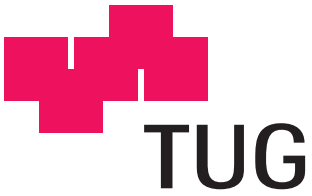
Diploma Thesis

Monaural Sound Localization

Anna Katharina Fuchs

Signal Processing and Speech Communication Lab
Graz University of Technology
Head: Univ.-Prof. Dipl.-Ing. Dr.techn. Gernot Kubin



Supervisor: Dipl.-Ing. Dr.techn. Christian Feldbauer

Graz, February 2011

Statutory Declaration

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

Graz, _____ _____

Place, Date Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, _____ _____

Ort, Datum Unterschrift

Danksagung

Diese Diplomarbeit entstand am Institut für Signalverarbeitung und Sprachkommunikation der Technischen Universität Graz.

Besonders bedanken möchte ich mich bei meinem Betreuer Dipl.-Ing. Dr.techn. Christian Feldbauer für die freundliche, geduldige und engagierte Unterstützung. Außerdem gilt mein Dank meinen Kollegen vom SPSC Lab, besonders Wolfgang. Sie hatten immer ein offenes Ohr für Probleme in allen Belangen und haben mir die Zeit während des Schreibens der Diplomarbeit versüßt.

Für das fleißige Korrekturlesen und die hilfreichen Anmerkungen bedanke ich mich vor allem bei meinem Bruder Clemens und meiner guten Freundin Lisa. Bedanken möchte ich mich auch bei den Personen, die mich während meiner Studienzeit begleitet und unterstützt haben, allen voran Christoph und Sieglinde, danke für eure Freundschaft.

Weiters möchte ich mich bei meinem Freund Michael bedanken. Danke für das Verständnis, für die Ermutigungen und für die Unterstützung in jeglicher Hinsicht.

Über allem steht natürlich meine liebe Familie, vor allem mein Vater, ohne den mein Studium nie möglich gewesen wäre.

Danke!

Graz, Februar 2010 Anna Katharina Fuchs

Kurzfassung

Die Lokalisation von Schallquellen ist in vielen Anwendungen von großer Wichtigkeit. Das menschliche Gehör wertet zur Ortung von Schallquellen sowohl binaurale als auch monaurale Unterschiede aus. Binaurale Informationen sind interaurale Laufzeit- und Pegeldifferenzen (ITD und ILD), die dadurch zustande kommen, dass ein Schallquellensignal ein Ohr vor dem anderen, und mit einem höheren Pegel erreicht. Monaurale Unterschiede hingegen werden durch die Außenohr-Übertragungsfunktion (head-related transfer function – HRTF) dargestellt. Diese Unterschiede beschreiben die richtungsabhängige, spektrale Verfärbung der einfallenden Schallwelle durch Kopf, Schulter, Torso und Pinna. HRTFs sind für jeden Einfallswinkel unterschiedlich. Dadurch kann von den einzelnen HRTFs auf die Richtung der Schallquelle rückgeschlossen werden.

Das Ziel dieser Arbeit ist es, mit nur einem Mikrofonsignal einen Sprecher in der Horizontalebene möglichst genau zu lokalisieren. Die entwickelte Lokalisationsmethode verwendet zum einen eine Datenbank mit gemessenen HRTFs und zum anderen statistische Modelle der richtungsunabhängigen Sprache. Dabei werden zunächst Mel-Frequenz Cepstral Koeffizienten (MFCCs) mit wenigen Dimensionen und Short-Time-Fourier Transformations Koeffizienten mit vielen Dimensionen als Merkmalsvektoren extrahiert. Anschließend wird einerseits mit dem Mindestabstandsklassifikator und andererseits mit Gausschen Mischmodellen (GMMs) und der Maximum Likelihood-Methode (ML) die Richtung der Schallquelle geschätzt. Die Evaluierung in einer synthetisch hergestellten Testumgebung erzielt ausgezeichnete Ergebnisse. Darum bietet diese Lokalisationsmethode einen guten Ansatz für weitere Untersuchungen in zukünftigen Arbeiten.

Abstract

Sound localization is an important task in many applications such as multimedia applications. The principles of human sound localization imply binaural (interaural level and time difference – ILD and ITD) as well as monaural cues. The latter are captured by the head-related transfer function (HRTF), which describes the direction-dependent, spectral shaping of the incident sound wave by head, shoulders, torso, and pinna. HRTFs are different for each direction of the incident sound wave. Therefore, they can be exploited to determine the direction.

The aim of this thesis is to develop an accurate talker localization strategy in the horizontal plane using the signal of only one microphone. Based on a set of HRTF measurements and statistical models of the direction-independent speech a sound localization method is developed. Low-dimensional mel-frequency cepstral coefficients (MFCCs) and high-dimensional spectral features (STFT coefficients) are taken and the direction of the sound source is evaluated with two different classificators, i.e. a minimum distance (MD)

classificator and Gaussian mixture models (GMMs) using a maximum likelihood (ML) framework. An evaluation of the developed method in a synthetic test environment yields excellent results and leads to a promising approach which can be further investigated in future research.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Sound localization is a task humans as well as animals are confronted with day-by-day. Since hundreds of years localization of enemies decides on survival. Localization is also important in many engineering applications such as localization of an active speaker or improvement of the signal-to-noise ratio (SNR) in hearing aids.

From the human auditory system it is known that the binaural cues, interaural time and level difference (ITD, ILD) and monaural spectral cues, presented by the head-related transfer function (HRTF), are used to localize a sound position. In applications the same cues are exploited. Therefore, often two or more microphones are used. Although, a higher number of microphones results in more accurate estimations of the sound source, the growing physical dimension is a huge disadvantage. Multiple microphones also result in higher costs.

The literature indicates that binaural cues are primarily important for the horizontal plane (azimuth) and monaural cues for the vertical plane (elevation). The role of monaural cues for the horizontal plane has not been clarified yet. In [1] the author states that monaural cues are more important even to determine the azimuth. This fact as well as the reduced physical size and cost, are good reasons to investigate a single-channel localization strategy.

1.2 Goals

In this thesis a localization strategy with a single microphone has been developed. Based on a set of measured head-related impulse responses (HRIRs) and a statistical model of speech, an estimation of the sound direction has been carried out. The estimation is performed in the spectral domain by use of a maximum likelihood (ML) based approach.

The resulting estimation strategy should be accurate and applicable in real-time applications. Furthermore, a speaker-independent strategy is desireable because an application should be useable for mass-market products. The advantages of a single-channel

sound source estimation are the lower costs for a single microphone and the possibility of developing very small gadgets which only contain a single microphone. Single-channel algorithms are especially useful in car environments or small-device based scenarios such as smart phones.

1.3 Organization of the Thesis

In *chapter 2* the fundamentals of localization are described. Binaural and monaural cues as well as their function in localization in the horizontal and vertical plane are discussed. Furthermore, mathematical tools are described which are important for the simulations in the following chapters. After the description of features which are often used in computational speech processing, commonly used classificators are introduced. Afterwards, the Gaussian mixture model and the most popular algorithm to estimate its parameters, the EM algorithm, as the main statistical model in this thesis, are explained in more detail. The last section in this chapter deals with related research which also perform single-channel localization.

In *chapter 3* the experimental setup including the speech and HRIR databases are described. Then approaches to a localization strategy are presented, whereby an adequate feature and classification method is chosen.

In *chapter 4* numerical results are presented. The problem, introduced by the adapted GMMs is solved by first constraining the area under investigation, then by training of direction-dependent GMMs. Afterwards, results of speaker-independent models and the influence of the segment length in the test scenario are shown.

Chapter 5 concludes this thesis and gives an outlook for future research. Definitions, abbreviations, and used symbols are explained in appendix A.

2 Fundamentals

2.1 Principles of Localization

Localization is something humans are confronted with day-by-day: When the street is crossed it has to be known whether a car comes along or not. When somebody calls a person's name, the person also needs to localize the sound source. There are hundreds of moments each day when the brain is forced to localize the source position of an event. But how humans are doing this? The main point is that humans have two ears. This fact, as well as the extraordinary shape of the outer ear (pinna) results in an amazing ability humans are able to localize sound sources.

2.1.1 Binaural and Monaural Cues

To define localization in the three-dimensional space, a coordinate system, as illustrated in figures 2.1 and 2.2, has to be introduced. The position of the sound source is defined relative to the head where the horizontal plane defines the left-right direction, also called azimuth θ. The vertical plane defines the up-down direction and is called elevation ϕ. In figure 2.2 the x- and y-axis span the horizontal plane whereas the y- and z-axis represent the vertical plane. Sound arrives at both ears, but the brain distinguishes between information gained from binaural signals and information gained from monaural signals. The signals arriving at the left and right ear deliver information about the interaural time difference (ITD) and interaural level (intensity) difference (ILD, IID). This means that the incident sound wave arrives at different times and with different levels at our ears. As long as the source is not exactly in front of the head, it takes longer for the sound to get to one ear than to the other one. This information is extracted from signals of both ears. Therefore, they are named binaural cues. Sound sources which originate direct in front or behind a human do not deliver ILD and ITD because they are the same for the left and the right ear. In this case monaural cues are important. Monaural cues are extracted from the signal of one ear, and represent the spectral shaping of the incident sound wave by head, shoulders, torso and, most important, the pinna.

The principle of localization based on binaural cues has been first established by Lord Rayleigh in his Duplex Theory [3]. Horizontal sound localization and vertical sound

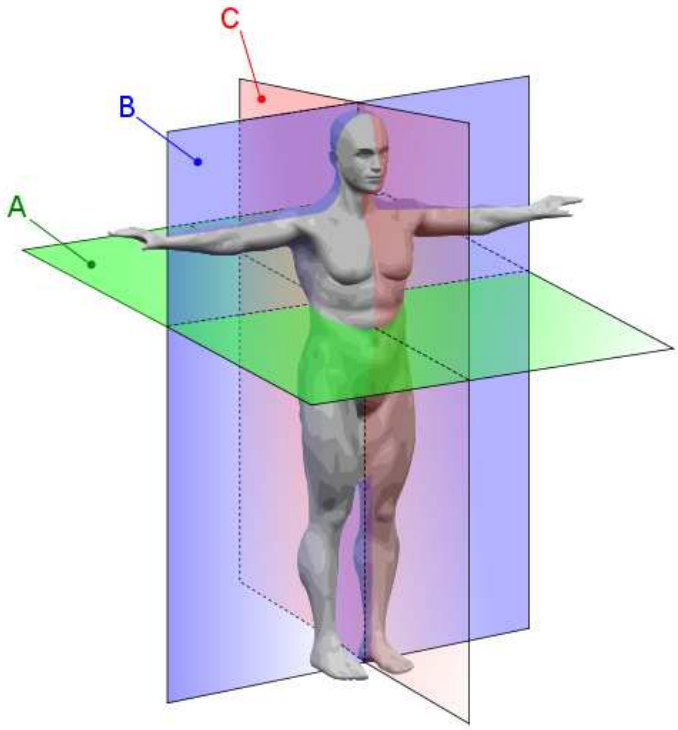


Figure 2.1: Coordinate system with different planes: (A) transverse, horizontal plane (B) coronal, frontal plane (C) sagittal, median, vertical plane [2]

localization work with different cues. Lord Rayleigh has not considered the impact of the shape of the pinna. This has been investigated in [4] and [5]. The spectrum of an incident sound wave is changed in a direction-dependent way. On the way from the sound source to the ear drum of a human, sound is shaped spectrally. This shaping can be seen as a linear time-invariant filter. Mathematically it is described by the so-called head-related transfer function (HRTF) and its time domain representation the head-related impulse response (HRIR). As a result an elementary familiarity with the sound is necessary because localization is carried out by comparing spectral patterns [6]. Monaural localization requires prior knowledge of the possible sounds because it is impossible to know whether a sound appears different because it is coming from a certain direction or if it was originally like that.

The HRTFs contain information about the spectral shaping of the sound on the way from the source position to the ear drum, as well as ITD and ILD information. $H_L(\theta, \phi, f)$ is the HRTF at the left ear and $H_R(\theta, \phi, f)$ at the right ear, respectively. If $s(n)$ is the source signal and $S(f)$ its Fourier transform representation, the input signals at the left

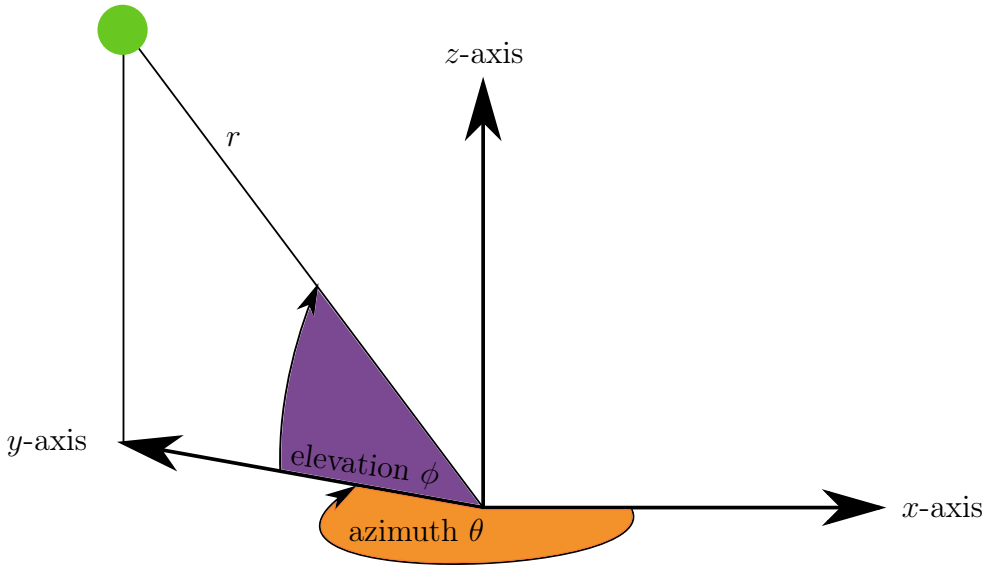


Figure 2.2: Coordinate system relative to the head with azimuth θ and elevation ϕ

and at the right ear of a human head (or an artificial head) can be written as

$$S_L(f) = H_L(\theta, \phi, f)S(f) \quad \text{and} \tag{2.1}$$

$$S_R(f) = H_R(\theta, \phi, f)S(f). \tag{2.2}$$

HRTFs depend on the frequency f and the source position including azimuth θ, elevation ϕ and range r. The ratio of $S_R(f)$ and $S_L(f)$ is the so-called interaural transfer function from which ILD and ITD can be detected. Binaural differences are independent of the source spectrum but they are inaccurate near the vertical plane because ILD and ITD are almost the same for both ears. The monaural relation $S_M(f)$ can be calculated as

$$S_M(f) = H_M(\theta, \phi, f)S(f) \tag{2.3}$$

with $S_M(f) = S_L(f)$ (or $S_R(f)$) and $H_M(\theta, \phi, f) = H_L(\theta, \phi, f)$ (or $H_R(\theta, \phi, f)$).

In [7] the influence of the pinna is investigated and a simple model is suggested. The impulse response consists of the direct path to the ear as well as an indirect path where the sound wave is reflected at the pinna. The author has been measured the impulse response of an artificial ear and found two significant echos. One echo is related to the azimuth and the other one to the elevation.

2.1.2 Cues for Localization in the Horizontal and Vertical Plane

Studies have shown that human beings can hear within a frequency range between 16 and 20000 Hz whereas the biggest sensitivity is between 2 and 5 kHz, which is reasoned in the

resonance frequency of the ear canal. Depending on the frequency range of the sound source different cues contribute to the localization in the horizontal and the vertical plane. Generally binaural cues are used to estimate the source in the horizontal plane whereas monaural cues are used in the vertical plane and in the cone of confusion where ITD and ILD are equal [8]. Depending on the frequency range of the sound source either ITD or ILD is used. This means that ITD and ILD are not equally effective at all frequencies. ITD is used in the frequency range where the wavelength is long compared to the dimensions of the head. Then the sound wave is diffracted around the head. This ratio of wavelength and head size is responsible for the type of cue which contributes mostly to the localization of the sound. At high frequencies the wavelength is short against the head size. Little diffraction occurs and therefore the shadowing effect of the head attenuates the amplitude of the incident sound wave.

ILD is neglectable below 500 Hz for sound sources which are far away. Things change when they are near the listener. Then ILD is high even at low frequencies [9]. In the frequency range between 1500 Hz – 3000 Hz sound localization is more critical than above and below these limits [4]. Cues for localization also vary depending on the nature of the sound.

Many studies have shown that spatial sound localization is still possible even if the test person has one complete blocked ear. Sound localization with only one ear suffers in the horizontal plane [10]. It is problematic to investigate the influence of the monaural spectral cues. To examine this influence, experiments are carried out where one ear of a test subjects is distorted [11] [12].

2.1.3 Distance Hearing

Localization consists of directional hearing as well as distance hearing. The cues for directional hearing are much better investigated than the cues for distance hearing. Furthermore, it is known that depending on the distance to the sound source, the loudness of the source plays an important role. In the free field case, the arriving sound wave can be seen as a plane wave. For such distances the air acts as a lowpass filter and high frequencies are more damped than low ones. The level of the sound decreases with -6 dB if the distance is doubled. This fact is exploited by the human hearing system to determine the distance [13]. If the sound source is near, the plane wave assumption has to be discarded. Then spectral cues are especially important. It is also known that ILD increases if the distance decreases below 1 meter whereas ITD is independent from the distance to the source [9]. In a reverberant environment the ratio between the direct and the reflected sound also offers a distance cue [14].

Moreover the fact whether the sound is familiar or not has an influence on distance hearing. Familiar sounds can be accurately estimated within 6 – 8 meters whereas

unfamiliar sound are only accurate within 2 – 4 meters [4].

2.1.4 Minimum Audible Angle

The values of ITDs reach from 0, when the sound source is located in front of the head, to approximately 690 μs, when the sound source is located beside the left or right ear [15]. Corresponding to the Duplex Theory, localization in the horizontal plane works well below 1500 Hz and above 3000 Hz but not in the frequency range in between.

The minimum values for resolution of ITD and ILD are about 10 μs and 0.5 dB, respectively [16]. The minimum audible angle (MMA) describes the accuracy of sound localization, i.e. MAA describes the smallest detectable change in angular position relative to the subject. For sinusoidal signals MMA is 1 ° in front of the head ($\theta = 0\,^\circ$) and 5 ° – 9 ° when the sound is located beside one of the two ears [17]. In case of broadband signals MAA is 5 ° in front and 20 ° beside one of the ears. Furthermore, head movements improve the localization task. In [18] MMA thresholds are determined for the horizontal and vertical plane. Test subjects have been allowed to move their heads. For broadband sources the mean MAA threshold in the horizontal plane is 0.97 ° whereas in the vertical plane it is 3.65 °. Sources which are in between, i.e. not directly on the horizontal or vertical plane, which is realistic, can be located accurately until 80 ° elevation with a mean MMA which is approximately the same as in the horizontal plane. Only near the vertical plane the mean MAA increases. Things change under noisy conditions or with present distractors [19]. In [20] an overview and own results are presented.

2.1.5 Summary

As mentioned before, the human hearing system and the ability to localize sound sources is amazing. The correctly estimated sound source depends on several aspects. The type of sound source can reach from very simple clicks, complex narrow- or broadband stimuli to noise or speech. Localization is well investigated in simple environments such as an anechoic environment with only little noise and no reflections. Localization cues act differently in different environments and also the importance of the different cues is changing depending on the type of the source.

Many localization strategies are based on binaural cues. Binaural source localization investigates the signals arriving on both ears. There are several computational models which are used to estimate the sound position. One of the most common principles is to calculate the cross-correlation between the ear input signals in order to estimate ITD and ILD. For example in [21] a localization method is proposed which localizes sound sources in the horizontal plane by jointly estimating ILD and ITD.

To summarize, localization is most accurate in the vertical plane (sound arrives from ahead) for sinusoidal tones. Changes in ITD cannot be detected for frequencies above 1500 Hz. Therefore, ITD is used for localization in the low frequency range whereas ILDs are detectable in the whole frequency range, but are of more importance in the higher frequency range. Monaural spectral cues are mainly responsible for localization in the vertical plane and in the cone of confusion.

2.2 Mathematical Tools – Statistical Models

In this thesis the angle of an incident sound wave in the horizontal plane is estimated. The localization method exploits data from a speech database with direction-independent speech, i.e. speech without direction information, and a database consisting of head-related impulse responses (HRIRs). This section introduces features to describe speech efficiently and statistical models for their classification.

2.2.1 Features

Often a large number of input data is not desirable in classification systems. Therefore, the input data will be transformed into a reduced presentation which contains the important properties. This smaller number is called feature [22]. Features should have values which are very similar for objects in the same class, and different for objects in a different class [23].

MFCCs

Speech production can be modeled by a linear source-filter model where the source signal (airstream from the lungs, which is either periodically interrupted by the vocal folds, or white noise produced by open vocal folds) is shaped by the vocal tract, which can be represented by a linear filter.

Mel-frequency cepstral coefficients (MFCCs) [24] are the most important and popular features for speech and speaker recognition because they describe the spectral envelope of speech segments in a compact way. MFCCs are short-term spectral-based features and they are an effective representation to retain useful direction-independent speech information. In figure 2.3 the process of creating MFCCs is explained.

First the speech signal is divided into segments. Typically a hamming window, which removes edge effects is used to produce overlapping segments. Then the DFT is calculated for each segment and the absolute value of the spectrum is taken. Afterwards,

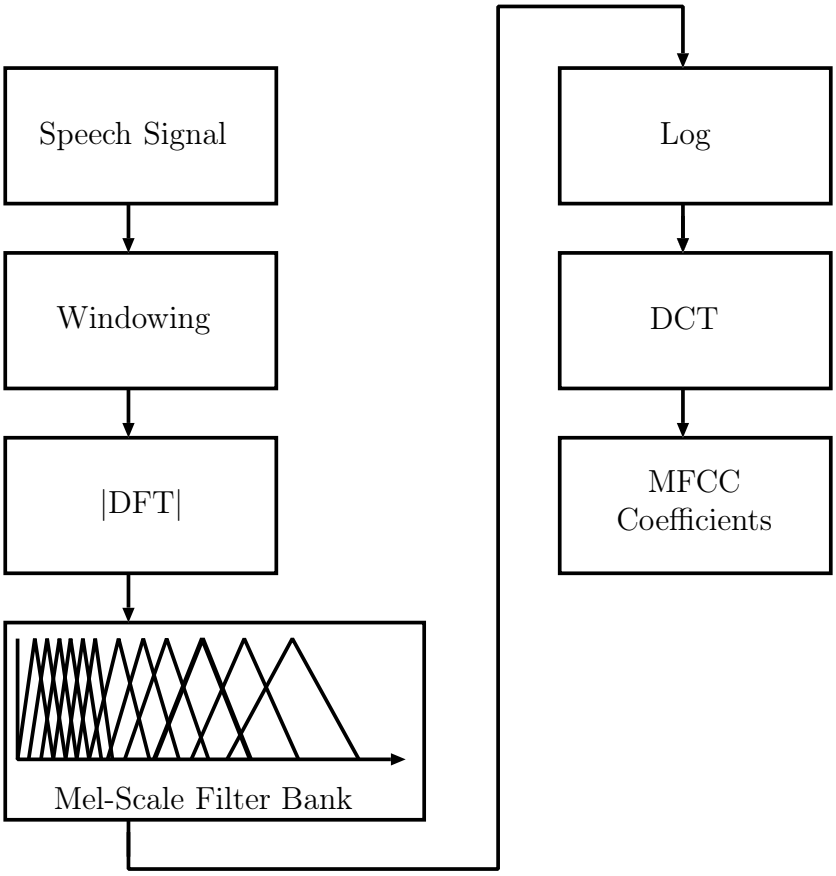


Figure 2.3: Calculation of the mel-frequency cepstral coefficients

the frequency bins are mapped on the mel scale to introduce human perceptual principles. The human perception of the frequency contents of a sound does not follow a linear scale. The mel scale is known to approximate the human auditory system more accurately. The scale maps each tone with frequency f to a subjective pitch m_{el} on the mel scale:

$$m_{el} = 2595 \log_{10} \left(1 + \frac{f}{700}\right). \tag{2.4}$$

The mel scale can be represented by a filter bank of e.g, triangular windows, where the bandwidth is approximately linear below 1 kHz and increases towards the higher frequencies.

Afterwards, the logarithm of the coefficients is taken. The last step is to apply a transform which reduces the feature dimensionality. Typically the discrete cosine transform (DCT) is applied. Afterwards, the higher frequency coefficients are discarded. The convolution of the source signal with the impulse response of the linear vocal tract filter is transformed into an addition by calculating the logarithm [25].

STFT Coefficients

The discrete Fourier transform (DFT) is one of the main tools in signal processing [26]. With the DFT a signal is transferred from the time domain into the frequency domain. Therein the spectral content of a signal can be analyzed. The DFT is often realized with the FFT, a fast and efficient implementation. There are several algorithms for an efficient calculation, but the most well-known is the algorithm of Cooley and Tukey [27], also known as Radix-2-Algorithm. For this algorithm the signal must have a length of 2^n (if not, the signal is zero-padded). Then for the calculation it is subdivided into small parts of length 2. The advantage is that by this decimation the number of operations decreases. Therefore, a fast and efficient calculation of the DFT is possible.

The spectral resolution, which means the minimal difference between two frequencies, is determined by the duration of the signal in time domain. If the duration of the signal in the time domain is long, also the spectral resolution is good. Equation

$$T = \frac{1}{\Delta f} \tag{2.5}$$

defines the relationship between the length T of the signal and the spectral resolution Δf. Zero-padding can not increase the time resolution because no additional information is introduced.

When the DFT is calculated the assumption is that the signal properties are constant. As long as the signals are stationary the calculation of the DFT is valid, but there are many signals, like speech, which are non-stationary signals. Therefore, usually speech signals are analyzed with the short-term Fourier transform (STFT) where the changes of frequencies are revealed over time. STFT extracts segments of the speech signal with a window in a way that the signal can be seen as wide-sense stationary within this segment. The windowed segments can be overlapping in a way that the boundaries are smooth. The window can be chosen arbitrary, i.e. Rectangular, Hann, Hamming, Kaiser. Usually the analysis is influenced by the window length. A smaller window leads to a better temporal resolution, but a worse spectral resolution and vice versa: the longer the window, the better the spectral resolution and the worse the temporal resolution. Then the DFT can be calculated for each of these segments [26]. The result is a time-depending Fourier transform, also called short-term Fourier transform (STFT). For each time index the N-point DFT of a time signal of length L which is shifted by R samples is calculated.

The segments are gained by multiplying the discrete time signal $x[n]$ with a window function $w[n]$ of length L. The window is not zero inside $0 \leq n \leq L-1$ and zero elsewhere. Then the STFT for the windowed signal can be written as shown in equation

$$X[rR, f] = \sum_{n=0}^{L-1} x[rR+n]w[n]e^{-j(2\pi/N)fn}. \tag{2.6}$$

N is the number of discrete frequency bins, r is an integer and R determines the position of the window in time domain. The window is shifted R samples. If $R < L$ the segments overlap and if $R > L$ some samples are discarded. To be able to reconstruct the signal, equation

$$N \geq L \geq R \tag{2.7}$$

should hold. Equation

$$x[rR + n] = \frac{1}{Nw[n]} \sum_{f=0}^{N-1} X[rR, f] e^{j(2\pi/N)fn} \tag{2.8}$$

allows the reconstruction of the time signal.

2.2.2 Classificators

MFCCs and STFT coefficients are features which are used in this thesis. Given speech material for training the corresponding features are extracted and saved as feature vectors. Now classificators are described which are able to decide from which direction an utterance arrives. Statistical pattern recognition is used to find models for data which belongs together. In [28] three different approaches to design a classificator are introduced, namely a classificator based on similarity, a classificator based on the probabilistic approach, and a classificator where the decision boundaries are optimized by an error criterion. The first two approaches shall be compared in this thesis. To make the equations better readable the used symbols are described in appendix A.

Basically, for a set of feature vectors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ where $\mathbf{x}_l$ represents a D-dimensional feature vector, a probability density function $P(\mathbf{x}_l)$ has to be estimated. Afterwards, the class-conditional density $P(\mathbf{x}_l|c_i)$ can be estimated for each class c_i. In this thesis, statistical models for each of the 72 directions are available. Each direction represents a class c_i. In [29] classificators are described with discriminant functions $g_i(\mathbf{x}_l)$. A feature vector $\mathbf{x}_l$ is assigned to a class c_i if

$$g_i(\mathbf{x}_l) > g_j(\mathbf{x}_l) \quad \text{for all} \quad j \neq i. \tag{2.9}$$

To minimize the probability of misclassification $g_i(\mathbf{x}_l)$ is taken to be equal to the posterior probability $P(c_i|\mathbf{x}_l)$. The posterior probability cannot be calculated directly, instead Bayes's theorem saying that

$$P(c_i|\mathbf{x}_l) = \frac{P(\mathbf{x}_l|c_i)P(c_i)}{P(\mathbf{x}_l)} \tag{2.10}$$

is used. Therein the posterior probability is calculated with the conditional probability $P(\mathbf{x}_l|c_i)$ and the prior probability $P(c_i)$. The prior probability $P(\mathbf{x}_l)$ is the probability that a feature vector occurs and thus a scaling factor.

With some simplifications the discriminant function can be written as

$$g_i(\mathbf{x}_l) = \ln P(\mathbf{x}_l|c_i) + \ln P(c_i) \tag{2.11}$$

The logarithm can be applied because it is a monotonically increasing function which does not influence the result.

In [22] three alternative approaches to density estimation are suggested. These are the parametric, the non-parametric and the semi-parametric method. In the parametric method a specific functional form of the density model is assumed. The density model contains a fixed number of parameters which are optimized given the feature vectors. The simplest used parametric model is the normal or Gaussian distribution which is fully described by the two parameters mean value μ and variance σ^2. The non-parametric method does not assume a specific functional form of the densities. In that case probability density functions are described in a basic definition. In the broadest sense, the probability density is defined by the probability that a feature vector falls inside the volume of some region. The K-nearest-neighbor method, which belongs to the non-parametric methods, determines this volume and thereby the probability density. For the classification, a hypersphere around a new feature vector which contains K points is determined. Then the new feature vector is assigned to the class which has the largest number of representatives inside the hypersphere. The nearest-neighbor rule is for the simple case where $K = 1$. Then the new feature vector is assigned to the class, the nearest neighbor belongs to. In the semi-parametric method a general form of the density model is assumed, where the parameters can be adapted and increased systematically. An example for this method are mixture distributions which will be explained later in more detail.

Gaussian Distribution

The normal or Gaussian distribution is the most widely used method to model probability densities. The analytic simplicity is the main reason why it is used so often. The probability density function of a normal distribution is given by equation

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{2.12}$$

in case of an univariate model and by equation

$$P(\mathbf{x}_l) = \frac{1}{\sqrt{\det(\boldsymbol{\Sigma})}(2\pi)^{D/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_l - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_l - \boldsymbol{\mu})\right) \tag{2.13}$$

in case of a multivariate model with D dimensions, where ^{T} is the transpose.

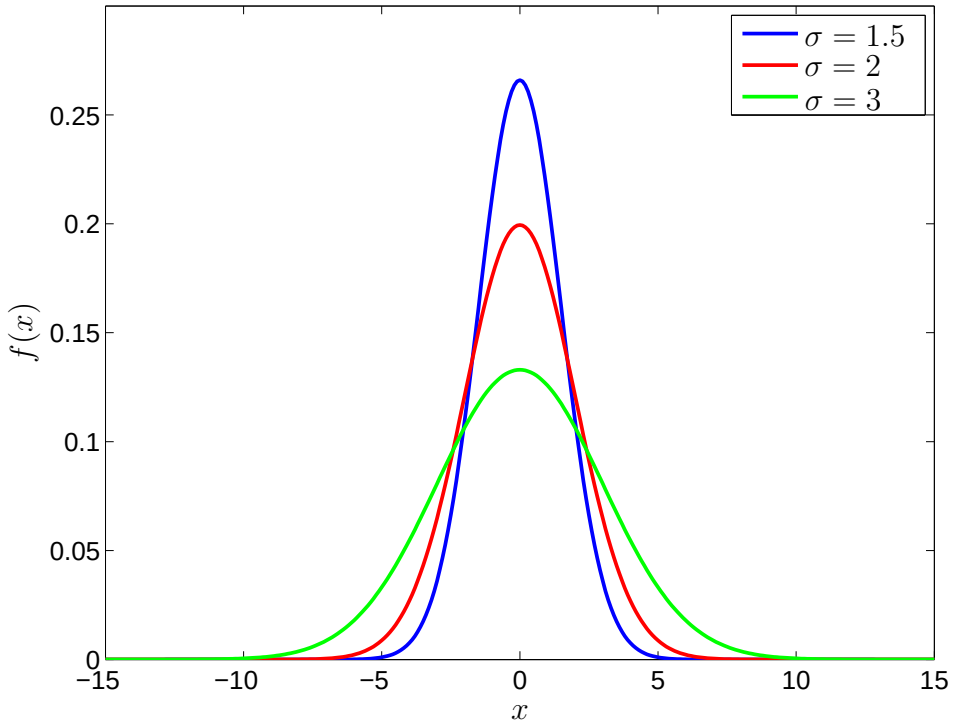


Figure 2.4: Univariate Gaussian distribution with $\mu = 0$ and varying σ

In figure 2.4 the univariate Gaussian distribution ($D = 1$) with $\mu = 0$ and varying σ is depicted. The mean value μ defines the point where the bell-shaped curve has its maximum and the variance σ^2 defines the width of the curve [30]. If $D = 2$ the multivariate Gaussian distribution looks like a cloud of dots where $\boldsymbol{\mu}$ defines the "center" of the cloud and $\boldsymbol{\Sigma}$ the "width". In the multivariate case $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)^T$ is a vector which contains the mean value for each dimension and $\boldsymbol{\Sigma}$ is a $D \times D$ matrix. This matrix is called covariance matrix and it indicates if the different dimensions (features) are independent of each other.

$\boldsymbol{\Sigma}$ can be full, diagonal or spherical. The different types are depicted for the two-dimensional case ($D = 2$) in figure 2.5. Full covariance matrices lead to a huge number of parameters for a high-dimensional input space, but this type of $\boldsymbol{\Sigma}$ fits the data best. On the other hand, the risk of over-fitting is given and it can be very costly in a high-dimensional feature space. Diagonal matrices, on the other hand, do not capture correlation among the variables. In this case the covariance matrix is $\boldsymbol{\Sigma} = \sigma_k^2 \mathbf{I}$ (with $\mathbf{I}$ the identity matrix). σ_k^2 is the k^{th} of D variances. The contours of constant probability densities are given by axis-aligned ellipsoids. They present a good compromise between quality and model size. Spherical covariance matrices, on the other hand, have a single value of σ^2 for the whole covariance matrix ($\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$). The contours in this case are concentric circles. This is the simplest form of $\boldsymbol{\Sigma}$ and is useful when only few data is

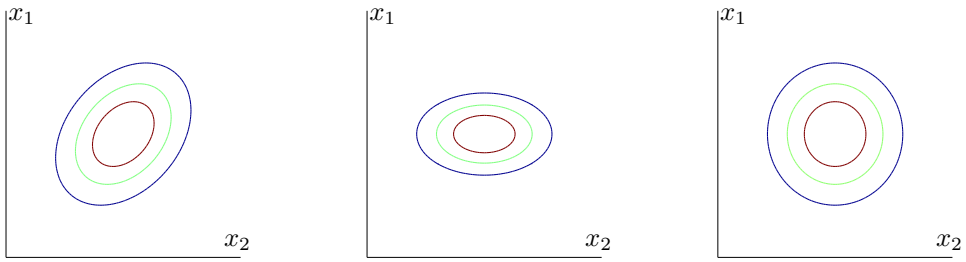


Figure 2.5: Contours of constant probability density for a covariance matrix which is (A) full, (B) diagonal and (C) spherical

available or computational speed and storage is an issue. Often $\mathbf{\Sigma}$ is taken to be diagonal because of simplicity. In this case an independent Gaussian distribution is in each dimension.

Minimum Distance Classificator

In this approach, pattern are considered to belong together when they are similar. The discriminant function (2.11) of the multivariate Gaussian distribution (2.13) can be written as

$$\begin{aligned} g_i(\mathbf{x}_l) &= \ln P(\mathbf{x}_l) = \\ &= -\frac{1}{2}\left(\mathbf{x}_l - \boldsymbol{\mu}_i\right)^T \mathbf{\Sigma}_i^{-1}\left(\mathbf{x}_l - \boldsymbol{\mu}_i\right) - \frac{D}{2}\ln 2\pi - \frac{1}{2}\ln\det(\mathbf{\Sigma}_i) + \ln P(c_i). \end{aligned} \tag{2.14}$$

In this case $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$ are the mean value and covariance matrix for the i^{th} class. If $\mathbf{\Sigma}_i$ is taken to be spherical, the terms which are independent from i can be neglected. Furthermore, the prior probabilities $P(c_i)$ are the same for all classes. Then (2.14) can be written as

$$g_i(\mathbf{x}_l) = -\frac{||\mathbf{x}_l - \boldsymbol{\mu}_i||^2}{2\sigma^2} \tag{2.15}$$

where $||\cdot||$ denotes the Euclidean distance.

Given some set of feature vectors of the training data, a prototype for each class is calculated. This prototype is the mean value vector of the feature vectors in this class [29]. The covariance matrix is assumed to be spherical and equal for all classes. Therefore, σ^2 in equation (2.15) can be neglected. If a new, unknown feature vector needs to be classified, the distance between this feature vector and the prototype of each class is calculated using the Euclidean distance

$$d = \sqrt{\sum_{k=1}^{D}(x_k - \mu_k)^2}. \tag{2.16}$$

This classificator is called minimum distance classificator.

Maximum Likelihood Classificator

In this thesis classification based on a maximum likelihood estimation is used. Because of the underlying approach the classification is termed maximum likelihood (ML) classificator. This is a probabilistic approach where a pattern is assigned to the class with the maximum posterior probability $P(c_i|\mathbf{x}_l)$. $P(c_i|\mathbf{x}_l)$ indicates the probability that a given feature vector $\mathbf{x}_l$ belongs to a class c_i. The feature vector is assigned to the class c_i with the highest posterior probability, i.e. to the class the feature vector most likely belongs to. As mentioned before, $P(c_i|\mathbf{x}_l)$ cannot be calculated directly, instead Bayes's theorem (2.10) is used. In this thesis, the conditional probability $P(\mathbf{x}_l|c_i)$ is modeled for each class with a Gaussian mixture model (GMM) which will be discussed later.

Maximum likelihood is a widely used estimator in which λ is set to the value that maximizes the likelihood function $P(\mathbf{X}|\lambda)$ [29]. A data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \ldots, \mathbf{x}_T\}$ is given, which is drawn from a distribution $P(\mathbf{x}_l|\lambda)$. $P(\mathbf{x}_l|\lambda)$ depends on the parameters specified by λ. Then the likelihood function can be written as

$$P(\mathbf{X}|\lambda) = \prod_{l=1}^{T} P(\mathbf{x}_l|\lambda) =: \mathcal{L}(\lambda|\mathbf{X}). \tag{2.17}$$

Often the so-called log-likelihood

$$\ln\left(\mathcal{L}\left(\lambda|\mathbf{X}\right)\right) = \ln \prod_{l=1}^{T} P\left(\mathbf{x}_l|\lambda\right) \tag{2.18}$$

is taken because it is easier to calculate. It can happen that there are numerical problems in calculating the likelihood function. Likelihood values might be too small which results in a numerical underflow. This problem can be solved by using the log-likelihood function instead. In the case of GMMs the likelihood function is a product of individual likelihoods, the log converts the product into a sum. As a result a sum of logarithmic sums has to be calculated. The inequality

$$\ln\left(x_1 + x_2\right) \neq \ln\left(x_1\right) + \ln\left(x_2\right) \tag{2.19}$$

describes that the logarithm of a sum is not the same as the sum of the logarithm. There is a trick to calculate $\ln\left(x_1 + x_2 + \cdots + x_N\right)$ from the log values $\ln\left(x_1\right), \ldots, \ln\left(x_N\right)$: The log values are divided by the largest term x_l and the scaled terms have to be converted to a linear term.

$$\ln\left(\sum_{n=1}^{N} x_n\right) = \ln\left(x_l\right) + \ln\left(\sum_{n=1}^{N} e^{\ln(x_n)-\ln(x_l)}\right) \tag{2.20}$$

To determine the model which matches best to the given features of a test sentence $\mathbf{X}$ the model is chosen to give the largest posteriori probability. For example, there are 10 models $\lambda_1, \lambda_2, \ldots, \lambda_{10}$. Each model is trained with speech material which arrives from one of 10 possible directions θ. Then a test sentence is spoken from one of the 10 positions and the directions θ have to be estimated. One gets

$$\hat{\theta} = \arg \max_{1 \leq i \leq 10} P(\lambda_i|\mathbf{X}) = \arg \max_{1 \leq i \leq 10} \frac{P(\mathbf{X}|\lambda_i) P(\lambda_i)}{P(\mathbf{X})} \tag{2.21}$$

with $P(\lambda_i) = 1/10$. This means that it is equally probable that the test sentence is spoken from one of the directions. $P(\mathbf{X})$, the prior probability that the sentence is spoken, is the same for all models and therefore a scaling factor. Now, (2.21) simplifies to

$$\hat{\theta} = \arg \max_{1 \leq i \leq 10} P(\mathbf{X}|\lambda_i) . \tag{2.22}$$

Again, maximization is done in the log domain. The estimation of $\hat{\theta}$ is given by the maximum likelihood estimation. In practice, estimates of distributions are used in place of the true densities.

Summary

To summarize, given the feature representation of an input signal a classificator maps this features to a set of labels, which consists of a number of known classes. In other words, an input $\mathbf{x}_l$ is assigned to one of C classes using a certain approach. The classification process always consists of two steps.

In the first step the classificator is trained to represent a class. The classes of the given training material can be known (supervised classification) or unknown (unsupervised classification). In this thesis either a prototype is saved for each class, or the parameters of a density function are estimated. The second step is the test scenario where a classificator is used to estimate results. Classification can be done by calculating the minimum distance or the maximum likelihood based on the posterior probability of the observed data $\mathbf{x}_l$ given the densities of the different classes. A class can be represented by a Gaussian mixture model (GMM). GMMs are the main concept in this thesis. Therefore, they will be described in more detail in the next section.

2.2.3 Gaussian Mixture Models

Mixture models are probabilistic models. They are useful to model arbitrary density distributions. A Gaussian mixture model (GMM) consists of a linear combination of

K multivariate Gaussian probability density functions given by equation (2.13). The mathematical representation is given by equation

$$P(\mathbf{x}_l|\lambda) = \sum_{m=1}^{K} b_m P_m(\mathbf{x}_l). \tag{2.23}$$

K is the number of Gaussian components, $\mathbf{x}_l$ is a D-dimensional random vector, b_m are the weights for each component and P_m is a single, multivariate Gaussian distribution. For b_m equation

$$\sum_{m=1}^{K} b_m = 1 \tag{2.24}$$

holds. With $\lambda = \{(b_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m); m = 1, 2, \ldots, K\}$ the whole mixture model is described.

The model can differ depending on the appearance of the covariance matrix:

- one $\boldsymbol{\Sigma}$ per Gaussian component,
- the same $\boldsymbol{\Sigma}$ for all Gaussian components $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K$.

Also we can use the same $\boldsymbol{\Sigma}$ for the model of each class. In this thesis there is a diagonal $\boldsymbol{\Sigma}$ for each Gaussian component of the mixture model.

GMMs have the power to evaluate a smooth and good approximation of an arbitrary formed density. This means, if there is an arbitrary shaped density, and the underlying density is not known, GMMs are a very good as approximation. Also if the features are not statistically independent it is allowed to take diagonal $\boldsymbol{\Sigma}$, because multiple Gaussians act together to build the overall probability distribution function.

The parameters λ of a model are estimated from a set of training data. There are several techniques available, but the most common is the expectation-maximization (EM) algorithm which iteratively determines the unknown model parameters. Within this algorithm the likelihood function, which is the probability of the training data given the parameters of the model, is maximized [31].

The EM algorithm

The EM algorithm maximizes the likelihood of a GMM when a training data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ is given. $\mathbf{X}$ is assumed to be independent and identically distributed. The training data set consists of T feature vectors. The log-likelihood function, which is often calculated, is given by (2.18) and in the GMM case turns into

$$\ln\left(\mathcal{L}\left(\lambda|\mathbf{X}\right)\right) = \ln \prod_{l=1}^{T} P\left(\mathbf{x}_l|\lambda\right) = \sum_{l=1}^{T} \ln\left(\sum_{m=1}^{K} b_m P_m\left(\mathbf{x}_l\right)\right). \tag{2.25}$$

The algorithm is a general solution to the problem of finding the optimal parameters in the maximum likelihood sense. The maximization is done iteratively because the likelihood function is a nonlinear function of the model parameter and it is not possible to estimate the parameters directly. The algorithm finds the maximum posteriori probability for a given observed feature vector. $P(\mathbf{x}_l|\lambda)$ is the probability that the vector $\mathbf{x}_l$ is produced by the model which is specified in λ. Furthermore, a model λ_i will be trained for each class i.

The EM algorithm works iterative which means that the algorithm has to be initialized with a start model λ. Then a new model $\tilde{\lambda}$ is estimated such that $P(\mathbf{X}|\tilde{\lambda}) \geq P(\mathbf{X}|\lambda)$. Then the new model $\tilde{\lambda}$ is used for the next iteration step. This is continued as long as a convergence threshold is not reached. Every dimension of the mixture model is a multivariate Gaussian distribution and has its own mean values $\boldsymbol{\mu}_m$ and covariance matrices $\boldsymbol{\Sigma}_m$. The problem is that it is not known which training vector contributes to which Gaussian.

- Estimation Step:

$$P(m|\mathbf{x}_l, \lambda) = \frac{P(m, \lambda)P(\mathbf{x}_l|m, \lambda)}{P(\mathbf{x}_l, \lambda)} = \frac{b_m P_m(\mathbf{x}_l)}{\sum_{m=1}^{K} b_m P_m(\mathbf{x}_l)} \tag{2.26}$$

- Maximization Step:

$$\tilde{b}_m = \frac{1}{T}\sum_{l=1}^{T} P(m|\mathbf{x}_l, \lambda) \tag{2.27}$$

$$\tilde{\boldsymbol{\mu}}_m = \frac{\sum_{l=1}^{T} P(m|\mathbf{x}_l, \lambda) \cdot \mathbf{x}_l}{\sum_{l=1}^{T} P(m|\mathbf{x}_l, \lambda)} \tag{2.28}$$

and

$$\tilde{\boldsymbol{\Sigma}}_m = \frac{\sum_{l=1}^{T} P(m|\mathbf{x}_l, \lambda) \cdot \mathbf{x}_l \mathbf{x}_l^T}{\sum_{l=1}^{T} P(m|\mathbf{x}_l, \lambda)} - \tilde{\boldsymbol{\mu}}_m \tilde{\boldsymbol{\mu}}_m^T \tag{2.29}$$

with

$$\tilde{\boldsymbol{\Sigma}}_m = \begin{bmatrix} \tilde{\sigma}_{1,1}^2 & \cdots & \tilde{\sigma}_{1,D}^2 \\ \vdots & \ddots & \vdots \\ \tilde{\sigma}_{D,1}^2 & \cdots & \tilde{\sigma}_{D,D}^2 \end{bmatrix} \tag{2.30}$$

In the estimation step the posteriori probability for the m^{th} Gaussian component is calculated by equation (2.26). Then the m^{th} Gaussian mixture component parameters $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are estimated in the maximization step by equations (2.27) - (2.29).

$P(m|\mathbf{x}_l, \lambda)$ is the probability for $\mathbf{x}_l$ to belong to the m^{th} component of the mixture model. Each training vector contributes to the new parameter but the probability of the mixture should be higher, to belong to one component m than to another. This

means that for all training vectors $\mathbf{x}_l$ there are corresponding probabilities $P(m|\mathbf{x}_l, \lambda)$ to belong to the components.

Within the EM algorithm the initialization is important because the results depend on the initial values. In this thesis mixture weights are initialized by $b_m = \frac{1}{K}$, $\boldsymbol{\mu}_m$'s are initialized randomly and $\boldsymbol{\Sigma}_m$'s are initialized with the variance of the input training data for each dimension and is equal for each component. The number of iterations is set to eight. If the number of components is too large and the number of training data is too small covariance matrices become singular or ill-conditioned [32]. Moreover, the optimal number of components strongly depends on the available training data.

Number of Components of the GMM

The goal of a statistical model is to describe the given training data set in an optimal way. On the other hand the model should also fit to unknown data, received from the same unknown distribution. The case of underfitting, where the statistical model does not match optimally to the training data, as well as the case of overfitting, where the model fits "only" or "too well" to the training data should be avoided.

Let the number of components in GMMs be K. For high values of K, the GMM can describe an arbitrary distribution, but we force the danger of overfitting which is easily reached. There are several methods to calculate the optimal number of components. If components are added to the GMM, the likelihood will always increase. Therefore, the maximum likelihood principle will score at the highest number of components.

Analytic methods, like the Akaike information criterion (AIC) [33] or the Bayesian information criterion (BIC) [34] investigate the value of the likelihood as well as the number of parameters and the amount of training data. In [35] the number of components in a text-independent speaker recognition system is investigated. Well known criteria, like the BIC and the AIC, are compared. Furthermore, the Goodness of Fit (GOF) is investigated and proved to reach comparable results. The Goodness of Fit (GOF) is a measure for the limit of the amount of components in the GMM. To calculate it, the data vectors which contribute to the mixtures must be known. Another method, which is also used in this thesis, is the empirical method, where the whole data set is divided into a training data set and a test data set. The model is trained using the EM algorithm and the performance is evaluated on the test data set.

2.3 Related Work

Sound source localization is a challenging task and many authors address this problem in their research. Typically sound localization algorithms are based on processing signals received by multiple, spatially separated sensors.

These approaches typically use either a beamformer and maximize its steered response power or use methods based on the time difference of arrival [36]. Only few tried to exploit HRTFs to reach a strategy for a single-channel method.

In [37] a sound localizer with two microphones, two artificial ears and a HRTF database is derived. HRTF measurements are taken and interpreted as linear time-invariant FIR (finite impulse response) filters. To use the HRTF efficiently and to fasten the algorithm of source localization the length is truncated using several techniques. A diffuse-field equalization, which shortens the HRTF from 512 to 128 samples, a balanced model truncation, which separates the FIR filter into a low-order IIR (infinite impulse response) filter, and a principal component analysis are carried out. The localization strategy is the following: First the inverse filters of the HRTF data set are calculated offline by a simple exchange of numerator and enumerator. The localization is made via the inverse filters of the reduced HRTFs. The original signal is filtered with the original HRTF in order to generate a direction-dependent signal. Afterwards, the resulting signal, which contains a certain direction information is convolved with the inverse, reduced HRTF to gain the original signal. These calculations are made for the left as well as for the right ear. The correlation between the left ear signal and the right ear signal is calculated and the filter which produces the maximal autocorrelation is taken as the direction of the source. Furthermore, to verify the result, also the minimum distance between the two signals is calculated. The diffuse-field equalized HRTF set reaches the best results with a localization accuracy of 96 %. As a next step the previous authors concentrate on the problem of filter inversion. A simple filter inversion makes a filter unstable. In [38] an inversion algorithm based on outer-inner factorization is presented.

In [39] again two microphones, one placed inside and one placed outside the ear canal are used to determine the sound localization in the three-dimensional space. In this paper the sought HRTFs are estimated using the self splitting competitive learning (SCCL) clustering technique, a well-known technique in image processing. The basic idea is to isolate the HRTFs from the two microphone signals. This is possible because the first signal contains the HRTF and the second does not. HRTFs depend on the source position. Ideally they are different for each position. Both signals are divided in the spectral domain. The result represents the effect of the pinna, head and torso and should cluster at each frequency around the corresponding complex HRTF value. Then the algorithm clusters the data in all frequency bins over several time frames separately. The last step is to find the HRTF from the database which corresponds to the estimated HRTF. From 100 measurements with a required speech signal length of 4.7 seconds, 62 % of the

azimuth angles and 71 % of the elevation angles are estimated correctly. The remaining 38 % and 29 % deviate 5.5 ° and 10.5 ° on average. The SSCL clustering technique is replaced in [40] by a simple correlation mechanism, using a set of HRTFs. The above mentioned division result is correlated with each HRTF from a database and outputs the normalized correlation coefficients. The following decision device gives an estimation of the sound source location by identifying the maximum correlation coefficient. From 1000 measurements 60 % of the speech sources are correctly estimated in the free space with 2.5 ° azimuth deviation and 5 ° elevation deviation.

Although the localization strategies mentioned above use monaural cues for localizing azimuth and elevation, two microphones are needed. The following procedures use only one microphone but, as a drawback, they are limited to a plane.

In [41] a system is introduced which localizes sound in the vertical plane with a single microphone. A so-called neuromorphic microphone is used which copies the structure of the outer ear. This means that it introduces time delays like the convoluted structure of the pinna. As a simplified model, the artificial pinna is designed as a parabolic structure with a single microphone. The microphone represents the ear canal. The recorded sound consists of the direct sound and an echo. Therefore, localization is possible because of the analysis of the echo times. Then the recorded signal is processed with a gamma filter. The original HRTFs are more complex than the simplified versions. The gamma filter approximates the recorded signal with a FIR filter. Different coefficients represent different angles of the incident sound wave. After that a multilayer perceptron neural network is trained to reach final localization results. Localization within a range of 8 ° is possible.

A later paper of the same authors is [42]. In this paper a localization method using one microphone is developed and implemented. A CMOS integrated circuit is designed and fabricated. The algorithm is inspired by the fact that direction is encoded in the time delay between the direct sound and the reflected sound. A microphone is located between a special reflection surface and the sound source. The time delay between the direct sound and the reflected sound is calculated based on an onset detecting circuitry which detects the energy change in the sound signal. For the testing scenario only pulses are allowed. The sound source is moved along a line and the time delays are computed. Then different LEDs are activated depending on the calculated time delays. Localization is made indirectly via the time delays.

Also in [43] localization with a single microphone is carried out. First a GMM is trained with clean speech. Afterwards, the acoustic transfer function is estimated by maximizing the likelihood with the clean speech GMM and test material uttered from each position. Then a GMM for each position is trained with the estimated acoustic transfer function. The last step is to find the GMM with the maximum likelihood among the estimated GMMs corresponding to each position. As features cepstral coefficients are taken because they are assumed to effectively represent clean speech information.

Experiments are taken with synthesized, reverberant speech. One, five and ten training sentences which are uttered from $10, 30, 50, 70, \ldots, 150$ and 170 degrees (nine directions) are used to train the acoustic transfer function GMMs. The evaluation is carried out on three directions (30, 90 and 130 degrees). With five sentences to estimate the acoustic transfer function and in the task where three different directions have to be estimated the direction accuracy is almost 100 %. Additionally, tests are carried out with GMMs trained with the observed speech instead of GMMs trained with the acoustic transfer functions. The accuracy of the estimation decreases. In the follow-up paper [44] the same approach is evaluated in (a) a simulated reverberant environment, (b) a simulated noisy reverberant environment using a speaker-independent speech model and (c) in a real environment using a speaker-dependent speech model. The localization accuracy increases when the number of mixtures, the amount of training data as well as the test segment length is increased. In (a) 3, 5, 7 and 9 positions are evaluated with a localization accuracy of 51.3 % for 9 positions. In (c) 5 positions are used for training and 2 for the test which results in a localization accuracy of 94.8 % for the first position and 79 % for the second.

In [45] sound localization is carried out with a single microphone and an "artificial pinna". In this paper the prior distribution of sound as well as the direction-dependent transfer function of the pinna are modeled. A signal, recorded by one microphone, is given as the convolution of the sound source and the direction-dependent transfer function plus additive white Gaussian noise. Then the source signal is modeled using a Hidden Markov Model (HMM) with training material of different sounds (speech, animal sounds and natural sounds). The transfer functions are estimated with standard noise excitation methods. The most likely value for θ is estimated by a ML framework. Four different "artificial pinnas" were constructed to record transfer functions that depend strongly on the direction of the sound. Then the transfer functions for angles between $0\,^\circ$ and $345\,^\circ$ in steps of $15\,^\circ$ are estimated and the values are interpolated for finer angles. The best results for mixed human speech is delivered by a pinna where a plastic-cast, that has smooth grooves build on it in various directions, is used. In this case an average error of $7.7\,^\circ$ occurs.

Finally, in [46] the distance of a sound source is estimated using a single microphone recording in a room environment. A number of statistical and source specific features are computed from speech signals. With pattern recognition techniques a robust distance estimator is developed. Features, based on the source signal which depend on the distance between source and receiver are introduced. These features consider the fact that temporal and spectral characteristics of speech in reverberant environments depend on the distance. The presented features are (1) the ratio of the 0.9 percentile of a linear prediction residual and the root mean square values, (2) the kurtosis of the linear prediction residual, (3) the skewness of the spectrum and (4) the skewness of energy differences. The first two features are calculated from the linear prediction residual (LP residual). The amplitudes of the LP residual give information about the reverberations in a signal.

Therefore, they are related to the distance. Reverberation increases when the distance to the microphone increases. Then the kurtosis of the LP residual takes lower values. The third feature takes into account that reverberation smears energy across frequencies. Additionally, the four features are also extracted from a bandlimited version of the signal. GMMs are trained and evaluated for each distance (0 m, 1 m, 1.5 m, 2.5 m, and 3.5 m). All features together deliver the best results with 85 % localization accuracy. For a close distance the system works best whereas the performance decreases with the distance. Moreover, a speaker-dependent scenario outperforms the speaker-independent case.

3 Numerical Experiments

In this chapter approaches to estimate a sound position in the horizontal plane based on a set of measured head-related impulse responses (HRIRs) are developed. For training and test a database with direction-independent speech samples and a database with 72 HRIRs is available. Therefore, first a statistical model is developed and afterwards a classificator is chosen to localize a speech sound. A signal, received at a single microphone, has to be classified to a certain azimuth θ.

3.1 Experimental Setup

3.1.1 The Speech Database

The database consists of utterances spoken by four female and four male speakers at a sampling frequency $f_s = 16\,\text{kHz}$. The utterances are direction-independent which means that they do not contain any direction information. It is divided into two data sets, one for training and one for the test. The different utterances of each speaker are not identical and as a result also the length changes. The duration of an utterance influences the performance of the estimation of the correct angle of the incident sound wave. In table 3.1 and 3.2 the durations of the utterances for each speaker are listed in the training as well as in the test set. The presumption is that the error rate depends on

	$\boldsymbol{Sp}_1$	$\boldsymbol{Sp}_2$	$\boldsymbol{Sp}_3$	$\boldsymbol{Sp}_4$	$\boldsymbol{Sp}_5$	$\boldsymbol{Sp}_6$	$\boldsymbol{Sp}_7$	$\boldsymbol{Sp}_8$
Gender	m	f	f	m	f	f	m	m
No. of utterances	65	67	65	65	66	65	70	66
Duration [min]	10.03	9.91	8.78	9.47	9.24	8.09	9.42	9.54

Table 3.1: Duration of all training utterances for female (f) and male (m) speakers

the length of the speech material used to train the statistical models, and on the length of the test utterance. If different test utterances are used and their length is variable, the performance will also change depending on the length of the utterance. Furthermore, as shown later, the model depends on the initial values of the EM algorithm.

	$\boldsymbol{Sp}_1$	$\boldsymbol{Sp}_2$	$\boldsymbol{Sp}_3$	$\boldsymbol{Sp}_4$	$\boldsymbol{Sp}_5$	$\boldsymbol{Sp}_6$	$\boldsymbol{Sp}_7$	$\boldsymbol{Sp}_8$
Gender	m	f	f	m	f	f	m	m
No. of utterances	15	15	15	15	15	15	15	15
Duration [min]	1.82	1.73	2.01	2.14	2.02	1.43	1.74	1.72

Table 3.2: Duration of all test utterances for female (f) and male (m) speakers

3.1.2 The HRIR Database

The HRIR database has been downloaded from the internet [47]. The recordings were made in an anechoic room with the KEMAR-mannequin from Bill Gardner and Keith Martin at MIT Media Lab. KEMAR is designed for making HRTF measurements. Its dimensions are motivated by a natural human being. The impulse responses were generated using maximum length pseudo-random binary sequences at a sampling frequency $f_s = 44.1\,\text{kHz}$. Impulse responses for 710 source locations were generated. For this thesis, the sampling frequency is reduced to $f_s = 16\,\text{kHz}$ and only the horizontal plane is considered. This means that at an elevation of $0\,^\circ$ all positions for the azimuth are taken which results in 72 HRIRs ($0\,^\circ$ – $355\,^\circ$ in steps of $5\,^\circ$). The range between the KEMAR and the sound source is 1.4 m. The 72 HRIRs are depicted in figure 3.1, where the values are calculated in a logarithmic scale. Figure 3.2 shows the different HRTFs. Here the values of each HRTF are also logarithmic. It is eye-catching that the main peak is around 2200 Hz for nearly all HRTFs. Figure 3.3 shows the correlation of HRTFs with each other. It can clearly be seen that HRTFs of some angles are correlated to other ones, e.g., $0\,^\circ$ and $5\,^\circ$. On the other hand, some are clearly uncorrelated, e.g., $110\,^\circ$. The assumption is that the estimation works good for HRTFs which are uncorrelated and bad for the others.

3.1.3 Pre-Processing, Training and Test

The experiments are carried out with Matlab. The basic scheme is depicted in figure 3.4. There is a pre-processing step where speech material for training and test is synthesized and features are extracted. After the pre-processing the training is carried out, where speech material from the training set is taken to train a statistical model. The statistical model can be speaker-dependent or speaker-independent. The main difference is that different input material is taken to train the statistical models. The principles which are explained in the next sections hold for the speaker-dependent and the speaker-independent case. Then the results are evaluated in the test scenario.

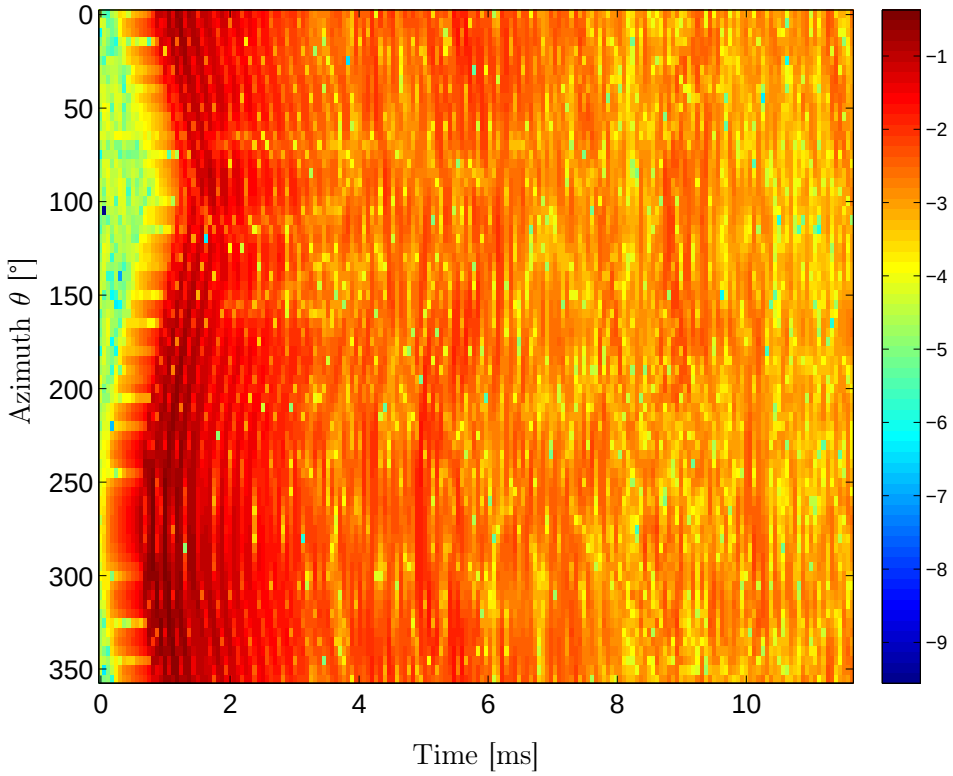


Figure 3.1: Head-related impulse responses for elevation $\phi = 0$ and azimuth θ between 0 ° and 355 ° in steps of 5 °

Synthesis of the Input

As shown in figure 3.4 and mentioned in the previous section, the speech database is divided into a training and a test set. The training set is used to train a statistical model. This can be done either by training of a direction-independent or a direction-dependent statistical model. The same data set can be used to train each of the two models. Speech input for the training of direction-dependent statistical models and the test is realized in a synthetic way whereas the input for the direction-independent statistical model remains unchanged. Utterances are synthesized with the speech database and the head-related impulse responses of a HRIR database. This is also shown in figure 3.4. Given a direction-independent speech signal $x_{di}(t)$ and one of the HRIRs $h_\theta(t)$, where θ defines the azimuth of the HRIR, a direction-dependent signal $x_\theta(t)$ can be calculated. $x_\theta(t)$ can be described mathematically by a linear convolution of the direction-independent speech signal and the HRIR.

$$x_\theta(t) = x_{di}(t) * h_\theta(t) \tag{3.1}$$

As explained in [26] the Fourier transform turns a convolution in the time domain into a multiplication in the frequency domain. Let $X_\theta(f)$, $X_{di}(f)$ and $H_\theta(f)$ be the Fourier transforms of $x_\theta(t)$, $x_{di}(t)$ and $h_\theta(t)$, respectively. Then equation (3.1) turns into equa-

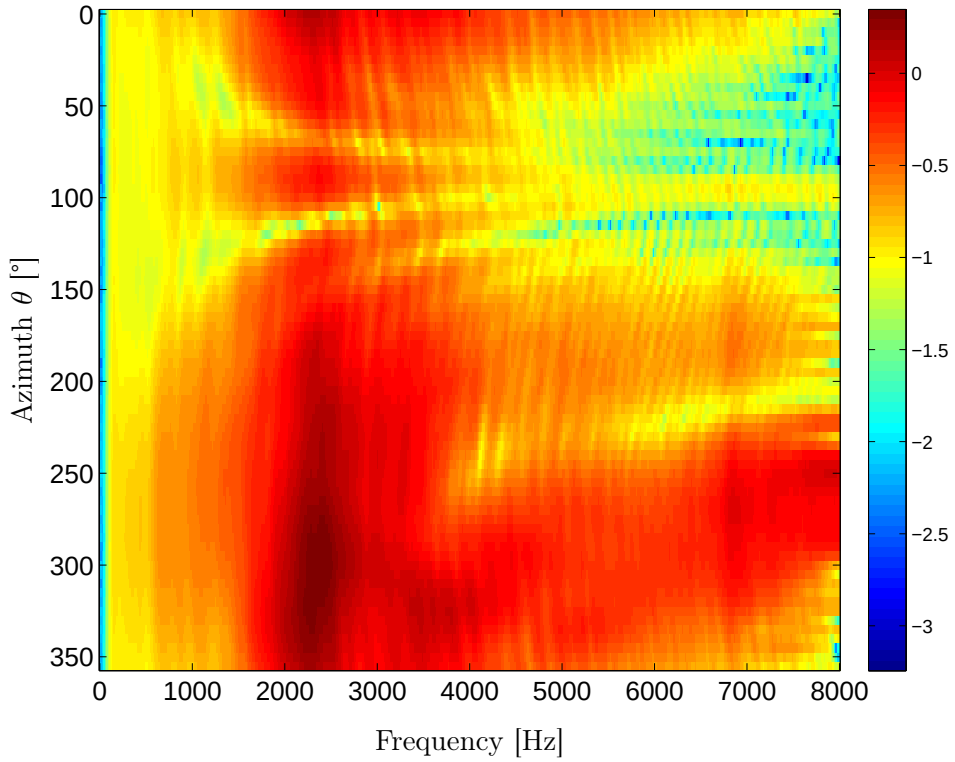


Figure 3.2: Head-related transfer functions for elevation $\phi = 0$ and azimuth θ between 0 ° and 355 ° in steps of 5 °

tion

$$X_\theta(f) = X_{di}(f) \cdot H_\theta(f). \tag{3.2}$$

In fact, if the DFT is used in equation (3.2) the multiplication is a circular convolution, whereas the convolution in the time domain is linear.

Feature Extraction

MFCCs and STFT coefficients are well known and widely used features in speech processing.

The calculation of MFCCs is depicted in figure 2.3. A hamming window with the length 256 is used. Furthermore, 29 mel filterbanks are taken. The feature vector finally consists of 20 coefficients per time step. STFT coefficients are calculated in two different ways. First (type 1) a hamming window with the block length of 512 samples, 50 % overlap and a 1024-point FFT is used. Afterwards the logarithm of the absolute values of the coefficients is taken. The second variant (type 2) is to replace the hamming window with a rectangular window with the length of 1024 samples. No overlapping is applied, but the 1024-point FFT is unchanged. Also the last step, to calculated the

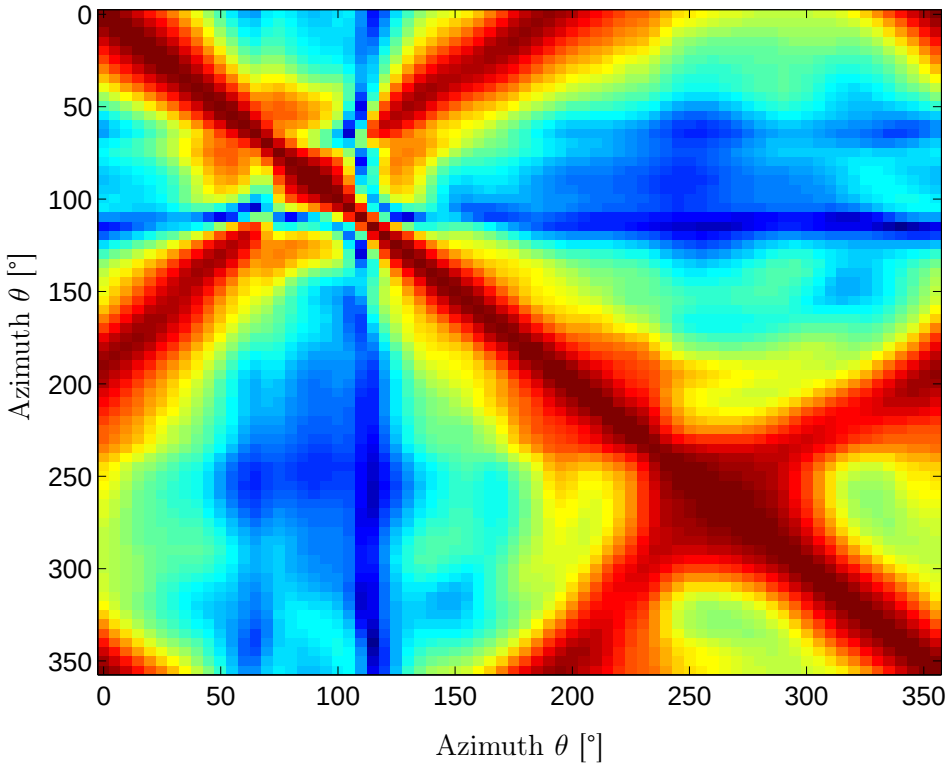


Figure 3.3: Autocorrelation of HRTFs

logarithm of the absolute values of the coefficients is carried out. This variant is shown in figure 3.5.

MFCCs as well as STFT coefficients are calculated for each time step. This means that the result is a matrix where each column represents a feature vector for a given period of time. This period of time is defined by the time window used.

Evaluation – the Absolute Error Angle $|\epsilon|$

After the training of the statistical models the test is carried out. Speech utterances from each direction are synthesized and tested in order to estimate the direction. Then the error angle ϵ is calculated and saved. To define ϵ a coordinate system, which is shown in figure 3.6, is defined. The head can be replaced by a microphone which captures the incident sound wave. The elevation is $0\,^\circ$ and the angles increase clockwise. $\theta = 0\,^\circ$ is in front of the subject, $\theta = 90\,^\circ$ beside the right ear and $\theta = 270\,^\circ$ beside the left ear. Based on this consideration the error angle always lies between $-175\,^\circ$ and $+180\,^\circ$, e.g., if the test utterance origins at $\theta = 5\,^\circ$ and the estimated angle $\hat{\theta} = 345\,^\circ$, the error angle ϵ is $-20\,^\circ$. Afterwards, the absolute value of ϵ is taken and the mean value $\mu_{|\epsilon|}$ and standard deviation $\sigma_{|\epsilon|}$ is calculated for the whole test set.

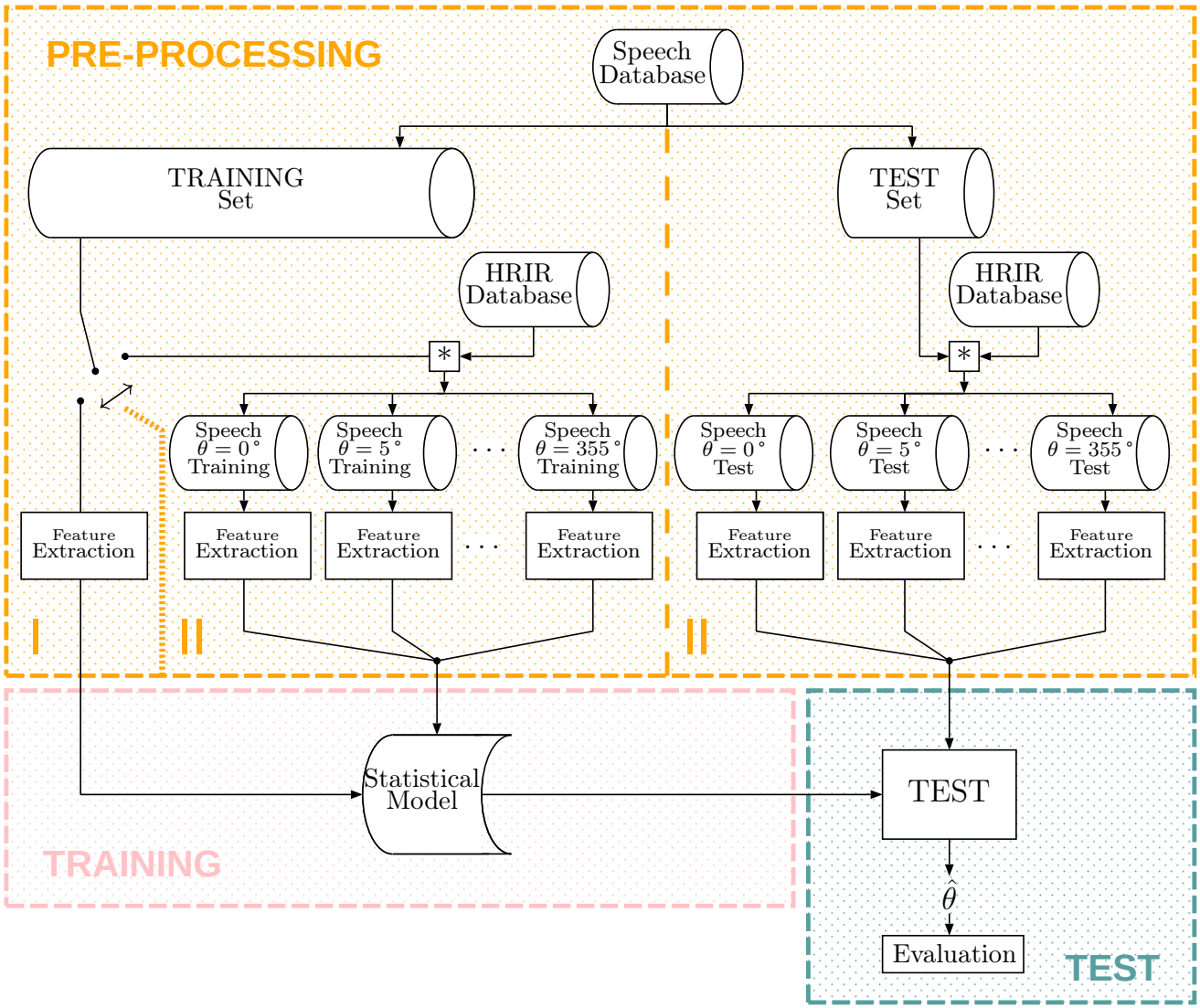


Figure 3.4: Block diagram of experimental setup: pre-processing with division of the speech database into training set and test set, synthesizing of direction-dependent speech utterances and feature extraction; (|) for direction-independent statistical model, (||) for direction-dependent statistical model and (||||) always valid; training of either direction-independent or direction-dependent statistical model and test (classification) and evaluation of the results

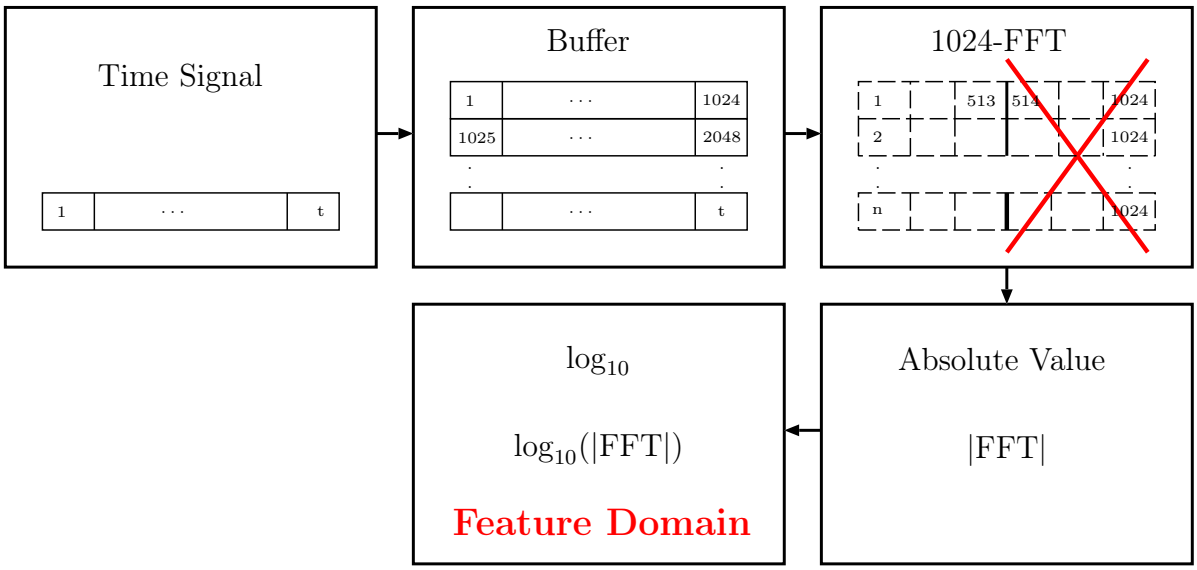


Figure 3.5: Calculation of the STFT feature vectors; type 2

3.1.4 Summary

To summarize, speech utterances from each of the 72 directions in the horizontal plane are synthesized by a convolution of the speech utterance with each of the 72 HRIRs from the database. This means that each sentence is uttered from each direction. Afterwards, features are extracted. Then either direction-independent utterances or synthesized direction-dependent utterances are used to train a statistical model. Finally the test is carried out and results are evaluated. This is shown in figure 3.4. For the test 15 test utterances from each of the 72 directions ($0\,^\circ$ – $355\,^\circ$ in steps of $5\,^\circ$) are synthesized and evaluated. The angle $\hat{\theta}$ is estimated and the error angle ϵ is saved. Then the absolute error angles $|\epsilon|$ for each direction, all 15 utterances and all eight speakers, are averaged and result in a mean absolute error angle $\mu_{|\epsilon|}$. Furthermore, the standard deviation $\sigma_{|\epsilon|}$ of the absolute error angles are calculated.

3.2 Approaches for Estimating the Angle

3.2.1 MFCCs with Minimum Distance

The production of human speech can be seen as a linear filtering of an excitation signal by the vocal tract. In the cepstral domain this convolution turns into an addition. Therefore, it is possible to separate the impulse response and the excitation signal. As mentioned before the HRIRs act as a filter on the direction-independent speech signal.

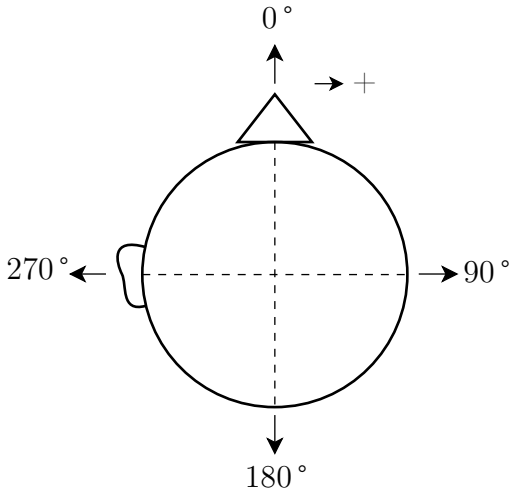


Figure 3.6: Absolute error angle $|\epsilon|$ relative to the head

Based on this consideration MFCCs are chosen to be the feature.

Training and Test

MFCC features are extracted from the direction-independent training set. Then the statistical model is trained with these features. For the training temporal averaging over the whole training data set of each speaker is applied to achieve a 20-dimensional prototype vector $\hat{\boldsymbol{\mu}}_s$ for each speaker s. These prototypes are saved. In the test scenario the direction-dependent test utterance is also converted into MFCCs with the same criteria as in the training scenario. Again, temporal averaging over all feature vectors are applied to gain a 20-dimensional feature vector $\hat{\boldsymbol{\mu}}_\theta$. Moreover, each direction can be seen as a different class. To determine the class, the test utterance belongs to, a minimum distance classificator is used. So as to do, the prototype feature vectors are subtracted from the feature vector of the test utterance. The result is assumed to yield the contribution of the HRIR. Then the result is compared to the MFCC representation of the HRIR database via minimal distance calculation as shown in figure 3.7. In other words, the nearest neighbor is assumed to be the filter with which the direction-independent speech signal is convolved. By knowing the filter, the direction of the sound source is also known.

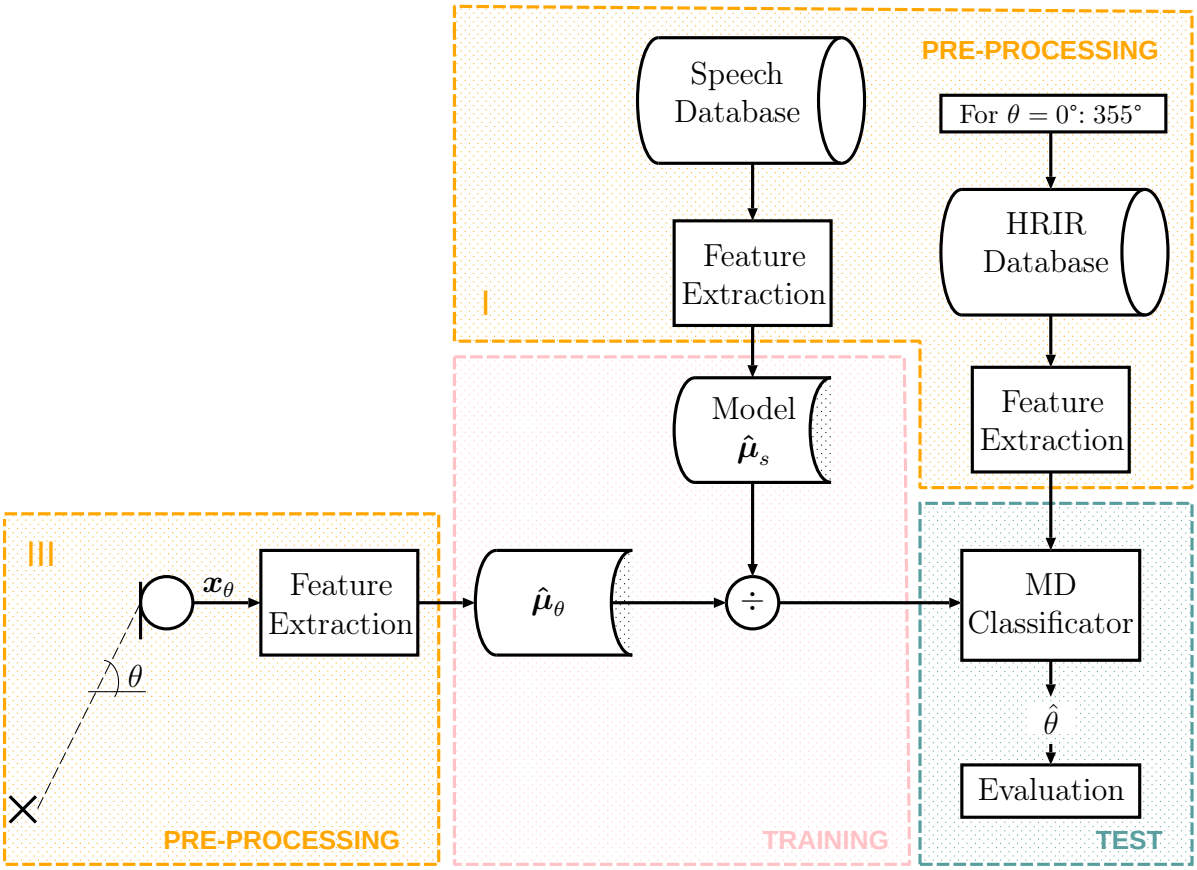


Figure 3.7: Experimental setup with minimum distance measure: pre-processing with synthesizing of test utterances and feature extraction of MFCCs (division operator in the training scenario converts into a subtraction operator) or STFT coefficients; training, test with minimum distance classificator and evaluation

3.2.2 STFT Coefficients with Minimum Distance

As a next step high-dimensional spectral features are used. As described in 3.1.3 STFT features of type 1 are calculated.

Training, Test and Conclusion

Again one statistical model for each speaker is trained by averaging the STFT coefficients over time which results in a 512-dimensional feature vector $\hat{\boldsymbol{\mu}}_s$. This vector is saved as a prototype for each of the eight speakers. The test scenario stays the same as before except that the STFT coefficients of the test utterance $\hat{\boldsymbol{\mu}}_\theta$ are *divided* by the prototype feature vectors. This is presented in figure 3.7.

As shown later this very simple statistical model delivers reasonable good results. To reach more accurate classification results another classification method has to be investigated.

3.2.3 STFT Coefficients using ML based single GMM

Again high-dimensional spectral features are taken. This time, STFT features of type 2 are taken. Moreover, the statistical model as well as the classification method are changed.

Training, Test and Conclusion

A GMM are used to model the speaker-dependent, direction-independent speech samples. Principles of GMMs are explained in chapter 2. A speaker-dependent GMM is trained using the EM algorithm. As a statistical model for each speaker the parameters of the GMM are saved, namely, a vector of weights $\mathbf{b}$, a matrix of mean values $\boldsymbol{\mu}$ (for each component one vector) and the covariance matrix $\boldsymbol{\Sigma}$ (each column represents the diagonal of the covariance matrix of one component). The feature vectors of the input data are calculated as shown in figure 3.5. The result is a $T \times D$ matrix (in this thesis $D = 513$) where each row represents a D-dimensional feature vector.

The GMM is trained with speech material which is direction-independent. Therefore, the GMM is direction-independent. To create direction-dependent GMMs from the direction-independent GMM the statistical models are adapted with each feature vector of the HRIR database. This is shown in figure 3.8 and is explained in more detail in section 3.2.3. Afterwards, classification is carried out via ML calculation.

Adaption of Speaker Model

It is assumed that a single direction-independent model is trained in the training scenario. Afterwards, direction-dependent models are created from the single direction-independent model. For the adaption the model parameters of the GMM are adapted. The direction information is brought in by shifting the vector $\boldsymbol{\mu}$ of each mixture component of the statistical model with the HRIRs. To be more specific, the feature coefficients of the HRIRs are added to the $\boldsymbol{\mu}$'s of each component of the GMM. The addition results from the fact that the calculations are done in the log domain. 72 GMMs *with* direction information can be calculated with little computational effort as shown in figure 3.8. The big advantage of this statistical model is that only a single statistical model for direction-independent speech has to be trained. The direction-dependent manipula-

tion will be calculated afterwards by a simple addition. Although the simple statistical model adaption is an advantage, it also brings along some problems. Due to the calculation of the Fourier transform with the DFT, the adaption of the direction-dependent model is consistent with a *circular* convolution, whereas the synthesized test utterances correspond to a *linear* convolution (3.1).

3.2.4 Matched and Mismatched Test Environment and Consequences

For the test scenario the direction-dependent test utterances are produced in a synthetic way. In the real world a recording of speech from a certain direction is assumed by a linear convolution of the direction-independent speech with the impulse response of the room (including the spectral shaping from the pinna). Mathematically it can be calculated by equation (3.1). The problem is that the adaption of the single, direction-independent model results in a circular convolution. This does not match the case where speech material for the test is synthesized with a linear convolution of the test utterance and the HRIRs. Therefore, a decreased performance is expected.

For the evaluation a matched environment as well as a mismatched environment are used to calculate results. In the matched environment speech material for the test is produced by a multiplication instead of a convolution. More precisely, the HRTFs and the spectra of the speech utterances are multiplied. Note that a DFT is used to get the HRTFs from the HRIRs and the spectra from the time domain speech signals. The matched case is depicted in figure 3.9. The mismatched environment is the real world scenario as described before where test utterances are created by a convolution. This is shown in figure 3.10. In the real case, where the signals are convolved in the time domain, the influence of the window creates serious problems which cannot be corrected. Because of the DFTs to calculate HRTFs and speech spectra, the hamming window is applied to both time domain signals. In the test scenario, on the other hand, the test utterance is convolved with one of the HRIRs and then the STFT features are extracted. The result is cut into segments by a hamming window and the DFT is carried out. In this case, the window is only applied once. Therefore, a mismatch is predictable. To get acceptable results nevertheless, the hamming window is replaced by a (inevitable) rectangular window. Therefore, in the GMM based approach features are calculated according to type 2 of section 3.1.3. As a consequence three different approaches are investigated:

1. adaption of HRTFs
2. restriction to a limited area
3. direction-dependent GMMs

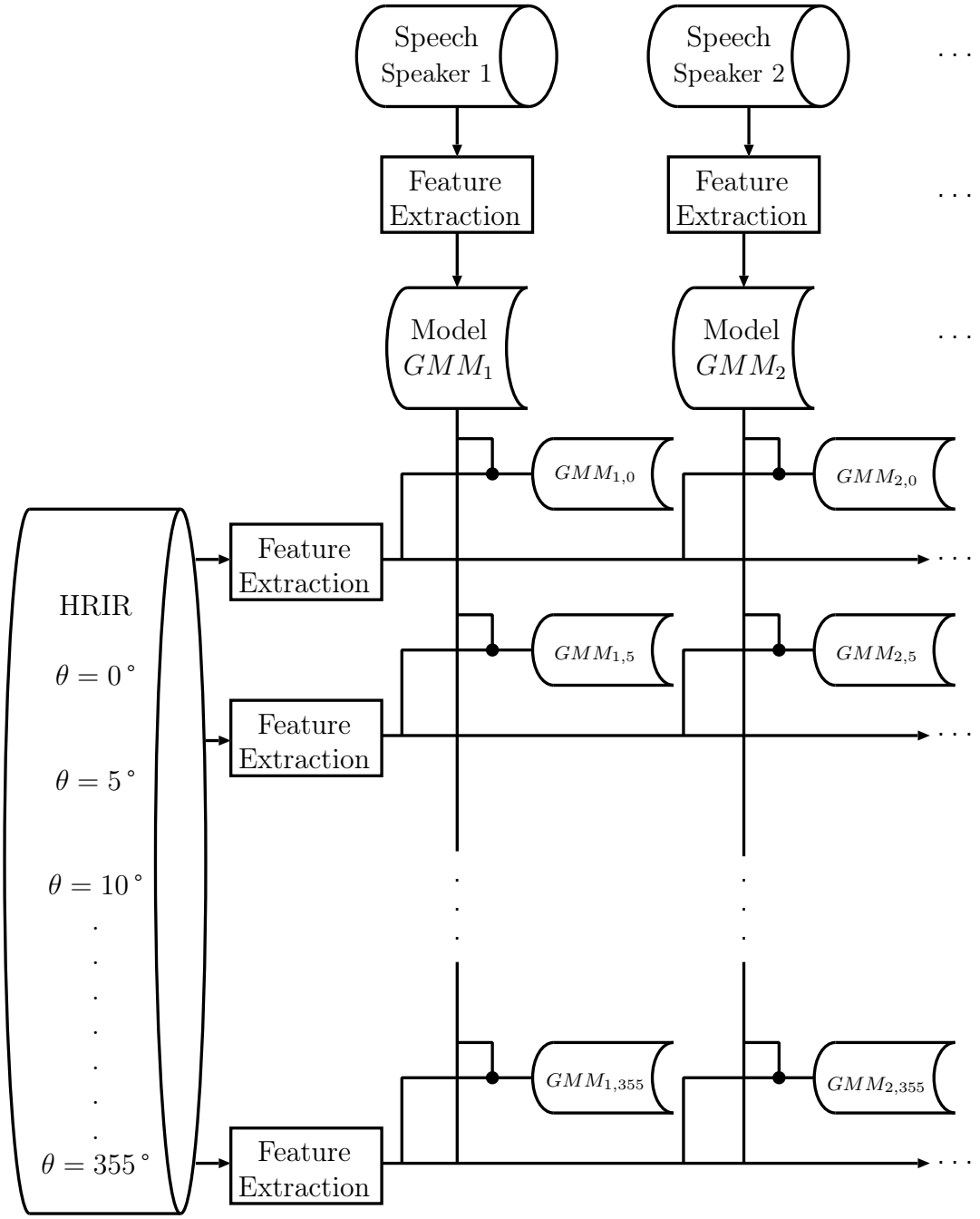


Figure 3.8: Adaption of the single, direction-independent GMM (Model – GMM_s for each speaker s) to yield direction-dependent GMMs ($GMM_{s,\theta}$ for each speaker s for direction θ)

These three approaches are introduced to handle the problems caused by the circular effects.

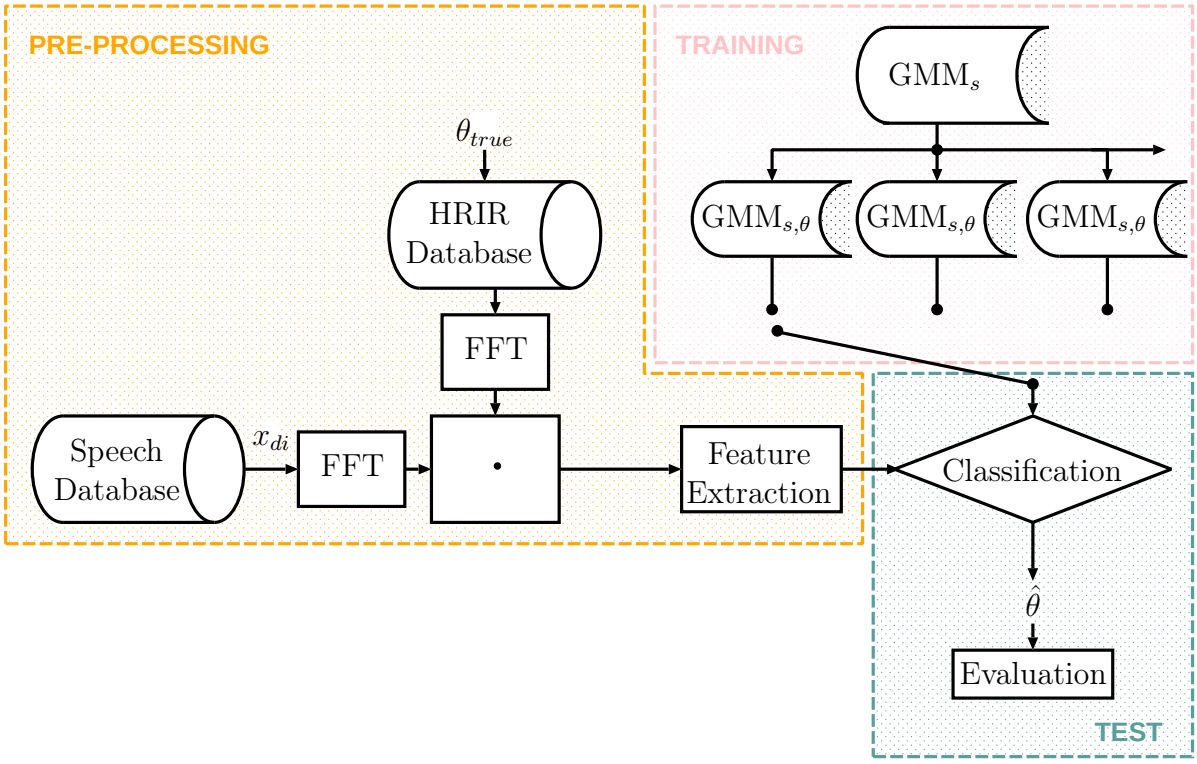


Figure 3.9: Pre-processing: *matched case* with multiplication (only possible in synthetic scenarios); training with adaption of single, direction-independent GMM and test with ML based classification and evaluation

Adaption of HRTFs

As mentioned before, the adaption of the GMM[1] to obtain direction-dependent GMMs from the direction-independent GMM prototypes introduces circular effects. The HRTFs in the feature domain are taken to adapt the GMMs. Therefore, the HRTFs need to be transformed in a way that a linear manipulation is applied instead of the circular manipulation. This is done with a system identification where each HRTF is linear approximated through a block adaptive method. In figure 3.11 the block diagram of the system identification is shown. The unknown coefficients of the linear adapted HRTF $h_{\theta,circ}$ are changed to match as closely as possible the desired signal d which is the DFT of the output of the linear system with the impulse response h_θ. The adapted HRTFs are calculated in the feature domain. Then these HRTFs are taken to manipulate the direction-independent GMM instead of the conventional HRTFs and tests are carried out with the same 15 utterances of all 8 speakers from each of the 72 directions.

[1]Note: Adaption of the GMMs and adaption of the HRIRs are two different things!

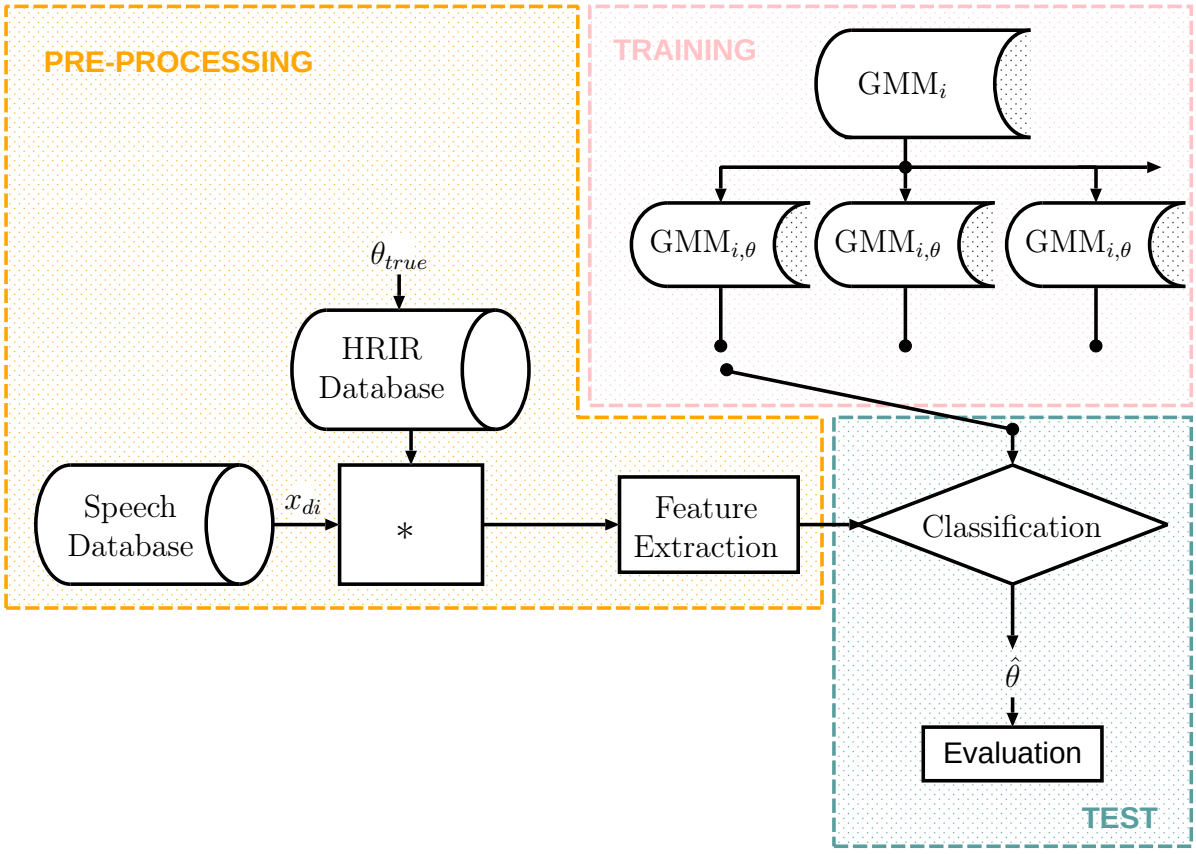


Figure 3.10: Pre-processing: *mismatched case* with convolution (inherent in real world scenarios); training with adaption of single, direction-independent GMM and test with ML based classification and evaluation

Limitation of the Models

A second consequence to eliminate the circular effects is to limit the area of investigation. First, the hamming window in the calculations of the feature vectors is discarded. The usage of a window has serious effects on the performance. The localization accuracy is expected to decrease in a mismatched case, but the circular influences are not assumed to effect all angels equally. The used HRIRs are very short in time domain. High energy peaks appear early in time domain and at low frequencies in frequency domain, respectively. Therefore, the HRTFs have a low-pass characteristic with high amplitudes in the low frequency domain and many values near zero at higher frequencies, which can also be seen in figure 3.2. HRIRs approximate minimum-phase filters. That is the reason why it is expected that the linear convolution with the speech signal is approximately the same as the circular multiplication of their spectra. To get around the problems only a limited area is investigated. This is acceptable because it depends on the application whether the whole horizontal plane is needed or not. Accurate localization in a half

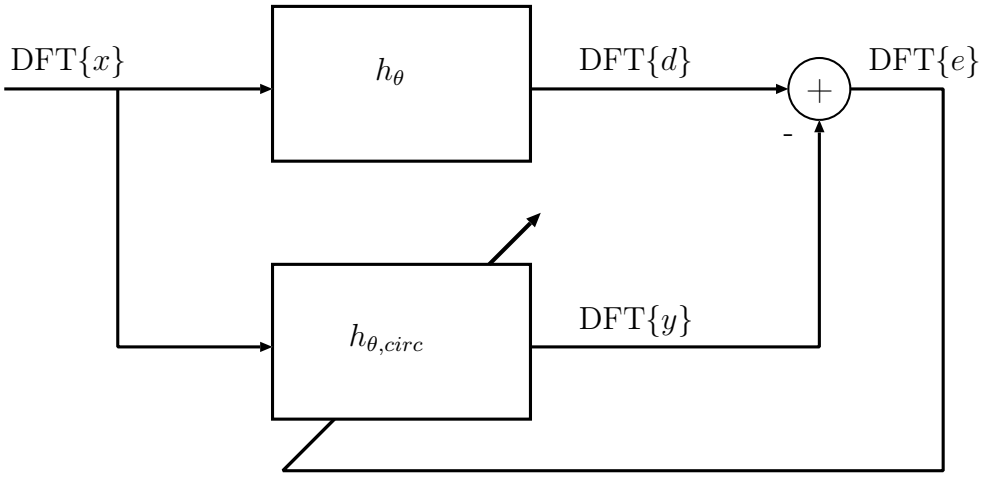


Figure 3.11: Block diagram of system identification

plane is also worth pursuing.

Direction-Dependent GMMs

A third consequence to eliminate the introduced circular influences is to train direction-dependent GMMs. The reason for the circular effects is the adaption of the direction-independent GMM to produce direction-dependent GMMs. Therefore, the adaption is discarded and replaced by a direct training of direction-dependent GMMs. This time synthesized speech material, which contains direction information (fig. 3.4), is used to train 72 direction-dependent GMMs (one GMM for each angle of θ). As a result localization should work in the whole range between $0\,^\circ$ and $355\,^\circ$.

Number of Components in the GMMs

The training data set is assumed to originate from a certain, unknown distribution. This distribution should be described in an optimal way by a statistical model. As described in the previous chapter, the optimal number of components in a GMM can be determined in different ways. The method, which will be used in this thesis, is the empirical method. For this method the model is trained using the EM algorithm and the performance is evaluated on the test data set. For different numbers K of components the absolute error angles $|\epsilon|$ are calculated. Afterwards, the mean value $\mu_{|\epsilon|}$ and standard deviation $\sigma_{|\epsilon|}$ of $|\epsilon|$ are determined. These values will give information on the performance of the statistical model depending on the numbers K of components. Above a certain number of K the performance changes insignificantly or even decreases. Looking at the results, a value for K can be chosen. The number of components in the GMM is important in

view of a correct evaluation and low computational complexity. If the model order is too low, the mixture model can not accurately model the system. On the other hand, too many components also influence the performance.

Speaker-Independent Models

It is also important to improve the speaker-dependent case to a more general speaker-independent case. As indicated earlier, the speech database consists of speech material of eight different speakers. In the speaker-dependent case statistical models are trained depending on the speaker, i.e. speech material from a speaker is taken to train a model for this speaker. Also in the test scenario speech material is taken from the corresponding speaker. Now, the speech material from five speakers (male as well as female) is taken to train one speaker-independent model. Then tests are carried out on the remaining three speakers to assure that the training and the test is not realized based on the same speech material. All methods mentioned above remain the same for speaker-dependent and speaker-independent models.

Influence of the Segment Length

Besides the amount of the training and test speech material and the choice and training of the statistical models, a crucial parameter for the performance is the segment length of the input utterance before a classification result is obtained. Especially for real-time applications it is important to shorten the input segments. On the one hand, the estimation of the direction of the source signal should be as accurate as possible. On the other hand, the update of the estimations should be as fast as possible. Although these two demands contradict each other, an optimal tradeoff has to be found. To examine the influence of the length of the input signal on the localization accuracy, the input speech utterances are cut into pieces. Then test and evaluation are carried out as explained above.

4 Discussion of the Numerical Results

In this chapter numerical results based on the previous considerations and developed approaches are presented. For the evaluation databases with direction-independent speech samples and with 72 HRIRs are used. Tests are carried out with 15 test utterances. Training utterances and test utterances are strictly separated in a way that no overlapping is possible and the test material is unknown by the model.

4.1 The Minimum Distance Approach

First MFCC features, then STFT features (type 1) are extracted and the direction of the sound source is estimated with a minimum distance approach as explained in sections 3.2.1 and 3.2.2.

For MFCC features the evaluation results are $\mu_{|\epsilon|} = 107.31\,^\circ$ and $\sigma_{|\epsilon|} = 49.54\,^\circ$ whereas $\mu_{|\epsilon|} = 22.93\,^\circ$ and $\sigma_{|\epsilon|} = 42.2\,^\circ$ for STFT features. In this case STFT coefficients clearly outperform MFCCs. In figure 4.13 $\mu_{|\epsilon|}$ and $\sigma_{|\epsilon|}$ is depicted as a function of the angle to estimate for the STFT coefficients. This means that for a speech utterance arriving from a fixed angle θ the error angle for all utterances and speakers is calculated and averaged, i.e. the histogram shows the error which is made at each sound source direction. It can be seen that the estimation works better for some angles than for others. Between $55\,^\circ$ and $70\,^\circ$ and between $115\,^\circ$ and $135\,^\circ$ the localization accuracy is very good, i.e. the mean absolute error angle is small. This can be explained by the MD classificator. Looking at figure 3.1 it can be seen that the HRTFs in this regions are unique. They are not likely to be confused with other HRTFs. This can also be seen in figure 4.1 where the hard-decision confusion matrix is shown. Along the x-axis there is the angle to estimate and on the y-axis the estimated angle, e.g., if the sound arrives from an angle of $0\,^\circ$ how often $0\,^\circ$, $5\,^\circ$, $10\,^\circ, \ldots$ is estimated. In other words, all absolute error angles of all speakers and all utterances are investigated. The absolute errors are a function of the true angle. All absolute errors for a true angle of, e.g., $0\,^\circ$, are taken into account where the amount of each estimated angle at a given true angle is counted. The occurrence of the estimated angles is indicated by the color. With 8 speakers and 15 utterances

from 72 directions, the number of estimations is $15 \cdot 8 \cdot 72 = 8640$. If all angles would be estimated correctly, this would result in a value of about 8640 (highest value, brightest color – deep red). In the best case the highest value is along the main diagonal.

Although MFCCs are used to separate a filter impulse response from the excitation signal this feature can not be used to identify HRIR in a minimum distance approach. MFCCs describe the envelope of a spectra, i.e. the vocal tract resonance frequencies (in speech production). HRIRs depend on the fine structure of a spectrum. Therefore, important information is lost. To further improve the estimation accuracy STFT features are evaluated with a ML framework.

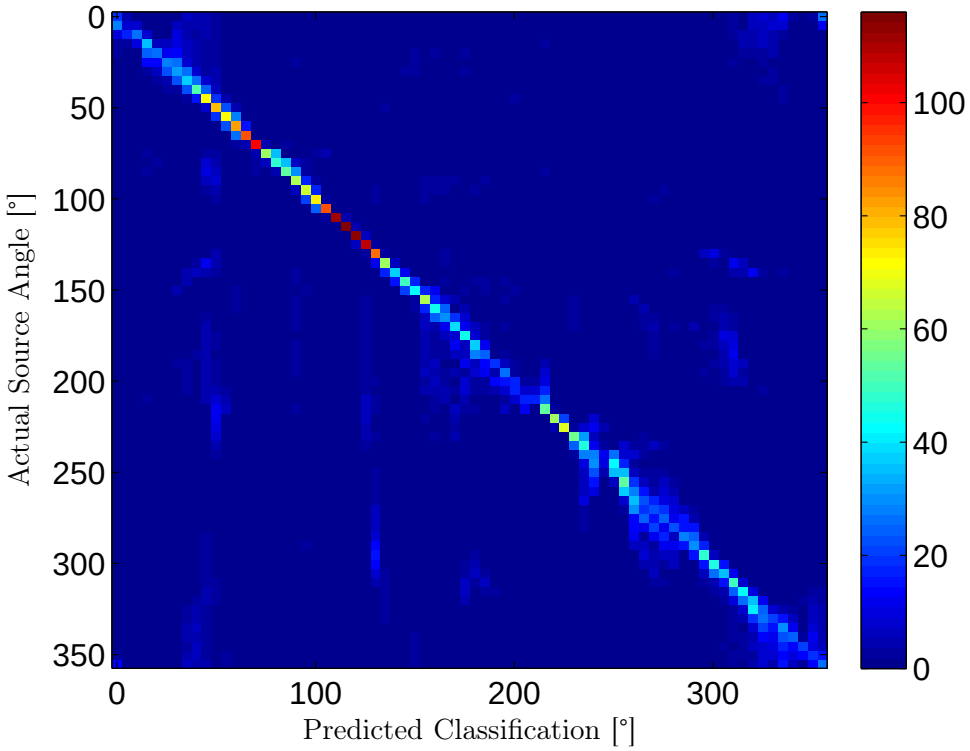


Figure 4.1: Hard-decision confusion matrix between true angles and estimated angles; STFT coefficients (type 1) with MD classificator

No. of Components	Matched Case		Mismatched Case	
	$\mu_{\|\epsilon\|}$	$\sigma_{\|\epsilon\|}$	$\mu_{\|\epsilon\|}$	$\sigma_{\|\epsilon\|}$
1	11.56	29.33	29.02	44.38
8	1.66	8.69	15.69	35.89
16	0.94	5.92	17.09	39.15
32	0.38	1.95	19.78	42.82
64	0.27	2.50	21.09	44.92

Table 4.1: $\mu_{|\epsilon|}$ and $\sigma_{|\epsilon|}$ for the absolute error angle $|\epsilon|$ for matched case and mismatched case and for different numbers of components; STFT coefficients (type 2) and ML classificator

4.2 The Maximum Likelihood Approach

4.2.1 Matched and Mismatched Case

As explained in section 3.2.3 test utterances are compared to adapted direction-independent GMMs via a ML framework. The adaption introduces circular effects. To minimize them first the hamming window is discarded and feature vectors of type 2 are taken.

First, to build a matched case with the circular effects due to the adaption of the GMMs, "synthetic" synthesized test utterances are taken to evaluate the method. These "synthetic" synthesized test utterances are produced by the multiplication of the spectrum of the test utterance and the HRTF. As presented in table 4.1 results are very accurate which is not a surprise because a matched case scenario has been simulated. In the mismatched case the estimation accuracy decreases. The results are depicted in figure 4.2. In the mismatched case the error increases when the number of used components increases. The number of used components also influences the amount of training material, which has to be larger the more components are used. If the amount of training material is too small, results are not reliable any more. To give insight, the hard-decision confusion matrix and the soft-decision confusion matrix for both cases are shown in figure 4.4. The soft-decision confusion matrix shows the likelihood values for a true angle and that one compared to each of the 72 adapted models. After calculating the likelihood values a decision is made. The angle with the maximum likelihood is taken as the estimated angle. The hard-decision confusion matrix is calculated and depicted as explained before.

In figure 4.4 most of the values are off the main diagonal. This means that there are similar models for different angles. In figure 4.3 it is depicted that the maxima are well suited on the main diagonal. This can also be followed from figure 4.13, 4.14, 4.15 and 4.16. The mean value and the standard deviation are very high for the true angles with

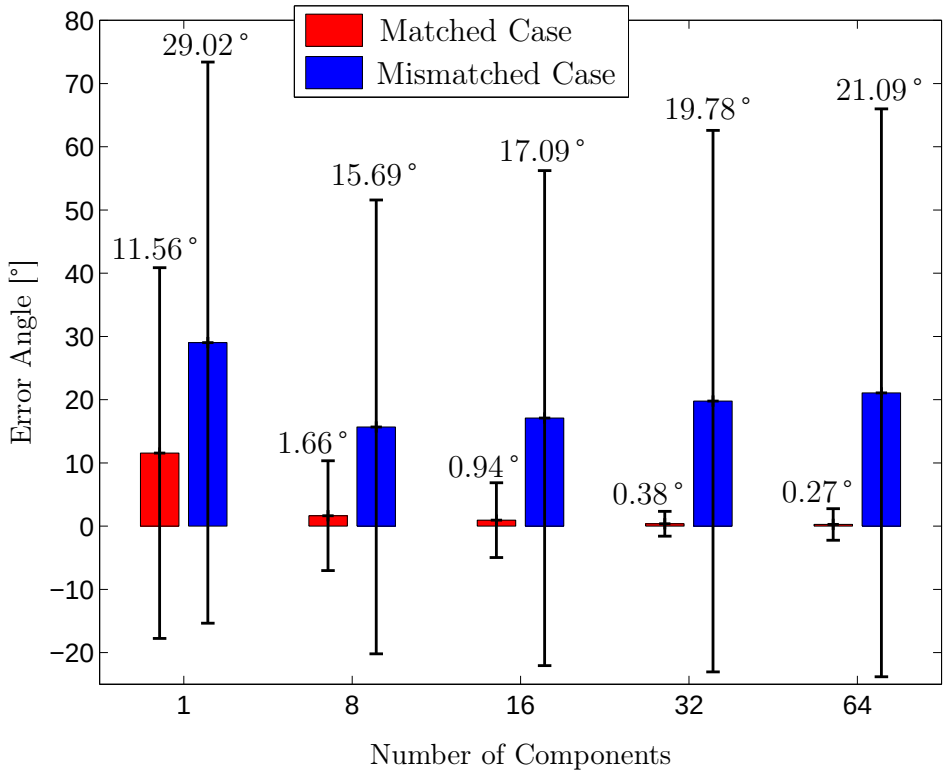


Figure 4.2: Comparison between matched case (red) and mismatched case (blue) as a function of the number of components; STFT coefficients (type 2) and ML classificator; the height of the bars indicates the $\mu_{|\epsilon|}$ and the black lines $\sigma_{|\epsilon|}$

incorrect estimations. The problem in the mismatched case is that a right-left confusion has been introduced. Angles between 5 ° – 45 ° get confused with angles in the range 315 ° – 355 ° (symmetrically around 0 °) (fig. 4.4).

4.2.2 Adaption of HRTFs

To eliminate the effects of the circular adaption of the direction-independent GMMs, HRTFs are adapted as explained in the last chapter. For this case the average error for all sentences and all speakers has $\mu_{|\epsilon|} = 15.69\,^\circ$ and $\sigma_{|\epsilon|} = 35.89\,^\circ$ for the non-adapted case and $\mu_{|\epsilon|} = 15.42\,^\circ$, $\sigma_{|\epsilon|} = 35.35\,^\circ$ for the adapted case.

In 7707 of the 8640 test cases the absolute error is the same for both models, which is 89.20 %. 84 times (0.97 %) the absolute error is the same but had a different sign. 444 times (5.12 %) the absolute error is smaller for the adapted model and 489 times (5.66 %) the absolute error of the non-adapted model is smaller. This results in the knowledge that an adaption is not necessary.

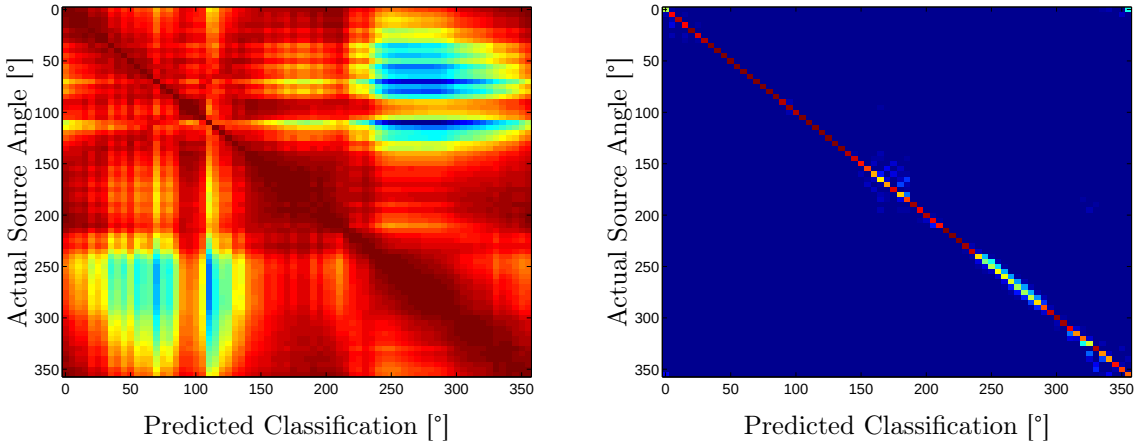


Figure 4.3: Confusion matrix of *matched case*: soft-decision (left) and hard-decision (right)

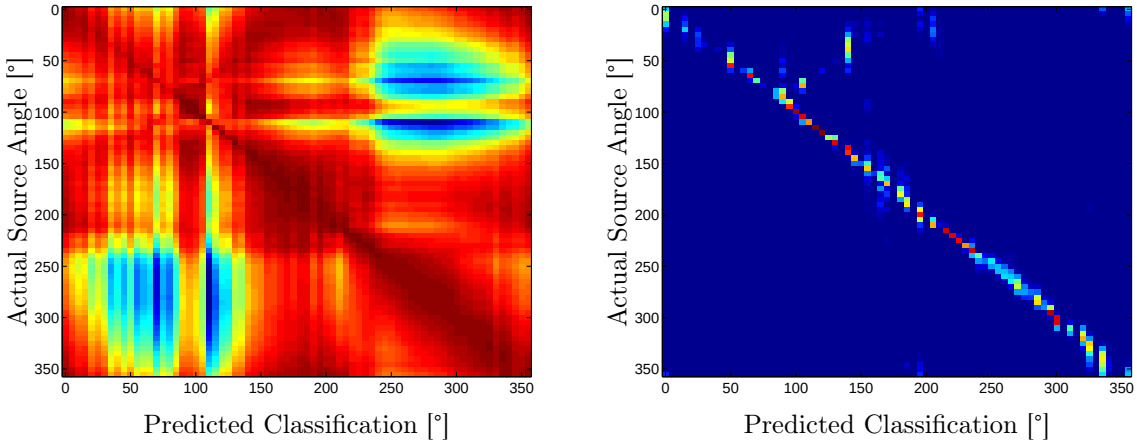


Figure 4.4: Confusion matrix of *mismatched case*: soft-decision (left) and hard-decision (right)

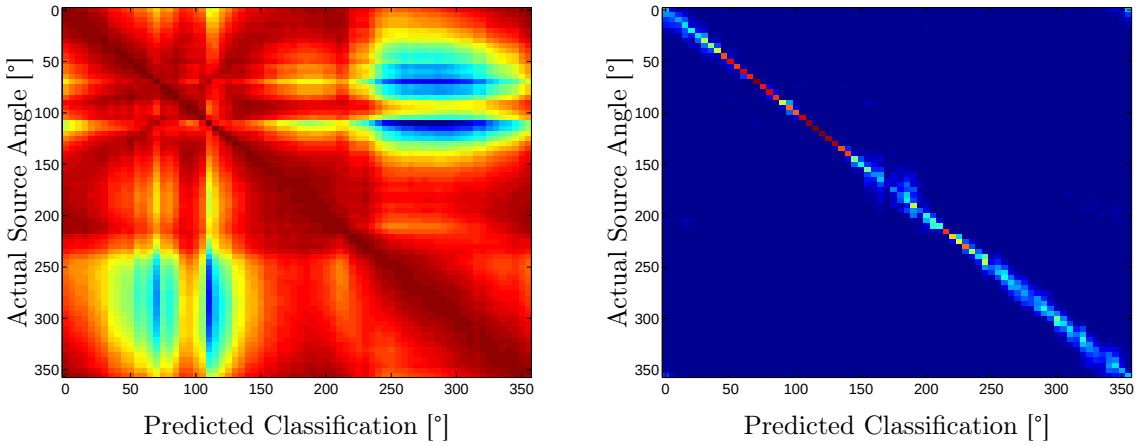


Figure 4.5: Confusion matrix of *direction-dependent GMMs (GMM+HRTF)*: soft-decision (left) and hard-decision (right)

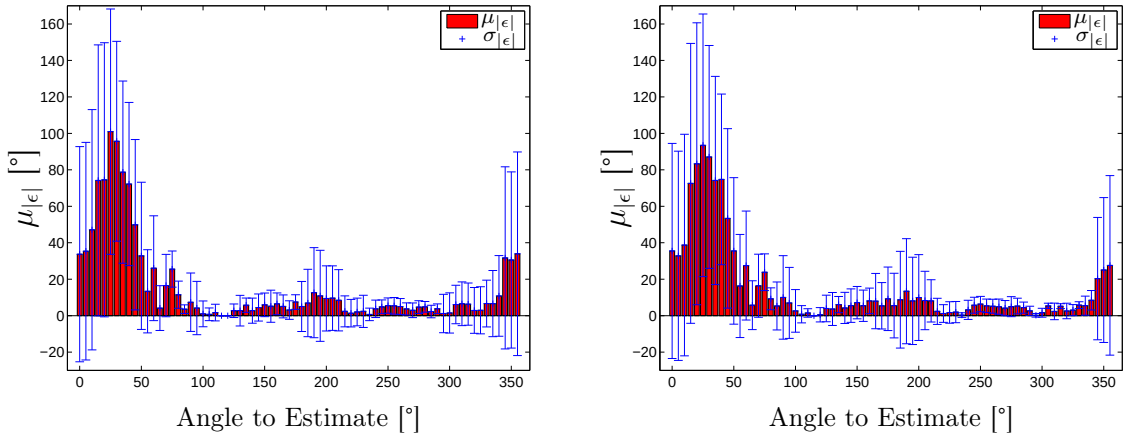


Figure 4.6: Mean value of absolute error angle $\mu_{|\epsilon|}$ and standard deviation $\sigma_{|\epsilon|}$ as a function of the angle to estimate; true HRTFs (left) and adapted HRTFs (right)

In figure 4.6 the error is shown as a function of the true angle for results with the conventional HRTFs and the adapted HRTFs. The conclusion is that a circular adaption of the HRTFs has no influence on the performance.

4.2.3 Limitation of the Model to 90° – 270°

As shown before, the absolute error of the mismatched case increases because of the adaption of the model. As a constraint only a half plane is taken into account. Thus, the critical area can be avoided. In figure 4.4 and 4.15 it can be seen that the problematic areas are approximately around $0\,^\circ$ – $70\,^\circ$ and $300\,^\circ$ – $355\,^\circ$. Many errors result in a confusion around $0\,^\circ$. Therefore, a restriction to the area in between is made. The evaluation of a limited area (LimArea) is carried out.

The trained GMMs are the same as before and also tests are carried out with the same adapted GMMs. The mean absolute error has $\mu_{|\epsilon|} = 4.03\,^\circ$ and $\sigma_{|\epsilon|} = 7.5\,^\circ$.

Selection of the Number of Components in the GMMs

As described in 3.2.4 the number of components in the GMM is chosen empirically. In figure 4.7 the results of the evaluation are plotted. The results are also depicted in table 4.2. The number K of components is varied from $1, 2, \ldots, 16, 32, 64$ components per model. Then the model is tested on the test data set, and $\mu_{|\epsilon|}$ and $\sigma_{|\epsilon|}$ for the absolute error are calculated. The results show that the performance only improves very little above two components. Because of computational advantages the number of components is chosen to be eight. It would also be possible to choose the number of components

individually for each speaker. In this way, the model optimally fits to a speaker.

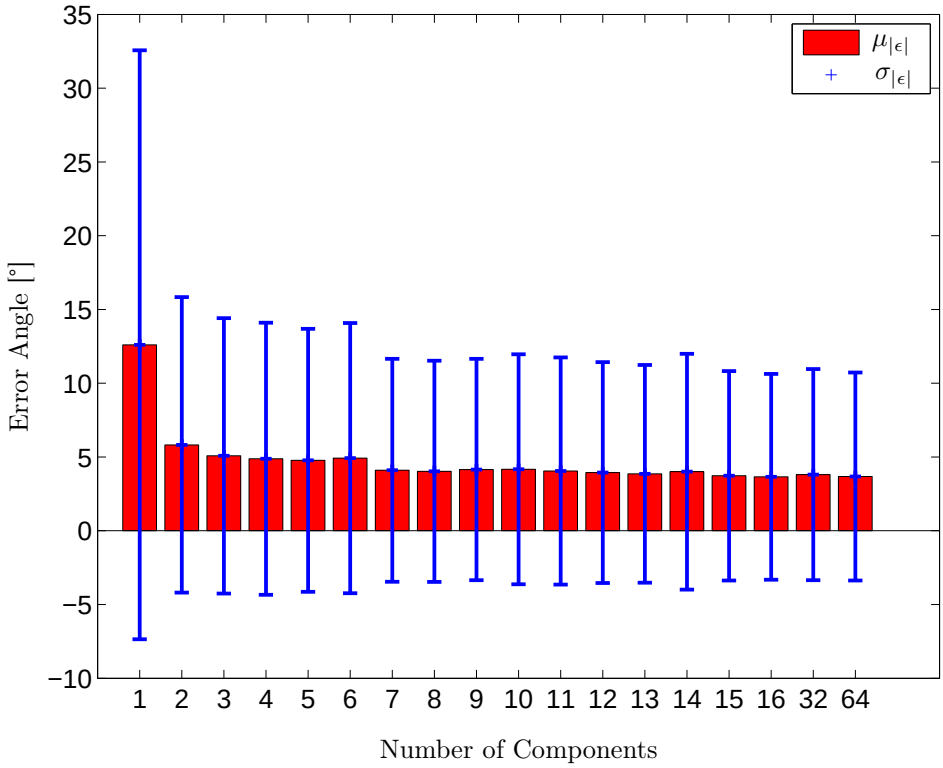


Figure 4.7: Comparison between different numbers of components in the GMM

4.2.4 Direction-Dependent GMMs

In the previous sections it is shown that the adaption of the direction-independent speech GMM introduces circular effects. GMMs for speech arriving from each of the 72 positions are available. GMMs are trained from speech uttered from a certain direction. The mean absolute error has $\mu_{|\epsilon|} = 5.86\,^\circ$ and $\sigma_{|\epsilon|} = 15.92\,^\circ$. In figure 4.5 the soft-decision and hard-decision confusion matrix substantiate this good result.

4.2.5 Speaker-Independent GMMs

To generalize the results a speaker-independent model is trained. In this case the training is carried out with speech material of speaker 1 to speaker 5 and the test with the remaining 3 speakers (case 1; fig. 4.8). Additionally, a model with speech material of speaker 3 to speaker 8 is trained where tests are carried out with speaker 1 to speaker 3 (case 2; fig. 4.8). For the limited area the mean absolute error angle has (case 1) $\mu_{|\epsilon|} =$

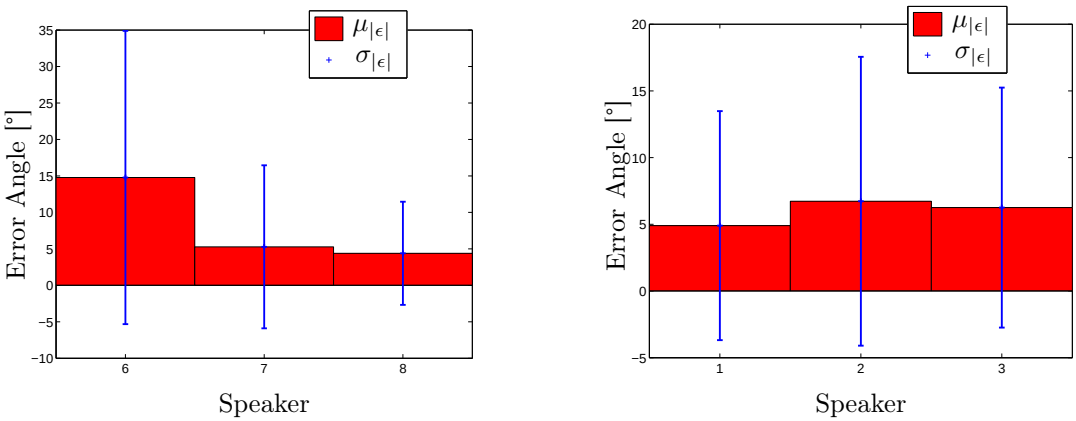


Figure 4.8: $\mu_{|\epsilon|}$ and $\sigma_{|\epsilon|}$ of $|\epsilon|$ for a limited area – on the left: tested on speaker 6 (female), 7 (male) and 8 (male); on the right: tested on speaker 1 (male), 2 (female) and 3 (female)

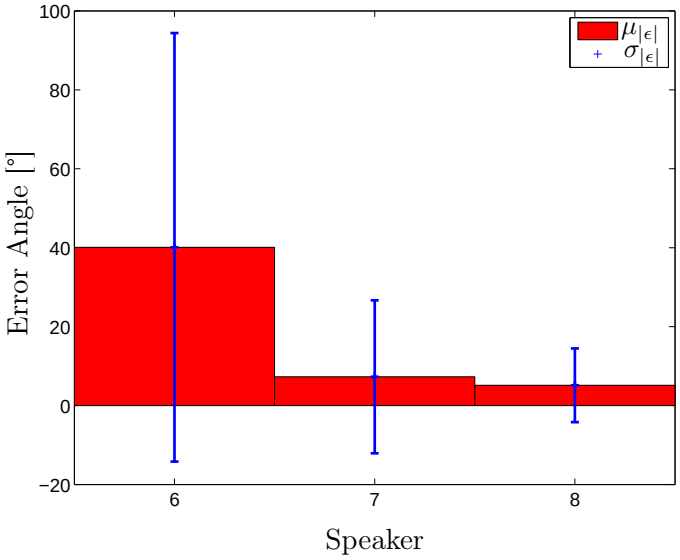


Figure 4.9: $\mu_{|\epsilon|}$ and $\sigma_{|\epsilon|}$ of $|\epsilon|$ for direction-dependent GMMs – tested on speaker 6 (female), 7 (male) and 8 (male)

No. of Components	**Mismatched Case**	
	$\mu_{\|\epsilon\|}$	$\sigma_{\|\epsilon\|}$
1	12.61	19.96
2	5.82	10.02
3	5.08	9.33
4	4.88	9.22
5	4.77	8.92
6	4.92	9.16
7	4.10	7.55
8	4.03	7.50
9	4.15	7.50
10	4.17	7.79
11	4.05	7.70
12	3.94	7.49
13	3.86	7.38
14	4.00	7.99
15	3.73	7.09
16	3.66	6.98
32	3.80	7.16
64	3.68	7.05

Table 4.2: $\mu_{|\epsilon|}$ and $\sigma_{|\epsilon|}$ for the absolute error angle $|\epsilon|$ for mismatched case and for different numbers of components

9.04° and $\sigma_{|\epsilon|} = 14.73^\circ$ and (case 2) $\mu_{|\epsilon|} = 2.24^\circ$ and $\sigma_{|\epsilon|} = 6.52^\circ$ (fig. 4.9). A speaker-independent model is also trained for the direction-dependent GMMs (GMM+HRTF) where the mean absolute error angle results in $\mu_{|\epsilon|} = 17.51^\circ$ and $\sigma_{|\epsilon|} = 37.31^\circ$. In this case the mean absolute error of speaker 6 is bigger than for the speakers 7 and 8. Therefore, also the averaged mean absolute error is high.

In figure 4.8 it is shown that the model works better for speaker 7 and 8, than for speaker 6. The speaker-independent model is trained with 3 female speakers and 2 male speakers. The speaker who reaches worse results is female. In figure 4.8 the training is carried out with speaker 4 to 8 (3 male and 2 female). In the test scenario again the male speaker reaches better results. In figure 4.8 (right) and 4.9, where the model is based on the same speakers, both times speaker 7 and 8 reach better results than speaker 6. The length of the training utterances as well as the test utterances are the shortest for speaker 6. The durations of test and training utterances are listed in table 3.1 and 3.2. This might be a reason why the other speaker outperform this speaker. Furthermore, as mentioned before, the performance of a GMM strongly depends on the initial values. Therefore, a GMM sometimes works better for some speakers than for others.

	$\mu_{\|\epsilon\|}$	$\sigma_{\|\epsilon\|}$
STFT-MD	22.93°	42.2°
STFT-ML (mismatched case)	15.69°	35.89°
STFT-ML (adapted HRTFs)	15.42°	35.35°
STFT-ML (GMM+HRTF)	5.86°	15.92°
SI (LimArea)	9.04°	14.73°
SI (GMM+HRTF)	17.51°	37.31°

Table 4.3: Comparison between different approaches: STFT coefficients with minimum distance classification (STFT-MD) and with maximum likelihood classification (STFT-ML) for the mismatched case, the case with adapted HRTFs and the case with direction-dependent GMMs (GMM+HRTF); speaker-independent (SI) model for limited area (LimArea) case and direction-dependent GMMs (GMM+HRTF)

4.2.6 Influence of the Segment Length

The performance is now evaluated as a function of the input segment length. So far, test utterances from each speaker arriving from each angle are taken as input. Now, all data is taken and cut into equally long segments of 1 second, 2.5 seconds, 5 seconds, 7 seconds and finally 10 seconds. Results are evaluated for the speaker-dependent as well as for the speaker-independent case and for the model with limited area (LimArea) and the direction-dependent GMMs (GMM+HRTF) (fig. 4.10, 4.11). The mean values are collected in table 4.4. Figure 4.12 compares both results. The error decreases when the segment length increases.

Some applications are delay sensitive. Therefore, an estimation of the direction is needed within a certain time. At a segment length of 2.5 seconds the mean error is 6.92° for the limited area case which is acceptable. For the direction-dependent GMMs the error is a bit higher with a value of 11.05°. The error increases further in the speaker-independent case to 11.07° (LimArea) and 21.09° (GMM+HRTF) (tab. 4.4).

4.2.7 Summary

In this chapter the different approaches are evaluated and compared. The matched case, where the spectrum of the speech signal is multiplied with the HRTF to synthesize speech from a certain direction, delivers optimal results in a "perfect" world scenario. Naturally the localization accuracy in this case is very high, but this method is not inherent in a real world scenario. In the mismatched case, where a linear convolution of the speech signal and the HRIR results in the direction-dependent utterance, the localization accuracy decreases dramatically. As a consequence the HRTFs are adapted,

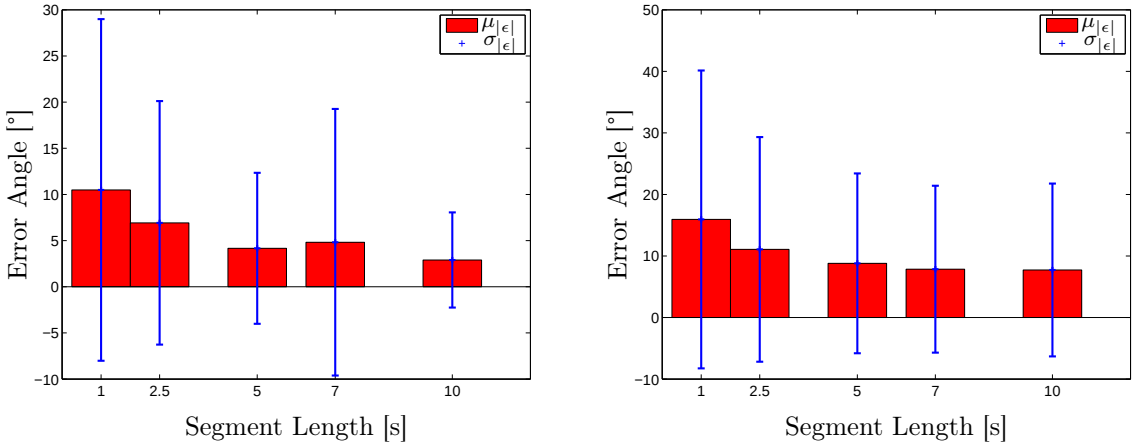


Figure 4.10: Error angle depending on the segment length (LimArea) – speaker-dependent case (left) and speaker-independent case (right)

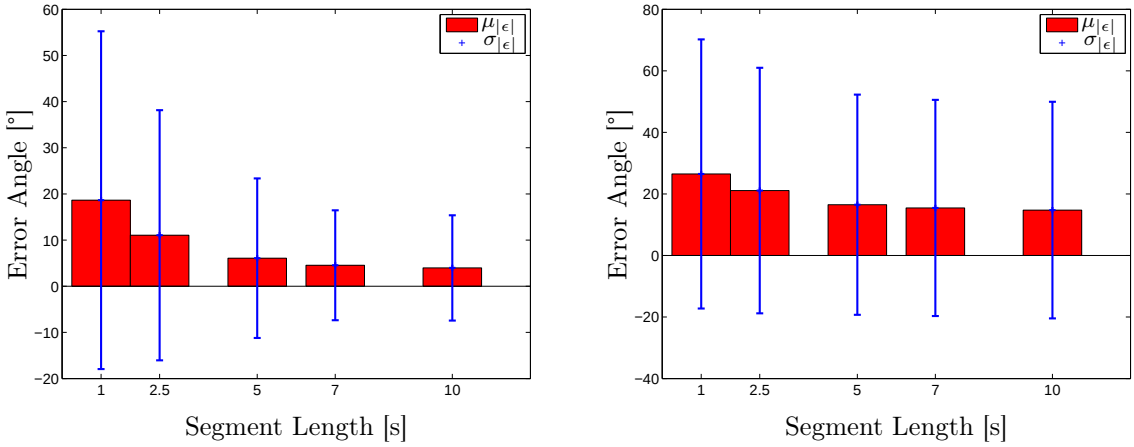


Figure 4.11: Error angle depending on the segment length (GMM+HRTF) – speaker-dependent case (left) and speaker-independent case (right)

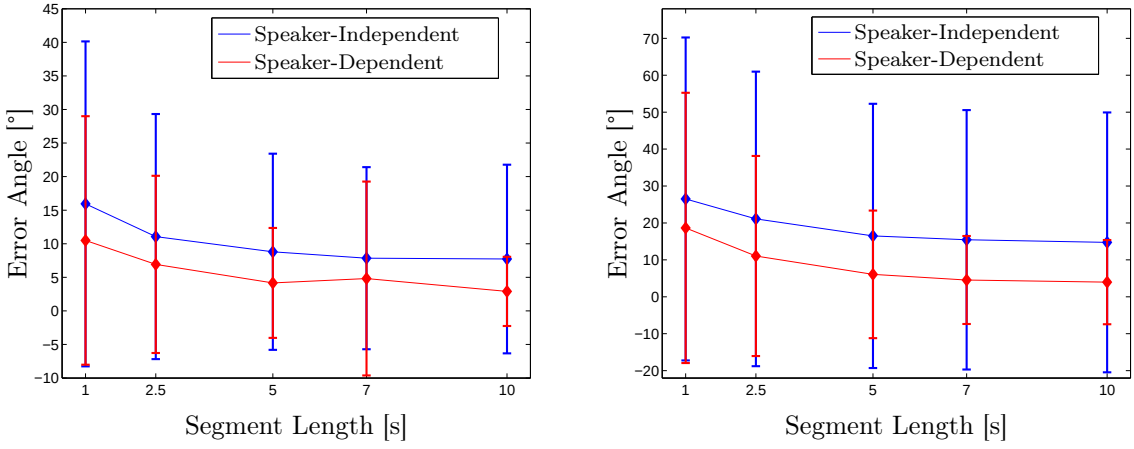


Figure 4.12: Comparison between speaker-independent and speaker-dependent case for LimArea (left) and GMM+HRTF (right)

SL	LimArea (SD)		GMM+HRTF (SD)		LimArea (SI)		GMM+HRTF (SI)	
	$\mu_{\|\epsilon\|}$	$\sigma_{\|\epsilon\|}$	$\mu_{\|\epsilon\|}$	$\sigma_{\|\epsilon\|}$	$\mu_{\|\epsilon\|}$	$\sigma_{\|\epsilon\|}$	$\mu_{\|\epsilon\|}$	$\sigma_{\|\epsilon\|}$
1	10.49	18.51	18.65	36.59	15.94	24.20	26.50	43.74
2.5	6.92	13.19	11.05	27.10	11.07	18.25	21.09	39.89
5	4.17	8.18	6.08	17.28	8.80	14.61	16.49	35.77
7	4.82	14.44	4.55	11.91	7.85	13.56	15.44	35.12
10	2.90	5.15	3.97	11.41	7.73	14.05	14.74	35.19

Table 4.4: $\mu_{|\epsilon|}$ and $\sigma_{|\epsilon|}$ depending on the segment length (SL) – comparison between model with limited area (LimArea) and model with direction-dependent GMMs (GMM+HRTF) for speaker-dependent (SD) and speaker-independent (SI) case

so as they neglect the circular effects. This does not increase the performance. In the next step the investigated area (LimArea) is restricted. In this case the problematic area is not considered. Therefore, also the localization accuracy increases. The disadvantage is that only a half plane is evaluated. Therefore, as a last consequence direction-dependent GMMs are trained (GMM+HRTF). In this case the localization accuracy increases, but the value is not directly comparable with the LimArea case because the investigated area differs. To generalize the results the tests are also carried out on the speaker-independent (SI) models of the limited area (LimArea-SI) and the direction-dependent GMMs (GMM+HRTF-SI).

The values for $\mu_{|\epsilon|}$ and $\sigma_{|\epsilon|}$ for all models are listed in table 4.3. Furthermore, a comparison is shown in figure 4.17 where the localization accuracy for each model is plotted. The localization accuracy is calculated by putting the number of investigations where the error is zero in relation to the total number of investigations. The different models are (1) the matched case, (2) the mismatched case, (3) the case with adapted HRTFs, (4) the case with direction-dependent GMMs (GMM+HRTF), (5) the case with limited area (LimArea) and the speaker-independent cases for (6) (GMM+HRTF) and (7) (LimArea). (1) – (4) use the same number of investigations, whereas in (5) the number is only the half (half of the area is investigated). Also in (6) and (7) the amount of investigations is smaller because in the speaker-independent case there are fewer test utterances. The two different curves in this figure represent different precisions. The red line indicates that an error lies within $\pm 2.5\,^\circ$, whereas the error is within $\pm 7.5\,^\circ$ for the blue line. Then the influence of the input segment length is investigated. The segment length is changed between 1 s, 2.5 s, 5 s, 7 s and 10 s. The localization accuracy of LimArea and GMM+HRTF is depicted in figure 4.18 for a resolution of $\pm 2.5\,^\circ$ and in figure 4.19 for a resolution of $\pm 7.5\,^\circ$. Furthermore, also the speaker-independent models are shown.

LimArea and GMM+HRTF are two different approaches to estimate the angle of

incident sound. Both models are promising depending on the type of application. The localization accuracy decreases when shorter input segments are taken, but still good results are reached for segments which are 1 second long. Furthermore, the performance decreases slightly when speaker-independent models are trained.

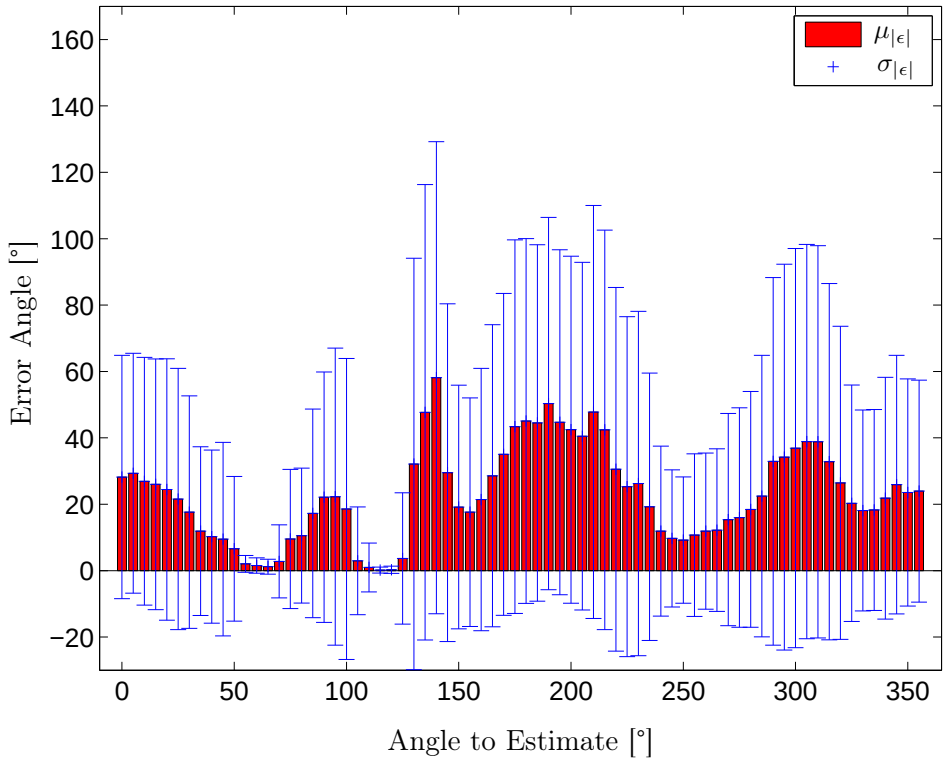


Figure 4.13: Mean value of absolute error angle $\mu_{|\epsilon|}$ and standard deviation $\sigma_{|\epsilon|}$ as a function of the angle to estimate; STFT coefficients with MD classificator

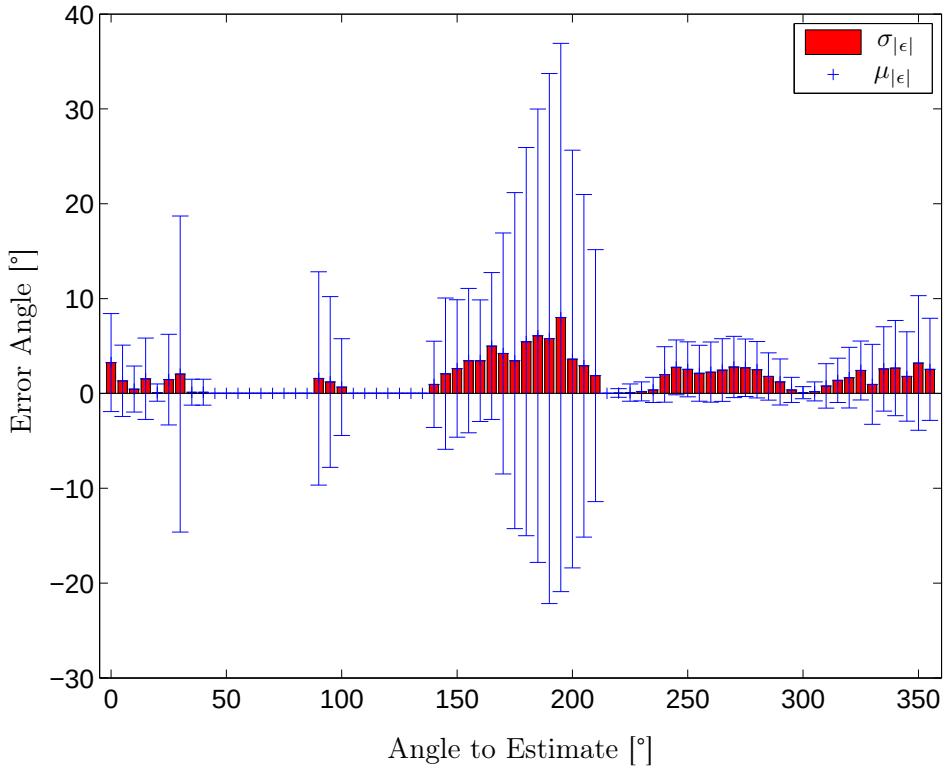


Figure 4.14: Mean value of absolute error angle $\mu_{|\epsilon|}$ and standard deviation $\sigma_{|\epsilon|}$ as a function of the angle to estimate; STFT coefficients with ML classificator – matched case

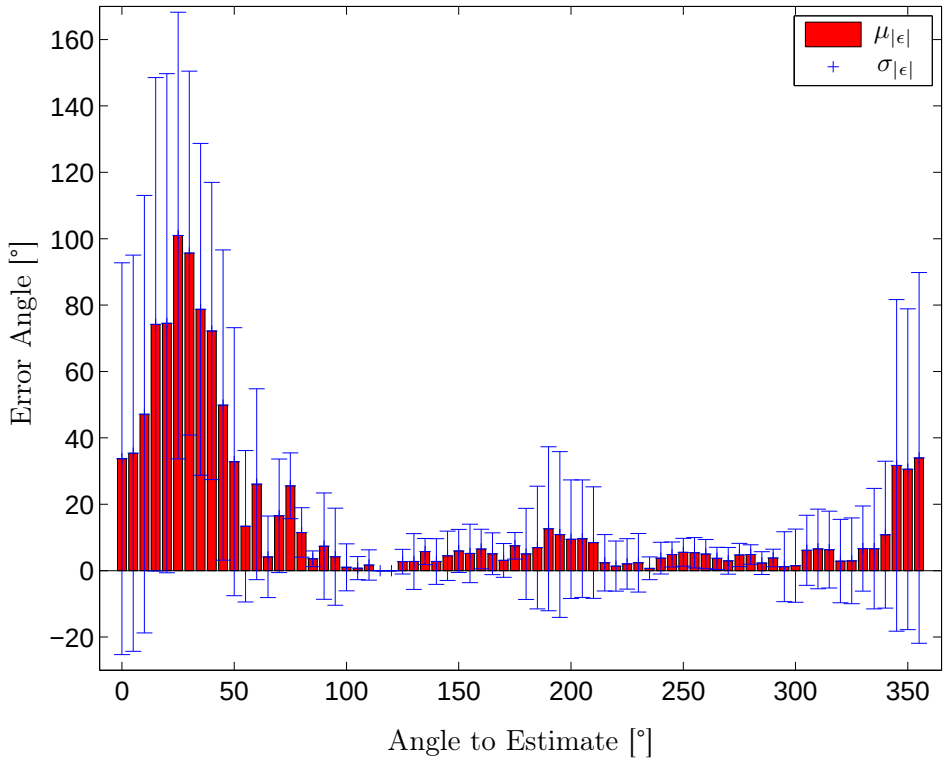


Figure 4.15: Mean value of absolute error angle $\mu_{|\epsilon|}$ and standard deviation $\sigma_{|\epsilon|}$ as a function of the angle to estimate; STFT coefficients with ML classificator – mismatched case

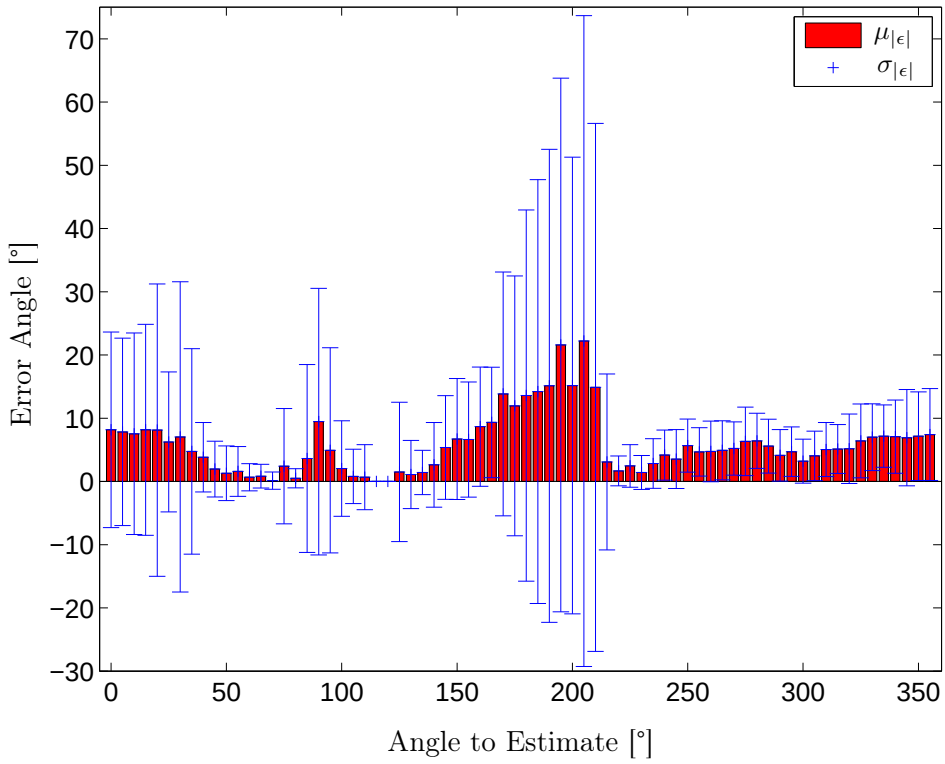


Figure 4.16: Mean value of absolute error angle $\mu_{|\epsilon|}$ and standard deviation $\sigma_{|\epsilon|}$ as a function of the angle to estimate; STFT coefficients with ML classificator – direction-dependent GMMs

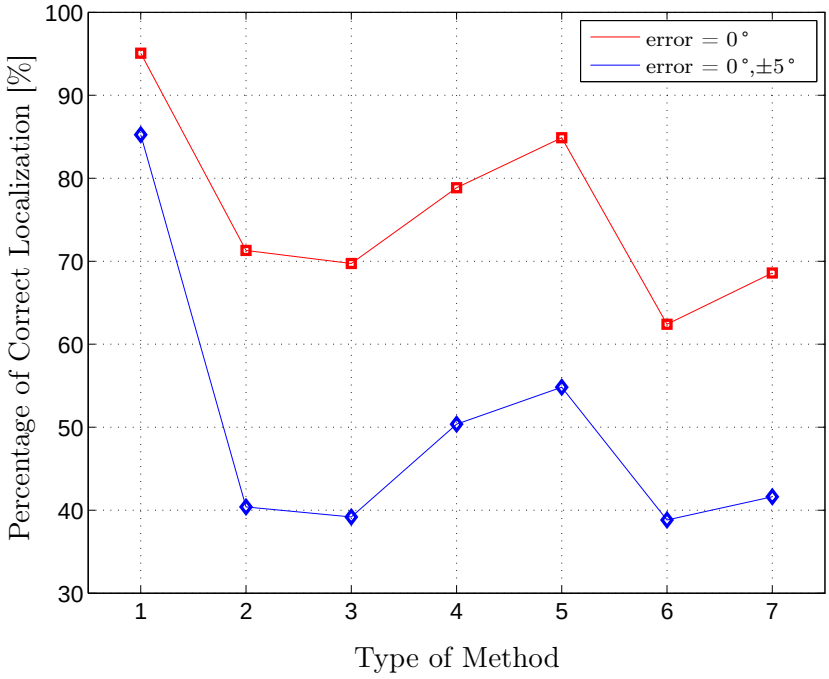


Figure 4.17: Comparison between all methods: (1) matched case, (2) mismatched case, (3) adapted HRTFs, (4) GMM+HRTF, (5) LimArea, (6) GMM+HRTF-SI, (7) LimArea-SI

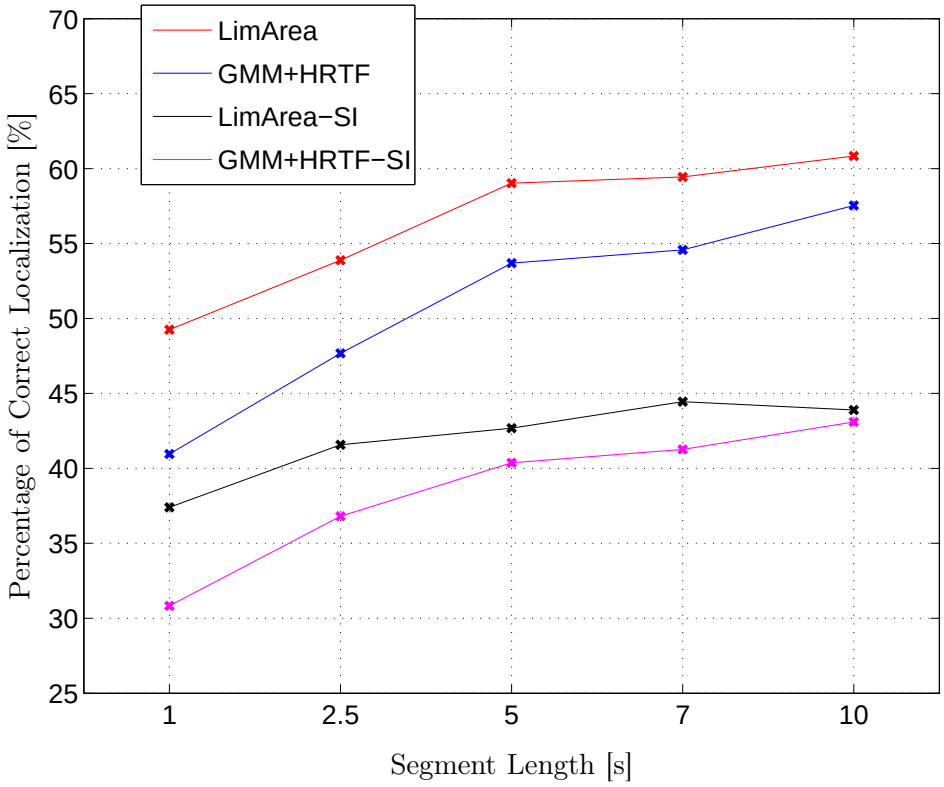


Figure 4.18: Comparison between all methods as a function of the segment length – localization accuracy (%) - resolution ±2.5°

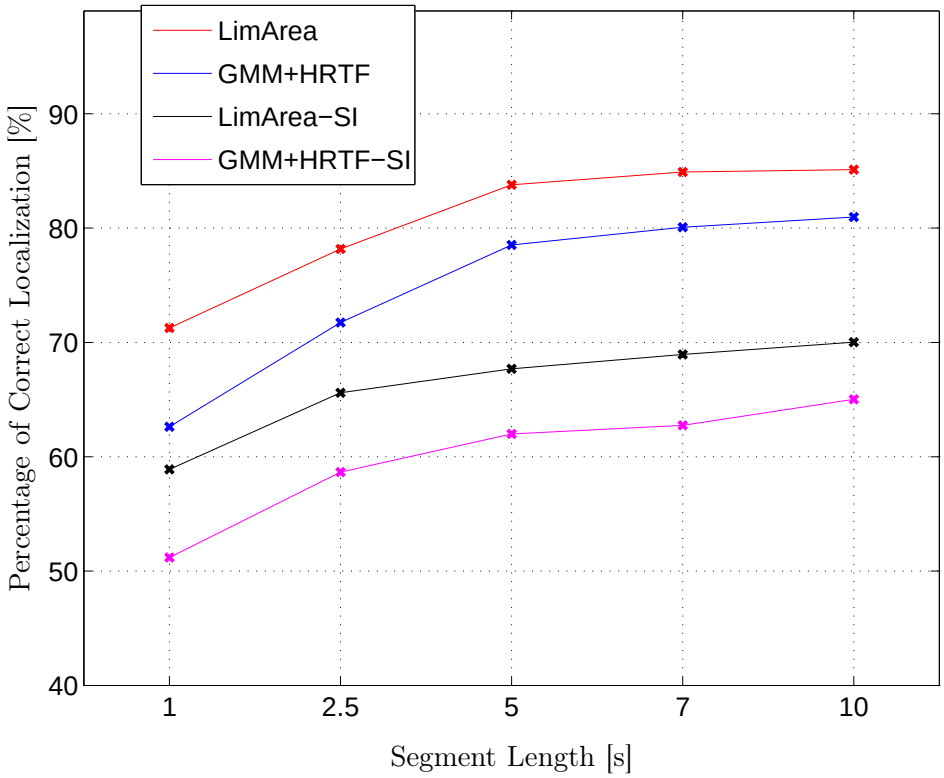


Figure 4.19: Comparison between all methods as a function of the segment length – localization accuracy (%) - resolution ±7.5°

5 Conclusion and Outlook

In this thesis a method to estimate the direction of an incident sound wave is investigated. Localization strategies based on microphone-arrays often are not suitable due to their cost and size. Therefore, as a principle constraint, only one receiver is allowed.

Localization is not only based on the time differences and level differences of the incident sound wave between the right and left ear. It also depends strongly on the spectral shaping due to the shape of the pinna. Localization is also possible with one ear. The underlying strategy is described by the so-called head-related impulse responses (HRIR) or their spectral representations, the head-related transfer functions (HRTF). Monaural sound localization is a difficult estimation problem because the incident angle of the sound is ambiguous. The source signal sounds different if it arrives from a different position. The problem is that for such a knowledge a basic familiarity with the sound is necessary.

Experiments for this thesis have been carried out with a set of HRIR measurements and a speech database. The HRIRs used have been recorded with a KEMAR dummy head microphone by Bill Gardner and Keith Martin at MIT Media Lab. The speech database consists of speech material of eight different speakers. MFCCs, the most popular feature in speech and speaker recognition, are a good representation of direction-independent speech. However, these features turned out to be not good enough for the investigations in a minimum distance approach. Monaural localization primarily depends on the fine structure of the HRIRs. Therefore, a higher resolution in the spectral domain is preferable. As features the logarithmic absolute STFT coefficients turned out to perform well. An incident sound wave from a certain direction is compared to GMMs in the feature domain. A hard decision obtained by the values of a log-likelihood calculation gives the estimation of the sound source. The GMMs are first trained with direction-independent speech. Then, in order to gain direction-dependent GMMs, the direction-independent speech models are adapted with the HRIRs, which contain the direction information. Thus, with a simple mathematical addition in the feature domain, the direction-dependent GMMs are calculated from the direction-independent GMMs. The negative consequence is that circular effects are introduced which destroy the localization accuracy. To get rid of these side effects three conclusions are drawn. First, HRTFs are adapted in a way that the circular effects are neglectable. This turned out to be insufficient to improve the localization accuracy. The second conclusion is to restrict the area of investigation. Problematic areas are skipped and the localization accuracy

increases, but with the drawback of the limit area. In case where a restriction is not allowed by an application, direction-dependent GMMs have to be trained. Then investigations in the whole horizontal plane are allowed. Additionally, speaker-independent models are trained for the limited area case and the case with the direction-dependent GMMs. Furthermore, the influence of the input segment is investigated for both cases as well as for speaker-dependent and speaker-independent models. For the limited area and the direction-dependent GMMs as well as for the speaker-independent cases excellent localization results, for a synthetic test scenario, can be presented.

The features used in this thesis are high-dimensional spectral features. Preliminary tests are carried out either with low-dimensional MFCCs (13 coefficients) or high-dimensional STFT coefficients (513 coefficients). To reduce the computational load, experiments with feature vectors with dimensions in between have to be carried out. Lower dimensions may perform as well as 513-dimensional STFT coefficients. MFCCs are only investigated with the minimum distance approach where they did not perform well. In additional evaluations it turned out that MFCCs (with low-dimensions) perform well in a ML framework. A comparison between MFCCs and STFT coefficients depending on the number of needed coefficients, is suggested. Although the number of components of the GMMs is 8, it is possible to increase or decrease this number to get more or less accurate results.

For further investigations in future research the localization strategies have to be verified in a real test environment. Reverberant rooms with echoes and noise complicate the task of localization. Additionally, a device with a single microphone and an "artificial pinna" has to be built. This pinna needs to be shaped irregularly and asymmetrically. The transfer function must strongly depend on the direction of the sound source. In [45] four designs were introduced. Maybe a suitable shape can be found based on the results presented in this paper.

In this thesis, sound localization is limited to the horizontal plane, i.e. the azimuth of the sound source. The focus and interaction of humans is primarily located on this plane. The used HRTFs only depend on the azimuth and elevation. The range is fixed at 1.4 meters. However, that is a crucial drawback because HRTFs also depend on the distance. This aspect should be taken into account in further studies. Nevertheless, the presented localization strategy outlines a promising approach.

A Definitions, Abbreviations and Symbols

A.1 Definitions and Abbreviations

AIC
: **A**kaike **I**nformation **C**riterion; criterion to determine the optimal number of mixture components in a GMM

azimuth
: Angle in the horizontal plane

BIC
: **B**ayesian **I**nformation **C**riterion; criterion to determine the optimal number of mixture components in a GMM

binaural
: Concerning both ears, i.e. binaural cues are exploited from the signals at both ears

CMOS
: **C**omplementary **M**etal-**O**xide-**S**emiconductor; technology for constructing integrated circuits

DCT
: **D**iscrete **C**osine **T**ransform

DFT
: **D**iscrete **F**ourier **T**ransform

Duplex Theory
: Theory about localization with ITD and ILD which was established by Lord Rayleigh; does not consider spectral cues

elevation
: Angle in the vertical plane

EM algorithm
: **E**xpectation-**M**aximization algorithm; algorithm to find the maximum likelihood in a statistical model

FFT
: **F**ast **F**ourier **T**ransform; efficient algorithm to compute the discrete Fourier transform

FIR
: **F**inite **I**mpulse **R**esponse; type of filter whose impulse response is of finite length

GMM
: **G**aussian **M**ixture **M**odel; parametric probability density function – consists of a linear combination of weighted multivariate Gaussian distributions

GMM+HRTF
: Approach with STFT coefficients, multiple, direction-dependent GMMs and ML classificator

GOF
: **G**oodness **O**f **F**it; criterion to determine the optimal number of mixture components in a GMM

HMM
: **H**idden **M**arkov **M**odel; generalization of a mixture model

horizontal plane
: Also called transverse plane; plane parallel to the ground; in this thesis: parallel to the ground and at the height of the ears

HRIR
: **H**ead-**R**elated **I**mpulse **R**esponse; path from a sound source to the ear drum is described by a linear filter (=HRIR – time domain)

HRTF
: **H**ead-**R**elated **T**ransfer **F**unction; path from a sound source to the ear drum is described by a linear filter (=HRTF – frequency domain)

IID
: See ILD

IIR
: **I**nfinite **I**mpulse **R**esponse; type of filter whose impulse response is of infinite length

ILD
: **I**nteraural **L**evel **D**ifference (also called interaural intensity difference); sound which is not located in front of the head are disturbed by the head and therefore have different levels at the two ears

ITD
: **I**nteraural **T**ime **D**ifference; sound which is not located in front of the head has different running times to one ear than to the other one

KEMAR
: **K**nowles **E**lectronics **M**annequin for **A**coustics **R**esearch; dummy head microphone

LED
: **L**ight-**E**mitting **D**iode; light source

Likelihood
: Conditional density as a function of its parameters

LimArea
: Approach where a limited area is investigated with STFT coefficients, adapted GMMs and ML classificator

Log-Likelihood
: Logarithm of likelihood due to a more simple calculation

LP
: **L**inear **P**rediction; mathematical operation

MD
: **M**inimum **D**istance classificator; classificator based on similarity

MFCC
: **M**el-**F**requency **C**epstral **C**oefficient; commonly used feature

MIT
: Media laboratory of **M**assachusetts **I**nstitute of **T**echnology

ML
: **M**aximum **L**ikelihood; procedure of finding the maximum of a known likelihood distribution for a given statistic

MMA
: **M**inimum **A**udible **A**ngle; smallest detectable change in angular position relative to the subject

monaural
: Concerning one ear, i.e. monaural cues are exploited from the signal at one ear

pinna
: Outer part of the ear; spectrally shapes incident sound waves

SCCL
: **S**elf **S**plitting **C**ompetitive **L**earning; clustering technique in image processing

SD
: **S**peaker-**D**ependent

SI
: **S**peaker-**I**ndependent

SL
: **S**egment **L**ength

SNR
: **S**ignal to **N**oise **R**atio

STFT
: **S**hort-**T**erm **F**ourier **T**ransform; frequency representation of a signal over time

STFT-MD
: Model with STFT coefficients and minimum distance approach

STFT-ML
: Model with STFT coefficients and maximum likelihood approach

vertical plane
: Also median or sagittal plane; plane perpendicular to the ground

A.2 Symbols

b	Mixture weights of a GMM
d	Euclidean distance
f	Frequency [Hz]
f_s	Sampling frequency
g_i	Discriminate function
$\mathrm{GMM}_{s,\theta}$	GMM of speaker s and direction θ
$h_\theta(t)$	HRIR from direction θ (time domain)
$h_{\theta,circ}$	Circular adapted HRIR from direction θ
$H_\theta(f)$	HRTF from direction θ (frequency domain)
$H_{L,R}(\theta, \phi, f)$	HRTF of the left/right ear (depends on θ, ϕ and f)
i	Index for classes $c_i \in C$
$\mathbf{I}$	Identity matrix
k	Index for dimensions of the feature vector; $1 \leq k \leq D$
l	Index for the feature vector; $1 \leq l \leq T$
$\mathcal{L}$	Likelihood function
m	Index for Gaussian components; $1 \leq m \leq K$
m_{el}	Subjective pitch [mel]
N	Number of frequency bins
$P(x)$	Probability density function of an univariate Gaussian distribution

$P(\mathbf{x})$	Probability density function of a multivariate Gaussian distribution
$P(c_i)$,$P(\lambda)$,$P(\mathbf{x}_l)$	Prior probability that c_i, λ respectively $\mathbf{x}_l$ occurs
$P(c_i\|\mathbf{x}_l)$	Posterior probability; probability that observation $\mathbf{x}_l$ belongs to class c_i
$P(\lambda\|\mathbf{x}_l)$	Posterior probability; probability that observation $\mathbf{x}_l$ belongs to GMM λ
$P(\mathbf{x}_l\|c_i)$	Conditional probability; probability that class c_i created observation $\mathbf{x}_l$
$P(\mathbf{x}_l\|\lambda)$	Conditional probability; probability that GMM λ created observation $\mathbf{x}_l$
r	Range: distance between ear drum and sound source; also: integer number
R	Number of samples a window is shifted
s	Index for the speaker; $1 \leq s \leq 8$
$s(n)$	Arbitrary source signal in time domain
$S_{L,R}(f)$	Arbitrary source signal in frequency domain which arrives at the left/right ear
Sp_s	Speaker s
$w[n]$	Window of length L
$x[n]$	Time discrete signal
$x_{di}(t)$	Direction-independent speech signal without direction information in time domain
$x_\theta(t)$	Direction-dependent speech signal from direction θ in time domain
$X_{di}(f)$	Direction-independent speech signal without direction information in frequency domain

$X_\theta(f)$	Direction-dependent speech signal from direction θ in frequency domain
$\mathbf{x}_l$	Feature vector
$\mathbf{X}$	Set of feature vectors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$
$X[rR, f]$	STFT of a windowed time discrete signal
Δf	Frequency resolution
ϵ	Error
$\lvert\epsilon\rvert$	Absolute error
θ, θ_{true}	Azimuth
$\hat{\theta}$	Estimated azimuth
λ	Parameter of a GMM $\lambda = \{(b_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m); m = 1, 2, \ldots, K\}$
μ	Mean value of an univariate Gaussian distribution
$\boldsymbol{\mu}$	Mean value vector of a multivariate Gaussian distribution
$\hat{\boldsymbol{\mu}}_s$	Mean value vector of feature vectors – prototype vector of speaker s
$\hat{\boldsymbol{\mu}}_\theta$	Mean value vector from STFT coefficients from direction θ
$\mu_{\lvert\epsilon\rvert}$	Mean value of $\lvert\epsilon\rvert$
σ	Standard deviation
σ^2	Variance
$\sigma_{\lvert\epsilon\rvert}$	Standard deviation of $\lvert\epsilon\rvert$
$\boldsymbol{\Sigma}$	Covariance matrix of a multivariate Gaussian distribution
ϕ	Elevation

e	Euler constant
^{T}	Transpose operator
$\lvert \cdot \rvert$	Absolute value
$\lvert\lvert \cdot \rvert\rvert$	Euclidean distance
$\div$	Division operator
$\cdot$	Multiplication operator
$*$	Linear convolution operator

Bibliography

[1] S. Carlile, "The auditory periphery of the ferret. II: The spectral transformations of the external ear and their implications for sound localization," *J. Acoust. Soc. Am.*, vol. 88, no. 5, pp. 2196 – 204, 1990.

[2] Wikipedia, "Human anatomy planes," June 2008, Information available online at http://en.wikipedia.org/wiki/File:Human_anatomy_planes.svg; visited on January 22nd 2011.

[3] L. Rayleigh, "On our perception of sound direction," *Philosophical Magazine*, vol. 13, pp. 214 – 232, 1907.

[4] J. Blauert, *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization.* Cambridge, USA: The MIT Press, 1996.

[5] R. Butler, *The Influence of the External and Middle Ear on Auditory Discriminations.* Berlin, Germany: Springer-Verlag, 1975.

[6] A. W. Mills, *Foundations of modern Auditory Theory.* New York, USA: Academic Press, 1972, ch. Auditory localization, pp. 303 – 346.

[7] D. W. Batteau, "The role of the pinna in human localization," *Proc. R. Soc. London*, vol. B168, pp. 158 – 180, 1967.

[8] J. C. Middlebrooks and D. M. Green, "Sound Localization by Human Listeners," *Annu. Review of Psychology*, vol. 42, no. 1, pp. 135 – 159, 1991.

[9] D. Brungart, N. Durlach, and W. M. Rabinowitz, "Auditory localization of nearby sources. II. Localization of a broadband source," *J. Acoust. Soc. Am.*, vol. 4, pp. 1956 – 68, 1999.

[10] P. M. Hofman, J. G. V. Riswick, and A. J. V. Opstal, "Relearning sound localization with new ears," *J. Nat. Neurosc.*, vol. 1, no. 5, pp. 417 – 421, 1998.

[11] F. L. Wightman and D. J. Kistler, "Monaural sound localization revisited," *J. Acoust. Soc. Am.*, vol. 101, pp. 1050 – 1063, 1997.

[12] P. Hofman and A. J. Van Opstal, "Binaural weigthing of pinna cues in human sound localization," *Exp. Brain Res.*, vol. 148, pp. 458 – 470, 2003.

[13] P. D. Coleman, "An analysis of cues to auditory depth perception in free space," *Psychological Bulletin*, vol. 60, no. 3, pp. 302 – 315, 1963.

[14] A. W. Bronkhorst, "Modeling auditory distance perception in rooms," in *Proc. EAA Forum Acusticum Sevilla*, Spain, 2002.

[15] B. C. J. Moore, *An Introduction to the Psychology of Hearing (5th Edition).* London, UK: Elsevier Academic Press, 2003, ch. Space Perception, pp. 233 – 269.

[16] A. W. Mills, "On the Minimum Audible Angle," *J. Acoust. Soc. Am.*, vol. 30, no. 4, pp. 237 – 246, 1958.

[17] R. M. Stern, G. J. Brown, and D. Wang, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Hoboken, USA: Wiley-IEEE Press, 2006, ch. Binaural Sound Localization, pp. 147 – 185.

[18] D. R. Perrott and K. Saberi, "Minimum audible angle thresholds for sources varying in both elevation and azimuth," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1728 – 1731, 1990.

[19] E. H. A. Langendijk, D. J. Kistler, and F. L. Wightman, "Sound localization in the presence of one or two distracters," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2123 – 2134, 2001.

[20] C. Lorenzi, S. Gatehouse, and C. Lever, "Sound localization in noise in normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 105, no. 3, pp. 1810 – 20, 1999.

[21] M. Raspaud, H. Viste, and G. Evangelista, "Binaural Source Localization by Joint Estimation of ILD and ITD," *Proc. IEEE Trans. Audio, Speech, Lang. Process. (TASSP)*, vol. 18, no. 1, pp. 68 – 77, 2010.

[22] C. M. Bishop, *Neural Networks for Pattern Recognition*, 1st ed. New York, USA: Oxford University Press, 1996, ch. Probability Density Estimation, pp. 33 – 76.

[23] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, USA: Wiley-Interscience, 2001, ch. Introduction, pp. 1 – 19.

[24] S. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357 – 366, 1980.

[25] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *J. Comput. Sci. Technol.*, vol. 16, no. 6, pp. 582 – 589, 2001.

[26] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 2nd ed. Dartmouth, USA: Prentice Hall Pearson Education, Inc., 1999.

[27] J. Cooley and J. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297 – 301, 1965.

[28] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4 – 37, 2000.

[29] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, USA: Wiley-Interscience, 2001, ch. Bayesian Decision Theory, pp. 20 – 83.

[30] M. J. Marin, K. Mengerson, and C. P. Robert, "Bayesian Modelling and Inference on Mixtures of Distributions," *Handbook of Statistics*, vol. 25, pp. 459 – 507, 2005.

[31] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. B, no. 39, pp. 1 – 38, 1977.

[32] C. Archambeau, J. A. Lee, and M. Verleysen, "On Convergence Problems of the EM Algorithm for Finite Gaussian Mixtures," in *Proc. 11th European Sym. Artificial Neural Networks*, Belgium, 2003, pp. 99 – 106.

[33] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proc. 2nd Int. Symp. Information Theory*, Budapest, Hungary, 1973, pp. 267–281.

[34] G. Schwarz, "Estimating the dimension of a model," *The Ann. of Statistics*, vol. 6, no. 2, pp. 461 – 464, 1978.

[35] M. F. A. El-Yazeed, M. A. E. Gamal, and M. M. H. E. Ayadi, "On the Determination of Optimal Model Order for GMM-Based Text-Independent Speaker Identification," *EURASIP J. Applied Signal Process.*, vol. 8, pp. 1078 – 1087, 2004.

[36] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001, ch. Robust localization in reverberant rooms, pp. 131 – 154.

[37] F. Keyrouz, Y. Naous, and K. Diepold, "A New Method for Binaural 3-D Localization Based on HRTFs," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, vol. 5, 2006, pp. 341 – 344.

[38] F. Keyrouz, K. Diepold, and P. Dewilde, "Robust 3D Robotic Sound Localization Using State-Space HRTF Inversion," in *Proc. Int. Conf. Robotics and Biomimetics*. IEEE Computer Society, 2006, pp. 245–250.

[39] F. Keyrouz, K. Diepold, and S. Keyrouz, "Humanoid Monaural Sound Localization Using Unsupervised Clustering," in *Proc. IEEE Int. Conf. Signal Process. and Communications*, 2007.

[40] F. Keyrouz, A. B. Saleh, and K. Diepold, "A Novel Approach to Robotic Monaural Sound Localization," in *Proc. 122nd Audio Engineering Society (AES) Conv.*, Vienna, Austria, 2007.

[41] C.-J. Pu, J. G. Harris, and J. C. Principe, "A neuromorphic microphone for sound localization," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics, Computat., Cybernetics and Simulation*, vol. 2, Orlando, USA, 1997, pp. 1469 – 1474.

[42] J. G. Harris, C.-J. Pu, and J. C. Principe, "A Monaural Cue Sound Localizer," *Analog Integr. Circuits Signal Process.*, vol. 23, no. 2, pp. 163 – 172, 2000.

[43] T. Takiguchi, Y. Sumida, and Y. Ariki, "Estimation of Room Acoustic Transfer Function using Speech Model," in *IEEE 14th Workshop on Statist. Signal Process.* Madison, USA: IEEE Computer Society, 2007, pp. 336 – 340.

[44] T. Takiguchi, Y. Sumida, R. Takashima, and Y. Ariki, "Single-channel talker localization based on discrimination of acoustic transfer functions," *EURASIP J. Advances in Signal Processing*, vol. 2009, pp. 1 – 9, 2009.

[45] A. Saxena and A. Y. Ng, "Learning sound location from a single microphone," in *Proc. IEEE Int. Conf. Robotics and Automation*, Japan, Kobe, 2009, pp. 1737 – 1742.

[46] E. Georganti, T. May, S. van de Par, A. Harma, and J. Mourjopoulos, "Single-Channel Sound Source Distance Estimation Based on Statistical and Source-Specific Features," in *Proc. 126th Audio Engineering Society (AES) Conv.*, Munich, Germany, 2009.

[47] B. Garnder and K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone," MIT Media Lab Perceptual Computing, Tech. Rep., 1994. [Online]. Available: http://sound.media.mit.edu/resources/KEMAR.html