented as the average recognition performance over a series of 20 simulations. The standard error of the reported recognition performance across all 20 simulations was below 2% for all experimental conditions. Results are presented for two types of recognizers, the MFCC-based recognizers (dashed lines) and the MD recognizers (solid lines). Black symbols indicate that the corresponding recognizers are based on UBM adapation. The two recognizers marked by crosses utilize *a priori* information about the noise spectrum.

a similar front-end for noise estimation. We believe that this performance gap can be explained by the conceptual difference between both recognizers and the way knowledge about the noise power is incorporated in both systems. Whereas the MD system only requires a binary decision about the reliability of individual T-F units, the noise reduction front-end needs an accurate estimation of the instantaneous SNR in order to design the

noise reduction filter, which is obviously much more difficult. Furthermore, errors in the estimated noise power spectrum will have different effects on speaker recognition performance for both systems. Whereas, e.g. an over-estimation of the noise will cause the estimated binary mask to be more sparse, leaving fewer T-F elements for classification, the noise reduction front-end will attenuate speech which can potentially distort the resulting MFCC feature vector. Our experimental results suggest that the errors induced by the noise estimation are more problematic for MFCC-based recognizers.

4.5 Discussion and conclusions

A robust speaker recognition system was presented, which combines missing data recognition with the adaptation of speaker models using universal background models. Compared to a GMM-based recognizer, the additional use of a UBM was shown to be especially beneficial in representing the spectral features in highly non-stationary noise conditions. The improvement was found to depend on the amount of available training material and was greatest for a small amount of speech material, which is often a constraint for practical applications.

The first experiment revealed that a MD recognizer in combination with a UBM was significantly more robust against false positive rates in the ideal binary mask and showed superior speaker recognition accuracy compared to a conventional GMM-based missing data recognizer. This benefit could be confirmed in the fourth experiment for conditions, where the IBM was estimated. One possible explanation is that the speaker models of a conventional GMM-based recognizer are initialized and trained independently. As a result, observations which were not well represented in the training stage can bias the likelihood computation across speaker models in very different ways. This problem is more likely to occur if only a small amount of training material is available or if T-F elements in the estimated binary mask are erroneously classified as reliable elements. When using a UBM, all speaker models are equally initialized using the UBM parameters and only those Gaussian components are adapted to the speaker-dependent speech material that show sufficient probabilistic alignment. Thus, the risk of a biased likelihood computation is reduced.

In the context of estimating the ideal binary mask using a local SNR criterion, several noise estimation techniques and gain functions were evaluated. Substantial speaker recognition accuracy differences were found across methods. Among all tested noise

estimation techniques, the minima-controlled recursive averaging (Cohen and Berdugo, 2002) and a modified version of the SNR-dependent recursive averaging (Lin *et al.*, 2003) consistently achieved the highest recognition performance across a variety of background noise scenarios. In combination with the aforementioned noise estimation techniques, the *MMSE log-STSA* gain function clearly outperformed spectral subtraction-based methods in terms of recognition accuracy in all noise conditions. The quality of the estimated binary mask was further analyzed by ROC statistics. This analysis revealed that best results are obtained by a conservative labeling of reliable T-F components, which is reflected in a low false positive rate.

Due to the sparsity of the estimated binary mask especially in conditions with low SNR, the speaker identification is only based on a fraction of the overall feature space. Further improvements can be expected if the decision about reliable and unreliable feature components is softened by using a fuzzy mask (Barker *et al.*, 2000). The current work focused on the recognition of speakers in noisy environments. Future work will also investigate the performance of the proposed system in the presence of reverberation and in scenarios with multiple target speakers.

This chapter is based on:

- (May *et al.*, 2011b): "Simultaneous localization and identification of speakers in noisy and reverberant environments," in *Proceedings of Forum Acusticum*, Aalborg, Denmark, pp. 2121–2126.
- (May *et al.*, 2011c): "Binaural detection of speech sources in complex acoustic scenes," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, pp. 241–244.
- (May *et al.*, 2012b): "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing* **20**(7), pp. 2016–2030.

5

A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation

In this chapter we present a binaural scene analyzer that is able to simultaneously localize, detect and identify a known number of target speakers in the presence of spatially positioned noise sources and reverberation. In contrast to many other binaural cocktail party processors, the proposed system does not require a priori *knowledge about the azimuth position of the target speakers. The proposed system consists of three main building blocks: binaural localization, speech source detection, and automatic speaker identification. First, a binaural front-end is used to robustly localize relevant sound source activity. Second, a speech detection module based on missing data classification is employed to determine whether detected sound source activity corresponds to a speaker or to an interfering noise source using a binary mask that is based on spatial evidence supplied by the binaural front-end. Third, a second missing data classifier is used to recognize the speaker identities of all detected speech sources. The proposed system is systematically evaluated in simulated adverse acoustic scenarios. Compared to state-of-the art MFCC recognizers, the proposed model achieves significant speaker recognition accuracy improvements.*

5.1 Introduction

While being constantly surrounded by a variety of different acoustic sources, among them concurrent speakers and environmental noise, the human auditory system is capable of recognizing a single target speaker and selectively following a conversation (Cherry, 1953, Bronkhorst, 2000). According to Bregman (1990), the underlying perceptual mechanisms that enable the human auditory system to perform *auditory scene analysis* (ASA) can be divided into two stages: First, the acoustic input is decomposed into a number of segments. In a second step, individual segments that are believed to belong to the same acoustic object are grouped to form a coherent stream.

The remarkable capabilities of the human auditory system to process an arbitrary target source in complex acoustic scenes have inspired a new field of research, termed *computational auditory scene analysis* (CASA), that attempts to achieve human performance with computational models by imitating the processing of the human auditory system (Wang and Brown, 2006). Despite extensive research efforts, computer algorithms based on binaural signals are not able to compete with the performance achieved by the human auditory system, and up to this point, computers are only able to perform a very restricted version of auditory scene analysis. In contrast to the human auditory system, which is remarkably robust against environmental noise and variations of acoustic conditions, computational models are usually trained for a particular acoustic scenario, and therefore, any mismatch between the training and the testing condition will decrease performance.

A powerful framework that attempts to overcome the aforementioned limitations by implementing concepts of ASA is the classification with missing, unreliable acoustic information (Cooke *et al.*, 2001), termed missing data (MD) classification, which is able to circumvent the mismatch between training and testing conditions. The acoustic input is first decomposed into individual time-frequency (T-F) units. Based on this two-dimensional segmentation, a so-called binary mask is used to identify whether an individual T-F unit is reliable (i.e. dominated by the target source) or unreliable (i.e. dominated by noise, interfering sources or reverberation). It has been shown that such an ideal binary mask (IBM), where the assignment of reliable and unreliable T-F units is known *a priori*, can substantially improve the recognition of speech (Cooke *et al.*, 2001) and the identification of speakers (Pullella *et al.*, 2008, May *et al.*, 2012a) in noisy conditions. Furthermore, it has been reported that applying an IBM to noisy speech can improve speech intelligibility in challenging acoustic scenarios for both normal

hearing listeners (Brungart *et al.*, 2006, Li and Loizou, 2007), as well as hearing-impaired subjects (Wang *et al.*, 2009). Consequently, the estimation of the ideal binary mask has been suggested to be the main goal of CASA (Wang, 2005).

An important aspect that is exploited by the human auditory system is the fact that it is provided with inputs from the left and the right ears. Given a complex acoustic scene, the human auditory system is able to benefit from the spatial separation between target and interfering sources (Hawley and Litovsky, 2004). By analyzing only the signals reaching both ears, humans can detect and localize a target in the presence of up to five competing sources (Simpson *et al.*, 2006).

There are a number of computational approaches that have used binaural cues in order to estimate the ideal binary mask, either to perform robust speech recognition (Palomäki *et al.*, 2004b, Harding *et al.*, 2006), or to segregate a target source from background noise (Roman *et al.*, 2003). However, an important drawback of these existing systems is that the location of the target source is assumed to be known *a priori*, which is a strong limitation for practical applications. A related area of research has focused on the binaural localization of multiple speech sources in the presence of reverberation (Christensen *et al.*, 2009, Woodruff and Wang, 2010b, May *et al.*, 2011a). In these studies all active sound sources in the acoustic scene were localized without further determining whether the source was speech or background noise, i.e. no inferences were possible about the nature of the sound sources. Thus, for a complex acoustic scene with multiple target speakers and interfering noise sources that are positioned at unknown spatial locations, it is not possible to simultaneously localize and recognize the target speakers with the aforementioned methods.

However, a wide range of applications such as hearing aids and teleconference systems require *a priori* knowledge about the azimuth location of the target sources, e.g. to steer a beamformer or to control processing parameters. Also the human auditory system is able to take advantage of *a priori* knowledge about the spatial configuration of sound sources in complex acoustic scenes. In multi-talker scenarios, a significant performance gain in speech recognition has been reported when the subject's attention was directed towards the spatial location of the target talker (Kidd *et al.*, 2005). Likewise, *a priori* knowledge about the locations of maskers in multi-talker mixtures has been shown to substantially reduce the localization error of speech for humans (Kopčo *et al.*, 2010). Thus, assistive systems which are able to retrieve information about the spatial position of target speakers can potentially be used to guide the attention of human listeners.

This chapter addresses the problem of jointly localizing and recognizing a known number of target speakers in adverse acoustic scenarios based on the analysis of binaural signals. For this purpose, a binaural scene analyzer is proposed that is able to simultaneously localize, detect and identify a predefined number of $\mathcal{S}$ speakers in the presence of reverberation and interfering noise sources that are placed at various spatial positions. As opposed to many other cocktail party processors, a speech detection module is proposed to link the localization and the recognition stage, thus allowing the system to operate without *a priori* knowledge about the azimuth position of the target speakers. The proposed system builds on the previously developed binaural front-end for robust sound source localization presented in Chapter 2, which is used to determine azimuth positions with relevant sound source activity. Based on this initial set of candidate positions, a speech detection module is presented to select azimuth positions that most likely correspond to speech sources. The final stage of the binaural scene analyzer recognizes the speaker identities of all detected speech sources. Therefore, the estimated azimuth position of the speech source is also used to select the *better ear* feature space for recognition, which aims at improving the signal-to-noise ratio (SNR) of the target speaker.

The performance of the proposed binaural scene analyzer is systematically evaluated in simulated multi-source scenarios. The estimated binary mask of the proposed system is compared with two formulations of the ideal binary mask and with a binaural system proposed by Palomäki *et al.* (2004a). Furthermore, speaker recognition experiments are conducted to compare speaker identification accuracy of the proposed system with the performance of MFCC-based recognizers.

The remainder of the chapter is organized as follows. The proposed binaural scene analyzer is described in the next section. Section 5.3 contains details about baseline systems and the evaluation procedure. The experimental results are shown in Section 5.4. Section 5.5 presents concluding remarks and summarizes the chapter.

5.2 Model architecture

The proposed binaural scene analyzer consists of three main building blocks, namely the binaural localization stage ①, the speech detection module ② and the speaker recognition stage ③. The system is shown in Fig. 5.1 and the individual processing stages will be described in detail in the following sections.

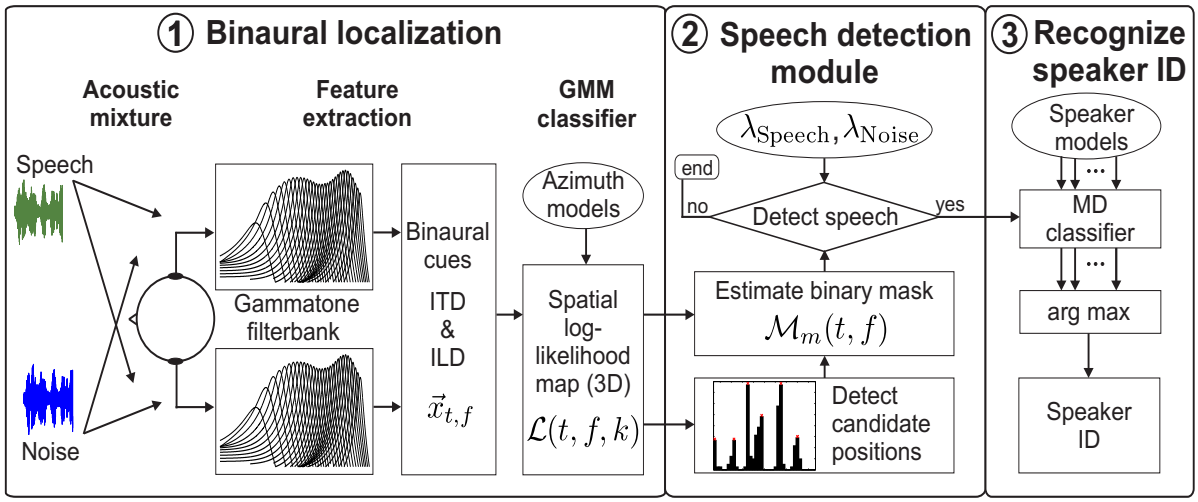


Figure 5.1: Schematic diagram of the proposed binaural scene analyzer. The system is divided into three main stages: binaural localization stage, detection of speech sources, and recognition of speaker identities. See Section 5.2 for details.

5.2.1 Binaural localization

The binaural localization stage ① is based on the previously developed auditory front-end for robust sound source localization described in Chapter 2. The acoustic input to the model is a binaural signal consisting of speech and noise sources that are randomly positioned at unknown spatial locations. The input (sampled at a rate of 16 kHz) is first split into auditory channels using a bank of $F = 32$ gammatone filters with center frequencies equally spaced on the equivalent rectangular bandwidth (ERB) scale (Glasberg and Moore, 1990) between 80 and 5000 Hz. More specifically, a fourth-order, phase-compensated gammatone filterbank (Brown and Cooke, 1994) is used to synchronize the binaural analysis across all gammatone channels at a common time instance. The neural transduction of the inner hair cells is simulated by half-wave rectification and square-root compression. Afterwards, interaural time (ITD) and level differences (ILD) are independently estimated for each auditory channel using overlapping frames of 20 ms with a 10-ms shift. The ITD is estimated by detecting peaks in the normalized cross-correlation function and the ILD is derived by comparing the energy between the left and the right ear signals. These two binaural cues are combined in a two-dimensional binaural feature space

$$\vec{x}_{t,f} = \left\{ \hat{\text{itd}}_f(t), \hat{\text{ild}}_f(t) \right\}, \tag{5.1}$$

where t is the frame number and f indexes the gammatone channel. As shown in Chapter 2, the joint analysis of both binaural cues facilitates the disambiguation of

the ITD cue by the ILD information, which is particularly beneficial in reverberant environments.

Similar to other binaural systems (Roman *et al.*, 2003, Harding *et al.*, 2006), the applied localization model is based on the supervised learning of ITDs and ILDs. A noticeable difference is that the model proposed in Chapter 2 employs a multi-conditional training stage which incorporates the uncertainties of binaural cues that are caused by a variety of acoustic conditions, including changes in the source/receiver configuration, the presence of competing sound sources and the impact of reverberation. In the present study, the localization model is extended to also include different radial distances between the source and the receiver (see Section 5.3.1 for more details regarding the training).

Based on the joint analysis of both binaural cues, the likelihood for each source location is determined by a Gaussian mixture model (GMM) classifier that has learned the azimuth-dependent distribution of ITDs and ILDs. Given a set of K sound source directions $\{\varphi_1, \ldots, \varphi_K\}$ that are modeled by a set of frequency-dependent GMMs $\{\lambda_{f,\varphi_1}, \ldots, \lambda_{f,\varphi_K}\}$, a three-dimensional spatial log-likelihood map can be computed that represents the likelihood that the kth sound source direction is active at time frame t and frequency channel f:

$$\mathcal{L}(t, f, k) = \log p(\vec{x}_{t,f} | \lambda_{f,\varphi_k}), \tag{5.2}$$

where $p(\vec{x}_{t,f} | \lambda_{f,\varphi_k})$ is a Gaussian mixture density consisting of $\mathcal{V}$ weighted component densities. As determined in Chapter 2, a constant GMM model complexity of $\mathcal{V} = 15$ Gaussian components for all gammatone channels and azimuth directions was found to give accurate localization performance. In the present study, a set of $K = 37$ azimuths spaced by 5° within the range of $[-90°, 90°]$ is considered.

5.2.2 Detection of speech sources

The task of the speech detection module ②, as shown in Fig. 5.1, is to use the spatial evidence supplied by the binaural front-end to find candidate positions with relevant sound source activity. From this initial set of candidate positions, a known number of $\mathcal{S}$ sources are selected that are most likely speech by exploiting the distinct spectral characteristics of speech and noise signals.

To this end, first, the evidence about a sound source location is integrated across all F

gammatone channels, and the most probable sound source position is used to reflect the azimuth estimate for each time frame:

$$\hat{\varphi}^{\mathsf{T}}(t) = \underset{1 \le k \le K}{\arg\max} \sum_{f=1}^{F} \mathcal{L}(t, f, k)\,. \tag{5.3}$$

Note that this across-frequency integration of log-likelihoods can be viewed as an implementation of the *straightness weighting* according to (Stern *et al.*, 1988, Shackleton *et al.*, 1992), which makes the detection of sound source positions that are consistently active across multiple frequency channels more likely.

To obtain a reliable estimation of active sources, all frame-based azimuth estimates $\hat{\varphi}^{\mathsf{T}}$ are pooled together over the entire mixture to form an azimuth histogram $H[k]$. This implies that the sound source positions are stationary throughout the time interval over which the histogram is calculated. $H[k]$ represents the number of azimuth estimates that are assigned to the kth sound source direction. Peaks within this histogram indicate azimuth directions with relevant sound source activity and the corresponding histogram bin indices are used to form an initial set of M speech source candidate positions $L = \{\ell_1, \ldots, \ell_M\}$. Each bin index ℓ_m corresponds to a local peak in the azimuth histogram.

Based on such a histogram, however, it is not possible to decide whether the detected activity corresponds to a speech source or to interfering noise. Nevertheless, assuming that all sources are spatially separated, the spatial information can be used to determine and isolate the contribution of individual sound sources on a T-F basis. To achieve this, the spatial log-likelihood map $\mathcal{L}(t, f, k)$ is used to estimate a binary mask $\mathcal{M}_m(t, f)$ for each of the M candidate positions by grouping T-F units according to common azimuth locations. More specifically, for each T-F unit the most likely position among all $m = 1, \ldots, M$ candidate positions is determined, and the individual T-F unit is added to the corresponding mask:

$$\mathcal{M}_m(t, f) = \begin{cases} 1 & \text{if } m = \underset{k \in L}{\arg\max}\, \mathcal{L}(t, f, k) \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

Rather than considering all K possible sound source directions, the candidate selection effectively reduces the number of alternatives per T-F unit, which results in a more dense binary mask.

Based on this mask, missing data classification (Cooke *et al.*, 2001) is performed to decide whether the corresponding source type is speech or noise. Our prior work has

shown that the mean absolute deviation of the smoothed envelope is a good descriptor for detecting speech sources in the presence of noise and reverberation (see Appendix 5.A for more details). In order to compute this feature, first a smoothed envelope e_f is obtained by low-pass filtering the half-wave rectified output of the fth gammatone channel with a time constant of 10 ms. Then, the mean absolute deviation of the smoothed envelope e_f is computed over B adjacent samples with a shift of O samples

$$\mathcal{F}_{\text{AD}}(t,f) = \frac{1}{B}\sum_{b=0}^{B-1} \left| e_f(tO+b) - \bar{e}_f \right|, \tag{5.5}$$

where $\bar{e}_f$ refers to the mean envelope of frame t. The feature reflects the amount of fluctuation and its magnitude is lower for speech-dominant T-F units compared to units that are corrupted by noise. Thus, it is possible to apply *bounded marginalization* where the true value of the unreliable feature components is bounded between zero and the observed feature magnitude (El-Maliki and Drygajlo, 1999, Cooke *et al.*, 2001, Raj and Stern, 2005). Signals of the left and the right ears are averaged prior to feature extraction. Similar to the binaural front-end, the processing is based on 20-ms frames with a shift of 10 ms.

For classification two GMMs with 32 Gaussian components and diagonal covariance matrices are trained to approximate the probability distribution of the feature space $\mathcal{F}_{\text{AD}}(t,f)$ that is extracted separately for speech and noise files. The first GMM, denoted as speech model λ_{Speech}, is trained with features based on a large pool of monaural speech files selected from the speech separation challenge (SSC) database (Cooke and Lee, 2006). The second GMM, termed noise model λ_{Noise}, reflects the feature distribution of all types of noise files drawn from the NOISEX database (Varga *et al.*, 1992). The GMMs are initialized by 20 iterations of the k-means clustering algorithm (Lloyd, 1982) and afterwards refined by the EM algorithm (Dempster *et al.*, 1977) using a stopping criterion of $1e^{-5}$ with a maximum of 300 iterations. About 29 minutes of training material is used for each GMM. To compensate for the mismatch between training (λ_{Speech} and λ_{Noise} are trained with clean signals) and testing (the speech detection module is applied in noisy and reverberant conditions), a missing data compatible normalization scheme is employed. Therefore, a frequency-dependent compensation factor is derived by computing the mean of the most intense feature values that are classified as being reliable by the estimated binary mask (Palomäki *et al.*, 2004a).

Given the binary mask $\mathcal{M}_m(t,f)$, the feature space $\mathcal{F}_{\text{AD}}(t,f)$ and the trained speech and noise models (λ_{Speech} and λ_{Noise}), the log-likelihood ratio p_m reflecting the evidence

for the mth speech source candidate can be determined by

$$p_m = \log\left(\frac{p\left(\mathcal{F}|\lambda_{\text{Speech}}\right)}{p\left(\mathcal{F}|\lambda_{\text{Noise}}\right)}\right). \tag{5.6}$$

A speech source is detected if the log-likelihood ratio is larger than a predefined threshold θ

$$p_m \begin{cases} \geq \theta & \text{accept} \quad \lambda_{\text{Speech}} \\ < \theta & \text{reject} \quad \lambda_{\text{Speech}}. \end{cases} \tag{5.7}$$

Selecting the optimal decision threshold is a non-trivial task because it is influenced by a variety of parameters; among them the number of speech and noise sources in the acoustic mixture, the SNR between all sources and the amount of reverberation. In preliminary experiments we found that a decision threshold of $\theta = 0$ performed well for a wide range of acoustic scenarios. Based on this criterion, all active sound sources are classified to be either speech or noise. After classification a set of $\hat{\mathcal{S}}$ log-likelihood ratios $\left\{p_1^{\text{Speech}}, \ldots, p_{\hat{\mathcal{S}}}^{\text{Speech}}\right\}$ is available that specifies the evidence of all detected speech sources. Moreover, a new set of histogram bin indices $L^{\text{Speech}} = \left\{\ell_1^{\text{Speech}}, \ldots, \ell_{\hat{\mathcal{S}}}^{\text{Speech}}\right\}$ is available, which is a subset of L, and reflects the individual bin positions of all detected speech sources in the azimuth histogram.

Reflections and the interaction of multiple competing sound sources can cause the azimuth histogram to have numerous local peaks. As a consequence, the number of detected speech sources $\hat{\mathcal{S}}$ might be larger than the number of *a priori* known speech sources $\mathcal{S}$. Thus, the final step is to select the $\mathcal{S}$ most likely speech sources. Instead of using the evidence from the missing data classifier directly for selection, we found that it is advantageous to apply an azimuth-dependent weight to the log-likelihood ratio of each detected speech source to account for the fact that speech sources that are more frequently represented in the azimuth histogram are more likely to reflect the real position of the speech sources (see Appendix 5.A for more details). The applied weight reflects the *a priori* probability that the corresponding source was active in the acoustic scene, and is approximated by the normalized azimuth histogram. The weighted log-likelihood ratio for the nth speech source is given by

$$p_n^{\text{Speech,W}} = p_n^{\text{Speech}} + \underbrace{\log\left(H\left[\ell_n^{\text{Speech}}\right] / \sum_k H\left[k\right]\right)}_{\text{azimuth weight}}. \tag{5.8}$$

Finally, the set of weighted log-likelihood ratios of all detected speech sources is rearranged

in descending order

$$\left\{ p_1^{\text{Speech,W}} \geq p_2^{\text{Speech,W}} \geq \ldots \geq p_{\hat{\mathcal{S}}}^{\text{Speech,W}} \right\} \tag{5.9}$$

and the azimuth locations corresponding to the highest $\mathcal{S}$ values are selected to represent the estimated speech source positions.

When using the frame-based azimuth estimates according to Eq. (5.3) for the initial selection of source candidate positions, it is required that a sufficiently large number of frames is dominated by the target sources which should be detected. However, in conditions where the SNR between the target speakers and the interfering noise sources is very low, or even negative, very few target source dominated frames may be found. Thus, when $\hat{\mathcal{S}} < \mathcal{S}$, the histogram of the frame-based azimuth estimates $\hat{\varphi}^{\mathsf{T}}$ was apparently dominated by locations corresponding to noise sources, and the histogram did not reflect the locations of all present speech sources. Indeed, it has been shown that spectro-temporal regions dominated by speech tend to be sparse in the presence of noise (Cooke, 2006). Thus, whenever $\hat{\mathcal{S}} < \mathcal{S}$, the azimuth histogram $H[k]$ is recomputed using the azimuth estimates on a T-F basis

$$\hat{\varphi}^{\mathsf{TF}}(t, f) = \underset{1 \leq k \leq K}{\arg\max}\, \mathcal{L}(t, f, k)\,. \tag{5.10}$$

Again, all local peaks within this histogram are considered as initial speech source candidates, and the missing data masks corresponding to these locations are estimated and fed to the missing data classifier to determine the most likely speech source positions (involving the aforementioned steps Eq. (5.4)-(5.9)). The rationale behind Eq. (5.10) is that speech source positions that were not resolved on a frame-by-frame basis can potentially be recovered when using azimuth estimates on a T-F level. In Section 5.4.1, the impact of alternative methods to select the set of speech source candidate positions is analyzed in order to justify the proposed frame-based selection of speech source candidates with the possibility to switch to time-frequency-based processing.

The proposed speech detection module is illustrated in Fig. 5.2 for two different acoustic scenes. Figure 5.2(A) shows the detection of one speech source in the presence of three factory noise sources in an anechoic room. Despite the presence of four competing sources, the estimated binary masks are quite similar to the ideal binary masks based on the *a priori* SNR. Figure 5.2(B) presents the detection of two speech sources in the presence of three factory noise sources in a reverberant room ($T_{60} = 0.29\,\text{s}$). In comparison to the anechoic scenario with one target source, the estimated binary masks of the two target

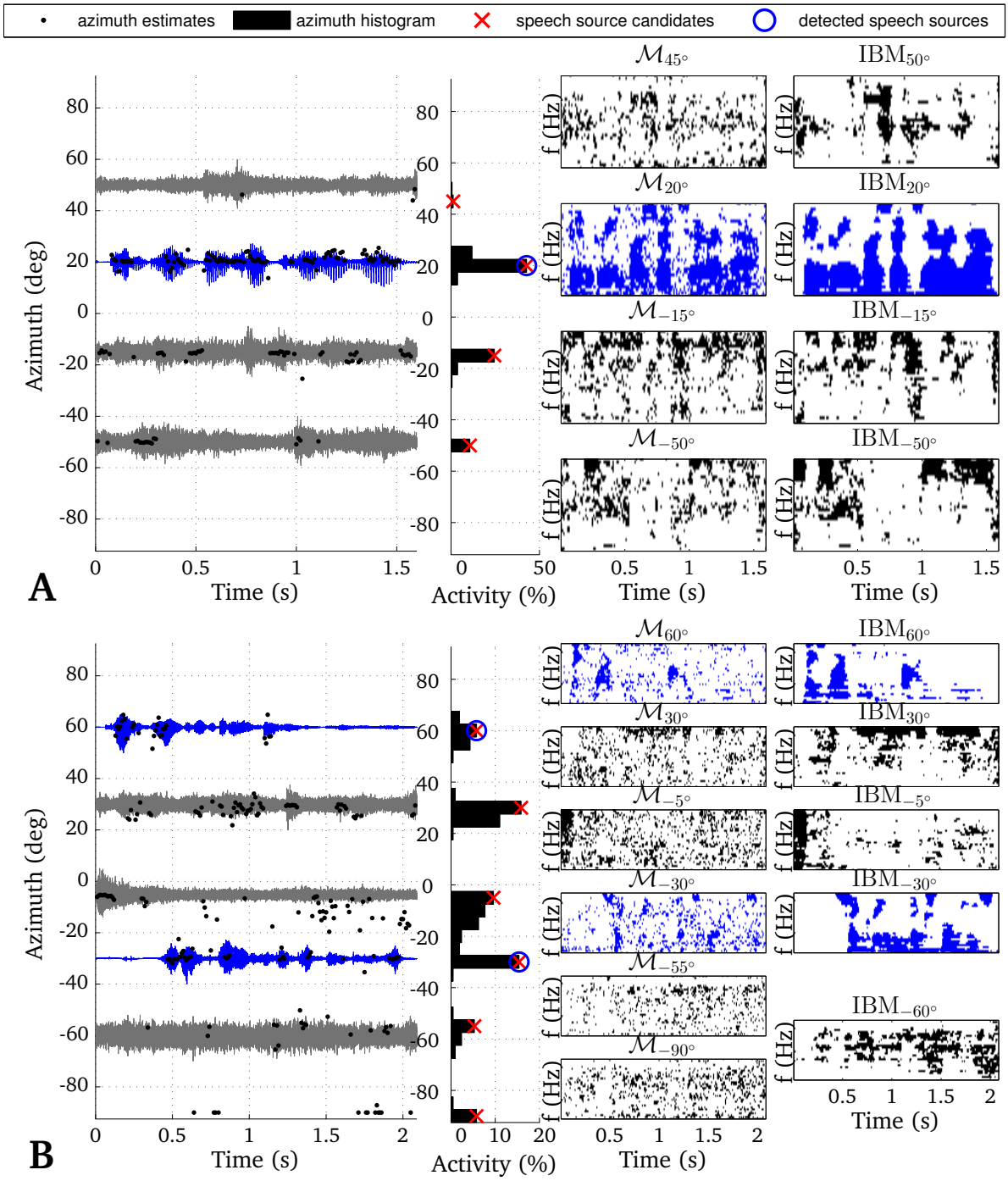


Figure 5.2: Demonstration of the speech detection module for two different acoustic scenes: (A) Detection of one speaker (20°) in the presence of three factory noise sources (50°, −15° and −50°, SNR = 0 dBA) in an anechoic room. (B) Detection of two speakers (60° and −30°) in the presence of three factory noise sources (30°, −5° and −60°, SNR = 0 dBA) in a reverberant room ($T_{60} = 0.29$ s). Each of the two subplots consist of three panels: The left panel shows the acoustic signals and the frame-based azimuth estimates, the middle panel depicts the azimuth histogram and the detected speech source candidates, and the right panels present the estimated binary masks ($\mathcal{M}$) of all candidate positions and the ideal binary masks (IBM) corresponding to all real sound source positions.

sources are more noisy due to the impact of reverberation.

5.2.3 Automatic speaker recognition

The final stage ③ of the proposed binaural scene analyzer (see Fig. 5.1) has the function of recognizing the speaker identity of the detected speech sources from a set of stored speaker models. For this purpose, a second missing data classifier based on bounded marginalization is supplied with the binary masks that were estimated by the speech detection module (see Section 5.2.2). The recognition of speakers is performed with spectral features reflecting the energy of individual frequency channels (Cooke *et al.*, 2001). Therefore, a map of auditory nerve firing rates, a so-called *ratemap*, is computed by averaging the smoothed envelope e_f (see Section 5.2.2) over B adjacent samples with a shift of O samples and subsequent cube-root compression

$$\mathcal{F}_{\mathcal{R}}(t,f) = \left(\frac{1}{B} \sum_{b=0}^{B-1} e_f(tO+b) \right)^{1/3} . \tag{5.11}$$

Speaker models are represented by 128-mixture GMMs with diagonal covariance matrices. In comparison to a conventional GMM-based missing data recognizer, Chapter 4 demonstrated that the combination of missing data recognition with universal background model (UBM)-based adaptation of speaker models (Reynolds *et al.*, 2000) yields substantial improvements in highly non-stationary noise scenarios. However, in order to apply this scheme to reverberant multi-source environments, a modification is required to account for the mismatch between the speaker models trained with monaural and anechoic speech, and the observed spectral features that are affected by HRTF filtering and reverberation. Similar to the speech detection module, a missing data compatible normalization scheme is required. The normalization scheme proposed in Palomäki *et al.* (2004a) was developed in the context of automatic speech recognition and applies a spectral normalization factor that is independently derived for each frequency channel. In contrast to automatic speech recognition, which is generally speaker-independent, an automatic speaker identification system exploits the frequency-dependent spectral variations across different speakers. We found that it is beneficial in terms of speaker recognition performance to average the normalization factor proposed in Palomäki *et al.* (2004a) across adjacent frequency channels prior to normalization. In this way, a similar normalization factor is applied to neighboring frequency channels and differences between adjacent channels will be related to speaker-specific variations. A sliding triangular window of size 7 is used for

averaging the normalization factor (these parameters were derived empirically based on pilot experiments).

For the adaptation of speaker models, two gender-dependent UBMs are used to represent the speaker-independent distribution of the ratemap feature. The two UBM models are initialized with 20 iterations of the k-means clustering algorithm (Lloyd, 1982) and further trained with the EM algorithm (Dempster *et al.*, 1977) using a stopping criterion of $1e^{-5}$ at a maximum of 300 iterations. Speaker-dependent models are obtained by adapting the well-trained UBM parameters to the speaker-dependent speech material. Therefore, first the gender selection is performed by selecting the UBM which shows the highest probabilistic alignment with the speaker-dependent material. Secondly, as suggested in Reynolds *et al.* (2000), only the mean vectors of the UBM are adapted using a relevance factor of 16.

In order to benefit from the fact that the binaural scene analyzer is provided with two acoustic signals from both ears, the estimated positions of speech sources are utilized to implement a better-ear selection of the feature space. Therefore, ratemaps are always computed for both the left $\mathcal{F}_{\mathcal{R}}^{\mathrm{L}}(t,f)$ and the right ear signals $\mathcal{F}_{\mathcal{R}}^{\mathrm{R}}(t,f)$. For recognition, the ratemap based on the ear signal that is closest to the estimated azimuth position of the corresponding speech source is selected individually for each detected speech source

$$\mathcal{F}_{\mathcal{R},n}(t,f) = \begin{cases} \mathcal{F}_{\mathcal{R}}^{\mathrm{L}}(t,f), & \varphi_{\ell_n^{\mathrm{Speech}}} < 0 \\ \mathcal{F}_{\mathcal{R}}^{\mathrm{R}}(t,f), & \varphi_{\ell_n^{\mathrm{Speech}}} > 0 \\ \frac{1}{2}\left(\mathcal{F}_{\mathcal{R}}^{\mathrm{L}}(t,f) + \mathcal{F}_{\mathcal{R}}^{\mathrm{R}}(t,f)\right), & \varphi_{\ell_n^{\mathrm{Speech}}} = 0. \end{cases} \tag{5.12}$$

This approach aims at increasing the SNR between the target and interfering sources, and the underlying effect is referred to as the *better ear* effect (Shinn-Cunningham *et al.*, 2001).

5.3 Evaluation setup

5.3.1 Acoustic mixtures

Acoustic sources were simulated by convolving monaural audio files with binaural room impulse responses (BRIRs). BRIRs were constructed by combining head related transfer functions (HRTFs) of a KEMAR artificial head taken from the MIT database (Gardner

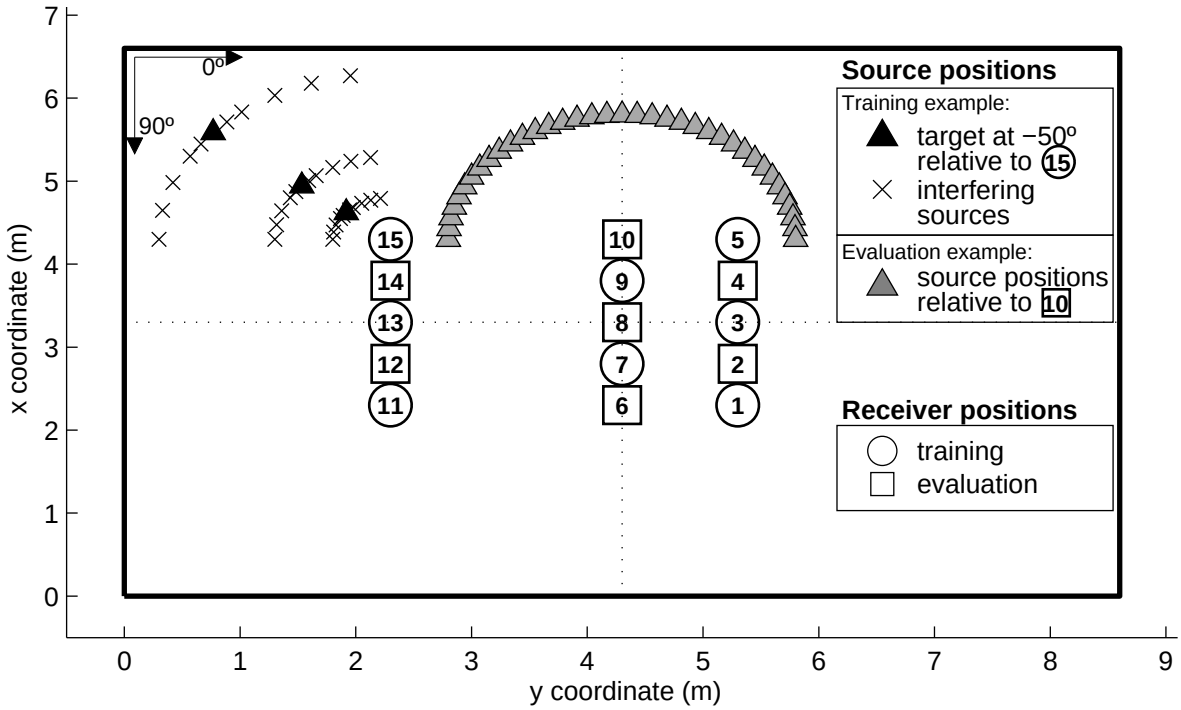


Figure 5.3: Schematic diagram of the room dimensions with all receiver positions used for training (circles) and evaluation (squares). Note that the training stage of the localization model incorporated three radial distances (0.5 m, 1 m and 2 m) between the receiver and the target positions as exemplarily shown for receiver position 15, which were different from the radial distance (1.5 m) used for evaluation. See Section 5.3.1 for details.

and Martin, 1994) with room impulse responses (RIRs) that were simulated according to the image-source model (Allen and Berkley, 1979). More specifically, the *roomsim* simulation software (Schimmel *et al.*, 2009) was used for that purpose. The receiver (KEMAR) was placed at various positions in a simulated room of dimensions 6.6 x 8.6 x 3 m at 1.75 m above the ground, as depicted in Fig. 5.3.

The binaural localization model was trained with BRIRs corresponding to eight training positions (different from those used for evaluation). In order to incorporate the same amount of uncertainty to the binaural cues of all gammatone channels, we intentionally chose a frequency-independent reverberation time of $T_{60} = 0.5$ sec for all training positions. It has been shown in Chapter 2 that this training enables the localization model to generalize to unseen absorption characteristics. Note that a different, frequency-dependent absorption characteristic is used for evaluation. The training of the binaural model also incorporated the effect of interfering sources. This is exemplarily shown in Fig. 5.3 for the training position (15). In order to train the binaural model for one particular sound source direction, the training consists of a target source placed at the corresponding azimuth (denoted by ▲) and an interfering source (indicated by ×) positioned at ±5°,

$\pm 10°$, $\pm 20°$, $\pm 30°$ and $\pm 40°$ with respect to the target azimuth. In addition to the training procedure described in Chapter 2, the multi-conditional training is extended to consider three different radial distances (0.5 m, 1 m and 2 m) between the source and the receiver. The corresponding direct-to-reverberation ratios (DRRs) are 7.3 dB, 1.4 dB and -4.8 dB.

For evaluation the receiver was randomly placed at seven evaluation positions using a radial distance of 1.5 m. To systematically evaluate the impact of reverberation, the surface *Acoustic plaster* was selected for all room boundaries within the room simulation software (Schimmel *et al.*, 2009) to create a specific, frequency-dependent absorption characteristic. Note that this absorption characteristic was different from the one used to train the localization model. Speech and noise sources (different from the material used to train the speech detection module) were randomly positioned within the azimuth range of $[-90°, 90°]$, while having an angular distance of at least $15°$ to the nearest source. A set of 1400, 4-source mixtures (one speech source) and 600, 5-source mixtures (two speech sources) were generated for each SNR condition. Mixtures had an average length of 1.83 s. The SNR was adjusted by comparing the A-weighted energy of all binaural speech sources with the A-weighted energy of all binaural noise sources. This weighting was applied to ensure that the SNR is adjusted in the frequency range that is relevant for speech. The design of the A-weighting filter was implemented according to (American National Standards Institute, 1983). To prevent that the energy of speech is underestimated due to silent parts, an energy-based voice activity detector (VAD) was used. A frame was considered to contain relevant speech activity, if its energy level was within 40 dB of the global maximum. The level between multiple speech or noise sources was always set equal.

For a given multi-source mixture, the localization error in degrees was evaluated by comparing the positions of the detected speech sources to their real positions.

Speaker recognition performance was evaluated on a closed set of speakers that were randomly selected from the SSC database (Cooke and Lee, 2006). The SSC database consists of 17000 clean utterances spoken by 34 speakers (18 males and 16 females). To ensure that there is no overlap between the speech material used for training and testing, the SSC database was randomly split into two equal sized sets consisting of 8500 files (250 sentences per speakers). The first half was used to train the two gender-dependent UBMs. Also, 950 sentences (about 29 minutes) were randomly selected from the first half to train the speech model λ_{Speech} of the speech detection module. The second half

of the SSC database was used to perform the speaker recognition experiments reported in Section 5.4, involving the training of speaker-specific models using UBM adaptation and the evaluation of the speaker identification accuracy. Because the amount of available speech material is often a limitation for practical applications, the speech material was restricted to 25 randomly selected sentences per speaker. For each speaker, 18 sentences were randomly chosen to train the speaker model and the remaining 7 sentences were used for evaluation. Because this randomized selection of training and testing material will to some extent influence the evaluated speaker identification accuracy, results are reported as the mean identification accuracy over a series of 20 simulations, each containing a new set of randomly selected speakers. Note that the speaker identification accuracy was measured on an utterance level.

5.3.2 Baseline systems

To serve as a baseline for recognition performance, a conventional robust speaker recognition system was trained with a feature vector consisting of 13 static MFCC coefficients including the 0th order coefficient and first order temporal derivatives (a total of 26 features). The static MFCC coefficients were computed using the RASTAMAT toolbox (Ellis, 2005). Parameters[1] were chosen to reproduce MFCC coefficients according to the hidden markov models toolkit (HTK). The delta coefficients were computed using a first-order orthogonal polynomial fit over a window of 5 frames (Soong and Rosenberg, 1988). For improved robustness, cepstral mean and variance normalization (CMVN) was performed, where the feature statistics are measured over the duration of one utterance (Openshaw and Mason, 1994, Tibrewala and Hermansky, 1997).

The first method, termed *MFCC Mono*, extracted the MFCC coefficients by averaging the signals of the left and the right ear. In addition, a recognizer was implemented that computed the MFCC feature vector, as described above, for both the left and the right ear signals. Based on the binaural analysis, the feature vector with the higher SNR (the *better ear*) was selected for recognition according to the estimated location of each target speech source, individually. This algorithm is referred to as *MFCC Binaural*. The third MFCC-based recognizer, denoted by *MFCC Binaural NR*, combined the previously described better-ear selection and a noise reduction (NR) stage. In Chapter 4, we

1 melfcc(in, fs, 'lifterexp', -22, 'nbands', 20, 'dcttype', 3, 'maxfreq', 8000, 'fbtype', 'htkmel', 'sumpower', 0, 'wintime', 20e-3, 'hoptime', 10e-3, 'numcep', 13). For details see http://labrosa.ee.columbia.edu/matlab/rastamat/mfccs.html

analyzed the impact of a variety of noise estimation and noise reduction schemes on speaker identification performance. Based on these findings, noise reduction is performed prior to MFCC extraction by recursively averaging the noise power spectrum (Hirsch and Ehrlicher, 1995) in combination with the *MMSE log-STSA* gain function (Ephraim and Malah, 1985). The noise estimation and the noise reduction is applied after the speech detection module and is performed independently for the left and the right ear signals.

Similar to the MD-based recognizer, speaker models of all MFCC-based recognizers are represented by 128-mixture GMMs and were adapted from two gender-dependent UBMs. Recognition of two target speakers is performed by accumulating the frame-based likelihoods over the entire test sequence and selecting the two most likely speaker identities.

Finally, a comparison is made with a recently proposed co-channel speaker identification system based on adapted GMMs (Saeidi *et al.*, 2010). This approach, denoted as *MFCC Co-channel*, combines frame-level likelihood scores and a Kullback-Leibler divergence (KLD) distance measure to find the most likely speaker identity on a frame-by-frame basis in co-channel scenarios. A UBM with 128 Gaussian components is trained with MFCC coefficients extracted from two-talker mixtures. For training, two-talker mixtures are created from the first half of the SSC database by mixing two sentences from different speakers at one of the following seven signal-to-signal ratio (SSR) levels: $\{-9, -6, -3, 0, 3, 6, \infty\}$ dB. For each SSR level, a set of 8500 co-channel mixtures is created, giving a total of 59500 audio files for UBM training. Speaker-dependent GMMs are adapted from the UBM by mixing 18 randomly selected training sentences for each speaker (see Section 5.3.1 for details) with 18 files from other speakers, again at different SSR levels. Because multiple talkers are always set to have equal power in the experiments, only the following SSR levels are considered for adaptation: $\{0, \infty\}$ dB.

5.3.3 Ideal binary mask

To evaluate the upper performance limit of missing data recognition systems, an ideal binary mask is commonly introduced that represents the ideal segmentation of the spectral feature space according to the contribution of all occurring sound sources. Studies related to recognition tasks in noisy conditions often utilize the ideal binary mask based on the *a priori* SNR between the target and the noise source (Cooke *et al.*, 2001). Note that the

SNR-based ideal binary mask, denoted as *IBM SNR*, is only able to segregate the target signal from the background noise, but the effect of reverberation is not taken into account. Another formulation of the ideal binary mask is to select only those T-F units where the spectral energy of the noisy and reverberated speech is within 3 dB of the spectral energy of the clean target source (Cooke *et al.*, 2001, Palomäki *et al.*, 2004b), taking into account both the effect of background noise and the impact of reverberation. The ideal binary mask based on this spectral criterion will be referred to as *IBM SPEC*. To create this mask for a given binaural mixture, the observed spectral energy of the noisy and reverberated speech signal is compared with the spectral energy of the clean speech signal, that has been convolved with the same HRTF but in anechoic conditions. In this way, the azimuth-dependent HRTF filtering does not bias the mask computation. Both definitions of the ideal binary mask will be used to evaluate the mask estimation performance of the proposed binaural front-end in the experimental section.

5.4 Experiments

We performed a series of localization and speaker recognition experiments to evaluate the proposed binaural scene analyzer in simulated adverse acoustic conditions. The first two experiments are aimed at evaluating the estimated localization information of the proposed system. Whereas the first experiment evaluates the ability of the speech detection module to localize the azimuth of speakers in multi-source scenarios, the second experiment analyzes the mask estimation performance of the binaural front-end. Therefore, the estimated binary mask is compared with two different formulations of the ideal binary mask and with the model proposed by Palomäki *et al.* (2004b). Both the estimated location of target speakers and the estimated binary masks are used in the third and fourth experiment to recognize the identity of speakers in complex acoustic scenes. More specifically, the third experiment is using a reduced set of 10 speakers to study the influence of the number of interfering noise sources on speaker recognition performance. Furthermore, the performance of the proposed system is compared with several baseline systems based on MFCC coefficients (see Section 5.3.2). Based on this comparison, the best performing baseline systems and the proposed method are used in the fourth experiment to recognize one and two simultaneously active target speakers using the full set of 34 speakers. In the last experiment, the final goal of the study is addressed by jointly analyzing the combined localization and speaker recognition performance of the proposed method using a confusion matrix.

5.4.1 Experiment 1: Speaker localization performance

The first experiment is analyzing the ability of the speech detection module to determine the azimuth position of a predefined number of speech sources from a set of candidate positions. Furthermore, the influence of the speech source candidate selection on speech detection performance is systematically investigated. Therefore, we compare the proposed candidate selection as described in Section 5.2.2 with two alternative approaches. The first alternative, indicated by $\hat{\varphi}^{\mathrm{TF}}$, determines candidate positions by detecting local peaks in the azimuth histogram which is solely based on azimuth estimates derived on a T-F level (according to Eq. (5.10)). The second alternative, denoted by $\hat{\varphi}^{\mathrm{T}}$, exclusively operates on frame-based azimuth estimates according to Eq. (5.3) to compute the azimuth histogram. Again, azimuth positions corresponding to all local peaks in the histogram are considered as candidate positions.

The SNR-dependent localization error in degrees of the detected speech sources is presented in Tab. 5.1 for all three candidate selection methods. It can be seen that the T-F-based selection of speech source candidates produces the highest error rates. When using the frame-based selection $\hat{\varphi}^{\mathrm{T}}$, a considerable improvement is achieved. This improvement can be attributed to the fact that the frame-based candidate selection integrates evidence of sound source activity across all frequency channels, which effectively increases the reliability of the resulting localization estimate. As a result, the number of candidate positions is reduced, which consequently reduces the number of alternatives per T-F unit in Eq. (5.4), thus increasing the density of the estimated binary mask which allows for a more accurate detection of the speech sources.

The proposed candidate selection, which combines both the frame-based and the T-F-based selection, can further improve the localization performance at low SNRs. In particular, in conditions with one interfering noise source, it appears that switching from frame-based to T-F-based processing can to some extent recover the position of speech sources, therefore, improving the localization performance by about 5 °. Regarding mixtures with one speaker, the average localization error is below 3 ° for SNRs as low as 0 dBA. When two speakers are simultaneously talking, the azimuth of both speakers is estimated within 5 ° accuracy for SNRs as low as 5 dBA. At lower SNRs, the error is noticeably increased.

The general trend that the localization error decreases with increasing number of noise sources can be attributed to the SNR definition described in Section 5.3.1, which compares

Table 5.1: SNR-dependent localization error of speech sources in degrees for binaural mixtures consisting of one and two speakers in the presence of interfering noise sources and reverberation.

$T_{60} = 0.29\,\mathrm{s}$	Candidate selection	SNR in dBA (factory noise) −5	0	5	10	20	Mean
1 speaker, 1 noise source	$\hat{\varphi}^{\mathrm{TF}}$	23.7	10.9	3.6	1.7	0.5	8.1
	$\hat{\varphi}^{\mathrm{T}}$	17.0	2.2	0.6	0.1	0.0	4.0
	Proposed	12.0	1.6	0.4	0.1	0.0	2.8
1 speaker, 2 noise sources	$\hat{\varphi}^{\mathrm{TF}}$	19.5	6.1	1.7	0.4	0.6	5.7
	$\hat{\varphi}^{\mathrm{T}}$	12.2	1.4	0.4	0.1	0.0	2.8
	Proposed	11.4	1.3	0.4	0.1	0.0	2.6
1 speaker, 3 noise sources	$\hat{\varphi}^{\mathrm{TF}}$	16.7	4.1	1.1	0.5	0.4	4.6
	$\hat{\varphi}^{\mathrm{T}}$	9.8	1.5	0.3	0.1	0.1	2.4
	Proposed	9.2	1.5	0.3	0.1	0.1	2.2
2 speakers, 1 noise source	$\hat{\varphi}^{\mathrm{TF}}$	31.0	24.2	15.5	9.4	3.4	16.7
	$\hat{\varphi}^{\mathrm{T}}$	32.3	12.0	3.1	1.5	0.9	10.0
	Proposed	27.8	12.1	3.0	1.5	0.9	9.1
2 speakers, 2 noise sources	$\hat{\varphi}^{\mathrm{TF}}$	31.5	21.2	13.5	7.3	3.6	15.4
	$\hat{\varphi}^{\mathrm{T}}$	29.4	12.7	4.6	1.5	1.0	9.8
	Proposed	27.4	12.6	4.6	1.5	1.0	9.4
2 speakers, 3 noise sources	$\hat{\varphi}^{\mathrm{TF}}$	29.2	18.1	11.7	6.9	4.1	14.0
	$\hat{\varphi}^{\mathrm{T}}$	26.2	11.5	4.4	2.3	0.4	9.0
	Proposed	25.8	11.4	4.4	2.3	0.4	8.9

the overall energy of all speech sources with the energy of all noise sources. With increasing number of noise sources, the noise energy is distributed across multiple directions, which increases the relative localization dominance of the speech source, thus allowing for a more accurate prediction of its azimuth position.

5.4.2 Experiment 2: Evaluation of the IBM estimated by the binaural front-end

The second experiment is used to verify the ability of the binaural front-end to estimate the ideal binary mask. The quality is systematically evaluated in terms of receiver operating characteristics (ROC) analysis (Fawcett, 2006). For this analysis, the estimated binary mask is compared with the ideal binary mask by calculating the percentage of correctly identified T-F units which are dominated by the target signal (true positive rate) and the percentage of misclassified T-F units which are dominated by interfering sources (false positive rate). For comparison, we also provide the difference between the true positive rate and the false positive rate, because it has been shown that this metric is highly

correlated with human speech intelligibility (Kim *et al.*, 2009). For the ROC analysis, we selected the *IBM SPEC* to represent the reference mask, which accounts for both the effect of interfering noise and reverberation. Moreover, the percentage of labeled T-F units is reported to reflect the amount of information that is available to the MD classifier. In addition to the ROC analysis, the corresponding speaker identification accuracy for a set of 10 speakers is provided to illustrate the implication of the true positive rate and the false positive rate on speaker recognition performance.

We compared the performance of the proposed mask estimation technique with two definitions of the ideal binary mask (see Section 5.3.3) and with a binaural front-end proposed by Palomäki *et al.* (2004b). In addition, to study the influence of the speech detection module on mask estimation performance, the proposed method is supplied with *a priori* knowledge about the azimuth positions of the target and the interfering sources.

The model proposed in Palomäki *et al.* (2004b) extracts both ITD and ILD cues in individual frequency channels. The ITD cue is warped by a table look-up to its corresponding azimuth and subsequently used to group T-F units according to common azimuth. The required azimuth locations of the target and interfering sources are provided by the speech detection module (see Section 5.2.2). The ILD cue is used to remove T-F units from the estimated binary mask where the ILD estimate is not consistent with the azimuth of the sound source, which is derived from the ITD analysis. The expected azimuth-specific ILD template is precomputed for all frequency channels above 2800 Hz.

We evaluated two variants of the model, first the combined ITD and ILD analysis denoted as *Palomäki ITD & ILD*, and *Palomäki ITD* which solely relies on ITD analysis. Note that the original model proposed in Palomäki *et al.* (2004b) includes an inhibition mechanism that emphasizes acoustic onsets. Whereas this inhibition might be advantageous for the azimuth estimation of sound sources on an utterance level (as indicated in Fig. 5 in Palomäki *et al.* (2004b)), preliminary tests revealed, however, that the resulting mask was very sparse and the model performed best in terms of speaker identification performance when the inhibition mechanism was switched off by setting the inhibition gain to zero. Apart from this modification, all other model parameters were chosen according to the recommendations of the authors (Palomäki *et al.*, 2004b).

The evaluation of all tested mask estimation methods is performed for binaural mixtures with one target speaker and one interfering factory noise source. The experimental results presented in Tab. 5.2 and Tab. 5.3 correspond to anechoic ($T_{60} = 0\,\mathrm{s}$) and

Table 5.2: Mask estimation performance (true positive (TP) rate, false positive (FP) rate, TP-FP rate and the number of labeled T-F units) and speaker identification (SID) accuracy in % for various methods in anechoic conditions ($T_{60} = 0\,\text{s}$).

T_{60} in s	Methods	%	SNR in dBA			Mean
			−5	0	5	
0	IBM SPEC	TP rate	100	100	100	100
		FP rate	0	0	0	0
		TP-FP rate	100	100	100	100
		T-F units	26.1	37.2	48.8	37.4
		SID accuracy	97.1	98	98.4	97.8
	IBM SNR	TP rate	64.9	72.3	78.2	71.8
		FP rate	0	0	0	0
		TP-FP rate	64.9	72.3	78.2	71.8
		T-F units	16.9	26.9	38.2	27.3
		SID accuracy	92.5	97.8	98.9	96.4
	Proposed *a priori*	TP rate	59	66.3	72.5	65.9
		FP rate	2.4	3.1	3.9	3.1
		TP-FP rate	56.6	63.2	68.6	62.8
		T-F units	17.2	26.6	37.4	27.1
		SID accuracy	90.5	97.8	98.9	95.7
	Proposed	TP rate	54.8	61.6	68.1	61.5
		FP rate	3.2	3.7	4.1	3.7
		TP-FP rate	51.6	57.9	64.0	57.8
		T-F units	16.6	25.2	35.4	25.8
		SID accuracy	84.8	94.9	97.6	92.4
	Palomäki ITD	TP rate	45	53.8	61	53.3
		FP rate	4.5	6.4	8.3	6.4
		TP-FP rate	40.5	47.4	52.7	46.9
		T-F units	15.1	24	34.1	24.4
		SID accuracy	79.5	93.4	97.6	90.2
	Palomäki ITD & ILD	TP rate	42.5	49.2	54.5	48.7
		FP rate	3.5	4.9	6.3	4.9
		TP-FP rate	39.0	44.3	48.2	43.8
		T-F units	13.7	21.4	29.8	21.6
		SID accuracy	67.1	80.9	90.3	79.4

reverberant ($T_{60} = 0.29\,\text{s}$) conditions, respectively. When comparing the model proposed by *Palomäki et al.* with and without ILD constraint, it can be seen that the model with ILD constraint produces a lower false positive rate (FP rate) in both anechoic and reverberant conditions. But at the same time, the true positive rate (TP rate) is noticeably lower, which consequently limits speaker identification performance. Especially in the reverberant condition, speaker identification accuracy of *Palomäki ITD & ILD* is on average 15.9% below *Palomäki ITD*. We believe that these results can be explained by the employed

Table 5.3: Mask estimation performance (true positive (TP) rate, false positive (FP) rate, TP-FP rate and the number of labeled T-F units) and speaker identification (SID) accuracy in % for various methods in reverberant conditions ($T_{60} = 0.29\,\text{s}$).

T_{60} in s	Methods	%	SNR in dBA			Mean
			−5	0	5	
0.29	IBM SPEC	TP rate	100	100	100	100
		FP rate	0	0	0	0
		TP-FP rate	100	100	100	100
		T-F units	18.4	27.8	38.1	28.1
		SID accuracy	95.3	97.5	97.7	96.8
	IBM SNR	TP rate	68.6	77.1	83.5	76.4
		FP rate	5.9	10.1	16	10.7
		TP-FP rate	62.7	67.0	67.5	65.7
		T-F units	17.5	28.7	41.7	29.3
		SID accuracy	85.4	94.8	97.3	92.5
	Proposed *a priori*	TP rate	58.7	64.5	69.9	64.3
		FP rate	12.6	14.7	17.8	15
		TP-FP rate	46.1	49.8	52.1	49.3
		T-F units	21.1	28.6	37.6	29.1
		SID accuracy	78.9	90.9	95.6	88.5
	Proposed	TP rate	52.5	55	58.4	55.3
		FP rate	12.2	12.2	13.5	12.7
		TP-FP rate	40.3	42.8	44.9	42.6
		T-F units	19.7	24.1	30.6	24.8
		SID accuracy	74.5	88.1	93.9	85.5
	Palomäki ITD	TP rate	37.5	41.9	45.7	41.7
		FP rate	12.2	13	14.7	13.3
		TP-FP rate	25.3	28.9	31.0	28.4
		T-F units	16.8	21	26.5	21.5
		SID accuracy	61.9	80.6	89.3	77.3
	Palomäki ITD & ILD	TP rate	34.7	37	38.8	36.8
		FP rate	11.2	11.6	13	11.9
		TP-FP rate	23.5	25.4	25.8	24.9
		T-F units	15.6	18.7	22.8	19
		SID accuracy	47.1	61.8	75.3	61.4

ITD look-up table and the precomputed ILD template[2], which are both trained with a single source in anechoic conditions. Such a training imposes very strict constraints on the expected ITDs and ILDs. However, it has been shown that binaural cues that are associated with a target source depend on the presence of interfering sources and their

2 Both the mapping function and the ILD template were derived for the same HRTFs used in our experiments. Note that training these functions with reverberant HRTFs did not improve the performance of the model.

relative strength to the target (Roman *et al.*, 2003). Also, reverberation has a severe effect on the ILD cue (Ihlefeld and Shinn-Cunningham, 2004, Shinn-Cunningham *et al.*, 2005), which will cause the ILD constraint to remove many T-F units from the mask, although the underlying template function does not match with the acoustic condition in which the model is applied.

The proposed method achieves significantly higher TP rates and lower FP rates in comparison with *Palomäki ITD*. Whereas speaker recognition performance of both methods is comparable in the anechoic condition, the proposed method substantially outperforms *Palomäki ITD* in the presence of reverberation. Especially at lower SNRs, the speaker identification performance of the proposed model is about 12% above the one by *Palomäki ITD*. In contrast to the model of Palomäki *et al.*, due to the multi-conditional training, the proposed binaural front-end is designed to operate in a variety of acoustic conditions, including reverberation and multi-source scenarios.

When replacing the speech detection module in the proposed model with *a priori* knowledge about the locations of the target and the interfering source, it can be seen that performance is quite similar in terms of TP rates and FP rates for SNRs as low as 0 dBA. This suggests that the speech detection module is able to robustly determine the location of the target source. Only for negative SNRs, more substantial differences can be observed, especially in terms of speaker identification accuracy. This difference can be explained by the increased error rate of the speech detection module at negative SNRs, as reported in Tab. 5.1.

Finally, we compare the two formulations of the ideal binary mask. It is interesting to note that while the mask produced by *IBM SPEC* generally contains more T-F units than *IBM SNR* in the anechoic condition, the mask is more sparse in reverberant conditions. Nevertheless, *IBM SPEC* consistently outperforms the IBM based on the *a priori* SNR in all experimental conditions. This implies that in order to improve on existing mask estimation techniques, the effect of reverberation should be taken into account to reduce the degrading effect of spectral variations that are caused by strong reflections.

5.4.3 Experiment 3: Speaker identification depending on the number of interfering noise sources

The third experiment compares the speaker identification performance of the proposed binaural scene analyzer with various MFCC-based recognizers using a reduced set of 10 speakers. Furthermore, the influence of the number of interfering noise sources is investigated.

The average speaker identification accuracy for one target speaker in reverberant conditions ($T_{60} = 0.29\,\mathrm{s}$) is presented in Fig. 5.4. Panels (A)-(C) show performance depending on the number of interfering sources, ranging from (A) one to (C) three simultaneously active factory noise sources that are randomly placed at different spatial locations. As expected, the speaker identification accuracy decreases with decreasing SNR for all methods. The performance of the MFCC-based recognizer *MFCC Mono* quickly deteriorates with decreasing SNR. The system *MFCC Binaural*, that selects the better ear feature space according to the estimated location of the target speaker, provides a substantial benefit over the monaural MFCC recognizer. This improvement can be in the range of 20% at lower SNRs. A possible explanation may be that the better ear signal has a better SNR than the monaural signal, in addition, there is less spectral distortion due to the head shadow. We found that the advantage of *MFCC Binaural* depends on the spatial separation between the target and the interfering noise sources and increases with increasing spatial separation. An additional performance gain is achieved by *MFCC Binaural NR*, where noise reduction is applied prior to MFCC extraction. This improvement is rather small for scenarios with one interfering noise source and moderately increases when two or three noise sources are present simultaneously.

In turn, the proposed system *MD Proposed* is outperforming the best MFCC-based recognizer in terms of speaker identification accuracy, especially at low SNRs. Regarding acoustic mixtures with one interfering noise source, the performance of the proposed system is close to the system *MD IBM SNR* that utilizes *a priori* SNR information, which indicates that the estimated binary mask that is provided by the binaural front-end is of high accuracy. The proposed system shows a stronger dependency on the number of interfering noise sources as compared to MFCC-based recognizers. In general, performance decreases with increasing number of noise sources, most noticeably when comparing results for scenarios with one and two interfering noise sources. This dependency might be related to the fact that the spatial separation between the target and the interfering sources effectively decreases with increasing number of noise sources. The average

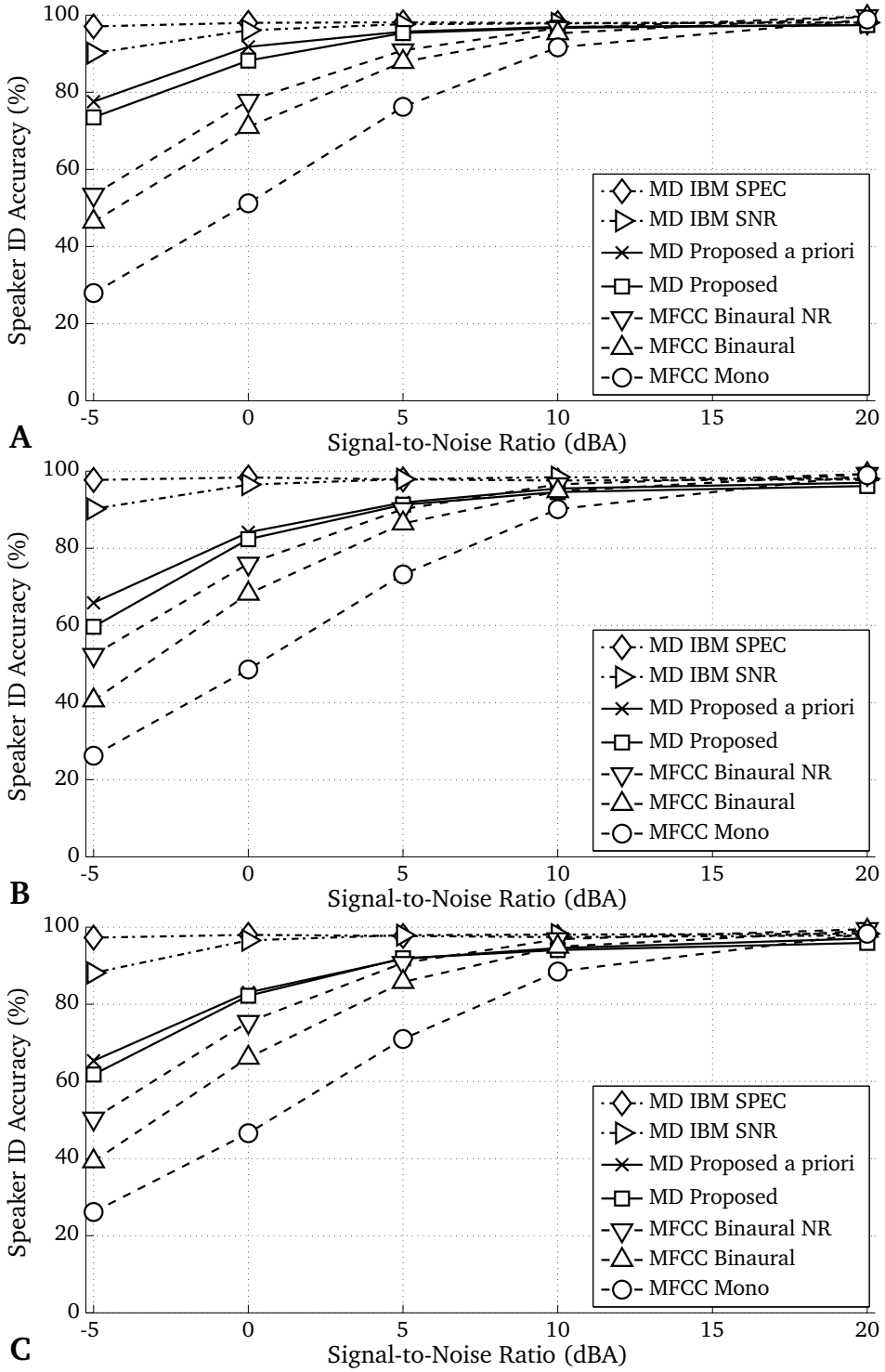


Figure 5.4: Experiment 3: Average speaker recognition performance in % for a set of 10 speakers in reverberant conditions ($T_{60} = 0.29\,\mathrm{s}$) in the presence of (A) one, (B) two and (C) three simultaneously interfering factory noise sources. The average recognition performance is plotted over a series of 20 simulations. Results are presented for three categories of methods, namely the IBM-based MD recognizers (dash-dotted lines), the proposed MD system (solid lines) and the MFCC-based recognizers (dashed lines). The standard error of recognition performance across all 20 simulations was below 3% for all experimental conditions.

azimuth spacing for mixtures with 1, 2 and 3 interfering noise sources is 70.4°, 52.3° and 41.1° respectively. It is reasonable to assume that the mask estimation is more challenging for mixtures with more closely spaced sound sources. The fact that no such dependency is observed for MD systems that utilize the ideal binary mask suggests that the mask estimation of the proposed method can potentially be improved, especially for multi-source scenarios with closely spaced sound sources.

At higher SNRs (20 dBA), MFCC-based recognizers provide some advantage over the proposed MD system, which is presumably caused by the fact that the distribution of MFCC features is more adequately modeled by Gaussian mixtures with diagonal covariance matrices compared to spectral features. This observation is consistent with results reported in previous studies (Palomäki *et al.*, 2004a, Srinivasan *et al.*, 2006).

In order to investigate the influence of the speech detection module on speaker identification accuracy, performance is also shown for the proposed system *MD Proposed a priori* that is employing *a priori* knowledge about the azimuth locations of the speech and noise sources. The distance between this method and *MD Proposed* can be interpreted as the error that is introduced by the speech detection module. As shown in Fig. 5.4, the performance of both methods is very similar, suggesting that the speech detection module is able to robustly detect the azimuth location of the target. This interpretation is also supported by the low localization error for mixtures with one target speaker, which is reported in Tab. 5.1.

Best results are achieved by the MD classifier that is using the ideal binary mask based on the spectral criterion *MD IBM SPEC*, because it considers both the masking effect of interfering noise and the deteriorating effect caused by reverberation.

5.4.4 Experiment 4: Multi-talker speaker identification

The fourth experiment compares the proposed method with the best performing baseline systems according to the third experiment using acoustic mixtures with one and two simultaneously active target speakers. Furthermore, the MFCC-based co-channel recognizer (Saeidi *et al.*, 2010) is evaluated. The full set of 34 speakers is used for this experiment.

The average speaker identification accuracy is depicted in Fig. 5.5 and Fig. 5.6 as a function of the SNR. Results are individually shown for mixtures with one and two

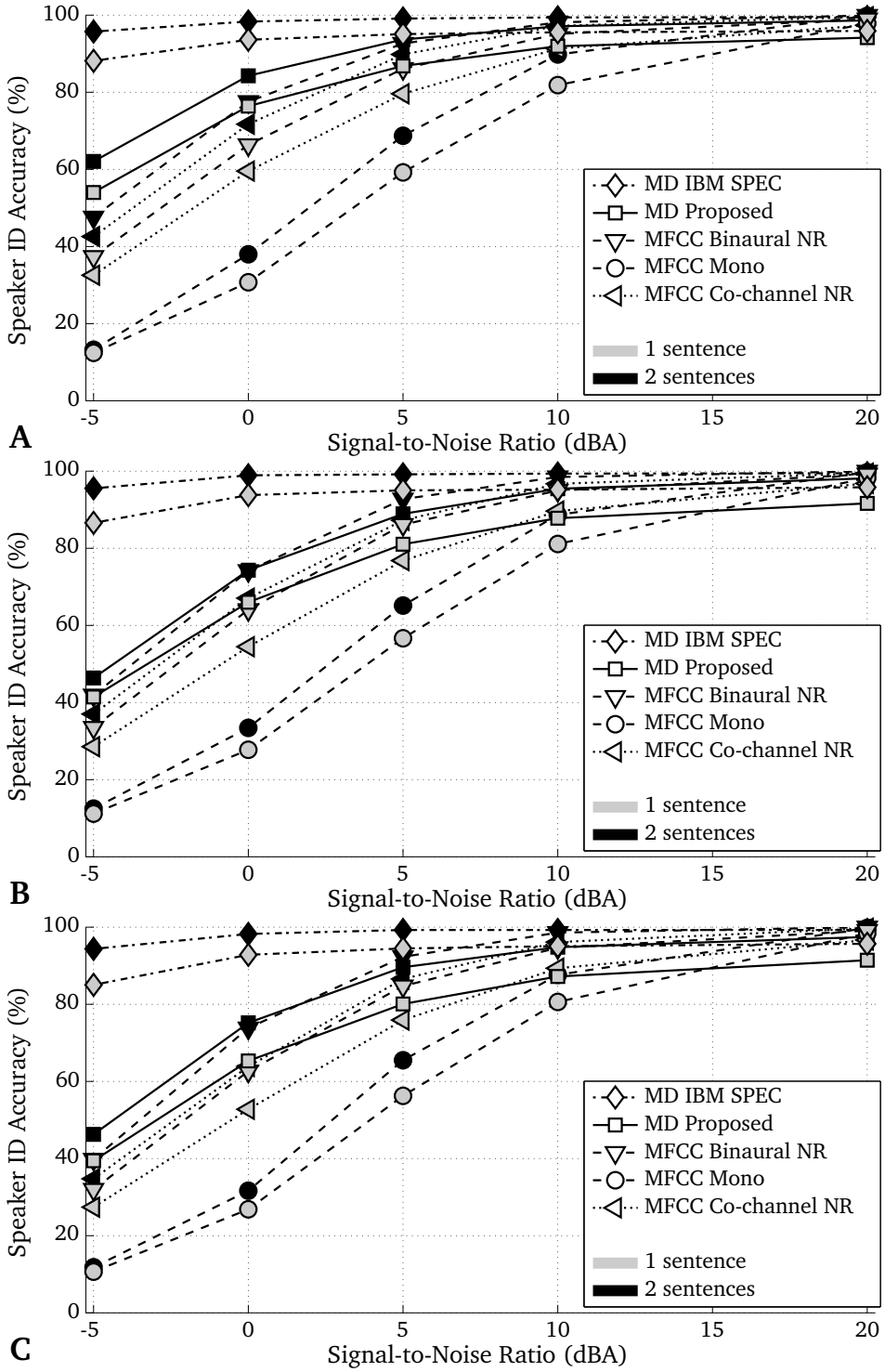


Figure 5.5: Experiment 4: Average speaker identification accuracy in % of one target speaker for a set of 34 speakers in reverberant conditions ($T_{60} = 0.29\,\mathrm{s}$) in the presence of (A) one, (B) two, and (C) three interfering factory noise sources. The gray and black symbols decode recognition performance based on one and two sentences, respectively. Results are presented for four categories of methods, namely the IBM-based MD recognizer (dash-dotted lines), the proposed MD system (solid lines), the MFCC-based recognizers (dashed lines) and the MFCC-based Co-channel recognizer (dotted lines).

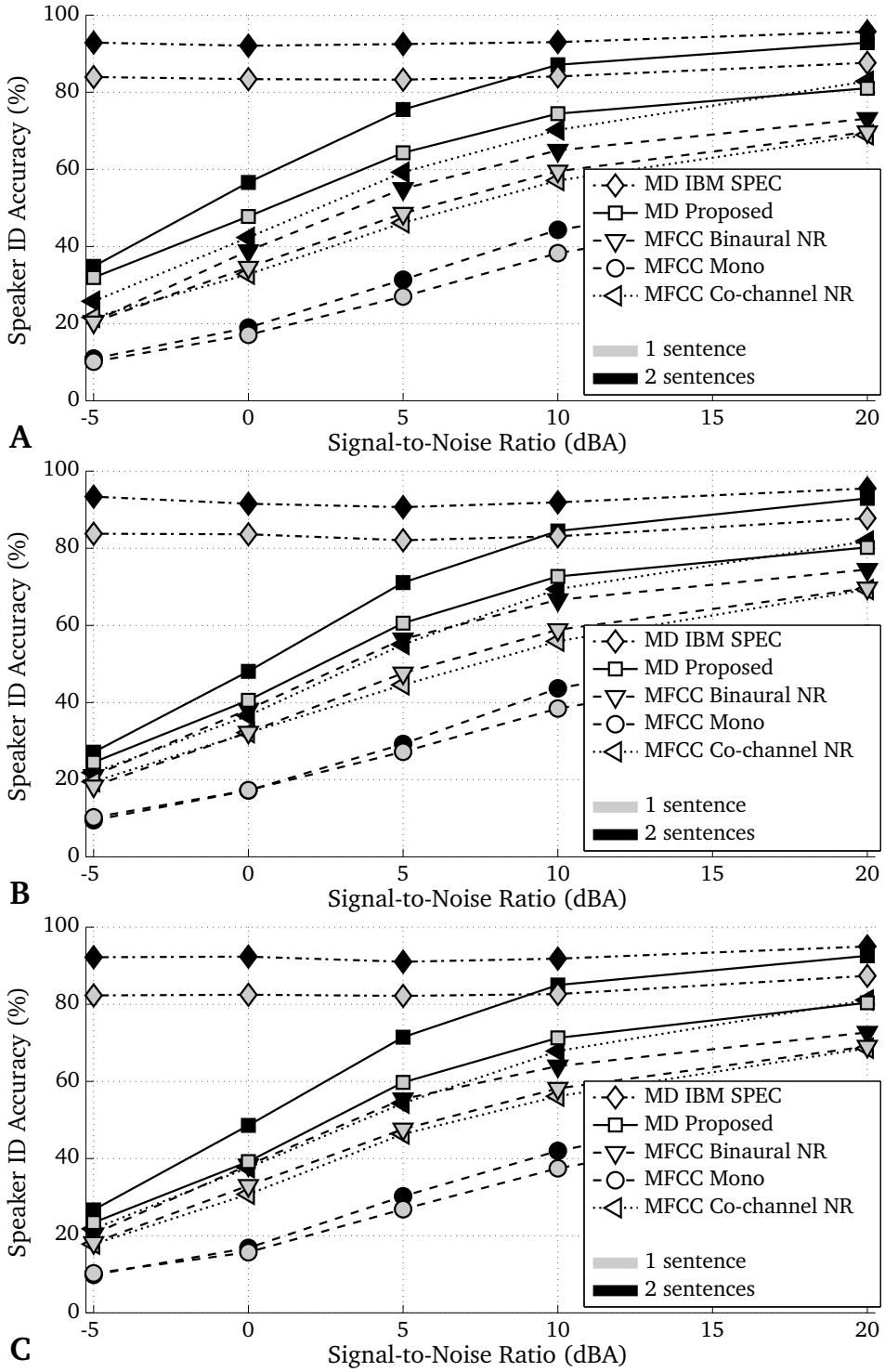


Figure 5.6: Experiment 4: Average speaker identification accuracy in % of two competing target speakers for a set of 34 speakers in reverberant conditions ($T_{60} = 0.29\,\mathrm{s}$) in the presence of (A) one, (B) two, and (C) three interfering factory noise sources. The gray and black symbols decode recognition performance based on one and two sentences, respectively. Results are presented for four categories of methods, namely the IBM-based MD recognizer (dash-dotted lines), the proposed MD system (solid lines), the MFCC-based recognizers (dashed lines) and the MFCC-based Co-channel recognizer (dotted lines).

competing target speakers (see Fig. 5.5 and Fig. 5.6). The MFCC-based recognizer *MFCC Binaural NR* that combines better-ear selection and noise reduction is working quite robust for mixtures with one target source and is by far superior to the conventional monaural MFCC-based recognizer. However, this benefit is noticeably reduced when two target speakers are simultaneously present. Because the front-end for MFCC feature extraction does not distinguish between target and interfering sources, the resulting feature vector reflects to some extent properties of all acoustic sources that are present in the acoustic scene. Apparently, the presence of a second target speaker, which is not ignored by the MFCC-based recognizer, creates a systematic bias that clearly limits speaker recognition performance.

The system *MFCC Co-channel NR*[3], which combines the multi-talker training with the noise reduction front-end described in Section 5.3.2, is able to alleviate the mismatch between the trained speaker models and the observed co-channel mixtures, thus substantially outperforming the monaural MFCC-based recognizers in two-talker mixtures. Note that a considerably larger amount of data is required for training the co-channel system. However, it can not reach the performance level of the proposed missing data recognizer, which aims at separating the contribution of both speakers.

In general, the proposed system *MD Proposed* shows a significant performance gain over all MFCC-based recognizers. For mixtures with one target speaker, this advantage is mostly found for very low SNRs. When two target speakers are simultaneously present, however, the benefit of the proposed method covers a wide range of SNRs and is especially pronounced at higher SNRs (starting at 5 dBA). This coincides with the SNR at which the speech detection module is still able to predict the azimuth of two speakers within 5 ° accuracy (see Tab. 5.1), thus the binary masks are estimated for the azimuth directions which correspond to the real positions of the speakers.

As expected, the highest speaker recognition accuracy is achieved by the MD classifier *MD IBM SPEC* that is based on *a priori* information about the reliable T-F units. Especially at lower SNRs, there is a substantial gap between the ideal and the proposed MD recognizer, suggesting that there is quite some room for improving the mask estimation.

3 Note that the co-channel approach is a monaural system. We tested several modifications of the co-channel system and selected the one with the best performance. It is conceivable that the co-channel approach would also benefit from binaural information.

We also investigated the effect of using two sentences to determine the speaker identities. For mixtures with one target speaker, using two sentences for recognition consistently improves performance for all methods. However, for mixtures with two simultaneously active target speakers, a smaller improvement is found for the MFCC-based systems when two sentences are concatenated. As already mentioned, the presence of a second target speaker is likely to cause a mismatch between training and testing, which can obviously not be reduced by increasing the observation time of the classifier. In contrast, the performance gain of *MD Proposed* can be as large as 15% at an SNR of 20 dBA when two sentences are used for recognition. Because the MD recognizer already operates on a restricted set of T-F units that is believed to contain reliable information about the target source only, a longer test sequence presumably supplies additional evidence about the speaker identity. Also, the co-channel system *MFCC Co-channel NR* shows a substantial performance improvement in the range of up to 15% when increasing the time interval used for recognition, most likely due to the reduced mismatch between the training and testing condition.

Finally, when comparing the overall speaker identification accuracy for the set of 10 speakers (third experiment) with the full set of 34 speakers (fourth experiment), it can be seen that the advantage of the MD recognizer over MFCC-based systems reduces as the set of speakers increases. A similar trend was reported in the context of speech recognition (Srinivasan *et al.*, 2006).

5.4.5 Experiment 5: Joint localization and speaker recognition

The last experiment evaluated the joint localization and speaker identification performance of the proposed method for multi-source mixtures consisting of three interfering factory noise sources, one or two simultaneously active target speakers, and reverberation ($T_{60} = 0.29$ s). Similar to the fourth experiment, the full set of 34 speakers is used and two sentences are concatenated for recognition.

In order to simultaneously compare the localization accuracy of the target speakers and the recognition accuracy of their identities, the performance of both tasks is jointly visualized by means of a two-by-two confusion matrix. The localization accuracy represents the percentage of correctly localized target speakers for which the estimated azimuth is within an absolute error margin of 5° compared to their real position.

The SNR-dependent confusion matrices are shown in Fig. 5.7 for mixtures with one and

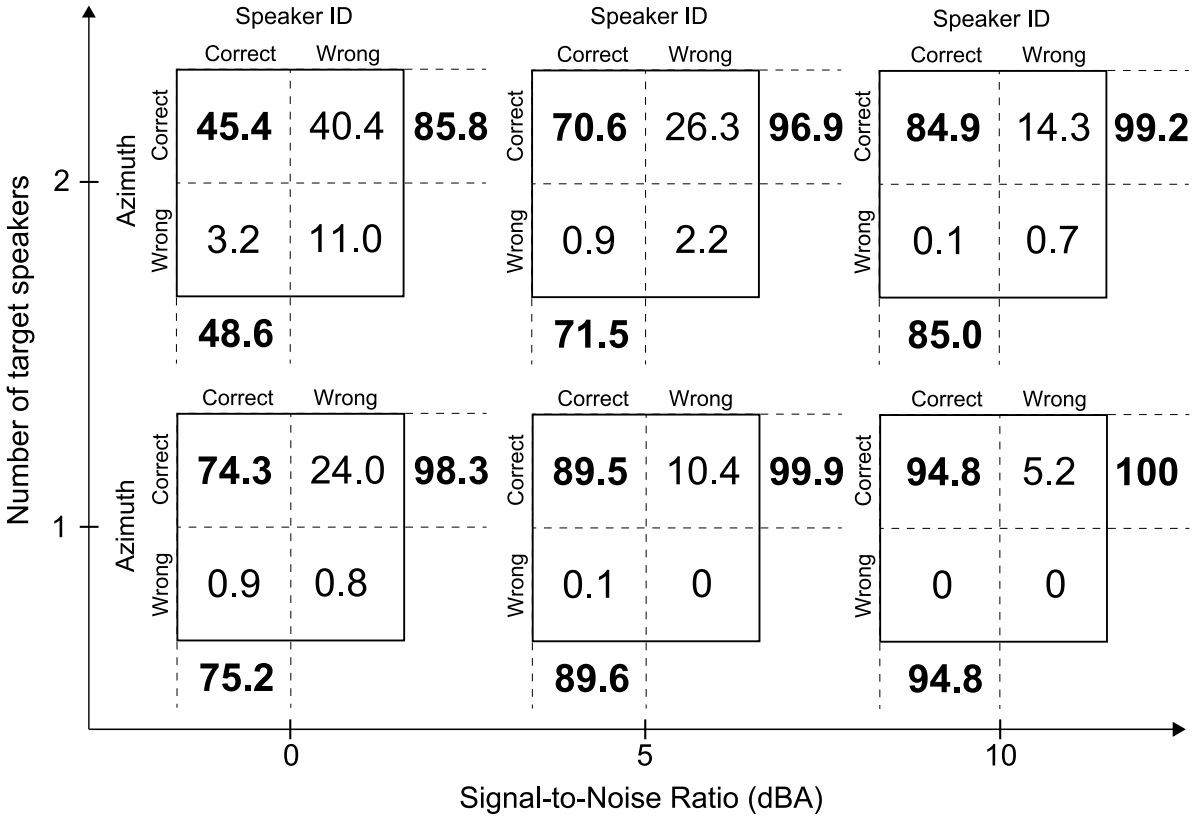


Figure 5.7: Experiment 5: SNR-dependent confusion matrices showing the joint localization accuracy and speaker recognition performance of the proposed binaural scene analyzer for mixtures consisting of one or two simultaneously active target speakers in the presence of three factory noise sources in a reverberant environment ($T_{60} = 0.29\,\mathrm{s}$).

two target speakers. For each individual confusion matrix, the sum along the first column represents the speaker identification accuracy, whereas the sum along the first row shows the localization accuracy. The first element of the main diagonal represents the joint localization and recognition performance, which signifies that both tasks were successfully accomplished by the proposed system.

It can be seen that the joint performance is very close to the overall speaker identification accuracy, which implies that most of the errors are induced by the speaker recognition stage. Indeed, the localization performance of the proposed model is very robust for a wide range of SNRs. Even for mixtures with two concurrent target speakers and three interfering noise sources, in 96.9% of the cases the azimuth of both speakers is correctly localized for SNRs as low as 5 dBA. However, there is a substantial discrepancy between the localization accuracy and the speaker identification accuracy, which is larger for mixtures with two competing speakers and generally increases with decreasing SNR. This gap may indicate that although the correct azimuth location of the target speaker is available, the accuracy of the estimated binary mask at very low SNRs is not sufficient to robustly determine the

identities of the detected speakers. Further research is required to improve the quality of the estimated binary mask for complex multi-source scenarios.

5.5 Discussion and conclusions

In this chapter we have presented a binaural scene analyzer that is able to jointly localize, detect and recognize a predefined number of target speakers in the presence of reverberation and interfering noise. The proposed system consists of three main building blocks: a binaural front-end for robust localization, a module for speech source detection and a stage for speaker identity recognition.

It was shown that the proposed speech detection module is able to robustly detect a predefined number of target speakers in multi-source scenarios. Based on this established link between the localization and the recognition stage, the proposed system is able to selectively focus on processing speech sources in the presence of interfering noise. The system does not require *a priori* knowledge about the azimuth position of the target sources, which is often a limitation for practical applications such as hearing aids.

A detailed ROC analysis was performed to compare the quality of the estimated binary mask of the proposed binaural front-end with the system proposed by Palomäki *et al.* (2004b). The analysis revealed that the proposed system produces binary masks that are closer to the ideal binary masks for both anechoic and reverberant conditions. The proposed front-end has two major advantages: it is designed to operate in reverberant, multi-source scenarios and it jointly analyzes both ITD and ILD cues.

The estimated azimuth position of the target speaker can be used to substantially improve performance of MFCC-based recognizers by selecting the *better ear* feature space for recognition. Regarding acoustic scenes with one target speaker, the improvement in terms of speaker identification accuracy was found to be in the range of 20% at low SNRs. An additional moderate improvement was achieved by applying a noise reduction scheme prior to extracting the MFCC coefficients. However, MFCC-based systems only perform well in acoustic scenes with one target source. This restriction is induced by the front-end for MFCC feature extraction which is not able to distinguish between target and interfering sources (e.g. the interfering noise sources and concurrent speakers). Whereas MFCC coefficients are, to some extent, able to cope with interfering noise, the

presence of a second target speaker clearly biases the resulting MFCC feature vector which consequently limits speaker identification performance. This sensitivity of MFCC-based recognizers to the presence of multiple target speakers can be significantly reduced by the co-channel approach (Saeidi *et al.*, 2010), which incorporates a multi-conditional training stage with two-talker mixtures to alleviate the mismatch between training and testing.

Overall, the proposed binaural scene analyzer is more robust compared to the MFCC-based systems, especially at lower SNRs. Considering acoustic mixtures with a single interfering noise source, the performance of the proposed binaural scene analyzer is close to the classifier that uses the ideal binary mask based on the *a priori* SNR. However, when increasing the number of interfering noise sources, the advantage of the proposed system decreases in comparison to the MFCC-based recognizers. Apparently, with decreasing spatial separation between target and interfering sources, it is more difficult to identify reliable T-F units of the target speaker by only exploiting binaural cues. In order to further improve the estimation of the binary mask, the analysis of binaural cues could be extended by additionally exploiting monaural cues such as pitch (Christensen *et al.*, 2009, Woodruff and Wang, 2010b).

Our experimental results indicate that there is a significant gap in speaker identification performance when comparing acoustic scenes with one and two simultaneously active target speakers. In the present study, multi-source scenarios consisted of a number of simultaneously active speech and noise sources that were completely overlapping. However, the amount of overlapping speech in a meeting or telephone conversation has been estimated to be in the range of 10% (Shriberg *et al.*, 2001). Also, the experimental results obtained in this study are based on simulations, and further tests with real recordings are required. Future research will focus on more realistic multi-source scenarios with natural overlap and turn-taking.

Furthermore, it was shown that the ideal binary mask, which considers both the effect of reverberation and interfering noise, outperformed the mask based on the *a priori* SNR. This suggests that two mechanisms may be required in order to further improve on existing mask estimation techniques: one that segregates the target from the background, and a second one that selects reliable T-F units that are not contaminated by reverberation. A task for future research is to investigate how to combine the segregation mask, as proposed in this chapter based on binaural cues, with a mask that assesses the reliability of individual T-F units in terms of reverberation, either based on modulation analysis

(Palomäki *et al.*, 2004a, 2006) or by exploring the effect of temporal masking (Kim *et al.*, 2011).

In this work, we assumed prior knowledge about the number of active target speakers that are present in the acoustic scene. An important aspect for future investigations is to automatically determine the number of active speech sources.

Finally, is was demonstrated that the proposed binaural scene analyzer is able to jointly localize and recognize two simultaneously active target speakers in the presence of reverberation and three interfering noise sources.

5.A Appendix: Speech detection module

In this appendix we justify the design of the speech detection module as described in Section 5.2.2. First, the ability of three different features to discriminate between speech and noise sources is evaluated. Secondly, it is shown that applying a weight to the log-likelihood ratio of the missing data classifier further improves the speech detection accuracy of the proposed speech detection module.

5.A.1 Feature extraction

In the context of speech recognition, missing data (MD) classification is usually performed with spectral features which reflect the energy of individual frequency channels (Cooke *et al.*, 2001). As the proposed speech detection module aims at discriminating between speech and noise, two alternative features are also evaluated in the framework of missing data classification.

The input signal (sampled at a rate of 16 kHz) is first split into auditory channels using a bank of $F = 32$ gammatone filters that cover the range of 80 to 5000 Hz. Note that signals of the left and the right ear are averaged prior to feature extraction. Features are based on 20 ms frames using a shift of 10 ms.

First, a smoothed envelope e_f is obtained by low-pass filtering the half-wave rectified output of the fth gammatone channel with a time constant of 10 ms. A map of auditory nerve firing rates, a so-called *ratemap*, is computed by averaging the smoothed envelope e_f over B adjacent samples with a shift of O samples and subsequent cube-root

compression

$$\mathcal{F}_{\mathcal{R}}(t,f) = \left(\frac{1}{B} \sum_{b=0}^{B-1} e_f(t\,O+b) \right)^{1/3}. \tag{5.A.1}$$

As an alternative, the mean absolute deviation of the envelope within a frame is used to reflect the amount of fluctuation

$$\mathcal{F}_{\mathrm{AD}}(t,f) = \frac{1}{B} \sum_{b=0}^{B-1} |e_f(t\,O+b) - \bar{e}_f|, \tag{5.A.2}$$

where $\bar{e}_f$ refers to the mean envelope of frame t. Similar to $\mathcal{F}_{\mathcal{R}}$, the feature magnitude of $\mathcal{F}_{\mathrm{AD}}$ is lower for speech-dominant T-F units compared to units that are corrupted by noise. Thus, for feature $\mathcal{F}_{\mathcal{R}}$ and $\mathcal{F}_{\mathrm{AD}}$ it is possible to use *bounded marginalization* (Cooke *et al.*, 2001) where the true value of the unreliable feature components is constrained to be between zero and the observed feature magnitude.

A third feature is based on the observation that speech-dominant areas typically show a higher amount of periodicity compared to noise (Atlas and Hengky, 1985). The harmonic structure of speech tends to excite a similar auto-correlation pattern across neighboring frequency channels. This synchrony is expected to be reduced due to the influence of noise. Following Shao *et al.* (2010), the synchrony is computed by correlating the normalized auto-correlation pattern A_f of the hair cell response h_f across neighboring frequency channels

$$\mathcal{F}_{\mathrm{S}}(t,f) = \frac{1}{\tau_{\max}} \sum_{\tau=0}^{\tau_{\max}-1} A_f(t,\tau)\, A_{f+1}(t,\tau+1), \tag{5.A.3}$$

where τ indexes the time lag and $\tau_{\max}$ refers to the maximum delay. The hair cell response h_f is obtained by applying half-wave rectification and square-root compression to output of the fth gammatone channel. Time lags corresponding to frequencies as low as 80 Hz are considered. Because no useful bounds can be defined for feature $\mathcal{F}_{\mathrm{S}}$, *unbounded marginalization* (Cooke *et al.*, 2001) is performed for this feature.

As described in Section 5.2.2, two GMMs, denoted as $\lambda_{\mathrm{Speech}}$ and λ_{Noise}, are trained with features based on monaural and anechoic speech and noise files. To compensate for the mismatch caused by reverberation, interfering noise and HRTF filtering, spectral normalization (Palomäki *et al.*, 2004a) is performed for feature $\mathcal{F}_{\mathcal{R}}$ and $\mathcal{F}_{\mathrm{AD}}$. As the synchrony feature $\mathcal{F}_{\mathrm{S}}$ is based on normalized auto-correlation patterns, its magnitude is limited to values between $[-1, 1]$. Therefore, no additional normalization is required.

5.A.2 Azimuth-weighted log-likelihood ratio

Based on Eq. (5.6) and Eq. (5.7), the speech detection module determines a set of $\hat{\mathcal{S}}$ log-likelihood ratios $\left\{p_1^{\text{Speech}}, \ldots, p_{\hat{\mathcal{S}}}^{\text{Speech}}\right\}$ which specifies the evidence of all detected speech sources. The final selection of the speech sources is performed by rearranging the set of likelihood ratios in descending order

$$\left\{p_1^{\text{Speech}} \geq p_2^{\text{Speech}} \geq \ldots \geq p_{\hat{\mathcal{S}}}^{\text{Speech}}\right\} \tag{5.A.4}$$

and selecting the azimuth locations corresponding to the $\mathcal{S}$ highest values.

In order to emphasize speech source candidates that are more frequently represented in the azimuth histogram, and therefore, are more likely to reflect the real position of speech sources, a weight is applied to each individual speech source candidate which reflects the *a priori* probability that this source was active in the acoustic scene. This weight is approximated by the normalized azimuth histogram and the azimuth-weighted log-likelihood ratio for each individual speech source candidate $p_n^{\text{Speech,W}}$ is computed according to Eq. (5.8). Similar to the first method, the resulting weighted log-likelihood ratios are ranked in descending order, and the azimuth positions corresponding to the $\mathcal{S}$ highest values reflect the most likely speech sources

$$\left\{p_1^{\text{Speech,W}} \geq p_2^{\text{Speech,W}} \geq \ldots \geq p_{\hat{\mathcal{S}}}^{\text{Speech,W}}\right\}. \tag{5.A.5}$$

5.A.3 Experimental results

Acoustic sources were simulated by convolving monaural audio files with binaural room impulse responses (BRIR). The receiver (KEMAR) was placed at seven evaluation positions in a simulated room of dimensions 6.6 x 8.6 x 3 m at 1.75 m above the ground, as shown in Fig. 5.3. For evaluation, the same acoustic setup is used as described in Section 5.3.1. A set of 600, 4-source mixtures (one speech source) and 600, 5-source mixtures (two speech sources) are generated for various SNR conditions. Speech and noise sources (different from the material used to train the speech detection module) were randomly positioned within the azimuth range of $[-90°, 90°]$, while having an angular distance of at least 15° to the nearest source. For a given multi-source mixture, performance is evaluated by comparing the positions of the detected speech sources to their real positions. A speech source is correctly detected only if the deviation from the real position is within an absolute error margin of 5 °.

Table 5.A.1: Speech detection accuracy in % of one and two speech sources in the presence of reverberation ($T_{60} = 0.29\,\mathrm{s}$) and three interfering noise sources for different features using the MD recognizer.

1 speech source 3 noise sources	SNR in dBA (factory noise) −5	0	5	10	20
MD $\mathcal{F}_{\mathcal{R}}$	68.9	86.2	94.1	97.6	98.9
MD $\mathcal{F}_{\mathrm{AD}}$	74.9	94.8	98.8	99.2	98.9
MD $\mathcal{F}_{\mathrm{S}}$	40.1	63.5	76.2	82.5	94.2
2 speech sources 3 noise sources	**SNR in dBA (factory noise) −5**	**0**	**5**	**10**	**20**
MD $\mathcal{F}_{\mathcal{R}}$	55.9	66.3	78.6	85	89.3
MD $\mathcal{F}_{\mathrm{AD}}$	56.5	69	82.8	88	87.4
MD $\mathcal{F}_{\mathrm{S}}$	43	52.3	62.5	70.3	79.7

First, the performance of the proposed system is evaluated using the three different features described in Section 5.A.1. The speech detection accuracy of the MD-recognizer based on Eq. (5.A.4) for mixtures with one and two speech sources is shown in Tab. 5.A.1. The synchrony feature $\mathcal{F}_{\mathrm{S}}$ showed the overall lowest performance. One possible explanation might be that the characteristic periodicity of speech is not only reduced by interfering noise but also due to the presence of multiple speech sources. Since all classifiers are trained with single-source mixtures, it seems that the classifier trained with the synchrony feature is least capable of generalizing to more complex acoustic scenes. The performance of feature $\mathcal{F}_{\mathcal{R}}$ and $\mathcal{F}_{\mathrm{AD}}$ is comparable at high SNRs. However, with decreasing SNR, feature $\mathcal{F}_{\mathrm{AD}}$ consistently outperformed the ratemap feature $\mathcal{F}_{\mathcal{R}}$. This advantage is observed for both scenarios with one and two target speakers and might be related to the fact that the mean average deviation is invariant to the overall signal level. In the following the MD-based classifier is based on feature $\mathcal{F}_{\mathrm{AD}}$.

In the second experiment, the effect of using the the azimuth-weighted log-likelihood ratio is analyzed. Therefore, the performance of the MD-based speech detection based on Eq. (5.A.4) is compared with the implementation according to Eq. (5.A.5). Furthermore, the speech detection is compared to two baseline systems. To demonstrate the added value of the speech detection module, the first baseline system, denoted as *Azimuth histogram*, solely relies on the estimated azimuth information. Speech sources are classified by selecting the $\mathcal{S}$ most prominent peaks in the azimuth histogram. Obviously, with decreasing SNR, the locations of the noise sources will dominate the azimuth histogram, and, consequently, the speech sources will not be seen. Secondly, a classifier based on 13 mel frequency cepstral coefficients (MFCCs) and their first-order dynamic coefficients with cepstral mean normalization is trained to discriminate between speech and noise.

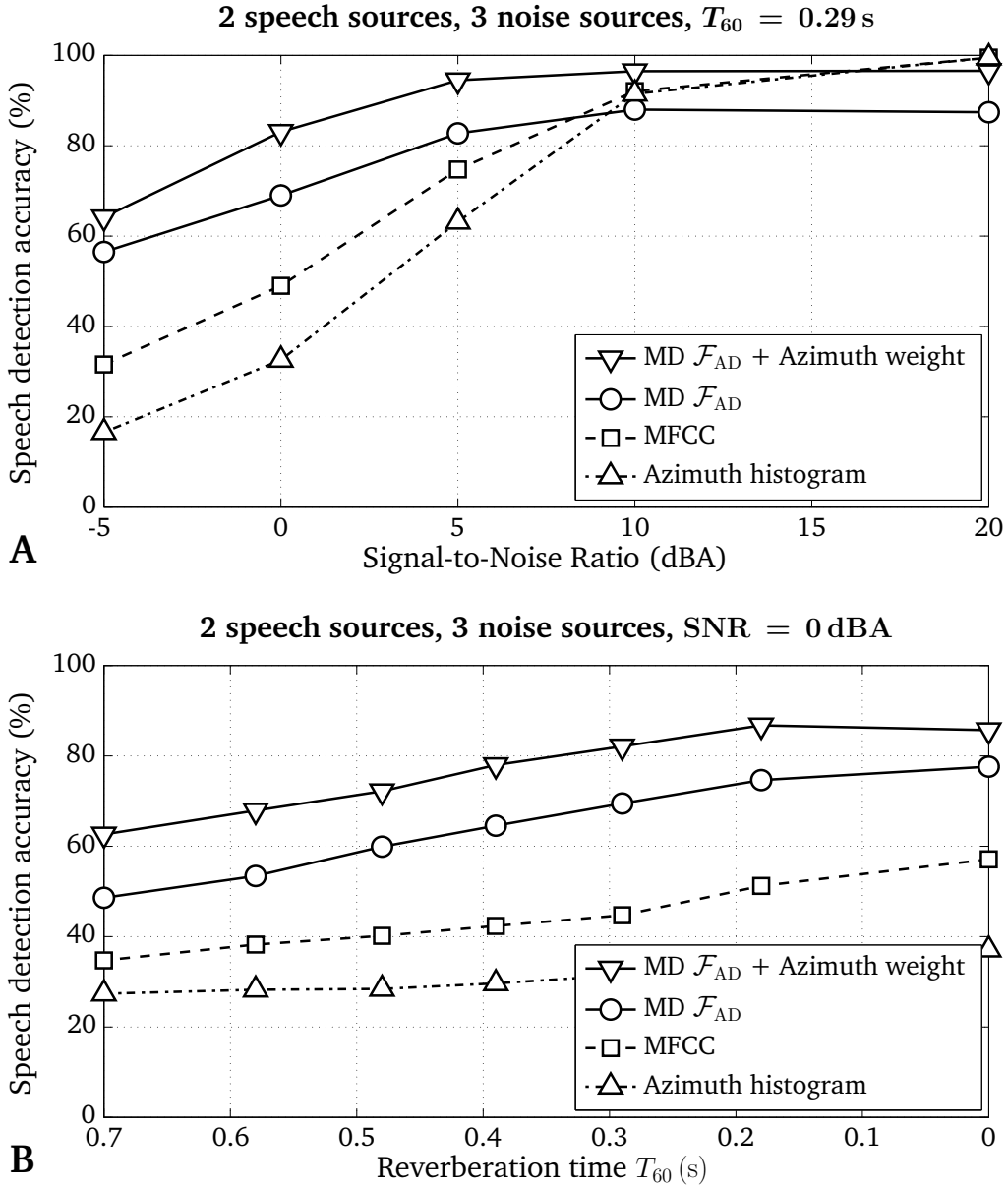


Figure 5.A.1: Speech detection accuracy in % of two speech sources in the presence of three factory noise sources and reverberation as a function of (A) the SNR and (B) the reverberation time T_{60}.

The classifier is trained with the same material that was used for the speech detection module. Based on a frame-by-frame decision, a modified azimuth histogram is computed by only using frames that were classified as being dominated by speech. Again, the $\mathcal{S}$ most dominant peaks reflect the detected speech sources. This second baseline system is referred to as *MFCC*.

The accuracy of detecting two simultaneously active speech sources in the presence of three non-stationary factory noise sources and reverberation is shown in Fig. 5.A.1. Panel (A) presents the speech detection accuracy as a function of the SNR for a fixed

reverberation time of $T_{60} = 0.29\,\mathrm{s}$. In panel (B), performance is shown for a fixed SNR of 0 dBA depending on the reverberation time T_{60}.

As expected the first baseline system is only able to reflect the positions of both speech sources in conditions with high SNR, but performance rapidly drops with decreasing SNR. The MFCC-based selection of speech-dominant frames substantially improved the speech detection accuracy. However, the MD-based system is significantly more robust, especially at low SNRs and conditions with strong reverberation. Furthermore, the selection of the most likely speech sources based on the azimuth-weighted likelihood values using Eq. (5.A.5) provides a distinct performance gain for all experimental conditions.

6

General Conclusions

This final chapter summarizes the main contributions of this dissertation and provides suggestions for further research.

6.1 Conclusions

The main objective of this thesis was to develop new algorithms that are capable of automatically retrieving information about complex acoustic scenes by means of binaural analysis and pattern recognition techniques. The primary focus of this work was the localization, detection and recognition of multiple speech sources in challenging acoustic environments, including interfering noise sources and reverberation.

In **Chapter 2** we have developed a probabilistic model for robust sound source localization based on the supervised learning of binaural cues. The basis of this model was the azimuth-dependent distribution of interaural time and level differences (ITDs and ILDs) which was approximated by a set of Gaussian mixture model (GMM) classifiers. In order to achieve robust sound source localization in reverberant multi-source scenarios, a multi-conditional training stage was employed to capture the uncertainties of binaural cues resulting from room reverberation, the presence of multiple competing sound sources,

and changes in the source-receiver configuration. To obtain a robust indication of the most dominant sound source azimuth per time frame, the evidence of sound source direction is integrated across frequency channels. A comprehensive analysis demonstrated the contribution of ITD and ILD cues and highlighted the benefit of jointly evaluating interaural time and level differences to achieve robust sound source localization. The experimental results showed that the proposed model is able to generalize for unknown source/receiver configurations and for absorption characteristics that were different from the ones used to train the model. A comparison with different auditory front-ends has shown that preserving the fine structure information of the ITDs at higher frequencies can be effectively exploited by the GMM-based classifier, which consequently improved sound source localization (cf. Section 2.5.2). The evaluation with up to four competing speech sources also revealed that the frame-based error rates of the proposed localization model are substantially below the error rates produced by binaural state-of-the-art localization methods. In contrast to other approaches, the performance of the proposed model, evaluated with up to four competing speakers, is almost independent of the number of speech sources.

Furthermore, a histogram analysis of the frame-based localization estimates has been proposed in order to predict the number of active sound sources that are present in the acoustic scene. Experimental results demonstrated that the number of active sound sources could be reliably estimated (above 90% accuracy) for mixtures with up to three competing speech sources in the presence of reverberation (up to a reverberation time of $T_{60} = 0.58\,\text{s}$).

The localization model presented in Chapter 2 was consequently used in **Chapter 3** to create a two-dimensional localization map by grouping individual time-frequency (T-F) units according to the spatial location of the most dominant sound source. In order to further increase the reliability of the estimated sound source direction of individual T-F units, a spectro-temporal integration stage was employed, which integrates spatial evidence across neighboring T-F units that are believed to belong the same spatial location. The integration stage was realized and controlled by a two-dimensional bilateral filter. Experimental results showed that this integration stage reduced the overall localization error and increased the amount of reliable T-F units in the estimated binary mask for acoustic mixtures consisting of two competing speech sources in the presence of reverberation.

In **Chapter 4**, a noise-robust framework for automatic speaker recognition (ASR) has been

presented. The approach builds on missing data (MD) classification where a so-called ideal binary mask (IBM) is utilized to divide the T-F representation of a noisy target signal into reliable (those being dominated by the target) and unreliable (those being dominated by background noise) T-F units. Noise-robustness is achieved by combining the MD-based recognition of speaker identities with the adaptation of speaker models using a universal background model (UBM). It was shown that the UBM-based adaptation of speaker models can effectively reduce the sensitivity of MD-based recognizers to erroneously labeled T-F units in the ideal binary mask. The combination was shown to be especially beneficial for acoustic scenarios with highly non-stationary background noise, which corroborates the practical relevance of this contribution.

Furthermore, the problem of estimating the ideal binary mask in noisy environments has been addressed. To accomplish this, the local signal-to-noise ratio (SNR) was assessed by comparing the estimated noise power with the estimated power spectrum of speech in individual T-F units. For this purpose, a comprehensive comparison of several noise estimation and noise reduction schemes was performed in a variety of different background noise scenarios. A detailed receiver operating characteristic (ROC) analysis revealed that erroneously labeling noise-dominated T-F units as being reliable (false positive) is most detrimental for the recognition of speaker identities. Consequently, the mask estimation for best recognition performance was achieved by a conservative labeling of T-F units, keeping the false positive rate as low as about 2%. Although there are quite some commonalities in the methods that are being applied in the field of noise reduction and ideal binary mask estimation, the experimental results clearly demonstrated the fundamental limitation of noise reduction algorithms to improve the performance of MFCC-based speaker recognition systems. The key advantage of MD-based speaker recognition over an MFCC-based recognizer with a noise reduction front-end is that it does not rely on a precise estimation of the instantaneous SNR which has fundamental limits but rather relaxes this requirement by utilizing a binary classification of reliable T-F units that are dominated by the target.

In **Chapter 5**, a binaural scene analyzer was proposed which is able to simultaneously localize, detect and recognize a predefined number of speakers in the presence of interfering noise sources and reverberation. The system combined the two previously developed building blocks, namely the binaural front-end for robust sound source localization presented in Chapter 2, and the framework for noise-robust automatic speaker recognition described in Chapter 4. In order to link the localization with the recognition stage, a speech detection module has been proposed which is able to detect a predefined

number of speech sources in the presence of interfering noise sources. Therefore, a set of speech source candidates is created first by selecting azimuth positions with relevant sound source activity. The speech detection module then estimates the ideal binary mask, which represents the contribution of each individual speech source candidate on a T-F basis. Based on this segmentation, an MD-based classifier exploits the distinct spectral characteristics of speech and noise signals in order to select the most likely speech source positions among all candidate positions. In this way, the binaural scene analyzer is able to automatically focus on processing speech sources without assuming *a priori* knowledge about the spatial location of the target speakers. Experiments demonstrated that the proposed binaural scene analyzer is capable of simultaneously localizing and recognizing the identity of multiple speakers in adverse acoustic conditions. Furthermore, it was shown that having access to the estimated position of the target source, which can be reliably estimated by the speech detection module, can significantly increase the performance of MFCC-based recognizers by selecting the *better ear* feature space for recognition. However, MFCC-based recognition is limited to acoustic scenes where only one target speaker is present. As soon as two target speakers are simultaneously active, the MFCC feature vector reflects the properties of both speakers, which substantially degrades performance. In contrast, the proposed system estimates the binary mask for each detected speech source based on spatial evidence that is provided by the binaural front-end. As a result, the system can recognize the identity of multiple competing target speakers.

6.2 Suggestions for future research

In this section we will provide an overview about possible future research directions.

The concept of missing data classification has been extensively used throughout this dissertation. For the required identification of reliable and unreliable elements in the T-F representation of complex acoustic scenes, we have exploited two different modalities: the estimation of the local signal-to-noise ratio presented in Chapter 4 and the grouping of T-F elements according to common spatial location in Chapter 5. It was shown in Chapter 5 that the accuracy of the estimated binary mask based on binaural cues decreases with increasing number of competing speech sources and interfering noise sources, which in turn limits the overall speaker identification accuracy. A plethora of other strategies might be employed to estimate the ideal binary mask. For example, Bregman's gestalt

principles provide a set of primitive grouping cues that could all potentially be exploited. It is likely that different cues may provide complementary information about the reliability of individual T-F units. A joint analysis of various cues should therefore allow for an improved mask estimation in contrast to exclusively relying on the analysis of one cue. However, only a few attempts have been devoted to jointly analyze the evidence that is provided by different cues. In the context of monaural speech source separation, classification approaches have been proposed to combine evidence from several monaural cues, such as pitch and modulation, in order to estimate the ideal binary mask in the presence of background noise (Seltzer *et al.*, 2004, Han and Wang, 2011). In comparison, e.g., with a local SNR criterion, improved performance was reported for the joint analysis of monaural cues. An alternative approach that combines monaural and binaural cues for robust sound source localization was recently presented by Woodruff and Wang (2010a,b, 2012). The monaural cue (pitch) was used to achieve sequential organization by grouping T-F units that are dominated by the same acoustic source. This grouping was then used to control the integration of binaural cues across contiguous T-F units for improved localization of speech sources in reverberation. In order to further improve the estimation of the ideal binary mask in binaural acoustic scenes with interfering noise sources and reverberation, the major challenge for future research is to successfully combine and integrate the evidence that is supplied by different cues.

On a related point, most of the approaches that aim at estimating the IBM assume that the decision about the reliability of an individual T-F unit is independent of the neighboring T-F units. However, it was shown in Chapter 3 that integrating the evidence of sound source location across adjacent time and frequency units that are believed to be dominated by one source can effectively increase the accuracy of localization estimates for multiple speech sources in reverberation. Further research is required to validate these experimental results for acoustic scenarios with background noise and to confirm that this benefit also relates to improvements in terms of speaker identification accuracy.

On a more fundamental level, the task of estimating the ideal binary mask can be formulated as a joint segregation and recognition problem, as performed by the speech fragment decoding (SFD) technique (Barker *et al.*, 2001, Barker and Coy, 2005, Barker *et al.*, 2005, Coy and Barker, 2005). The system tries to simultaneously find the optimal segmentation of reliable and unreliable T-F units and to determine the most likely output sequence. This approach has the potential advantage over a conventional missing data recognizer that it can account for errors in the estimated binary mask. Therefore, it could be worthwhile to apply the SFD technique in the context of binaural scene recognition.

In addition, the SFD idea could be further extended to also consider a variety of different grouping cues for T-F segmentation.

As discussed in Chapter 4, MD-based recognizers require a correspondence between the estimated binary mask and individual feature components, therefore, are constrained to the usage of spectral features. In contrast to more compact feature representations, such as the MFCCs, spectral features show a stronger dependency on the vocabulary size (see Section 4.4.5 and Section 5.4.4), presumably because of the covariance between the feature components. Further improvements can be expected when the spectral feature space is preprocessed in order to decorrelate the filterbank energies while still preserving the local meaning of individual feature components. In the context of speech recognition, a simple filtering technique was used to decorrelate the sequence of filterbank energies, and the resulting feature vector was shown to outperform cepstral coefficients in the task of speech recognition (Nadeu *et al.*, 1995, Paliwal, 1999, Nadeu *et al.*, 2001). When incorporating such a preprocessing stage in MD-based recognizers, a more compact feature space representation might be achieved that better matches the assumption of independent feature components that is implicitly implied by the choice of diagonal covariance matrices in the GMM classifier. In addition, dereverberation techniques (Park and Stern, 2007, Jeub *et al.*, 2010, Habets, 2011) may be applied as a preprocessing step in order to further enhance the spectral features prior to recognition and to reduce the mismatch between the training stage and the recognition stage.

Since the developed binaural localization model in Chapter 2 is based on a Gaussian mixture model framework, the model is scalable and can be easily extended to include time and level differences derived from a microphone array with more than two microphones. Furthermore, the current two-dimensional feature space consisting of ITDs and ILDs can be naturally extended with additional features that are depending on the sound source direction.

Currently, the localization model is either trained with binaural cues corresponding to one radial distance of 1.5 m (see Chapter 2 and Chapter 3) or with a set of three different radial distances (0.5 m, 1 m and 2 m, see Chapter 5). It was shown in Chapter 2 that the localization performance of the proposed binaural front-end depends on the radial distance between the source and the receiver and decreases with increasing radial distance. In order to further improve the performance of the localization model at larger radial distances, several localization models could be trained for specific radial distances, and a distance estimation technique (either using monaural (Georganti *et al.*, 2009, 2011) or

binaural approaches (Vesa, 2009)) could be employed to properly select the appropriate localization model for a particular radial distance.

The binaural scene analyzer presented in Chapter 5 requires that the number of target speech sources that should be detected and recognized in the acoustic scene is known *a priori*. Is was shown in Chapter 2 that a histogram of the frame-based azimuth estimates can be used to reliably predict the number of active sound sources for mixtures with up to three competing speakers in reverberation. But detecting the number of target speech sources in a complex binaural mixture with multiple target speakers, various interfering noise sources and reverberation is significantly more difficult than finding the overall number of active sound sources in an acoustic scene. Therefore, this point remains an important task for future investigations.

Finally, it is acknowledged that all the experimental results that are presented in this dissertation are based on simulations. Further experiments with real recordings are required to corroborate the effectiveness of the proposed binaural scene analyzer in real acoustic environments. In addition, the spatial position of sound sources was stationary and did not change over time. An important question is how to deal with moving sound sources. In an attempt to approach this difficult problem for anechoic conditions, a Hidden Markov Model (HMM) framework was proposed to track the azimuth of moving sound sources (Roman and Wang, 2003, 2008).

Binaural Scene Analysis

Localization, Detection and Recognition of Speakers in Complex Acoustic Scenes

Summary

The human auditory system has the striking ability to robustly localize and recognize a specific target source in complex acoustic environments while ignoring interfering sources. Surprisingly, this remarkable capability, which is referred to as auditory scene analysis, is achieved by only analyzing the waveforms reaching the two ears. Computers, however, are presently not able to compete with the performance achieved by the human auditory system, even in the restricted paradigm of confronting a computer algorithm based on binaural signals with a highly constrained version of auditory scene analysis, such as localizing a sound source in a reverberant environment or recognizing a speaker in the presence of interfering noise. In particular, the problem of focusing on an individual speech source in the presence of competing speakers, termed the *cocktail party problem*, has been proven to be extremely challenging for computer algorithms.

The primary objective of this thesis is the development of a binaural scene analyzer that is able to jointly localize, detect and recognize multiple speech sources in the presence of reverberation and interfering noise. The processing of the proposed system is divided into three main stages: localization stage, detection of speech sources, and recognition of speaker identities. The only information that is assumed to be known *a priori* is the number of target speech sources that are present in the acoustic mixture. Furthermore, the aim of this work is to reduce the performance gap between humans and machines by improving the performance of the individual building blocks of the binaural scene analyzer.

First, a binaural front-end inspired by auditory processing is designed to robustly determine the azimuth of multiple, simultaneously active sound sources in the presence of reverberation. The localization model builds on the supervised learning of azimuth-dependent binaural cues, namely interaural time and level differences. Multi-conditional training is performed to incorporate the uncertainty of these binaural cues resulting from reverberation and the presence of competing sound sources.

Second, a speech detection module that exploits the distinct spectral characteristics of speech and noise signals is developed to automatically select azimuthal positions that are

likely to correspond to speech sources. Due to the established link between the localization stage and the recognition stage, which is realized by the speech detection module, the proposed binaural scene analyzer is able to selectively focus on a predefined number of speech sources that are positioned at unknown spatial locations, while ignoring interfering noise sources emerging from other spatial directions.

Third, the speaker identities of all detected speech sources are recognized in the final stage of the model. To reduce the impact of environmental noise on the speaker recognition performance, a missing data classifier is combined with the adaptation of speaker models using a universal background model. This combination is particularly beneficial in non-stationary background noise.

Binaural Scene Analysis

Localization, Detection and Recognition of Speakers in Complex Acoustic Scenes

Samenvatting

Het menselijk auditief systeem is opmerkelijk goed in staat om in een complexe akoestische omgeving, in aanwezigheid van verschillende interfererende bronnen, zowel de richting als de aard van één specifieke bron te bepalen. Dit opmerkelijk vermogen, wat bekend staat als *auditieve scene analyse*, komt tot stand enkel door de twee geluidsgolven die onze oren bereiken, te analyseren. Computers zijn op dit moment niet in staat om deze prestaties van het menselijk auditief systeem te evenaren, zelfs niet voor een vereenvoudigde taak waar het computer algoritme werkt op basis van binaurale signalen in een sterk beperkte versie van auditieve scene analyse, zoals het localiseren van een geluidsbron in een omgeving met nagalm of het herkennen van een spreker met stoorgeluiden. Meer specifiek is aangetoond dat het focusseren op één individuele spreker in de aanwezigheid van meerdere concurrerende sprekers, bekend als het *cocktail party probleem*, een extreem grote uitdaging vormt voor computeralgoritmen.

Het belangrijkste doel van dit proefschrift is het ontwikkelen van een binaurale scene-analysator die in staat is om tegelijkertijd meerdere sprekers te detecteren, te herkennen en te lokaliseren in de aanwezigheid van nagalm en stoorgeluiden. De signaalverwerking in het voorgestelde systeem is verdeeld in drie modulen: een localisatie module, het detecteren van spraakbronnen, en het herkennen van de identiteit van de sprekers. De enige informatie die *a priori* als bekend wordt verondersteld is het aantal doelsprekers dat aanwezig is in het akoestische signaal. Daarnaast beoogt dit proefschrift het onderscheid in prestatie tussen mensen en machines te verkleinen door de prestaties van de individuele bouwstenen van de binaurale scene-analysator te verbeteren.

Ten eerste is er een binaurale voorverwerking ontworpen, gebaseerd op de menselijke auditieve verwerking, welke in staat is om betrouwbaar de azimut te bepalen van meerdere gelijktijdig actieve geluidsbronnen in de aanwezigheid van nagalm. Het localisatiemodel is gebaseerd op het trainen van een patroonherkenner (supervised learning) met azimutafhankelijke binaurale timing- en niveauverschillen. Om de onzekerheden in deze binaurale timing- en niveauverschillen, welke het gevolg zijn van nagalm en de aanwezigheid van stoorgeluiden, te incorporeren wordt er gebruik gemaakt van multi-conditionele training.

Ten tweede is er een spraak-detectie-module ontwikkeld die gebruik maakt van de onderscheidende spektrale eigenschappen van spraak en ruis om automatisch die azimutposities te selecteren waarvan het aannemelijk is dat ze corresponderen met spraakbronnen. Door de zo verkregen verbinding tussen de localisatiemodule en de herkenningsmodule, die gerealiseerd is door de spraak-detectie-module, is de binaurale scene-analysator in staat om selectief te focusseren op een van te voren bepaald aantal spraakbronnen die over onbekende posities verdeeld zijn, waarbij interfererende stoorbronnen op andere posities genegeerd worden.

Ten derde wordt in de laatste module de identiteit van de sprekers herkend voor alle spraakbronnen die eerder werden gedetecteerd. Om de invloed van omgevingsgeluid op de sprekerherkenning te verkleinen wordt een classificator gebruikt die kan werken met ontbrekende data (missing data classifier), gecombineerd met een adaptatie van de sprekermodellen op basis van een universeel achtergrondmodel (universal background model). Deze combinatie is vooral een voordeel bij niet-stationaire stoorgeluiden.

Binaural Scene Analysis

Localization, Detection and Recognition of Speakers in Complex Acoustic Scenes

Zusammenfassung

Das menschliche auditorische System ermöglicht es uns, eine spezifische akustische Quelle in komplexen akustischen Szenen zu erkennen und zu lokalisieren und dabei zusätzlich andere Störquellen auszublenden. Diese bemerkenswerte Fähigkeit beruht allein auf der Analyse der akustischen Signale beider Ohren und wird unter dem Begriff der auditorischen Szenenanalyse (engl. auditory scene analysis) zusammengefasst. Gegenwärtig existierende Algorithmen, welche auf der Analyse von binauralen Signalen basieren, sind bisher nicht in der Lage, diese Leistungsfähigkeit des auditorischen Systems zu erreichen. Selbst bei einer stark vereinfachten Aufgabe, wie beispielsweise der Lokalisation einer Schallquelle in verhallter Umgebung oder der Identifikation eines Sprechers im Störgeräusch, können Algorithmen nicht an die Leistungsfähigkeit des auditorischen Systems anknüpfen. Insbesondere die Konzentration auf einen Zielsprecher in Anwesenheit von weiteren konkurrierenden Sprechern, was oft auch als *Cocktail Party Problem* beschrieben wird, stellt für Algorithmen nach wie vor eine besondere Herausforderung dar.

Diese Dissertation präsentiert einen Algorithmus zur Analyse von binauralen Szenen, welcher eine simultane Lokalisation, Detektion und Identifikation von mehreren Sprechern in akustisch komplexen Situationen, bestehend aus mehreren Sprechern und mehreren Störgeräuschquellen, ermöglicht. Die Signalverarbeitung des entwickelten Systems besteht aus den folgenden drei Abschnitten: 1. Die Lokalisation von Schallquellen, 2. Die Detektion von Sprachquellen und 3. Die Identifikation von Sprechern. Die einzige Annahme, die der Algorithmus benötigt, ist die Anzahl der Zielsprecher, die in der akustischen Szene aktiv sind. Ein weiteres Ziel dieser Arbeit besteht darin, den Unterschied der Leistungsfähigkeit zwischen Menschen und Maschinen durch Weiterentwicklung der einzelnen Bausteine des binauralen Szenenanalysators zu verringern.

Die erste binaurale Vorverarbeitungsstufe, die teilsweise auf der menschlichen auditorischen Verarbeitung basiert, ermöglicht eine robuste Lokalisation von mehreren konkurrierenden Schallquellen in stark verhallten Umgebungen. Dieses Lokalisationsmodell besteht aus einem Mustererkenner, welcher mit richtungsabhängigen interauralen Zeit- und Pegelunterschieden systematisch für alle Zielrichtungen trainiert wird. Um die Unsicherheiten dieser binauralen Merkmale zu berücksichtigen, welche durch starken Nachhall und durch

die Anwesenheit von mehreren konkurrierenden Schallquellen hervorgerufen werden können, wird der Mustererkenner mit einer Vielzahl von verschiedenen akustischen Szenarien trainiert (engl. multi-conditional training).

Die zweite Stufe des Systems besteht aus einem Spracherkennungsmodul, welches die unterschiedlichen spektralen Eigenschaften von Sprach- und Rauschsignalen auswertet. Auf diese Weise kann das System automatisch die Richtungen selektieren, die am wahrscheinlichsten zu Sprachquellen gehören. Durch diese Verbindung zwischen der Lokalisation und der Identifikation ist der Algorithmus in der Lage, sich auf eine definierte Anzahl von Sprachquellen mit unbekannter räumlicher Position zu konzentrieren und zusätzlich konkurrierende Störgeräuschquellen aus anderen Richtungen zu ignorieren.

In der dritten Stufe erfolgt die Sprecheridentifikation aller detektierten Sprachquellen. Um den negativen Einfluss von Störgeräusch auf die Identifikationsleistung zu verringern, wird ein Klassifikator verwendet, der in der Lage ist mit nichtverlässlichen Datenpunkten zu arbeiten (engl. missing data classifier). Dieser Klassifikator wird mit der Adaptation von Sprechermodellen mithilfe eines universellen Hintergrundmodells (engl. universal background model) kombiniert. Diese Kombination ist insbesondere in Umgebungen mit nicht-stationärem Störgeräusch vorteilhaft.

Bibliography

Allen, J. B. and Berkley, D. A. (**1979**), "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America* **65**(4), pp. 943–950. (Cited on pages 21 and 106)

American National Standards Institute (**1983**), "American national standard specification for sound level meters," *ANSI/ASA S1.4-1983 (R2001)* . (Cited on pages 72 and 107)

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**), "Determination of the potential benefit of time-frequency gain manipulation," *Ear and Hearing* **27**(5), pp. 480–492. (Cited on page 6)

Atal, B. S. (**1974**), "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America* **55**(6), pp. 1304–1312. (Cited on pages 4, 62, and 74)

Atlas, L. and Hengky, L. (**1985**), "Cross-channel correlation for the enhancement of noisy speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Tampa, Florida, USA, pp. 724–727. (Cited on page 128)

Barker, J., Cooke, M., and Ellis, D. P. W. (**2001**), "Combining bottom-up and top-down constraints for robust ASR: The multisource decoder," in *Proceedings of the Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis (CRAC)*, Aalborg, Denmark. (Cited on page 137)

Barker, J., Cooke, M. P., and Ellis, D. P. W. (**2005**), "Decoding speech in the presence of other sound sources," *Speech Communication* **45**(1), pp. 5–25. (Cited on page 137)

Barker, J. and Coy, A. (**2005**), "Towards solving the cocktail party problem through primitive grouping and model combination," in *Proceedings of Forum Acusticum*,

Budapest, Hungary, pp. 1551–1556. (Cited on page 137)

Barker, J., Josifovski, L., Cooke, M., and Green, P. (**2000**), "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, pp. 373–376. (Cited on page 91)

Bernstein, L. R. and Trahiotis, C. (**1996**), "The normalized correlation: Accounting for binaural detection across center frequency," *Journal of the Acoustical Society of America* **100**(6), pp. 3774–3784. (Cited on page 33)

Bernstein, L. R., van de Par, S., and Trahiotis, C. (**1999**), "The normalized interaural correlation: Accounting for NoSπ thresholds obtained with Gaussian and "low-noise" masking noise," *Journal of the Acoustical Society of America* **106**(2), pp. 870–876. (Cited on page 33)

Berouti, M., Schwartz, R., and Makhoul, J. (**1979**), "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Washington, District of Columbia, USA, vol. 4, pp. 208–211. (Cited on pages 4 and 70)

Blauert, J. (**1997**), *Spatial hearing - The psychophysics of human sound localization*, The MIT Press, Cambride, MA, USA. (Cited on page 17)

Bodden, M. (**1993**), "Modeling human sound-source localization and the cocktail-party-effect," *Acta Acustica* **1**(1), pp. 43–55. (Cited on pages 8, 9, 14, and 27)

Boll, S. F. (**1979**), "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **27**(2), pp. 113–120. (Cited on pages 4 and 70)

Braasch, J. (**2005**), "Modelling of binaural hearing," in *Communication Acoustics*, edited by J. Blauert, Springer, Berlin, Germany, chap. 4, pp. 75–108. (Cited on page 15)

Bregman, A. S. (**1990**), *Auditory scene analysis: The perceptual organization of sound*, The MIT Press, Cambridge, MA, USA. (Cited on pages 1, 14, 48, and 94)

Bronkhorst, A. W. (**2000**), "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acustica* **86**, pp. 117–128. (Cited on pages 1, 8, and 94)

Bronkhorst, A. W. and Plomp, R. (**1992**), "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *Journal of the Acoustical Society of America* **92**(6), pp. 3132–3139. (Cited on page 14)

Brookes, M. (**2009**), "VOICEBOX: Speech processing toolbox for MATLAB," URL http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, accessed on 27th November 2009. (Cited on page 71)

Brown, G. J. and Cooke, M. (**1994**), "Computational auditory scene analysis," *Computer Speech and Language* **8**(4), pp. 297–336. (Cited on pages 18 and 97)

Brown, G. J., Harding, S., and Barker, J. (**2006**), "Speech separation based on the statistics of binaural auditory features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, vol. 5, pp. 14–19. (Cited on pages 15, 16, 20, 22, and 23)

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**), "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *Journal of the Acoustical Society of America* **120**(6), pp. 4007–4018. (Cited on pages 6, 67, and 95)

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2009**), "Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers," *Journal of the Acoustical Society of America* **125**(6), pp. 4006–4022. (Cited on page 6)

Brungart, D. S., Durlach, N. I., and Rabinowitz, W. M. (**1999**), "Auditory localization of nearby sources. II. Localization of a broadband source," *Journal of the Acoustical Society of America* **106**(4), pp. 1956–1968. (Cited on page 35)

Brungart, D. S. and Rabinowitz, W. M. (**1999**), "Auditory localization of nearby sources. Head-related transfer functions," *Journal of the Acoustical Society of America* **106**(3), pp. 1465–1479. (Cited on page 35)

Campbell, D. R., Palomäki, K. J., and Brown, G. (**2005**), "A MATLAB simulation of "shoebox" room acoustics for use in research and teaching," *Computing and Information Systems* **9**(3), pp. 48 – 51. (Cited on pages 20, 29, 35, and 55)

Cappe, O. (**1994**), "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing* **2**(2), pp.

345–349. (Cited on page 70)

Carter, G. C., Nuttall, A. H., and Cable, P. G. (**1973**), "The smoothed coherence transform," *Proceedings of the IEEE* **61**(10), pp. 1497–1498. (Cited on page 27)

Champagne, B., Bedard, S., and Stephenne, A. (**1996**), "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing* **4**(2), pp. 148–152. (Cited on pages 16 and 30)

Chen, J., Benesty, J., and Huang, Y. A. (**2005**), "Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments," *EURASIP Journal of Applied Signal Processing* **1**, pp. 25–36. (Cited on pages 16, 28, and 44)

Cherry, E. C. (**1953**), "Some experiments on the recognition of speech, with one and two ears," *Journal of the Acoustical Society of America* **25**(5), pp. 975–979. (Cited on pages 1, 8, and 94)

Christensen, H., Ma, N., Wrigley, S. N., and Barker, J. (**2007**), "Integrating pitch and localisation cues at a speech fragment level," in *Proceedings of Interspeech*, Antwerp, Belgium, pp. 2769–2772. (Cited on pages 8, 45, 47, 48, and 52)

Christensen, H., Ma, N., Wrigley, S. N., and Barker, J. (**2008**), "Improving source localisation in multi-source, reverberant conditions: Exploiting local spectro-temporal location cues," *Journal of the Acoustical Society of America* **123**(5), pp. 3294 (A). (Cited on pages 45 and 48)

Christensen, H., Ma, N., Wrigley, S. N., and Barker, J. (**2009**), "A speech fragment approach to localising multiple speakers in reverberant environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, pp. 4593–4596. (Cited on pages 45, 48, 54, 95, and 126)

Cohen, I. and Berdugo, B. (**2002**), "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters* **9**(1), pp. 12–15. (Cited on pages 69 and 91)

Cooke, M. (**2005**), "Making sense of everyday speech: A glimpsing account," in *Speech Separation by Humans and Machines*, edited by P. Divenyi, Kluwer Academic, Dordrecht, The Netherlands, chap. 21, pp. 305–314. (Cited on page 6)

Cooke, M. (**2006**), "A glimpsing model of speech perception in noise," *Journal of the*

Acoustical Society of America **199**(3), pp. 1562–1573. (Cited on pages 6 and 102)

Cooke, M., Green, P., and Crawford, M. (**1994**), "Handling missing data in speech recognition," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Yokahama, Japan, pp. 1555–1558. (Cited on page 5)

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**), "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication* **34**, pp. 267–285. (Cited on pages 5, 7, 62, 63, 65, 67, 68, 94, 99, 100, 104, 109, 110, 127, and 128)

Cooke, M. and Lee, T.-W. (**2006**), "Speech separation and recognition competition," URL http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm, accessed on 12th October 2010. (Cited on pages 71, 100, and 107)

Coy, A. and Barker, J. (**2005**), "Recognising speech in the presence of a competing speaker using a 'speech fragment decoder'," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Pennsylvania, USA, pp. 425–428. (Cited on page 137)

Darwin, C. J. (**1997**), "Auditory grouping," *Trends in Cognitive Sciences* **1**(1), pp. 327–333. (Cited on page 8)

Darwin, C. J. (**2008**), "Spatial hearing and perceiving sources," in *Auditory perception of sound sources*, edited by W. A. Yost, R. R. Fay, and A. N. Popper, Springer Science+Business Media, New York, NY, USA, chap. 8, pp. 215–232. (Cited on pages 8 and 48)

Davis, S. B. and Mermelstein, P. (**1980**), "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **28**(4), pp. 357–366. (Cited on pages 3 and 62)

Dempster, A., Laird, N., and Rubin, D. (**1977**), "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B* **39**(1), pp. 1–38. (Cited on pages 25, 66, 100, and 105)

Denbigh, P. N. and Zhao, J. (**1992**), "Pitch extraction and separation of overlapping speech," *Speech Communication* **11**(2-3), pp. 119–125. (Cited on page 8)

Devore, S., Ihlefeld, A., Shinn-Cunningham, B. G., and Delgutte, B. (**2007**), "Neural and behavioral sensitivities to azimuth degrade with distance in reverberant environments," in *Hearing - From Sensory Processing to Perception*, edited by B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey, Springer, Berlin, Germany, pp. 505–516. (Cited on page 44)

DiBiase, J., Silverman, H., and Brandstein, M. (**2001**), "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, edited by M. Brandstein and D. Ward, Springer, Berlin, Germany, chap. 8, pp. 157–180. (Cited on page 14)

Doblinger, G. (**1995**), "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proceedings of Eurospeech*, Madrid, Spain, pp. 1513–1516. (Cited on page 69)

Drygajlo, A. and El-Maliki, M. (**1998**a), "Speaker verification in noisy environment with combined spectral subtraction and missing data theory," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, Washington, USA, vol. 1, pp. 121–124. (Cited on pages 7, 63, and 68)

Drygajlo, A. and El-Maliki, M. (**1998**b), "Use of generalized spectral subtraction and missing feature compensation for robust speaker verification," in *Proceedings of RLA2C Workshop on Speaker Recognition and its Commercial and Forensic Applications*, Avignon, France, pp. 80–83. (Cited on pages 63 and 71)

El-Maliki, M. and Drygajlo, A. (**1999**), "Missing features detection and handling for robust speaker verification," in *Proceedings of Eurospeech*, Budapest, Hungary, pp. 975–978. (Cited on page 100)

Ellis, D. P. W. (**2005**), "PLP and RASTA (and MFCC, and inversion) in Matlab," URL http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/, accessed on 11th October 2011. (Cited on pages 73 and 108)

Ephraim, Y. (**1992**), "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Transactions on Signal Processing* **40**(4), pp. 725–735. (Cited on pages 4 and 71)

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 **32**(6), pp. 1109–1121. (Cited on pages 4, 70, and 71)

Ephraim, Y. and Malah, D. (**1985**), "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **33**(2), pp. 443–445. (Cited on pages 4, 70, 71, and 109)

Faller, C. and Merimaa, J. (**2004**), "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *Journal of the Acoustical Society of America* **116**(5), pp. 3075–3089. (Cited on page 14)

Fawcett, T. (**2006**), "An introduction to ROC analysis," *Pattern Recognition Letters* **27**(8), pp. 861–874. (Cited on pages 79 and 112)

Figueiredo, M. A. T. and Jain, A. K. (**2002**), "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(3), pp. 381–396. (Cited on pages 16, 26, and 28)

Flores, J. A. N. and Young, S. J. (**1994**), "Continuous speech recognition in noise using spectral subtraction and HMM adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Adelaide, South Australia, Australia, vol. 1, pp. 409–412. (Cited on page 68)

Gaik, W. (**1993**), "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling," *Journal of the Acoustical Society of America* **94**(1), pp. 98–110. (Cited on page 8)

Gardner, W. G. and Martin, K. D. (**1994**), "HRTF measurements of a KEMAR dummy-head microphone," Technical report #280, MIT Media Lab, Perceptual Computing, Cambridge, MA, USA. (Cited on pages 20 and 105)

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (**1993**), "TIMIT acoustic-phonetic continuous speech corpus," Technical report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, USA. (Cited on pages 21, 30, and 55)

Georganti, E., May, T., van de Par, S., Härmä, A., and Mourjopoulos, J. (**2009**), "Single channel sound source distance estimation based on statistical and source specific features," in *Proceedings of the 126th Audio Engineering Society (AES) Convention*, Munich, Germany. (Cited on page 138)

Georganti, E., May, T., van de Par, S., Härmä, A., and Mourjopoulos, J. (**2011**), "Speaker distance detection using a single microphone," *IEEE Transactions on Audio, Speech, and Language Processing* **19**(7), pp. 1949–1961. (Cited on page 138)

Glasberg, B. R. and Moore, B. C. J. (**1990**), "Derivation of auditory filter shapes from notched-noise data," *Hearing Research* **47**(1-2), pp. 103–138. (Cited on pages 18, 67, and 97)

Griffiths, L. J. and Jim, C. W. (**1982**), "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation* **30**(1), pp. 27–34. (Cited on page 2)

Habets, E. A. P. (**2011**), "Speech dereverberation using statistical reverberation models," in *Speech dereverberation*, edited by P. A. Naylor and N. D. Gaubitch, Springer, London, UK. (Cited on page 138)

Han, K. and Wang, D. L. (**2011**), "An SVM based classification approach to speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, pp. 4632–4635. (Cited on pages 7 and 137)

Harding, S., Barker, J., and Brown, G. (**2006**), "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Transactions on Audio, Speech, and Language Processing* **14**(1), pp. 58 – 67. (Cited on pages 14, 15, 20, 68, 95, and 98)

Härmä, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., Hiipakka, J., and Lorho, G. (**2004**), "Augmented reality audio for mobile and wearable appliances," *Journal of the Audio Engineering Society* **52**(6), pp. 618–639. (Cited on page 14)

Hawley, M. L. and Litovsky, R. Y. (**2004**), "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *Journal of the Acoustical Society of America* **115**(2), pp. 833–843. (Cited on pages 8 and 95)

Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (**1999**), "Speech intelligibility and localization in a multi-source environment," *Journal of the Acoustical Society of America* **105**(6), pp. 3436–3448. (Cited on pages 8 and 14)

Haykin, S. and Chen, Z. (**2005**), "The cocktail party problem," *Neural Computation* **17**(9), pp. 1875–1902. (Cited on page 1)

Hermansky, H. and Morgan, N. (**1994**), "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing* **2**(4), pp. 578–589. (Cited on pages 4 and 62)

Hirsch, H. G. and Ehrlicher, C. (**1995**), "Noise estimation techniques for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, Michigan, USA, vol. 1, pp. 153–156. (Cited on pages 69 and 109)

Hu, G. and Wang, D. L. (**2001**), "Speech segregation based on pitch tracking and amplitude modulation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, pp. 79–82. (Cited on page 7)

Hu, G. and Wang, D. L. (**2004**), "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks* **15**(5), pp. 1135–1150. (Cited on pages 7 and 68)

Hu, G. and Wang, D. L. (**2007**), "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, and Language Processing* **15**(2), pp. 396–405. (Cited on pages 7, 59, and 68)

Ianniello, J. P. (**1982**), "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **30**(6), pp. 998–1003. (Cited on page 30)

Ihlefeld, A. and Shinn-Cunningham, B. G. (**2004**), "Effect of source location and listener location on ILD cues in a reverberant room," *Journal of the Acoustical Society of America* **115**(5), pp. 2598. (Cited on page 116)

Jacovitti, G. and Scarano, G. (**1993**), "Discrete time techniques for time delay estimation," *IEEE Transactions on Signal Processing* **41**(2), pp. 525–533. (Cited on page 19)

Jeffress, L. A. (**1948**), "A place theory of sound localization," *Journal of Comparative and Physiological Psychology* **41**(1), pp. 35–39. (Cited on page 15)

Jeub, M., Schäfer, M., Esch, T., and Vary, P. (**2010**), "Model-based dereverberation preserving binaural cue," *IEEE Transactions on Audio, Speech, and Language Processing* **18**(7), pp. 1732–1745. (Cited on page 138)

Kidd, G., Jr., Arbogast, T. L., Mason, C. R., and Gallun, F. J. (**2005**), "The advantage of knowing where to listen," *Journal of the Acoustical Society of America* **188**(6), pp.

3804–3815. (Cited on page 95)

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," *Journal of the Acoustical Society of America* **104**(1), pp. 422–431. (Cited on page 8)

Kim, C., Kumar, K., and Stern, R. M. (**2011**), "Binaural sound source separation motivated by auditory processing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, pp. 5072–5075. (Cited on page 127)

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**), "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America* **126**(3), pp. 1486–1494. (Cited on page 113)

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L. (**2009**), "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *Journal of the Acoustical Society of America* **126**(3), pp. 1415–1426. (Cited on page 6)

Klumpp, R. and Eady, H. (**1956**), "Some measurements of interaural time difference thresholds," *Journal of the Acoustical Society of America* **28**(5), pp. 859–860. (Cited on page 44)

Knapp, C. H. and Carter, G. C. (**1976**), "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **ASSP-24**(4), pp. 320–327. (Cited on pages 15 and 27)

Kollmeier, B. and Koch, R. (**1994**), "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *Journal of the Acoustical Society of America* **95**(3), pp. 1593–1602. (Cited on page 8)

Kollmeier, B., Peissig, J., and Hohmann, V. (**1993**), "Real-time multiband dynamic compression and noise reduction for binaural hearing aids," *Journal of Rehabilitation Research and Development* **30**(1), pp. 82–94. (Cited on page 8)

Kopčo, N., Best, V., and Carlile, S. (**2010**), "Speech localization in a multitalker mixture," *Journal of the Acoustical Society of America* **127**(3), pp. 1450–1457. (Cited on page 95)

Kühne, M., Pullella, D., Togneri, R., and Nordholm, S. (**2008**), "Towards the use of full covariance models for missing data speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, pp. 4537–4540. (Cited on page 63)

Li, N. and Loizou, P. C. (**2007**), "Factors influencing glimpsing of speech in noise," *Journal of the Acoustical Society of America* **122**(2), pp. 1165–1172. (Cited on page 95)

Li, N. and Loizou, P. C. (**2008**a), "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *Journal of the Acoustical Society of America* **123**(3), pp. 1673–1682. (Cited on page 6)

Li, N. and Loizou, P. C. (**2008**b), "Effect of spectral resolution on the intelligibility of ideal binary masked speech," *Journal of the Acoustical Society of America* **123**(4), pp. EL59–EL64. (Cited on page 6)

Li, Y. and Wang, D. L. (**2008**), "On the optimality of the ideal binary time-frequency masks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, pp. 3501–3504. (Cited on page 7)

Li, Y. and Wang, D. L. (**2009**), "On the optimality of ideal binary time-frequency masks," *Speech Communication* **51**, pp. 230–239. (Cited on page 7)

Lin, L., Holmes, W., and Ambikairajah, E. (**2003**), "Adaptive noise estimation algorithm for speech enhancement," *Electronic Letters* **39**(9), pp. 754–755. (Cited on pages 69 and 91)

Lindemann, W. (**1986**a), "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," *Journal of the Acoustical Society of America* **80**(6), pp. 1608–1622. (Cited on pages 8 and 15)

Lindemann, W. (**1986**b), "Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front," *Journal of the Acoustical Society of America* **80**(6), pp. 1623–1630. (Cited on page 15)

Lippmann, R. P. (**1997**), "Speech recognition by machines and humans," *Speech Communication* **22**(1), pp. 1–15. (Cited on pages 2 and 62)

Lloyd, S. (**1982**), "Least squares quantization in PCM," *IEEE Transactions on Information*

Theory **28**(2), pp. 129–137. (Cited on pages 26, 66, 100, and 105)

Loizou, P. C. (**2007**), *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, FL, USA. (Cited on pages 4 and 69)

Lu, L., He, J., and Palm, G. (**1997**), "A comparison of human and machine in speaker recognition," in *Proceedings of Eurospeech*, Rhodes, Greece, pp. 2327–2330. (Cited on page 3)

Lyon, R. F. (**1983**), "A computational model of binaural localization and separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Boston, Massachusetts, USA, pp. 1148–1151. (Cited on page 8)

Ma, N., Green, P., Barker, J., and Coy, A. (**2007**), "Exploiting correlogram structure for robust speech recognition with multiple speech sources," *Speech Communication* **49**, pp. 874–891. (Cited on pages 7 and 68)

Martin, R. (**2001**), "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing* **9**(5), pp. 504–512. (Cited on page 69)

Martin, R. (**2006**), "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing* **86**(6), pp. 1215–1229. (Cited on page 69)

May, T., van de Par, S., and Kohlrausch, A. (**2009**), "A probabilistic model for robust acoustic localization based on an auditory front-end," in *Proceedings of the NAG/DAGA*, Rotterdam, The Netherlands, p. 254 (A). (Cited on pages 10 and 13)

May, T., van de Par, S., and Kohlrausch, A. (**2010**), "The effect of spectro-temporal integration in a probabilistic model for robust acoustic localization," in *2nd International Symposium on Auditory and Audiological Research (ISAAR 2009)*, Helsingør, Denmark, pp. 125–134. (Cited on pages 10 and 47)

May, T., van de Par, S., and Kohlrausch, A. (**2011**a), "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing* **19**(1), pp. 1–13. (Cited on pages 10, 13, and 95)

May, T., van de Par, S., and Kohlrausch, A. (**2011**b), "Simultaneous localization and identification of speakers in noisy and reverberant environments," in *Proceedings of Forum Acusticum*, Aalborg, Denmark, pp. 2121–2126. (Cited on pages 11 and 93)

May, T., van de Par, S., and Kohlrausch, A. (**2011**c), "Binaural detection of speech sources in complex acoustic scenes," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, pp. 241–244. (Cited on pages 11 and 93)

May, T., van de Par, S., and Kohlrausch, A. (**2012**a), "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *IEEE Transactions on Audio, Speech, and Language Processing* **20**(1), pp. 108–121. (Cited on pages 7, 10, 61, and 94)

May, T., van de Par, S., and Kohlrausch, A. (**2012**b), "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing* **20**(7), pp. 2016–2030. (Cited on pages 11 and 93)

McLachlan, G. and Peel, D. (**2000**), *Finite Mixture Models*, John Wiley & Sons, New York, NY, USA. (Cited on page 26)

Meyer, B. T., Brand, T., and Kollmeier, B. (**2011**), "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes," *Journal of the Acoustical Society of America* **129**(1), pp. 388–403. (Cited on page 3)

Moore, B. C. J. (**2003**), *An introduction to the psychology of hearing*, Academic Press, San Diego, California, USA, 5th ed. (Cited on page 6)

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**), "A model for prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society* **45**(4), pp. 224–240. (Cited on page 18)

Nabney, I. T. and Bishop, C. M. (**2001-2004**), "NETLAB package," URL http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/, accessed on 11th October 2011. (Cited on pages 28 and 73)

Nadeu, C., Hernando, J., , and Gorricho, M. (**1995**), "On the decorrelation of filter-bank energies in speech recognition," in *Proceedings of Eurospeech*, Madrid, Spain, pp. 1381–1384. (Cited on page 138)

Nadeu, C., Macho, D., and Hernando, J. (**2001**), "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication* **34**(1-2), pp. 93–114. (Cited on page 138)

Nix, J. and Hohmann, V. (**2006**), "Sound source localization in real sound fields based on empirical statistics of interaural parameters," *Journal of the Acoustical Society of America* **119**(1), pp. 463–479. (Cited on pages 16, 25, and 32)

Openshaw, J. P. and Mason, J. S. (**1994**), "On the limitations of cepstral features in noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Adelaide, South Australia, Australia, pp. 49–52. (Cited on pages 4, 74, and 108)

Paliwal, K. K. (**1999**), "Decorrelated and liftered filter-bank energies for robust speech recognition," in *Proceedings of Eurospeech*, Budapest, Hungary, pp. 85–88. (Cited on page 138)

Palomäki, K. J., Brown, G. J., and Barker, J. P. (**2004**a), "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Communication* **43**(1-2), pp. 123–142. (Cited on pages 7, 63, 96, 100, 104, 119, 127, and 128)

Palomäki, K. J., Brown, G. J., and Barker, J. P. (**2006**), "Recognition of reverberant speech using full cepstral features and spectral missing data," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, pp. 289–292. (Cited on pages 7 and 127)

Palomäki, K. J., Brown, G. J., and Wang, D. L. (**2001**), "A binaural auditory model for missing data recognition of speech in noise," in *Proceedings of the Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis (CRAC)*, Aalborg, Denmark. (Cited on page 8)

Palomäki, K. J., Brown, G. J., and Wang, D. L. (**2004**b), "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication* **43**(4), pp. 361–378. (Cited on pages 8, 14, 27, 68, 95, 110, 113, 116, and 125)

Park, H.-M. and Stern, R. M. (**2007**), "Missing feature speech recognition using dereverberation and echo suppression in reverberant environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, vol. 4, pp. 381–384. (Cited on page 138)

Pullella, D., Kühne, M., and Togneri, R. (**2008**), "Robust speaker identification using combined feature selection and missing data recognition," in *Proceedings of the IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, Nevada, USA, pp. 4833–4836. (Cited on pages 63 and 94)

Raj, B. (**2000**), "Reconstruction of incomplete spectrograms for robust speech recognition," Ph.D. thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. (Cited on page 5)

Raj, B. and Stern, R. M. (**2005**), "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine* **22**(5), pp. 101–116. (Cited on page 100)

Rangachari, S. and Loizou, P. C. (**2006**), "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication* **48**, pp. 220–231. (Cited on page 69)

Rangachari, S., Loizou, P. C., and Hu, Y. (**2004**), "A noise estimation algorithm with rapid adaptation for highly nonstationary environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Quebec, Canada, pp. 305–308. (Cited on page 69)

Renevey, P. and Drygajlo, A. (**2000**), "Statistical estimation of unreliable features for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul, Turkey, vol. 3, pp. 1731–1734. (Cited on page 71)

Renevey, P. and Drygajlo, A. (**2001**), "Detection of reliable features for speech recognition in noisy conditions using a statistical criterion," in *Proceedings of the Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis (CRAC)*, Aalborg, Denmark. (Cited on page 7)

Reynolds, D. and Rose, R. (**1995**), "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing* **3**(1), pp. 72–83. (Cited on pages 24, 63, 64, and 66)

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (**2000**), "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing* **10**, pp. 19–41. (Cited on pages 63, 66, 73, 104, and 105)

Ris, C. and Dupont, S. (**2001**), "Assessing local noise level estimation methods: Application to noise robust ASR," *Speech Communication* **34**, pp. 141–158. (Cited on pages 7 and 68)

Rohdenburg, T., Goetze, S., Hohmann, V., Kammeyer, K.-D., and Kollmeier, B. (**2008**), "Objective preceptual quality assessment for self-steering binaural hearing aid microphone arrays," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, pp. 2449–2452. (Cited on page 3)

Roman, N. and Wang, D. L. (**2003**), "Binaural tracking of multiple moving sources," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, China, vol. 5, pp. 149–152. (Cited on pages 9 and 139)

Roman, N. and Wang, D. L. (**2006**), "Pitch-based monaural segregation of reverberant speech," *Journal of the Acoustical Society of America* **120**(1), pp. 458–469. (Cited on page 7)

Roman, N. and Wang, D. L. (**2008**), "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Processing* **16**(4), pp. 728–739. (Cited on page 139)

Roman, N., Wang, D. L., and Brown, G. J. (**2002**), "Location-based sound segregation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, Florida, USA, pp. 1013–1016. (Cited on page 9)

Roman, N., Wang, D. L., and Brown, G. J. (**2003**), "Speech segregation based on sound localization," *Journal of the Acoustical Society of America* **114**(4), pp. 2236–2252. (Cited on pages 9, 15, 16, 17, 18, 20, 27, 48, 68, 95, 98, and 116)

Roman, N. and Woodruff, J. (**2011**), "Intelligibility of reverberant noisy speech with ideal binary masking," *Journal of the Acoustical Society of America* **130**(4), pp. 2153–2161. (Cited on pages 6 and 7)

Rosenberg, A. E., Lee, C.-H., and Soong, F. K. (**1994**), "Cepstral channel normalization techniques for HMM-based speaker verification," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Yokohama, Japan, pp. 1835–1838. (Cited on page 4)

Saeidi, R., Mowlaee, P., Kinnunen, T., Tan, Z.-H., Christensen, M. G., Jensen, S. H., and Fränti, P. (**2010**), "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, pp. 4565–4568.

(Cited on pages 109, 119, and 126)

Schimmel, S. M., Müller, M. F., and Dillier, N. (**2009**), "A fast and accurate "shoebox" room acoustics simulator," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, pp. 241–244. (Cited on pages 106 and 107)

Schmidt-Nielsen, A. and Crystal, T. H. (**2000**), "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data," *Digital Signal Processing* **10**, pp. 249–266. (Cited on page 3)

Seltzer, M. L., Raj, B., and Stern, R. M. (**2004**), "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication* **43**(4), pp. 379–393. (Cited on page 137)

Shackleton, T. M., Meddis, R., and Hewitt, M. J. (**1992**), "Across frequency integration in a model of lateralization," *Journal of the Acoustical Society of America* **91**(4), pp. 2276–2279. (Cited on pages 25 and 99)

Shamsoddini, A. and Denbigh, P. N. (**2001**), "A sound segragation algorithm for reverberant conditions," *Speech Communication* **33**(3), pp. 179–196. (Cited on page 8)

Shao, Y., Srinivasan, S., Jin, Z., and Wang, D. L. (**2010**), "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Speech Communication* **24**, pp. 77–93. (Cited on page 128)

Shao, Y. and Wang, D. L. (**2003**), "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, China, pp. 205–208. (Cited on page 63)

Shao, Y. and Wang, D. L. (**2006**), "Robust speaker recognition using binary time-frequency masks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, pp. 645–648. (Cited on page 63)

Shinn-Cunningham, B. G., Kopčo, N., and Martin, T. J. (**2005**), "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *Journal of the Acoustical Society of America* **117**(5), pp. 3100–3115. (Cited on pages 15 and 116)

Shinn-Cunningham, B. G., Schickler, J., Kopčo, N., and Litovsky, R. (**2001**), "Spatial

unmasking of nearby speech sources in a simulated anechoic environment," *Journal of the Acoustical Society of America* **110**(2), pp. 1118–1129. (Cited on page 105)

Shriberg, E., Stolcke, A., and Baron, D. (**2001**), "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proceedings of Eurospeech*, Aalborg, Denmark, pp. 1359–1362. (Cited on page 126)

Simpson, B. D., Brungart, D. S., Iyer, N., Gilkey, R. H., and Hamil, J. T. (**2006**), "Detection and localization of speech in the presence of competing speech signals," in *Proceedings of the International Conference of Auditory Display (ICAD)*, London, UK, pp. 129–133. (Cited on page 95)

Soong, F. K. and Rosenberg, A. E. (**1988**), "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **36**(6), pp. 871–879. (Cited on pages 74 and 108)

Srinivasan, S., Roman, N., and Wang, D. L. (**2006**), "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication* **48**(11), pp. 1486–1501. (Cited on pages 63, 87, 119, and 123)

Srinivasan, S., Samuelsson, J., and Kleijn, W. B. (**2007**), "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Audio, Speech, and Language Processing* **15**(2), pp. 441–452. (Cited on pages 4 and 71)

Sroka, J. J. and Braida, L. D. (**2005**), "Human and machine consonant recognition," *Speech Communication* **45**(4), pp. 401–423. (Cited on page 3)

Stern, R. M., Zeiberg, A. S., and Trahiotis, C. (**1988**), "Lateralization of complex binaural stimuli: A weighted-image model," *Journal of the Acoustical Society of America* **84**(1), pp. 156–165. (Cited on page 99)

Tibrewala, S. and Hermansky, H. (**1997**), "Multi-band and adaptation approaches to robust speech recognition," in *Proceedings of Eurospeech*, Rhodes, Greece, pp. 2619–2622. (Cited on pages 4, 74, and 108)

Tomasi, C. and Manduchi, R. (**1998**), "Bilateral filtering for gray and color images," in *Proceedings of the International Conference on Computer Vision (ICCV)*, Bombay, India, pp. 839–846. (Cited on page 52)

Varga, A. P., Steeneken, H. J. M., Tomlinson, M., and Jones, D. (**1992**), "The

NOISEX-92 study on the effect of additive noise on automatic speaker recognition," Tech. rep., Speech Research Unit, Defence Research Agency, Malvern, UK. (Cited on pages 72 and 100)

Vesa, S. (**2009**), "Binaural sound source distance learning in rooms," *IEEE Transactions on Audio, Speech, and Language Processing* **17**(8), pp. 1498–1507. (Cited on page 139)

Viikki, O. and Laurila, K. (**1998**), "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication* **25**(1-3), pp. 133–147. (Cited on page 4)

Virag, N. (**1999**), "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing* **7**(2), pp. 126–137. (Cited on page 70)

Viste, H. and Evangelista, G. (**2004**), "Binaural source localization," in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx'04)*, Naples, Italy, pp. 145–150. (Cited on page 15)

Vizinho, A., Green, P., Cooke, M., and Josifovski, L. (**1999**), "Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study," in *Proceedings of Eurospeech*, Budapest, Hungary, pp. 2407–2410. (Cited on pages 7, 68, and 69)

Wang, D. L. (**2005**), "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi, Kluwer Academic, Dordrecht, The Netherlands, chap. 12, pp. 181–197. (Cited on pages 7, 56, 67, and 95)

Wang, D. L. (**2007**), "Computational scene analysis," in *Challenges for Computational Intelligence*, edited by W. Duch and J. Mandziuk, Springer, Berlin, Germany, pp. 163–191. (Cited on page 2)

Wang, D. L. and Brown, G. (Eds.) (**2006**), *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, John Wiley & Sons, Hoboken, NJ, USA. (Cited on pages 2, 14, and 94)

Wang, D. L., Kjems, U., Pedersen, M. S., and Boldt, J. B. (**2009**), "Speech intelligibility in background noise with ideal binary time-frequency masking," *Journal of the Acoustical Society of America* **125**(4), pp. 2336–2347. (Cited on page 95)

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2008**), "Speech perception of noise with binary gains," *Journal of the Acoustical Society of America* **124**(4), pp. 2303–2307. (Cited on page 6)

Wilson, K. and Darrell, T. (**2005**), "Improving audio source localization by learning the precedence effect," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Pennsylvania, USA, vol. 4, pp. 18–23. (Cited on page 14)

Wittkop, T. and Hohmann, V. (**2003**), "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Communication* **39**(1-2), pp. 111–138. (Cited on page 3)

Wohlmayr, M. and Képsi, M. (**2007**), "Joint position-pitch extraction from multichannel audio," in *Proceedings of Interspeech*, Antwerp, Belgium, pp. 1629–1632. (Cited on page 8)

Woodruff, J. and Wang, D. L. (**2010**a), "Integrating monaural and binaural analysis for localizing multiple reverberant sound sources," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, pp. 2706–2709. (Cited on pages 8 and 137)

Woodruff, J. and Wang, D. L. (**2010**b), "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Transactions on Audio, Speech, and Language Processing* **18**(7), pp. 1856–1866. (Cited on pages 8, 95, 126, and 137)

Woodruff, J. and Wang, D. L. (**2012**), "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing* **20**(5), pp. 1503–1512. (Cited on page 137)

Wu, M., Wang, D. L., and Brown, G. J. (**2003**), "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing* **11**(3), pp. 229–241. (Cited on page 7)

Zakarauskas, P. and Cynader, M. S. (**1993**), "A computational theory of spectral cue localization," *Journal of the Acoustical Society of America* **94**(3), pp. 1323–1331. (Cited on page 44)

Zhang, L. and Wu, X. (**2005**), "On cross correlation based discrete time delay estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP), Philadelphia, Pennsylvania, USA, vol. 4, pp. 981–984. (Cited on pages 19, 25, and 27)

Zhao, D. Y., Kleijn, W. B., Ypma, A., and de Vries, B. (**2008**), "Online noise estimation using stochastic-gain HMM for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing* **16**(4), pp. 835–846. (Cited on page 71)

Zwislocki, J. and Feldman, R. (**1956**), "Just noticeable differences in dichotic phase," *Journal of the Acoustical Society of America* **28**(5), pp. 860–864. (Cited on page 44)

Acknowledgments

It feels very good to be able to write these final pages. I want to acknowledge the help and support of many people that I received during the last years.

First of all, I would like to thank Steven van de Par and Armin Kohlrausch for accepting me as a Ph.D. candidate at the Eindhoven University of Technology and for creating a very nice and inspiring working environment. It's been a pleasure for me to work on this project and I really consider myself lucky to get the support and the supervision of both of you. I would like express my deepest gratitude to Steven for his irreplaceable support and his tireless patience. Whatever question I asked, his guidance never failed, and without him and our regular discussions, this work would not have been possible. Armin, thank you so much for your continuous support, also during the last period of the project when I moved back to Oldenburg.

I am also grateful to Prof. Dr. ir. Simon Doclo, Prof. Dr. ir. Jan W. M. Bergmans and Dr. Sriram Srinivasan for giving me valuable feedback on my thesis. Furthermore, I would like to thank Prof. Dr. ir. Simon Doclo, Prof. Dr. ir. Jan W. M. Bergmans, Dr. Sriram Srinivasan, Prof. Dr.-Ing. Timo Gerkmann and Prof. Dr. ir. Jean-Bernard Martens for participating in my thesis commitee.

I owe my gratitude to Eleftheria Georganti and John Mourjopoulos for the great collaboration and for many enjoyable evenings at conferences and meetings. I also want to thank my former colleagues Aki Härmä and Sriram Srinivasam for the nice and fruitful discussions. Furthermore, I would like to thank all the members of the Acoustics group at the University of Oldenburg and my former office mates at Philips Research, Janto Skowronek, Tommi Määttää, Nicolas Le Goff, Tom Goossens, Othmar Schimmel, Alberto Novello and Michael Bruderer for the great time we shared at the High Tech Campus. The financial support of Philips Research is very much appreciated.

A special thanks goes to all my friends, in particular Niklas Harlander and Gunnar Dröge, who helped me formatting my brain whenever it was necessary. Darrin, thank you so much for your superb proofreading service and for the great BBQ sessions. Moreover, I want to thank my parents and my brother for always supporting me throughout this adventure. Finally, I want to thank Doro, for being the most amazing partner I can imagine, and for reminding me that there is so much more than research.

Curriculum Vitae

Personal:	
name	Tobias May
date of birth	September 17th, 1980
place of birth	Oldenburg, Germany
nationality	German
Education:	
2010 - 2012	**Ph.D. candidate** at the University of Oldenburg, Germany, and the Eindhoven University of Technology, The Netherlands
2007 - 2010	**Ph.D. candidate** at the Eindhoven University of Technology, and Philips Research Laboratories, Eindhoven, The Netherlands
2005 - 2007	**Master of Science (M.Sc.)** in Hearing Technology and Audiology, University of Oldenburg, Germany - Thesis: *Analysis and Improvement of 3D Headphones Sound Reproduction* (Advisor: Prof. Dr. Dr. B. Kollmeier and Prof. Dr. R. Aarts)
2001 - 2005	**Dipl.-Ing. (FH)** in Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany - Thesis: *Analysis and Development of Multichannel Noise Reduction Schemes for Hearing Aids* (Advisor: Prof. Dr.-Ing. J. Bitzer and Prof. Dr. M. Hansen)