ACOUSTIC SENSOR NETWORK GEOMETRY CALIBRATION AND APPLICATIONS

Dissertation

zur Erlangung des Grades eines

Doktors der Ingenieurwissenschaften

der Technischen Universität Dortmund
an der Fakultät für Informatik

von

AXEL PLINGE

Dortmund

2017

*To my friends and the sea for giving me space,
perspective and a place to return to.*

TAG DER MÜNDLICHEN PRÜFUNG:	14.06.2017
DEKAN:	Prof. Dr.-Ing. Gernot A. Fink
GUTACHTER:	Prof. Dr.-Ing. Gernot A. Fink Prof. Dr.-Ing. Rainer Martin

AKRONYMS

ANC adaptive noise canceler
ASA auditory scene analysis
ASN acoustic sensor network
ATF acoustic transfer function

BM blocking matrix
BoF bag-of-features
BSS blind source separation

C-SRP cumulative steered response power
CASA computational auditory scene analysis
CGM corpus geniculatum medium
CN cochlear nucleus
CNN convolutional neural network

DCT discrete cosine transform
DNN deep neural network
DoA direction of arrival
DRR direct-to-reverberation ratio

EM expectation-maximization
ERB equivalent rectangular bandwidth
EVD eigenvalue decomposition

FA false alarm
FBF fixed beamformer
FFT fast Fourier transform
FN false negative
FOV field of view
FP false positive
fwSNRseg frequency weighted segmental SNR

GCC generalized cross-correlation
GCC-PHAT generalized cross-correlation with phase transform
GFCC Gammatone frequency cepstral coefficient
GMM Gaussian mixture model
GPC graphics processing unit
GSC generalized sidelobe canceler

HMM hidden Markov model
HoG histograms of oriented gradients

IC inferior colliculi
ICA independent component analysis
IID interaural intensity difference
IIR infinite impulse response
ILD interaural level difference
ISM image source model
ITD interaural time difference

LCMV linearly constrained minimum variance
LPC linear prediction coefficient
LS least squares
LSO lateral superior olive

MD missed detection
MDS multidimensional scaling
MFCC mel frequency cepstral coefficient
ML maximum likelihood
MMSE minimum mean square error
MoG mixture of Gaussians
MSO medial superior olive
mTDoA maximum time difference of arrival
MUSIC multiple signal classification
MVDR minimum variance distortionless response

NMF non-negative matrix factorization

p.d.f. probability density function
PCA principal component analysis
PoAP peak over average position

RANSAC random sampling consensus
RBF radial basis function
RBM restricted Bolzmann machine
ReLU rectified linear unit
RIR room impulse response
RMS root mean square
RTF relative transfer function

SAI stabilized auditory images
SIF spectral image features
SNR signal-to-noise ratio
SOC superior olivary complex
SRP-PHAT steered response power with phase transform
STFT short time Fourier transform
SVD singular value decomposition
SVM support vector machine

TDoA time difference of arrival
ToA time of arrival
ToF time of flight
TP true positive
TTL time to live

UBM universal background model

VAD voice activity detection

WASN wireless acoustic sensor network

NOTATION & SYMBOLS

$\imath$	imaginary unit ($\imath = \sqrt{-1}$)
$\otimes$	convolution $a \otimes x(t) := \sum_k a(k)x(t-k)$
$\oplus$	cross-correlation $(x \oplus y)(\tau) := \sum_t x(t)y(t+\tau)$
$\odot$	Hamacher fuzzy t-norm
$\circ\!\!-\!\!\bullet$	Fourier transform correspondence from frequency to time domain, $H(f) \circ\!\!-\!\!\bullet h(t)$ symbolizes that $H(f)$ is the frequency domain representation of the time domain signal $h(t)$ after applying a Fourier transform.
$\propto$	proportional
$.^T$	transposed matrix or vector
$\hat{\cdot}$	an estimated quantity
$\tilde{\cdot}$	a measured quantity
$\bar{\cdot}$	an average of a quantity
$\mathcal{E}$	expected value
μ	mean
σ	standard deviation
$\mathcal{N}$	normal distribution
f_s	sampling frequency
T_{60}	reverberation time
c	speed of sound
x	source signal, i.e., dispersed by a human speaker
y	received signal, i.e., at a microphone
$\mathbf{z}$	feature vector computed from the received signal
t	discrete-time sampling index
k	frame index
K	frame size
Ω_c	class or cluster with index c
$\mathbf{s}_n$	source position with index n, mostly two-dimensional with respect to the ground floor
n	source position index
N	number of source positions
$\mathbf{m}_i$	position of microphone i
M	number of microphones
$\mathbf{r}_i$	receiver node position, the center of microphone array i
o_i	azimuthal orientation of microphone array i with respect to the global world coordinate system
R	number of microphone arrays or nodes
$\theta_{n,i}$	azimuthal direction of arrival (DoA) at microphone array i in degrees relative to the node orientation o_i for speech event n

$\tau_{n,(i,j)}$	time difference of arrival (TDoA) between signals y_j and y_i at positions $\mathbf{m}_j$ and $\mathbf{m}_i$ for sound event n from position $\mathbf{s}_n$.
ϵ_a	error of DoA measurements $\tilde{\theta}$
ϵ_τ	error of TDoA measurements $\tilde{\tau}$
ϵ_v	error of visual localizations $\hat{\mathbf{s}}$
ϵ_l	error of acoustic two-dimensional localizations $\hat{\mathbf{s}}$
ϵ_o	error of estimated orientations $\hat{o}$
ϵ_r	error of estimated positions $\hat{\mathbf{r}}$

CONTENTS

There is something common in the orientations in a city and in any scientific area: from every given point we must be able to reach any other one.

Polya and Szego

1 INTRODUCTION

In the modern world, we are increasingly surrounded by computation devices with network access in everyday situations. While smartphones are the most popular sign of this change, tablets, laptops, hearing aids, television sets, and game consoles are other examples of such devices. The availability of these devices offers new possibilities. When the devices are equipped with one or more microphones, they can work together collaboratively and become nodes in an acoustic sensor network (ASN). For the collaboration to be powerful for practical applications, the ASN needs knowledge about itself and the acoustic scene around it. This thesis provides methods that enable the nodes to automatically acquire such knowledge. The nodes can learn their geometric arrangement and the type and position of sound sources around them.

1.1 SCENARIOS

A small existing example for an ASN is the wireless combination of a hearing aid with a smartphone. The hearing impaired person can use the smartphone as a "radio microphone". It can be held by or pointed at others and transmit their speech into the hearing aid. This requires manual positioning.

More sophisticated solutions are possible if more devices are included in the network. Figure 1.1 shows an example of persons with heterogeneous devices that can work together as an ASN. These are a smartphone, tablet, laptop, and hearing aids. The assembly in this case can be considered "ad hoc", meaning that the devices come together in

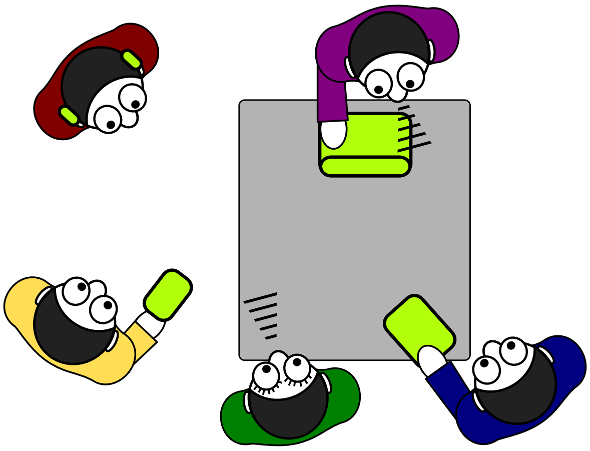

Figure 1.1: ASN example: The nodes (bright green: hearing aids, smartphone, laptop, and tablet) communicate wirelessly and process the acoustic scene together.

the moment they are to be used as a network. Let us assume, for example, that the person with the hearing aids wants to listen to the one person it faces, which does not carry a device of her own. An enhanced speech signal could be provided by the devices held by persons standing next to the speaker working together. In order to make this possible, the devices would have to be aware of their relative positioning and the speaker in question. As we can see from this example, when the nodes become automatically aware of the position and type of sound sources around them, more advanced solutions become possible. To provide robust real-time methods that provide such information is the goal of this thesis.

There are other scenarios in which this will be vital. Consider the same situation in a conference room equipped with cameras. The five persons are having a business meeting where remote participants will be joining. The cameras' positions are known in advance, as they are part of a fixed installation at the walls or ceiling. As the acoustic nodes can localize themselves as well as the active speakers, this can be used to select and control cameras for transmission and again provide an enhanced speech signal. If the participants of the meeting are known, the identity of the active speaker can be transmitted as well.

These examples show the focus of this thesis, ASNs in close proximity, typically indoors and assembled ad hoc. While the methods were developed with such situations in mind, they are applicable in a broader range of scenarios including automation, assisted living, urban planing, wildlife surveillance and more.

1.2 INFORMED ASN

The goal of this thesis is to provide robust methods for automatic and autonomous application of ASNs. What is generally missing for the collaboration of the nodes to be useful and effective, is knowledge about the auditory scene. Therefore, they should acquire this knowledge automatically.

Definition of tasks

Two types of information can be distinguished: First, the identification of types of auditory objects present over time. Second, the relative positioning of these objects in the ASN. For this, the relative geometry of the nodes has to be inferred as well.

The acquisition of information can be divided in three tasks: The first task is the detection and classification of sounds. Foremost speech has to be identified correctly for both localization and enhancement. The identification of other ambient sounds can guide the processing. Classifying the type of noise distorting the speech signal allows to chose and control the enhancement algorithm. The detection of human actions such as footsteps or chair movement helps to interpret the scene. The second task is the geometric calibration of the ASN. In order to localize speakers and other sound sources, the nodes have to know their relative arrangement with respect to each other. The third task is said localization. Once the geometry of the network is established, the nodes can collaboratively find the position of sound sources.

Challenges

To be useful for practical applications, the methods developed in this thesis have to overcome several challenges. The first challenge is reverberation. As the ASN will often be situated indoors, where the sound is reflected by the walls and other surfaces, all methods have to be robust against an unknown level of reverberation. The second challenge concerns the computations itself. It has to be realized in parallel computation distributed over the nodes. As the network can be set up in an ad hoc manner, the computations should be fast enough to be applied in real-time. As the scene may not be static, it is favorable to use adaptive online algorithms. A third challenge stems from the fact that the network connections will often be wireless. This imposes bandwidth and timing constraints. Therefore, the amount of information shared between the nodes should be low and the exchange itself be tolerant against delays. A fourth challenge is the independence of the ASN. As the methods have to be autonomous to be truly useful in practice, no additional devices or information should be necessary for them to work.

Existing approaches

The systematic approach for automated gathering of information in the ASN is a novel perspective. However, the three tasks identified are not completely new in themselves and already addressed by existing methods. Several state-of-the-art approaches for the individual tasks of sound classification, geometry calibration and speaker tracking exist. These do not cope with all the challenges just described. Several shortcomings can be identified that will be overcome with the novel approach in this thesis.

The detection and classification of auditory objects has been investigated for decades. Recently, the classification of overall scenes and the event detection has become the focus of renewed attention. This is probably also partially due to the advent of ASNs. Challenges still not fully overcome in event detection are robustness and online applicability with limited resources. Another open issue is the reliable speech detection in noise for enhancement.

Geometry calibration of distributed microphones is addressed by existing methods, but many of them require additional constraints and means that are not required otherwise. Such means are often speakers playing dedicated signals in a calibration phase, either on the devices themselves or required additionally. Rather than using a dedicated calibration sequence with special sounds, it is desirable to perform the calibration online form speech alone. Another shortcoming in existing techniques is the fact that they do not explicitly consider nodes with multiple microphones. Such nodes can estimate the angle towards sound sources, allowing the ASN to pinpoint them by triangulation. In order for this to be feasible, the relative orientation has to be calibrated with high accuracy.

Acoustic speaker tracking is a well-researched field. The collaborative online tracking by an ASN has become a new focus as this imposes new challenges. Only few existing methods consider the information exchanged between the nodes. It should be of low bandwidth and the exchange tolerant to transmission delays and errors. Similarly, how to combine this information to handle concurrent speakers is a challenging research question.

1.3 CONTRIBUTION

The author's contributions to research developed in this thesis are novel methods for sound classification, sensor calibration, and speaker tracking in ASNs. The individual methods are described in the following, along with the corresponding publications. As these are collaborative works, the author's individual contributions will be pointed out.

Detection and classification of auditory objects

A novel application of the bag-of-features (BoF) paradigm on acoustic event classification and detection is introduced. By using soft quantization and supervised training for the BoF model, superior accuracy is achieved. The method is working online and can be computed in a fraction of real-time on a consumer computer. It can be used for speaker identification as well. The method was first published at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) held in May 2014 in Florence, Italy [PGF14]. The author's contributions to the joint paper were the selection and adaptation of the BoF method, the selection of features with the introduction of Gammatone and loudness features, and the organization of the recordings. A joint in-depth investigation of the method an further extensions is presented in an article in the IEEE/ACM Transactions on Speech, Audio and Language Processing [GPF17].

Based on the method's good speech detection ability, it was extended in order to provide control for a beamformer, thus performing blind speech enhancement. The method can handle multiple concurrent noises from different directions. The classification of noise in stationarity levels allows to predict the achieved improvement. It was published at the IEEE Sensor Array and Multichannel processing workshop (SAM) held in July 2016 in Rio de Janeiro, Brazil [PG16]. The author's contributions were the design of the classifier and the introduction of a dedicated training strategy based on levels of stationarity, as well as the conduction of the recording and experiments.

Speaker tracking

The author's diploma thesis [Pli10] introduced a neuro-biologically inspired speaker localization method for microphone arrays. This single node approach proved very robust against reverberation and is able to automatically determine the number of active speakers.

Within this PhD thesis, two methodical improvements were made: In order to be applicable independently without manual adjustments, an automatic gain estimation was added. In order to better handle the unknown number of concurrent speakers, the author introduced an application of the expectation-maximization (EM) algorithm that realizes probabilistic clustering according to auditory scene analysis (ASA) principles. The refined method was first published at the European Signal Processing Conference (EUSIPCO) held in September 2013 in Marrakesh, Morocco [PF13].

Based on this approach, a system for Euclidean tracking in ASNs was designed, first published at the ICASSP conference held in May 2014 in Florence, Italy [PF14a]. The author designed the system to be online while using little bandwidth since only sparse information has to be exchanged. It is robust against jitter and transmission errors. The author added the association based on spectral similarity and introduced a special triangulation scheme that incorporates the expected accuracy.

Calibration of the nodes geometry

As the topic is of increased interest but few systematic reviews were available at the time, a dedicated survey of the field was performed. It was published in an article in the IEEE Signal Processing Magazine in July 2016 [PJHUF16]. The author contributed the general idea of the survey and the organization of state-of-the-art approaches according to an application oriented taxonomy. He also contributed the evaluation framework, data recordings, and experiments with his methods for off-line calibration as well as multidimensional scaling (MDS) based techniques.

For video conferencing scenarios, it is important to provide integration of acoustic and visual sensors. The first method developed is a multimodal approach using the visual localization of a speaker at a small number of fixed positions. By matching the positions to the direction of arrival (DoA) estimates of the microphone arrays, their absolute position and orientation are derived. The method was devised by the author and first published in the EUSIPCO conference held in September 2014 in Lisbon, Portugal [PF14b].

The second method is using acoustic measurements only. It works with speech events from distinct unknown positions. The author introduced a single target function that combines DoA and time difference of arrival (TDoA) measurements. This allows for offline calibration with dedicated recordings and achieves high orientation accuracy. The method was first published in the International Workshop on Acoustic Signal Enhancement (IWAENC) held in September 2014 in Antibes, France [PF14c].

The method was later refined for online application. The author introduced an evolutionary algorithm that is able to solve the combined target function in real-time. The use of incremental measurements makes it possible to calibrate online. By using a sparse spike representation computed by the neuro-biologically inspired speaker localization model for the DoA estimation, it is robust and requires lower bandwidth for sharing information between the nodes. This version is presented in an article in the IEEE Signal Processing Letters [PFG17].

Method connections

The overall goal is the development of an automatically informed ASN. The novel methods developed for this goal complement each other in multiple ways, as illustrated in Figure 1.2: The event classification is employed as pre-filter for the calibration and tracking, as it helps to exclude non-speech sounds. The single array tracking method is employed in turn for acoustic sensor network geometry calibration. Together with visual speaker localization, it is used for the multimodal approach. Once the geometry is calibrated, the sensor network can work together for Euclidean speaker tracking. Both the sound classification and tracking methods can control a beamformer for speech enhancement.

1.4 OUTLINE

As this thesis proposes novel methods for three complementary tasks, the remaining text of this thesis is organized in a way that the overlapping aspects can be addressed together while the individual methods are discussed separately. As many of the developed methods are based on shared assumptions, common principles and fundamentals of the novel methods will be introduced in a background chapter. Thereafter, the tasks of sound classification, sensor localization and person tracking will each be discussed in individual chapters. Each chapter contains an overview of the state-of-the-art related work and an in depth description of the developed method. As the methods are used in conjunction with each other in overlapping scenarios, a detailed evaluation of all three methods will be described in a common chapter. A final conclusion chapter summarizes the main results and contributions of the thesis. The individual chapters' contents are:

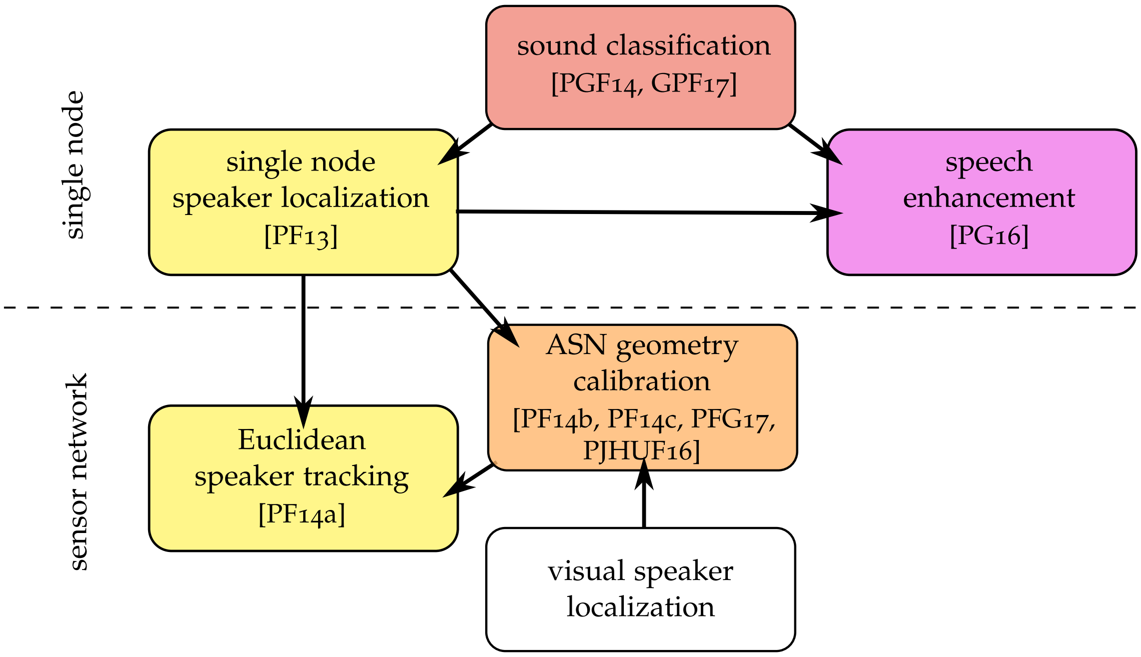
single node
sensor network
sound classification [PGF14, GPF17]
single node speaker localization [PF13]
speech enhancement [PG16]
ASN geometry calibration [PF14b, PF14c, PFG17, PJHUF16]
Euclidean speaker tracking [PF14a]
visual speaker localization

Figure 1.2: Dependencies of the novel methods. Sound classification is used as pre-filter for tracking and geometry calibration. It is also used to provide control information for speech enhancement. The central calibration of an acoustic sensor network enables Euclidean speaker tracking.

- The following Chapter 2 introduces common notation and fundamental background knowledge for the work. The basic mathematical description of the geometric setup and signals involved will be introduced first. A short overview of the physiology of human hearing and the main concepts of the influential auditory scene analysis theory will be given. Thereafter the basic principle and notation of the maximum likelihood estimation used for both classification and localization will be introduced. The chapter is concluded by a brief outline of the common concepts for speech enhancement.

- The state-of-the-art in sound classification will be described in Chapter 3 with a focus on the different classifiers and features used in acoustic event detection. The novel method applying the bag-of-features approach is explained in detail. The chapter is concluded by a description of the blind speech enhancement method based on the classification of speech.

- The acoustic person localization task will be addressed in Chapter 4. First, common state-of-the-art approaches will be described, including the neurobiologically inspired models developed by the author prior to this thesis. Then, the novel method for single array speaker localization and tracking will be explained. The chapter is concluded by a description of the extension of this method to multiple distributed microphone arrays.

- The problem of acoustic sensor geometry calibration will be introduced in depth in chapter 5. Related methods for the calibration of individual microphones and microphone arrays will be discussed. The author's contributions will be presented, starting with the novel audiovisual method. Then the novel approach to solve the more difficult problem of calibration from acoustic data only will be explained.

- In Chapter 6, all proposed methods will be evaluated. First the common metrics and data used for the evaluation is described. Then the sound classification will be evaluated for event detection and speech enhancement. The evaluation results for single node speaker localization will be presented next. The following sections present evaluation results for the calibration of distributed microphone arrays using simulated data or results from the single node localization. Then the results for tracking with multiple distributed microphone arrays are described. The effect of calibrating the arrays with the novel calibration method will be investigated. An experiment employing all methods in conjunction is presented at the end of the evaluation.

- Chapter 7 concludes this text with a discussion of the results achieved. A perspective towards further research will be given and the impact of the new methods for ASN applications will be outlined.

Here and elsewhere we shall not obtain the best insight into things until we actually see them growing from the beginning.

Aristotle, Politics

2 BACKGROUND

For methods in an acoustic sensor network (ASN) there are some common fundamental principles. In this chapter, the mathematical formulation of the sound propagation towards microphones will be introduced. The basic measurements and geometric relations will be described in the next section. As insights of the neuro-biological principles of human sound perception are used in several methods, the basic terms and ideas will be introduced in Section 2.2. Next, the common framework of maximum likelihood (ML) based clustering used in both the classification and localization methods will be explained in Section 2.3. Concluding this chapter is Section 2.4 which introduces basic principles of speech enhancement.

2.1 SOUND AND GEOMETRIC RELATIONS

In order to introduce notation and give some general background, the basic model of sound propagation will be introduced here. As most of the work is concerning indoor scenarios, a characterization of reverberation and its effect will be given. The mathematical notation will be introduced starting with the measurements on the example of a pair of microphones. Then it will be extended to the measurements in an network of nodes and the quantities to be estimated in its geometric calibration.

2.1.1 *Sound*

Sound is propagated as density variation in the acoustic medium. It consists of longitudinal compression waves in fluids and air. Whether it originates from a loudspeaker or a human mouth, it starts as almost spherical wave. The expansion pattern is bent by the head or loudspeaker, more so at high frequencies. When multiple sources are present, their signals add linearly.

Propagation

The propagation speed is a function of the medium through which the wave travels. Air forms an almost homogenic medium, so the speed can be assumed constant. Since the variation is small for indoor temperatures, the speed it is often approximated by the constant value 343 m/s or linearly as function of the temperature K in Kelvin [Kut00]:

$$c = 167.61 + 0.6K = 331.5 + 0.6(K - 273.15) \tag{2.1}$$

This simplification does not hold for other media. In sea water, for example, the speed has to be modeled as a function of the local salinity as well as the local temperature. Similarly, when sound is conducted through floor or furniture, the medium is hardly homogenic and the propagation therefore far from linear.

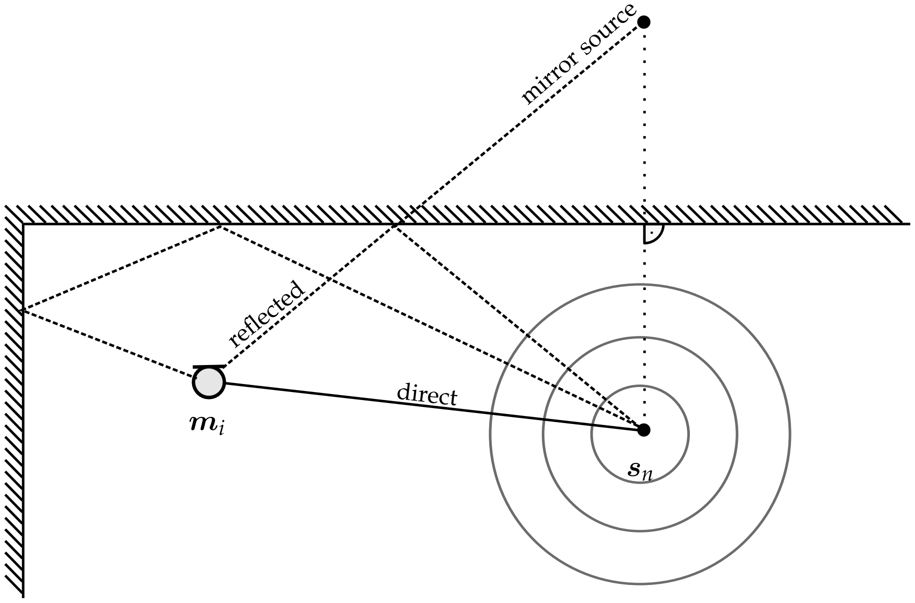

Figure 2.1: Indoor sound propagation model: The spherical wave from the source position s_n is propagated to the sensor m_i. Reverberation adds different paths which lead to the signal arriving later at the sensor. These can be modeled as mirror sources as shown for the first order reflection.

Reverberation

In indoor environments, the sound does not only reach the sensor on the direct path, but via indirect paths by reflection on walls and objects in the room as illustrated in Figure 2.1. It is notable that the signal is attenuated at each reflection, which can be quantified by an absorption coefficient. The reverberations can also be seen as image sources with a lower amplitude and time delay coming from the mirrored directions [AB79].
One way to quantize the amount of reverberation is the reverberation time T_{60}, defined as the time it takes the total reverberation of a signal to attenuate by 60 dB after the source is turned off.
The distance at which the reverberation and the direct path signal have equal power is called critical distance. It can be approximated with the reverberation time T_{60} [s] and room volume V $[m^3]$ as [Kut00, p. 317]:

$$r_D \approx 0.057\sqrt{\frac{V}{T_{60}}}. \tag{2.2}$$

Signal model

As the processing is digital, we are dealing with discrete-time signals with a sampling index t with respect to the sampling frequency f_s. All time measurements will be expressed accordingly.
Because of the additive and linear properties of sound, all paths can be added and expressed as a single linear filter a. This filter is called the acoustic transfer function (ATF). In this case it is also referred to as room impulse response (RIR). The microphone signal is modeled as the signal filtered by the ATF with some additive noise n:

$$y_j(t) = a_j \otimes x_j(t) + n(t) \tag{2.3}$$

where j is the microphone index. Often, the processing is done in the short time Fourier transform (STFT) domain. The Fourier transform is applied to overlapping time frames k of, e.g., $K = 1024$ samples yielding a frequency representation $Y_j(k, f)$ of the time signal. Each vector Y_j is comprised of $K = F$ frequency bins with index f.

$$Y_j(k, f) \circ\!\!-\!\!\bullet\, y_j(t = kK, kK+1, \ldots, kK+K-1) \tag{2.4}$$

Here, the convolution in (2.3) becomes a multiplication [Smi99]:

$$Y_j(k, f) = A_j(k, f) X_j(k, f) + N(k, f). \tag{2.5}$$

By stacking the vectors $\boldsymbol{Y} = [Y_0, Y_1, \ldots Y_{M-1}]^T$ of the microphone signals and the corresponding transfer functions $\boldsymbol{A} = [A_0, A_1, \ldots A_{M-1}]^T$, the recording of one or more acoustic sensor nodes can be written as

$$\boldsymbol{Y}(k, f) = \boldsymbol{A}(k, f) X(k, f) + \boldsymbol{N}(k, f). \tag{2.6}$$

2.1.2 *Sensor node measurements*

For the proposed methods, we assume each sensor node is equipped with a non-linear planar microphone array. A microphone array can acquire three basic types of measurements with respect to a sound source. They will be defined in the following. As each can be estimated using cross-correlation, this will be formalized at the end of this section.

ToA

The time $t_{n,i}$ it takes to reach a sensor at position $\boldsymbol{m}_i$ from a source position $\boldsymbol{s}_n$ is linearly dependent on the distance. This time is referred to as the time of arrival (ToA) or time of flight (ToF),

$$t_{n,i} = \|\boldsymbol{s}_n - \boldsymbol{m}_i\| f_s / c. \tag{2.7}$$

In Figure 2.2 the geometric properties for a pair of microphones are illustrated. Since the propagation of sound is spherical, the distance $d_{n,(i,j)}$ in the direction towards source $\boldsymbol{s}_n$ is equal to the difference of the individual distances of each microphone towards the source,

$$d_{n,(i,j)} = \|\boldsymbol{s}_n - \boldsymbol{m}_i\| - \|\boldsymbol{s}_n - \boldsymbol{m}_j\| \,. \tag{2.8}$$

TDoA

If signals y_i,y_j received at two microphones with index i, j were free of reflections and noise, they would be identical but for attenuation and the time offset $\tau_{i,j}$, called the time difference of arrival (TDoA).

$$y_j(t) \propto y_i(t + \tau_{i,j}) \tag{2.9}$$

It is given in samples by the difference of distances from the microphone positions $\boldsymbol{m}_i$, $\boldsymbol{m}_j$ to the source position $\boldsymbol{s}_n$ multiplied with the given sampling rate f_s and divided by the speed of sound c,

$$\tau_{(i,j)}\left(\mathbf{s}_n\right) = \tau_{n,(i,j)} = d_{n,(i,j)} f_s / c = (\|\boldsymbol{s}_n - \boldsymbol{m}_j\| - \|\boldsymbol{s}_n - \boldsymbol{m}_i\|) f_s / c = t_{n,j} - t_{n,i} \,. \tag{2.10}$$

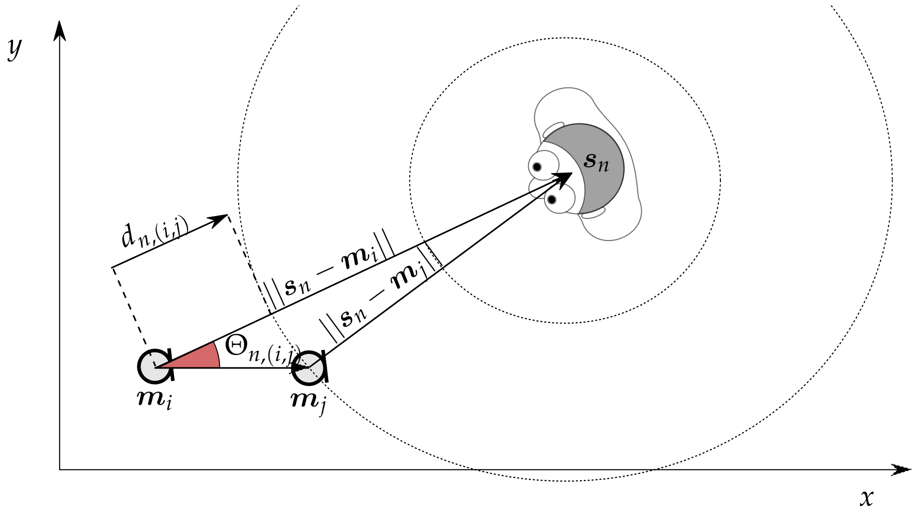

Figure 2.2: Geometric relations between two microphones and a speaker. The sound propagation from the speaker is illustrated as dotted circles.

A constant TDoA for one microphone pair corresponds to a hyperboloid surface in the room. The positions in line with a microphone pair are termed the "endfire" positions, the perpendicular ones are termed "broadside" positions. We can see that the TDoA is zero at the broadside positions and reaches its maximum value at the endfire positions.

DoA

If the distance of the source is large compared to the microphone distance, the hyperboloid becomes an angular line segment. This situation is referred to as the far field assumption. In this case, the circle intersecting the position $\boldsymbol{m}_j$ of the right microphone is approximately a line perpendicular to $\boldsymbol{m}_j - \boldsymbol{s}_n$, cf. Figure 2.2. The TDoA becomes a function of the angle $\Theta_{n,(i,j)}$ towards the speaker

$$\tau_{n,(i,j)} \approx \cos\left(\theta_{n,(i,j)}\right) \|\boldsymbol{m}_i - \boldsymbol{m}_j\| f_s / c \tag{2.11}$$

Therefore, only this angle, the direction of arrival (DoA) can be estimated from the TDoA:

$$\theta_{n,(i,j)} \approx \arccos\left(\frac{\tau_{n,(i,j)} c}{\|\boldsymbol{m}_i - \boldsymbol{m}_j\| f_s}\right). \tag{2.12}$$

The DoA can also be expressed as a unit vector. In the two dimensional case, the following vector $\boldsymbol{\alpha}$ can be used given the angle θ relative to the coordinate origin.

$$\boldsymbol{\alpha}(\theta) := \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix} \tag{2.13}$$

Cross-correlation

A typical way to derive these measurements is the cross-correlation. For discrete-time signals, this is the element-wise multiplication of the two signals with a given time offset

τ, sometimes also denoted as expected value over time. In practice, this is evaluated over a short time frame of K samples.

$$(x_i \oplus x_j)(k,\tau) := \mathcal{E}_t\left(x_i(t)x_j(t+\tau)\right) = \sum_{t=kK}^{kK+K-1} x_i(t)x_j(t+\tau) \tag{2.14}$$

The time lag where the function reaches its maximum value is the time offset between two non-periodic signals.

$$\hat{\tau}_{k,(i,j)} = \operatorname*{argmax}_{\tau}\,(x_i \oplus x_j)(k,\tau). \tag{2.15}$$

All three measurements can be estimated by cross-correlation. The ToA is computed by correlating the signal received at one microphone with the source signal. The TDoA for a pair of microphones can be estimated by correlating their signals. The DoA is derived from the TDoA as described above (2.12).

When correlating microphone signals in order to estimate the TDoA, spectral distortion and reverberation can change the signals in a way that the cross-correlation's maximum is not always found at the time lag resulting form the length difference of the direct paths. In order to counter these effects, compensating filters $\tilde{h}_i$, $\tilde{h}_j$ with $x_i(t) \approx \tilde{h}_i y_i(t)$ can be applied before computing the cross-correlation. Then the TDoA estimate is computed as

$$\hat{\tau}_{k,(i,j)} = \operatorname*{argmax}_{\tau}\,\underbrace{\mathcal{E}_t\left\{(\tilde{h}_i \otimes y_i(t))(\tilde{h}_j \otimes y_j(t+\tau))\right\}}_{r_{y_iy_j}(k,\tau)}\,. \tag{2.16}$$

The function r is known as the generalized cross-correlation (GCC) [KC76]. It is often computed in the STFT domain in the following way: Rather than computing the cross-correlation in the time domain, the cross-power spectral density $\Phi_{y_iy_j}(k,f)$ is computed as the product of one signal's spectrum with the complex conjugate of another signal's spectrum. For deterministic signals, this is identical with the Fourier transform of the cross-correlation [Smi99]:

$$\Phi_{y_iy_j}(k,f) = Y_i(k,f)Y_j^*(k,f) \;\bullet\!\!-\!\!\circ\; (y_i \oplus y_j)(k,\tau)\big|_{\tau=-K/2}^{K/2} \tag{2.17}$$

Therefore the GCC can be computed in the Fourier domain from the cross-power spectral density multiplied with a compensation $G(k,f)$ given by the conjugate product of the spectral compensations:

$$r_{y_iy_j}(k,\tau) \;\circ\!\!-\!\!\bullet\; \frac{1}{F}\sum_{f=-F/2}^{F/2} \underbrace{\tilde{H}_i(k,f)\tilde{H}_j^*(k,f)}_{G(k,f)}\,\Phi_{y_iy_j}(k,f)e^{\mathrm{j}2\pi f/F\tau} \tag{2.18}$$

In practice the distortions are unknown, so this $G(k,f)$ has to be estimated as well. A successful approach is the use of the phase transform that normalizes the spectral amplitudes.

$$G_{y_iy_j}^{PHAT}(k,f) = \frac{1}{\left|\Phi_{y_iy_j}(k,f)\right|} \tag{2.19}$$

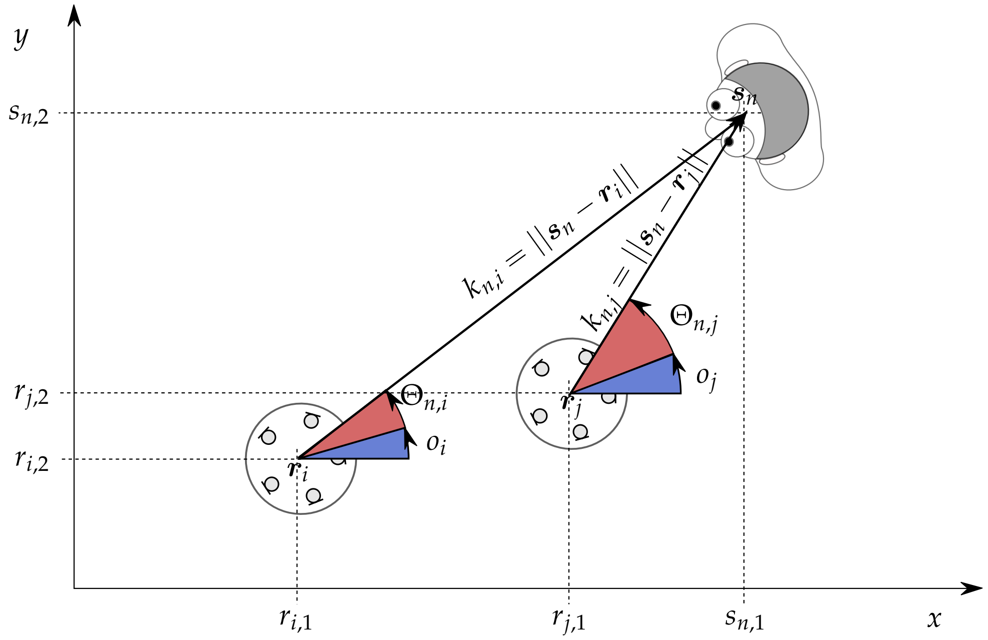

Figure 2.3: ASN geometry and triangulation using two nodes.

By substituting the unknown $G(k,f)$ with this function, we arrive at the function $r^{PHAT}(k,\tau)$ known as the generalized cross-correlation with phase transform (GCC-PHAT):

$$r^{PHAT}_{y_i y_j}(k,\tau) \circ\!\!-\!\!\bullet \frac{1}{F} \sum_{f=-F/2}^{F/2} \frac{\Phi_{y_i y_j}(k,f)}{\left|\Phi_{y_i y_j}(k,f)\right|} e^{\mathrm{j}2\pi f/F\tau} \tag{2.20}$$

The TDoA estimate is computed as its maximum over possible TDoAs according to (2.16).

2.1.3 *Sensor network geometry*

When using more than one sensor node for spatial processing, their relative geometry is relevant. For the target scenarios, the two-dimensional projection to the floor is used. We denote the absolute position of a node with index i as $\boldsymbol{r}_i$. The node is equipped with multiple microphones arranged in a planar array. We define its orientation as o_i as the angle of the line from the center of the node to the first microphone to the abscissa. Using these definitions, we can formalize the relation between the geometry of one or more nodes and one source position. Figure 2.3 shows these quantities and illustrates how the speaker position relates to the ASN geometry and the DoA measurements.

Triangulation

Using the known geometry of two nodes and simultaneous DoA measurements, the speaker can be localized in absolute coordinates by triangulation – except when the speaker and the two nodes form a line, in this case the direction but not the distance can be derived.

Given the orientations $o_{i,j}$ and positions $r_{i,j}$ of two acoustic sensor nodes, and given the DoAs $\theta_{i,t}$ and $\theta_{j,t}$ the speaker position is computed by solving

$$\hat{s}_{n,(i,j)} = \boldsymbol{r}_i + k_{n,i}\boldsymbol{\alpha}\left(o_i + \theta_{n,i}\right) = \boldsymbol{r}_j + k_{n,j}\boldsymbol{\alpha}\left(o_j + \theta_{n,j}\right) \ . \tag{2.21}$$

This is easily solved with some vector algebra. Given the definition of α it is already a unit vector, its perpendicular vector can be defined as:

$$\boldsymbol{\alpha}_{\perp}(\theta) := \begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \end{pmatrix} \tag{2.22}$$

As the distance of $\boldsymbol{s}_n$ to the ray starting in $\boldsymbol{r}_j$ is zero, it follows that

$$\boldsymbol{\alpha}_{\perp}\left(o_j + \theta_{n,j}\right)^T \left(\boldsymbol{s}_n - \boldsymbol{r}_j\right) = 0. \tag{2.23}$$

We can formulate a simple vector equality

$$\boldsymbol{s}_n - \boldsymbol{r}_j = (\boldsymbol{r}_i - \boldsymbol{r}_j) + k_{n,i}\boldsymbol{\alpha}\left(o_i + \theta_{n,i}\right). \tag{2.24}$$

We can now substitute (2.24) in (2.23) and it follows that

$$\boldsymbol{\alpha}_{\perp}\left(o_j + \theta_{n,j}\right)^T \left((\boldsymbol{r}_i - \boldsymbol{r}_j) + k_{n,i}\boldsymbol{\alpha}\left(o_i + \theta_{n,i}\right)\right) = 0, \tag{2.25}$$

which is solved for $k_{n,i}$. For $k_{n,j}$ obviously the solution is the same but for swapping i and j. The distances can be computed as scalar product of the connecting vector of the array positions and the vector in the perpendicular DoA direction divided by the scalar product of the perpendicular direction and the other DoA direction:

$$k_{n,i} = \frac{(\boldsymbol{r}_i - \boldsymbol{r}_j)\boldsymbol{\alpha}_{\perp}\left(o_j + \theta_{n,j}\right)^T}{\boldsymbol{\alpha}\left(o_i + \theta_{n,i}\right)\boldsymbol{\alpha}_{\perp}\left(o_j + \theta_{n,j}\right)^T} \text{ and } k_{n,j} = \frac{(\boldsymbol{r}_j - \boldsymbol{r}_i)\boldsymbol{\alpha}_{\perp}\left(o_i + \theta_{n,i}\right)^T}{\boldsymbol{\alpha}\left(o_j + \theta_{n,j}\right)\boldsymbol{\alpha}_{\perp}\left(o_i + \theta_{n,i}\right)^T} \tag{2.26}$$

When both distances $k_{n,i}$ and $k_{n,j}$ are positive, the rays starting at $\boldsymbol{r}_i$ in direction $o_i + \theta_{n,i}$ and $\boldsymbol{r}_j$ in direction $o_j + \theta_{n,j}$ have an intersection at $\boldsymbol{s}_n$ as stated in (2.21).

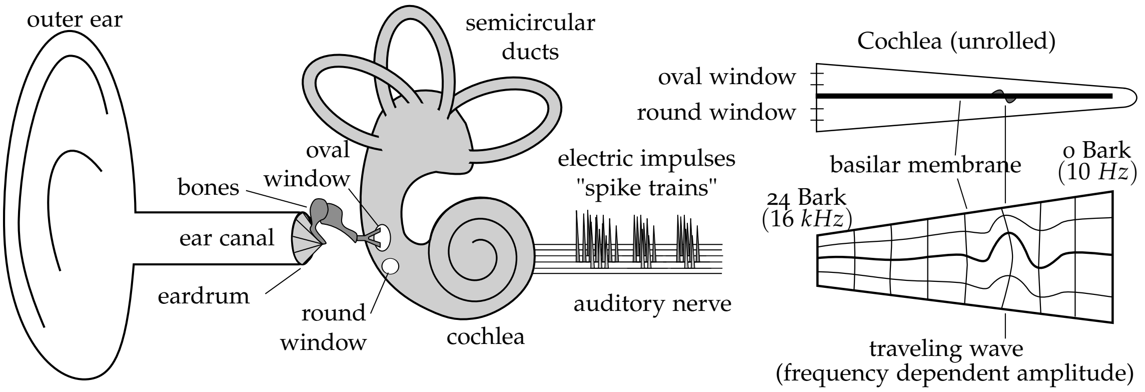

Figure 2.4: Mammalian hearing – From sound waves to electrical impulses

2.2 HUMAN AND MACHINE HEARING

The biologically inspired processing has been shown to improve the performance of different methods in practical conditions. The approach of using neurological and psychological insights in acoustic signal processing is also known as "machine hearing" [Lyo10, Lyo17]. As the proposed localization methods are also based on insights in and theories on human hearing, the fundamental principles will be described in this section. The early stages of sensory processing are well researched by neurological experiments, its basic functionality and corresponding computational models will be described. The higher stages of neural processing are less clearly understood by neural studies. Here psychological experiments allow understanding of the overall function. The influential theory of auditory scene analysis (ASA) offers rich interpretations and will be outlined.

2.2.1 *From sound waves to electrical impulses*

The transformation from sound waves into electrical impulses in mammalian hearing is illustrated in Figure 2.4. The sound waves cause movement of the eardrum, which is amplified mechanically by three assides and transmitted to the cochlea via the oval window. Inside the cochlea, a traveling wave in an incompressible fluid is set into motion. Since the volume of the fluid changes along the spiral, the amplitude of the wave is a function of its frequency and place, realizing a frequency-to-place transformation. The basilar membrane is placed along the spiral. Between the membrane and the bone, the organ of Corti resides which contains outer and inner hair cells. The former change the flexibility of the membrane, acting as an active filter. The latter generate electrical impulses when sheared by the movement. The auditory nerve continuously transports these impulses to the brain. This signal is figuratively referred to as "spike trains" [Han89].

2.2.2 *Cochlear models*

A multitude of computational models exists for this process [WB06]. One of the first models was proposed by Lyon in 1982 [Lyo82] and later refined to binaural processing [Lyo83]. His models use a Gammatone filterbank to model the frequency selectivity and frequency-to-place encoding of the cochlea. For encoding of the band amplitude, halfway rectification and square-root compression is used. For phase encoding, zero crossing detection in the band filtered signals is employed. Many later computational auditory scene analysis (CASA) models are derived from this approach. Most use a

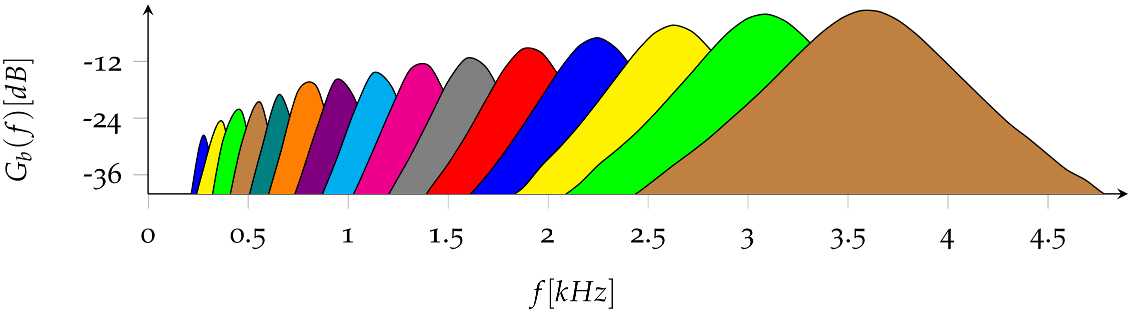

Figure 2.5: Gammatone filterbank, filter amplitudes in dB for ERB spaced center frequencies in the range of 0.3 to 3.6 kHz

Gammatone filterbank, often with the infinite impulse response (IIR) filter approximation derived by Slaney [Sla93]. Both halfway rectification and zero crossings are still widely used. According to some researchers [Gro03], the human hearing actually employs signal maxima for the phase locking process.
In this thesis, a Gammatone filterbank is used for the event detection and localization. It is computed from the STFT as introduced in [PHF10]: Rather than using IIR filters, the filterbank is computed with the fast Fourier transform (FFT) overlap-add method [Smi99]. The microphone signals y_i are each transformed into the spectral domain with the STFT. Input frames with 50% overlap are multiplied with a cisoid (sine-shaped) window before the FFT is computed. For each band b, the resulting frame spectrum Y_i is multiplied with a specific frequency response G_b to yield a bandfiltered spectrum $Z_{i,b}$.

$$Z_{i,b}(k,f) = G_b(f)Y(k,f) \tag{2.27}$$

The center frequencies f_b are distributed equidistantly on the equivalent rectangular bandwidth (ERB) scale [GM90]:

$$ERBS(f) = 24.1 log_{10}\left(1 + 4.37 \cdot 10^{-3} \cdot f\right) \tag{2.28}$$

The filters are defined using a Gammatone approximation introduced by Unoki et al. [UA99], with $\imath$ denoting the imaginary unit and w_b the Glasberg-Moore bandwidth [GM90]:

$$\hat{G}_b(f) = \left(1 + \frac{\imath(f - f_b)}{w_b}\right)^{-4} \quad \text{with} \quad w_b = 24.7(4.37 f_b + 1). \tag{2.29}$$

The frequency responses of the filters are illustrated in Figure 2.5. After the filtering, the signal for each band is transformed back into the temporal domain by the inverse Fourier transform. The frames are added together with 50% overlap to yield the resulting continuous time signal $z_{i,b}(t)$.

2.2.3 *Early neural processing*

The auditory nerve first reaches the cochlear nucleus (CN). The outer ear has a specific shape that leads to cancellation of specific frequencies depending on the DoA of the signal. This is evaluated in the CN, allowing for estimation of the sources' elevation.

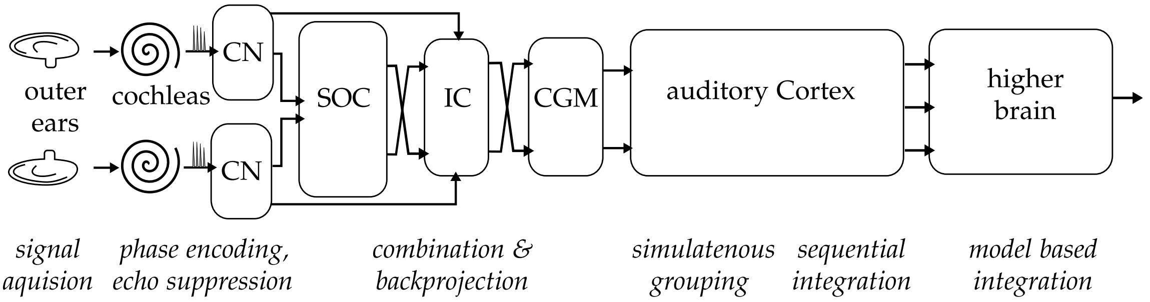

Figure 2.6: Auditory pathways and ASA stages of processing.

From the CN, the signal reaches the superior olivary complex (SOC) in the auditory midbrain, where the main binaural cues are computed. Due to the spatial distance of the ears, the sound arrives with a direction dependent TDoA. In the medial superior olive (MSO), this interaural time difference (ITD) is derived. According to the Jeffress-Colburn model [Jef48] cross-correlation of the band-filtered signals is performed. Multiple more advanced models have been proposed (see [Bla96] for an overview).

In Lyons binaural model [Lyo83] zero crossings are used for phase encoding of the signals. The zero-crossings are then used for computation of the ITD. Rather than computing the correlation of signals derived from two microphones, the ITD can be calculated more efficiently as time offset between zero crossing positions [HOS95, PS06, KAK06].

Due to head shadowing effects, the signal also has a direction dependent level difference. In the lateral superior olive (LSO), the interaural intensity difference (IID) is derived.

In the inferior colliculi (IC), the sounds are back-projected to three-dimensional space. This is done using ITD and IID as well as the elevation estimates from the CN.

The next stage is the corpus geniculatum medium (CGM), where frequency, intensity and binaural information are combined in a frequency and spatially dependent representation. Temporal changes are analyzed within this representation in the middle region [Han89, p. 521].

2.2.4 *Echo suppression*

Many physiological studies have shown the so called "precedence effect", sometimes also referred to as the "law of the first wave-front", in mammalian hearing: In localization, the first wave front of a signal is evaluated with higher weighting [Bla96]. This is a mechanism for echo suppression, because the first wave-front reaching the ear is the one that took the shortest, i.e., direct path. The reflections arrive later due to their longer path through the room. The effect was also shown in neurological research, where it is mostly referred to as "onset dominance" [DIH$^+$09]. The precedence of onsets in the neural processing is said to be found both in the monaural and binaural stages of the neural pathway.

Monaural echo suppression may already be done in the CN [BvH07]. A basic way to realize this is to compare the signal to an average shifted in time and suppress any signal below that average to mimic neural saturation [PBW04, PHF10].

2.2.5 *Auditory scene analysis*

The term of ASA was coined by Albert Bregman in the 1990s. Influence by Gestalt theory, he was one of the first to point out the analogies of the process of building an abstract representation from hearing with the interpretation of a visual scene [Bre90].
Using extensive psychological studies, the auditory perception is interpreted as a bottom-up process in which various cues are extracted from the auditory input. Important monaural cues are the timbre, i.e., the spectral distribution, and pitch, i.e., the fundamental frequency, of the sounds. Studies also show the importance of the direct-to-reverberation ratio (DRR) and other indirect cues. The most important binaural cues are the ITD and IID as they provide localization of sounds.

Processing stages

According to ASA, these cues are used to group together individual parts of the sounds as belonging to the same source and forming an auditory object. The process of combining related acoustic features is called "simultaneous grouping". Strong cues for this process are common onset, ITD and IID as well as amplitude and frequency modulation.
The next stage is the integration of these groups over time. In the "sequential integration", consecutive groups are connected when they are close by at least one cue. Strong cues for this process are pitch, rhythm and spatial proximity. In a larger time contexts, there can be multiple cues suggesting contradicting associations. Here, the one with highest proximity is chosen.
The bottom-up information acquired by the auditory cortex is not necessarily sufficient. In higher brain regions, different learned models are used to interpret the scene. These include a model of the room and its acoustic properties as well as speech and speaker models. This stage is termed "model-based integration". The "glimpsing model" suggests that human speech perception in adverse conditions is based upon sparse clear events with high signal-to-noise ratio [Coo06] and filled by top-down model-based processing.

CASA

The computational implementation of ASA principles has become increasingly popular in the last decade. CASA approaches often target the separation of concurrent speakers by means of time-frequency masks, often in the most difficult single channel case [WB06]. However, several auxiliary tasks are addressed by CASA, such as speaker localization and identification from the signals of an artificial human head [MPK13].

2.3 MAXIMUM LIKELIHOOD CLUSTERING

ML estimation via the expectation-maximization (EM) algorithm is used in a variety of applications. It estimates parameters for statistical models in an iterative manner in cases when an analytical solution is impossible or very hard to obtain. Some of the most common case are the estimation of Gaussian mixture model (GMM) or hidden Markov model (HMM) parameters [Bil98, Fin14]. The estimation of a mixture of Gaussians (MoG) is also considered as probabilistic clustering, and can be seen as an extension of Lloyds algorithm [LBG80, Llo82] which is often used for its initialization. Since the EM algorithm is employed in many of the proposed methods, the basic principle and the MoG estimation will be introduced here along with the notation used within this thesis.

2.3.1 *Maximum likelihood estimation*

The ML estimation is common practice for deriving parameters of statistical models. Its general principle can be formalized as follows: The observation is given as a set $\boldsymbol{Z}$ of N independent and identically distributed (i.i.d.) data points $\mathbf{z}_n$. These can be, e.g., feature vectors or spatial likelihood values. The set can be described by a probability density of $\mathbf{z}_n$ with parameters $\boldsymbol{\Omega}$, denoted as $P(\mathbf{z}_n|\boldsymbol{\Omega})$, and the probability density function (p.d.f.) is written as

$$f_{\boldsymbol{\Omega}}\left(\boldsymbol{Z}\right) = P\left(\boldsymbol{Z}|\boldsymbol{\Omega}\right) = \prod_{n=1}^{N} P\left(\mathbf{z}_n|\boldsymbol{\Omega}\right) =: \mathcal{L}\left(\boldsymbol{\Omega}|\boldsymbol{Z}\right), \tag{2.30}$$

the function $\mathcal{L}$ is referred to as the likelihood function. It expresses how well the model corresponds to the data. The goal of the ML estimation is to find parameters $\boldsymbol{\Omega}^*$ that maximize $\mathcal{L}$. In many cases, e.g., when working with exponential distributions, it is easier to maximize the log-likelihood. This leads to the same parameters as log is a strictly monotonically increasing function.

$$\boldsymbol{\Omega}^* = \underset{\boldsymbol{\Omega}}{\operatorname{argmax}}\, \mathcal{L}\left(\boldsymbol{\Omega}|\boldsymbol{Z}\right) = \underset{\boldsymbol{\Omega}}{\operatorname{argmax}}\, \log \mathcal{L}\left(\boldsymbol{\Omega}|\boldsymbol{Z}\right) = \underset{\boldsymbol{\Omega}}{\operatorname{argmax}} \sum_{n=1}^{N} \log P(\mathbf{z}_n|\boldsymbol{\Omega}) \tag{2.31}$$

In many practical cases, there is either no closed form solution or it is hard to obtain.

2.3.2 *The EM algorithm*

The general EM algorithm is iteratively optimizing a statistical model in such cases [DLR77]. It is applied to a model that contains some hidden or latent variables whose parameters are unknown. The observation might not inform us about these or the formulation of the model can be greatly simplified by introducing them. The data set $\boldsymbol{Z}$ is considered *incomplete* in the sense that the observation contains only a fraction of the complete information. The maximized expression then becomes the *complete-data log-likelihood* with latent or hidden random variables $\boldsymbol{v}$:

$$\log \mathcal{L}\left(\boldsymbol{\Omega}|\boldsymbol{Z}, \boldsymbol{v}\right) = \log P\left(\boldsymbol{Z}, \boldsymbol{v}|\boldsymbol{\Omega}\right) \ . \tag{2.32}$$

The EM algorithm iteratively performs two steps in order to maximize this likelihood. The first is the expectation or E-step that estimates the hidden variables based on the

current model parameters and the observed data. The second is the maximization or M-step, that computes new model parameters using these hidden variable values.
In the E-step, the expected value of the complete-data log-likelihood $\log P\left(\boldsymbol{Z}, \boldsymbol{v}|\boldsymbol{\Omega}\right)$ is computed with respect to the hidden $\boldsymbol{v}$ given the observed data $\boldsymbol{Z}$ and the current parameter estimates $\boldsymbol{\Omega}^{(\ell)}$:

$$Q(\boldsymbol{\Omega}, \boldsymbol{\Omega}^{(\ell)}) = \mathcal{E}_{\boldsymbol{v}}\{\log p\left(\boldsymbol{Z}, \boldsymbol{v}|\boldsymbol{\Omega}\right)|\boldsymbol{Z}, \boldsymbol{\Omega}^{(\ell)}\} \tag{2.33}$$

As $\boldsymbol{v}$ is a random variable with the p.d.f. $f(\boldsymbol{v}|\boldsymbol{Z}, \boldsymbol{\Omega}^{(\ell)})$, the expectation can be computed as:

$$\mathcal{E}_{\boldsymbol{v}}\{\log P\left(\boldsymbol{Z}, \boldsymbol{v}|\boldsymbol{\Omega}\right)\Big|\boldsymbol{Z}, \boldsymbol{\Omega}^{(\ell)}\} = \int_{\boldsymbol{v}} \log P\left(\boldsymbol{Z}, \boldsymbol{v}|\boldsymbol{\Omega}\right) f\left(\boldsymbol{v}|\boldsymbol{Z}, \boldsymbol{\Omega}^{(\ell)}\right) d\boldsymbol{v} \tag{2.34}$$

In the M-step, the model parameters $\boldsymbol{\Omega}$ are changed in order to maximize the expected value of the complete-data log-likelihood.

$$\boldsymbol{\Omega}^{(\ell+1)} = \underset{\boldsymbol{\Omega}}{\operatorname{argmax}}\, Q\left(\boldsymbol{\Omega}, \boldsymbol{\Omega}^{(\ell)}\right) \tag{2.35}$$

Each iteration of the E- and M-step increases the likelihood.

$$\mathcal{L}\left(\boldsymbol{\Omega}^{(\ell+1)}|\boldsymbol{Z}\right) \geq \mathcal{L}\left(\boldsymbol{\Omega}^{(\ell)}|\boldsymbol{Z}\right). \tag{2.36}$$

Thus, the algorithm converges to a stationary point or local maximum of $\mathcal{L}$ [Wu83]. This also holds true when the E- and M-steps are executed in an incremental fashion, i.e., updating per sample and not in batch over the full observed data, c.f. [NH93].

2.3.3 *Application to mixture of Gaussians estimation*

Given this framework, it is possible to derive the required equations for a MoG [Bil98]. The MoG consists of a fixed number of C Gaussians with index $c = 1...C$ that are combined linearly. The contribution of each Gaussian is expressed by its prior

$$P(\Omega_c) = \eta_c \quad \text{with} \quad \sum_{c=1}^{C} \eta_c = 1. \tag{2.37}$$

Assuming normal distributions without covariance, the parameters Ω_c for the Gaussian with index c are the mean $\boldsymbol{\mu}_c$ and variance $\boldsymbol{\sigma}_c$ and the mixture weights η_c. The full set of parameters searched for in this case is

$$\boldsymbol{\Omega} = \{\Omega_1, \ldots \Omega_C\} \quad \text{with} \quad \Omega_c = \{\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c, \eta_c\} \tag{2.38}$$

The model based probability for a observed sample $\mathbf{z}_n$ is

$$P\left(\mathbf{z}_n|\boldsymbol{\Omega}\right) = \sum_{c=1}^{C} \eta_c \mathcal{N}\left(\mathbf{z}_n|\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c\right). \tag{2.39}$$

The observed or *incomplete-data log-likelihood* in this case becomes:

$$\log \mathcal{L}(\boldsymbol{\Omega}|\boldsymbol{Z}) = \sum_{n=1}^{N} \log\left(\sum_{c=1}^{C} \eta_c \mathcal{N}\left(\mathbf{z}_n|\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c\right)\right) \tag{2.40}$$

E-STEP We suppose unobserved hard assignment of data points to the mixture components. The hidden data v is defined as the selection of mixture components for each observed data point $\mathbf{z}_n$. So $\boldsymbol{v} = \{v_1, \dots v_N\}$ where $v_n \in \{1, \dots, C\}$ assigns $\mathbf{z}_n$ to the mixture component with that index. In order to compute the distribution of the hidden data, choose a fixed set of the parameters $\boldsymbol{\Omega}^{(\ell)}$. The mixture weights can be seen as priors $\eta_c = P(c)$ for the selection of the mixture component. With this, Bayes's rule can be applied (2.41) followed by the law of total probability (2.42) to obtain

$$P\left(\Omega_{c=v_n}|\mathbf{z}_n, \boldsymbol{\Omega}^{(\ell)}\right) = \frac{P\left(\mathbf{z}_n|\Omega_c, \boldsymbol{\Omega}^{(\ell)}\right)}{P\left(\mathbf{z}_n|\boldsymbol{\Omega}^{(\ell)}\right)} \tag{2.41}$$

$$= \frac{P\left(\mathbf{z}_n|\Omega_c, \boldsymbol{\Omega}^{(\ell)}\right)}{\sum_{c'} P\left(\Omega_c|\boldsymbol{\Omega}^{(\ell)}\right) P\left(\mathbf{z}_n|\Omega_c, \boldsymbol{\Omega}^{(\ell)}\right)} \tag{2.42}$$

$$= \frac{\eta_c \mathcal{N}\left(\mathbf{z}_n|\boldsymbol{\mu}_c^{(\ell)}, \sigma_c^{(\ell)}\right)}{\sum_{c'} \eta_{c'} \mathcal{N}\left(\mathbf{z}_n|\boldsymbol{\mu}_{c'}^{(\ell)}, \sigma_{c'}^{(\ell)}\right)}. \tag{2.43}$$

Therefore, the E-Step is computing the log-likelihood of the current model as

$$\log \mathcal{L}\left(\boldsymbol{\Omega}^{(\ell)}|\boldsymbol{Z}\right) = \sum_{n=1}^{N} \log \sum_{c=1}^{C} \eta_c \mathcal{N}\left(\mathbf{z}_n|\boldsymbol{\mu}_c^{(\ell)}, \sigma_c^{(\ell)}\right). \tag{2.44}$$

Thus it is sufficient to evaluate (2.41).

M-STEP Equation (2.41) allows to express the probability of a fixed instance $\boldsymbol{v}$ of hidden data as

$$P\left(\boldsymbol{v}|\boldsymbol{Z}, \boldsymbol{\Omega}^{(\ell)}\right) = \prod_{n=1}^{N} P\left(v_n|\mathbf{z}_n, \boldsymbol{\Omega}^{(\ell)}\right). \tag{2.45}$$

The *complete-data log-likelihood* (2.33) becomes a sum over all choices $\mathbf{Y}$ of $\boldsymbol{v}$

$$Q(\boldsymbol{\Omega}, \boldsymbol{\Omega}^{(\ell)}) = \sum_{\boldsymbol{v} \in \mathbf{Y}} \log\left(\mathcal{L}(\boldsymbol{\Omega}|\boldsymbol{Z}, \boldsymbol{v})\right) P\left(\boldsymbol{v}|\boldsymbol{Z}, \boldsymbol{\Omega}^{(\ell)}\right) \tag{2.46}$$

which can be expressed as (cf. [Bil98, p. 4] for the derivation)

$$Q(\boldsymbol{\Omega}, \boldsymbol{\Omega}^{(\ell)}) = \sum_{n=1}^{N} \sum_{c=1}^{C} P(\Omega_c|\mathbf{z}_n, \boldsymbol{\Omega}^{(\ell)}) \log\left(\eta_c P\left(\mathbf{z}_n|\Omega_c\right)\right) \tag{2.47}$$

$$= \sum_{n=1}^{N} \sum_{c=1}^{C} P\left(\Omega_c|\mathbf{z}_n, \boldsymbol{\Omega}^{(\ell)}\right) \log \eta_c + \sum_{n=1}^{N} \sum_{c=1}^{C} P\left(\Omega_c|\mathbf{z}_n, \boldsymbol{\Omega}^{(\ell)}\right) \log \mathcal{N}\left(\mathbf{z}_n|\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c\right). \tag{2.48}$$

As (2.48) is a sum of positive expressions, both sides can be maximized independently in order to maximize the likelihood in the M-step. This takes some lengthy derivation using Lagrange multipliers, which can be found in, e.g., [Bil98, pp. 5-7]. By maximizing the left side of the sum in (2.48), the new mixture weights are given:

$$\eta_c^{(\ell+1)} = \frac{1}{N} \sum_{n=1}^{N} P\left(\Omega_c | \mathbf{z}_n, \boldsymbol{\Omega}^{(\ell)}\right) \ . \tag{2.49}$$

Then, (2.48) is used to find the new parameters according to (2.35). By maximizing the right side of the sum in (2.48), the parameters of the individual Gaussians are computed. Using these, the new means are given as

$$\boldsymbol{\mu}_c^{(\ell+1)} = \sum_{n=1}^{N} \underbrace{\frac{P\left(c | \mathbf{z}_n, \boldsymbol{\Omega}^{(\ell)}\right)}{\sum_{n'=1}^{N} P\left(c | \mathbf{z}_{n'}, \boldsymbol{\Omega}^{(\ell)}\right)}}_{\rho_c(\mathbf{z}_n)} \mathbf{z}_n . \tag{2.50}$$

and the variance is re-estimated using the new means

$$\boldsymbol{\sigma}_c^{(\ell+1)} = \sqrt{\sum_{n=1}^{N} \underbrace{\frac{P\left(c | \mathbf{z}_n, \boldsymbol{\Omega}^{(\ell)}\right)}{\sum_{n'=1}^{N} P\left(c | \mathbf{z}_{n'}, \boldsymbol{\Omega}^{(\ell)}\right)}}_{\rho_c(\mathbf{z}_n)} \left(\mathbf{z}_n - \boldsymbol{\mu}_c\right)^2} \ . \tag{2.51}$$

As both equations contain a term that expresses the contribution of each mixture to each sample, contribution weights $\rho_c(\mathbf{z}_n)$ can be precomputed.

TERMINATION Both the E- and M-step are ideally repeated until the distribution parameters converge. In practice, the relative change of the likelihood can be computed as

$$\Delta\mathcal{L}^{(\ell+1)} = \frac{\log \mathcal{L}\left(\boldsymbol{\Omega}^{(\ell+1)} | \boldsymbol{Z}\right) - \log \mathcal{L}\left(\boldsymbol{\Omega}^{(\ell)} | \boldsymbol{Z}\right)}{\log \mathcal{L}\left(\boldsymbol{\Omega}^{(\ell+1)} | \boldsymbol{Z}\right)} . \tag{2.52}$$

The loop is terminated when the log-likelihood does no longer change more that a threshold, $\Delta\mathcal{L}^{(\ell+1)} < \Delta\mathcal{L}_{\text{min}}$, or a maximum number of iterations is reached, $\ell = \ell_{\text{max}}$.

2.4 SPEECH ENHANCEMENT

One of the most important applications that can benefit from the methods developed in this thesis is the enhancement of speech signals. A multitude methods exist for speech enhancement using multiple microphones. An in-depth description is beyond the scope of this work. More details can be found in survey articles [VB88, GVMGO17] or books dedicated on this topic [BMC05, BW01]. The relevant fundamental principles will be introduced in this section, including their realizations in ASNs. First three basic approaches will be described. Then the different control mechanisms required for practical application will be discussed.

2.4.1 *Approaches*

The speech enhancement methods employ a variety of optimization criteria in order to derive different types of filters. From a constructive standpoint, it is possible to distinguish between three basic types: Data-independent beamforming, data-dependent beamforming [VB88] and blind source separation (BSS) [MLS07], cf. Figure 2.7.

Data-independent beamforming

One basic, but yet robust, type are data-independent beamformers, namely delay-and-sum or filter-and-sum beamformers. The principle of the former is to delay the signal of each microphone with a fixed time delay in order to compensate the TDoAs in the direction of the source [DM03, PT13]. The delay-and-sum method is very robust as the signals are only shifted in time. The gain in signal-to-noise ratio (SNR) is achieved by the fact that the desired signal in the source direction is unchanged while other directions are mitigated by uncorrelated combination. The filter-and-sum beamformer is an extension designed for multipath environments, i.e., reverberant enclosures. The time delays are replaced by a more general matched filter [JF95].

It was shown that a generalized delay-and-sum beamformer can be applied in ASNs using asynchronous communication. A distributed solution converges to the centralized one over a number of communication iterations [ZH14].

Data-dependent beamforming

Better enhancement can be gained by the so-called data-dependent beamformers. One such spatial filter is the minimum variance distortionless response (MVDR) beamformer that steers a "beam" towards the desired source while minimizing sounds from all other directions [Ows85, VB95, HBC$^+$10]. This can be split in two parallel processing paths in the well-established generalized sidelobe canceler (GSC) implementation [GJ82]: A fixed beamformer (FBF) focusing on the source and a blocking matrix (BM) that blocks it and provides noise reference signals to the subsequent adaptive noise canceler (ANC), cf. Figure 2.7 middle.

In an ASN realization based on message passing, the MVDR principle was applied [HZH$^+$12]. By setting a trade-off parameter, it is possible to increase convergence speed at the cost of reduced performance. For the fastest convergence, only the delay-and-sum solution as in [ZH14] is reached.

The linearly constrained minimum variance (LCMV) is an extension of the MVDR as the constraint of a constant gain for the desired source is combined with further linear constraints that maximize or minimize the response for other directions [EC83, VB88]. In an ASN realization it was shown that transmission of a compressed signal from each

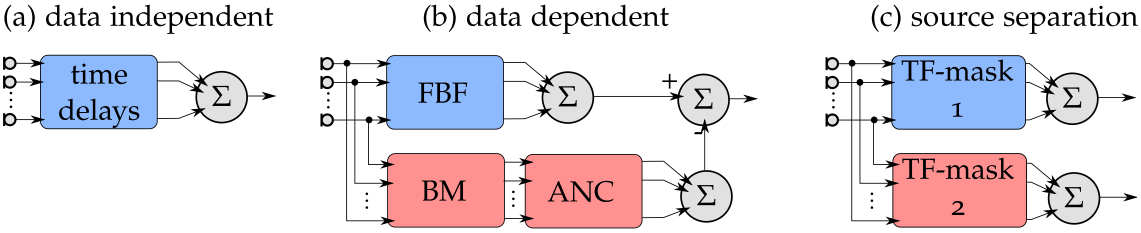

Figure 2.7: Types of multi-microphone speech enhancement: (a) data-independent beamforming using time delays, (b) data-dependent adaptive optimization in a GSC structure and (c) source separation based on time-frequency masks.

node is sufficient to achieve the same performance as the centralized solution where all signals are available [BM12].

For both MVDR and LCMV, the fixed response for the desired source is often a delay-only steering vector, thus only using the direct path. In reverberant enclosures, the room impulse response (RIR) consists of many reflections. Hence, methods using an estimate of the entire RIR provide better speech quality [GBW01, MGGC09].

Spatially based filters can not fully mitigate diffuse noise as it will be part of the signal from the look direction. Speech can be further enhanced by a time-varying spectral filter. A post filter can be added to the beamformer that suppresses noise while keeping the time-varying speech signal unchanged. Either adding a single-channel Wiener post-filter to an MVDR beamformer or a multichannel Wiener filter can achieve an optimal solution with respect to the minimum mean square error (MMSE) criterion. More advanced approaches incorporate perceptual measures in order to improve the intelligibility, e.g. the speech distortion weighted multi channel Wiener filter [DSWM07].

Blind source separation

Another family of beamformers is based on BSS concepts that aim at imposing independence between the sources [BAK04, MLS07]. The contribution of individual sources in the received signal is estimated blindly without information on the source signal. Arising problems are the unknown gain and association of time-frequency components for each source, referred to as scaling and permutation ambiguity [PLKP08]. High reverberation limits the performance of these approaches, as the reflections will be found in many time-frequency bins not related to the direct path of a source [AMM$^+$03].

One type of solution is the use of independent component analysis (ICA) that uses the assumption that the signals are statistically independent. This is often implemented by using higher than second order statistics [NK11].

Another type only employs the weaker sparsity assumption. It states that it is highly likely that each time-frequency bin of the STFT of the signals is dominated by the signal of at most one speaker. Thus it is estimated which time-frequency bins belong to which source and the signal is decomposed accordingly. This can be obtained by subspace decomposition with additional assumptions like the non-Gaussianity of the sources [PS03].

2.4.2 *Control mechanisms*

Applying any of the methods described blindly without control information is of limited practical use. The delay-and-sum method requires the microphone array to point at the desired source if no estimate of the direction is provided. The data-dependent methods require information about the activity of the desired source in order to estimate the filters. The BSS methods benefit from additional information that allows to distinguish between sources.

Speech activity

One fundamental control information is the activity of speakers in the time domain. This can be used in the data-dependent approach. The different paths of the GSC can be adapted in a straightforward manner [GBW01]. When the desired source is active, the FBF focusing on it can be estimated. The BM is given by the orthogonal subspace. When the source is inactive, the ANC is updated from the signal. Leakage of speech in the blocking signal produces speech distortion. This can be mitigated by adding speech signal information to the ANC estimator [DSWM07], which also requires a reliable speech detection.

Classification-based approaches have shown to be more robust than simpler measures for the detection of speech in noise [BKA10]. For this, acoustic event detection methods can be used. Despite recent progress, overlapping events, especially speech with noise, remain a challenge for them [SGB^{+}15]. An event detection based approach will be presented in chapter 3 (pp. 37–38).

Speaker position

The direction of the desired source can be utilized in delay-and-sum beamfomers by setting the time delays accordingly [PT13]. In both MVDR and LCMV beamformers, the DoA can be used as a constraint for the filter estimation [TTH14]. In an ASN comprised of nodes with small microphone arrays, Euclidean localization is possible based on triangulation of DoAs, cf. chapter 4. This can be used to derive localized filters in an ASN [TH13]. Directional information can also be incorporated into BSS beamformers in order to resolve the permutation ambiguity [PA02, SMAM04, NOS08, RMGB^{+}13].

Time-frequency masks

The time-frequency mask for each speaker can be used to separate speakers. After estimation of this mask, it is applied to the signal in the STFT domain. This is a common concept in both BSS [PLKP08] and CASA [WB06].

CASA approaches often target the difficult problem of separating speakers in a single-channel signal by means of ASA cues and top-down models [JW09, HW10]. Binaural realizations make use of spatial cues [RWB03].

Multi-microphone BSS methods often employ straightforward ML clustering based on location cues, e.g., the phase difference between microphone pairs [ANS10]. Rather than hard assignment of bins to speakers, soft masks using fractional values achieve better results in practice. After localizing speakers probabilistically, such a soft mask can be inferred from the probability of each time-frequency bin belonging to a given source [MM11]. A simple but effective ASN realization chooses the closest node after localization and employs a soft mask derived in this way to its signals [DCG15].

Time-frequency masking can induce some distortion to the signals. Madhu et al. [MM11] showed that a GSC realization of an LCMV beamformer driven by the estimated mask achieves good enhancement with less signal degradation than using the mask directly. Another interesting hybrid approach combines ML mask estimation with a multichannel Wiener filter in a joint algorithm in order to minimize distortion [AN11].

One person's data is another person's noise. K.C. Cole

3 ACOUSTIC EVENT DETECTION

The question what is happening in a given environment by acoustic means is addressed by two similar tasks, acoustic scene classification and acoustic event detection. Acoustic scene classification is answering the question what the overall scene is, i.e., if the recording takes place in an office, a bus, or on the street. Acoustic event detection answers the question which events are happening in particular and at what time, i.e., a person speaking, a printer working, a phone ringing, or a siren howling. The latter task will be addressed in this work. The objective is to develop an online method that is robust and implemented in real-time while providing state-of-the-art results. It should be able to generalize well from limited training data.

Within this thesis, it is important to have a method to distinguish speech from other sounds. This is required for the other methods developed in this thesis in the following ways: For speaker tracking, non-speech sounds should be excluded in order to allow a correct estimation of the speakers activity and position. Even more, the method developed here can help distinguish different speakers. For geometry calibration, it is vital to exclude sounds not transmitted solely by air, such as chairs moving or footsteps, since the distance measurements derived from sound are assuming the propagation to be with the speed of sound in air. For speech enhancement, reliable control information on whether the signal contains speech or certain kinds of noise is required.

The classification is difficult because of the diversity of the acoustic events. Human speech is comprised of sounds of different phone classes, i.e. vowels, plosives and fricatives that have individual spectrum and time characteristics. Other sound types are also complex because they are comprised of a variety of individual sounds, e.g. chair movement can produce knocking and rubbing sounds, handling paper can include rustling and knocking on the table and so on. Sounds like footsteps are individually different depending on the person and kind of shoes. It is desirable for a sound classification method to be able to handle the diverse composition and generalize in a way to cover different, possibly unheard realizations of the sound types.

In this chapter, first the state-of-the-art will be introduced. The different types of features and classifiers will be explained and the existing methods will be introduced according to these two aspects. Thereafter, the proposed method for acoustic event classification and detection be explained. Finally, the method providing speech and noise type detection as control information for a beamformer is described.

3.1 STATE-OF-THE-ART

Over the last decades, a large number of different approaches for acoustic event detection have been proposed [TMZ+07, MHEV10, SGB+15]. They are used in a large variety

of applications, from security and surveillance [SSKP16, MK16, CFP+13, YS01] to urban planning and wildlife monitoring [SKG13, KSJM09, ZSB13].
Two main aspects that these differ in are the features used and the classification scheme. In the following, first the different features and then the range of classification approaches will be introduced on examples of state-of-the-art methods.

3.1.1 *Features*

For sound and especially speech processing, the mel frequency cepstral coefficients (MFCCs) are one of the most widely used features. Along with MFCCs, a variety of technical features such as zero crossing rate, and linear prediction coefficients (LPCs) are used. Some of researchers tend to include features more oriented towards human perception, such as the "perceptual" feature set introduced by Temko et al. [TN06].
Since considerable progress has been made by applying insights from human perception in the field of computer or machine vision, similar approaches have been advocated for acoustics [Lyo10]. Two recent approaches show such an application. One uses auditory images computed from the output of an auditory filterbank over time. Another uses an MFCC like feature based on Gammatone filterbanks [SSW07].

Mel frequency cepstral coefficients

The MFCCs are the most common and successful features in speech recognition. To compute MFCCs, the input signal is filtered by a mel frequency filterbank, from the logarithm of its magnitude the discrete cosine transform (DCT) is computed. Typically the second to 13th coefficient is used [HAH01]. The first and second temporal derivative is often included to capture transient features, resulting in a de-facto standard 39 dimensional feature vector in speech recognition systems.

Prediction coefficients

Although originally designed for speaker independent speech recognition, the MFCCs are also commonly used in speaker identification [KL10, TP11]. The LPCs or perceptual linear prediction (PLP) coefficients are theoretically more speaker dependent, as they approximate the vocal tract. But spectral features as the MFCCs have become very popular in speaker identification in the last two decades. Some approaches combine both PLP and MFCC [TCHJH12]. One recent strategy is to use not the Gaussian mixture model (GMM) or hidden Markov model (HMM) classification of the MFCCs, but rather build a universal speaker model referred to as universal background model (UBM), and look at the differences after re-training the model with the data in question. The differences in model parameters are concatenated into a large feature vector that is then used for classification of the speaker [KL10, TP11].

Perceptual features

Nadeu et al. introduced filter-band energies as an alternative for speech recognition. The design allowed for decorrelation and equalization of the variance [NHG95]. Temko et al. [TN06] combined the frequency band features with zero-crossing rate, short time energy, four sub-band energies, spectral flux, calculated for each of the defined subbands, spectral centroid, spectral bandwidth and pitch. The so called 'perceptual' feature set has been used in several approaches [TN09, PMMM14].

Gammatone features

The long history of psychoacoustic research has been complemented by computational modeling of the human hearing process [WB06] where ERB-spaced Gammatone filterbanks are used (see section 2.2.2 on page 17). From that the Gammatone frequency cepstral coefficients (GFCCs) were derived [SSW07]. First, a time-frequency representation is computed by a filterbank composed of 64/128 Gammatone filters between 50 and 4,000/8,000 Hz, respectively. The cubic root of the magnitude of the filter outputs is used as Gammatone feature over 10 ms frames. It should be noted that the term "cepstral" is used here even though no log operation is performed, just cubic compression. The GFCCs were shown to be more robust against noise than MFCCs in the task of speaker identification.

By combining GFCCs with the estimation of a time-frequency mask, Zhao et al. [ZSW12] constructed a speaker identification system with improved noise robustness. In parallel to the Gammatone features, a computational auditory scene analysis (CASA) based binary time-frequency mask of speech presence is computed. Both marginalization and reconstruction are applied in parallel to cope with noisy and missing time-frequency bins. The reconstruction module uses a universal speech model to estimate the missing Gammatone features. Thereafter, the DCT is applied in order to compute 22 GFCCs. A GMM is used to identify the speaker based on these features. The bounded marginalization is computed in the spectral domain on the Gammatone features. The resulting spectral features are also classified by a GMM. The reconstruction only works better than marginalization for lower noise conditions of 12 dB signal-to-noise ratio (SNR) or more. However, the combination of the two classifiers is better than reconstruction alone, even at -6 dB SNR.

Recently, the combination of GFCCs and MFCCs for sound recognition was investigated. A scream detection system proposed by Lei et al. [LM14] uses both MFCC and GFCC features. First, high energy segments are extracted from the input signal. Then, 12 MFCC and GFCC coefficients are computed along with their first and second derivatives. The resulting 72 dimensional feature vector is then reduced to 36 dimension by regularized principal component analysis (PCA). The mean and standard deviation over several time slices of equal length are computed. These values are then classified as scream or non-scream by a support vector machine (SVM). The evaluation shows a decrease in error for the combination of MFCC and GFCC features, although the GFCCs perform worse than the MFCCs, the combination outperforms each feature used individually.

Auditory and spectral images

When the output of a cochlear filterbank is captured over a number of consecutive time frames, and the energy of each band at each frame is color coded, the result is a cochleogram similar to a spectrogram. As the resulting patterns exhibit periodic time variations with pitch, they are sometimes stabilized by setting a trigger point and centering the sliding window on that. These so called stabilized auditory images (SAI) [PRH+92] can be used with image processing techniques.

Rehn et al. used such images for text-based retrival of sounds [RLB+09]. A large sound database composed from over eight thousand sound files was used. About half the files were from a commercial sound effects database, the others were collected from various websites. For training and evaluation, the sound files were tagged with keywords. The user was entering a keyword and presented with an ordered list of sound files, the retrival set.

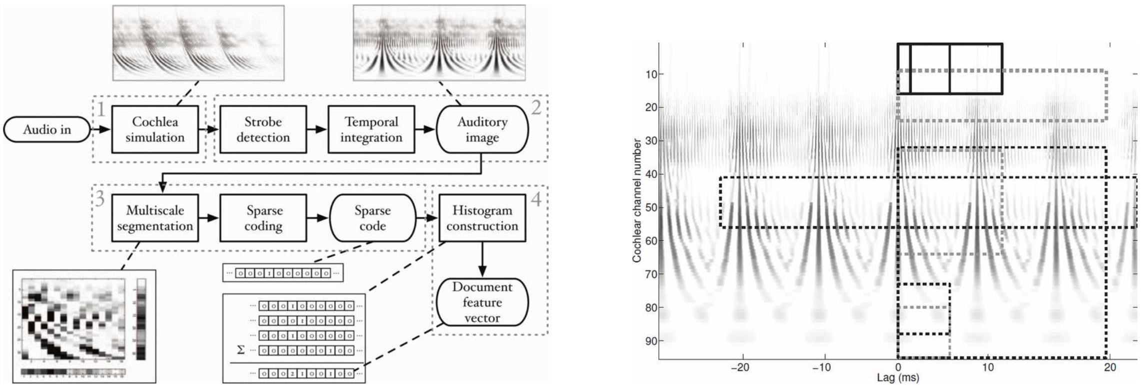

Figure 3.1: Stabilized auditory image features: Computation pipeline (left) and areas used for vector quantization (right). Halfway rectified band signals in multiple Gammatone filtered bands form a cochleogram over time. The image is stabilized by setting a trigger point based on the envelope. The cochleogram is used as two dimensional image, in which different areas are selected. Each area is quantized and the concatenated codebook entries provide the final feature vector used in classification or retrieval. Images from [LPC11], © 2011 IEEE.

Two methods for computation of the retrival set were used. First, standard MFCC features along with the first and second derivative were computed. These were clustered using a large unsupervised codebook with eight thousand centroids. Second, a SAI was computed. The image was encoded by a sparse coding technique and the codes were summed over time. Both features were compared in retrival performance using the PAMIR [GB08] retrival system. In a three fold cross-validation, the auditory features outperformed the MFCCs with 70% vs. 60% precision for the top 1-5 ranked sounds, i.e., the top elements of the retrival set.

The system was used for acoustic event classification by McLoughlin et al. [MZX$^+$15]. The task was to classify events with and without added noise as described in [DTC13]. Sound samples of 50 different classes were taken from a sound scene database. Training was done with the clean signals only. The test was done on the clean signals and mixed with noises at 0, 10, and 20 dB SNR. As the use of vector quantization diminished the performance, this step of the PAMIR system was omitted. The sounds were classified using a SVM with a linear kernel. A total of ten 35 ms frames was used, which improved the results compared to using only a single frame. Another small improvement in performance in noise conditions was achieved by replacing the classifier with a deep neural network (DNN). Compared to using MFCC features with the SVM, the SAI shows slightly better performance, especially in low SNR. Similarly, the SAI with temporal context and the DNN slightly outperform an MFCC-HMM baseline system.

The specially constructed spectral image features (SIF) were introduced by Dennis et al. [DTL11]. A grayscale spectrogram image is mapped into several monochrome images for different dynamic ranges. These are partitioned into small blocks, whose central moments are computed and concatenated into a feature vector. On the task described above [DTC13], these features used with a DNN outperform both the MFCC-HMM and SAI-DNN by far in noisy conditions. With the additional use of a denoising algorithm and energy weighting, the DNN with SIF achieves 85% to 95% accuracy over different noise levels.

3.1.2 *Classifiers*

The classifiers used for acoustic event detection span a range of pattern recognition methods. One of the first are GMMs and HMM approaches that were originally used for speaker and speech recognition. These are still common, but today general purpose classifiers such as SVMs and random forests are also applied successfully. In the last years, using two-stage approaches and bag-of-features (BoF) methods have been shown to be advantageous over single-stage classification. Recently, DNNs have been applied. These can also be considered multi-stage systems that perform automatic feature computation and classification.

Gaussian mixture models

A common approach to model different sounds is to use a set of GMMs that are individually trained for each class. The GMMs scores are summed over all frames and the class with the highest likelihood is chosen. Since the summation discards any temporal information, the method is sometimes termed "Bag-of-Frames" [ADP07, GSB$^+$13]. For speaker identification, this approach has been used with MFCCs features successfully for decades [Rey95]. One common extension is the use of differences to an universal background model (UBM) [TP11]. The task of speaker diarization is determining "who spoke when" with no prior speaker models. In many diarization applications, a clustering algorithm determines the different speakers using individual GMMs or their differences to an UBM [MBE$^+$12]. Given its continued success in speaker classification, it is not surprising that the combination of MFCC features with GMMs is also used in acoustic event [ADP07, VBK$^+$13] and scene classification [BGSP15].

Hidden Markov models

Hidden Markov models (HMMs) are the most widely used classifier in speech and handwriting recognition [Fin14]. They are also applied for speaker identification and diarization [MBE$^+$12] and acoustic event detection [SMS$^+$13]. Their main advantage is the inherent modeling of dynamic temporal alignment. When Gaussian mixtures are used to model the emissions, the HMM is a generalization of the GMM to multiple temporal states with different models of the distribution of the features over time. There are several architectures employed for acoustic event detection. The most basic architecture is to use a fixed number of states for each event in a sub-model, and then to connect them as parallel alternatives. The detected events are then determined by the Viterbi path going through the states of the corresponding sub-model. This is the same architecture as a typical HMM for word recognition. One recent application of this approach was used by Diment et al. [DHV13] in the 2013 D-CASE challenge. They constructed a HMM using sub-models with three states per class and an additional one state submodel for the background class. The sub models were using a mixture of eight Gaussians to model the MFCC features. In order to detect overlapping events, multiple passes of the Viterbi algorithm are computed. After each pass, the used states are forbidden to enter again, so the next pass is forced to estimate a different event.

Deep neural networks

In recent years, DNNs have been able to outperform the state-of-the-art in pattern recognition tasks [LBH15]. One of the key applications that showed advantages of the nonlinear manifold learning ability of DNNs was the use for output modeling in speech recognition [MDH11, HDY$^+$12]. Gaussian-Bernoulli restricted Bolzmann machines (RBMs)

are pre-trained as generative models for spectrograms and then later re-trained together with the HMM. It was shown that an increase of performance is achieved by replacing both the MFCC features and the GMM modeling with a neural network.
So it is not surprising that DNNs are also applied for acoustic event recognition. By applying a moving window as input, a DNN classifier can be used for online recognition. As typically one-of-k coding is used in the output layer, concurrent events may be trained and recognized as well. One of the remaining disadvantages of the DNN approaches is the considerable training time of days or weeks, even when implemented on graphics processing units (GPUs) that provide large parallel processing power. Another disadvantage is the need for a large set of training material. When the amount of data is limited, as is quite reasonable in real life scenarios, the DNN approaches may be outperformed by the much simpler GMM as this is able to generalize better from limited data [KSWP16, SAG16].
One approach is to use a deep belief like architecture of stacked RBMs to learn a representation from feature data. The different RBM layers can be trained individually in order to avoid vanishing gradients in back-propagation. The input layer is often a Gaussian-Bernoulli RBM as in speech recognition applications. Subsequent layers can be Bernoulli-Bernoulli. The output layer can be a one-of-k coding layer, encoding each possible event class with a single output neuron. This architecture was used by McLoughlin et al. [MZX+15] on the task already described in the stabilized auditory images (SAI) section 3.1.1. When using MFCC features, the DNN outperforms the SVM but not the HMM for event detection in noise. It is likely that the abstraction done in the feature computation is too strong for the Gaussian-Bernoulli RBM in the DNN configuration to learn a robust representation. When using the SAI features, the DNN performs slightly better than the HMM. The spectral image features (SIF) are more suited as they allow the DNN to learn a feature representation from the preprocessed spectrogram, which results in notably better performance in noisy conditions.
Rather than using stacked RBMs, it is also possible to use a DNN composed of several fully connected layers. The rectified linear unit (ReLU) activation function reduces the vanishing gradient problem and benefits the training speed. Typically the input and hidden layers are used with dropout, meaning that randomly selected neurons are omitted in the training stage [SHK+14]. The last fully connected layer applies one-of-k coding and is followed by a softmax layer for classification. Hertel et al. [HPM16] used such an architecture for acoustic event classification. In order to allow the DNN to learn a feature representation, the raw audio data or the spectrum was used as input to the network. Interestingly enough, the DNN was able to correctly classify most events even from time domain data. The spectrum data performed better. Kong et al. [KSWP16] used a similar structure with mel band energies as input feature. Rather than using a softmax layer, they used a sigmoid function in the one-of-k coding. By setting a threshold for the minimum output activation, this network is able to detect concurrent events.
A common architecture in image processing are the convolutional neural networks (CNNs). Here, a small filter is applied to an input layer by computing the weighted sum of filter coefficients and adjacent input values. As the filter is learned but fixed for the whole input, this performs a two dimensional convolution. After such a convolutional layer, a subsampling is performed by a max-pooling layer that computes the maximum value of a small input region. By stacking pairs of convolutional and pooling layers, the size of the layers is subsequently reduced. After this, several fully connected layers with dropout are used. The final fully connected layer again implements one-of-k

coding followed by a softmax decision. This type of architecture was also used by Hertel et al. [HPM16] for acoustic event classification. In comparison to the DNN architecture described before, better results were achieved. The successive reduction in layer size benefits the abstraction and provides a level of shift invariance. Using the spectrum as input, the performance was state-of-the-art.

Bag of features

The bag-of-features (BoF) approach originated in text retrieval [BYRN99]. Here, histograms of word occurrence are used to retrieve text documents relevant to a textual query. Since the histograms are discarding any information on word order, the approach was named "bag of words". The idea was transferred to general retrieval and classification tasks, first in the field of computer vision [SZ03]. In this approach, first the feature values are clustered using vector quantization. Typically, Lloyds' algorithm [LBG80, Llo82] is used, which minimizes the quantization error by iterative re-estimation of the codebook entries as best representatives of the training data. The codebook entries are referred to as "vocabulary". Second, a histogram is computed over the number of input frames assigned the individual codebook entries. Third, these histograms are then used for classification by, e.g., a multiclass SVM. The augmentation of the second step, the quantization, has become an active field of investigation. More information is encoded by using soft quantization and supervised training [CLVZ11].

One of the first applications of BoF to acoustic event classification was the approach of Pancoast et al. [PA12]. In their implementation, a large number of MFCCs and their deltas are used along with an overall energy estimate of the time window. These are then classified by a multiclass SVM. They used a histogram intersection kernel, which was shown to outperform a pure linear classification. The histogram intersection kernel computes the vectors scalar product of two vectors $\boldsymbol{a}, \boldsymbol{b}$ by the component-wise minimum:

$$k_{\mathrm{HI}}(\boldsymbol{a}, \boldsymbol{b}) = \sum_{l=1}^{L} \min\{a_l, b_l\} \; . \tag{3.1}$$

A related idea was implemented by Phan et al. [PM14]. So called "superframes" are computed as mean and deviation of each feature over all frames in an 0.1 s time window. This is used to train a random forest classifier to recognize the event corresponding to each superframe. For pre-segmented events, the histogram of the classifications is computed and used in turn to train an SVM. The SVM performed best when using a χ^2 or histogram intersection kernel. The histogram approach was later superseded by an integrated random regression forest framework that also uses the forest to detect the event on- and offsets [PMMM14].

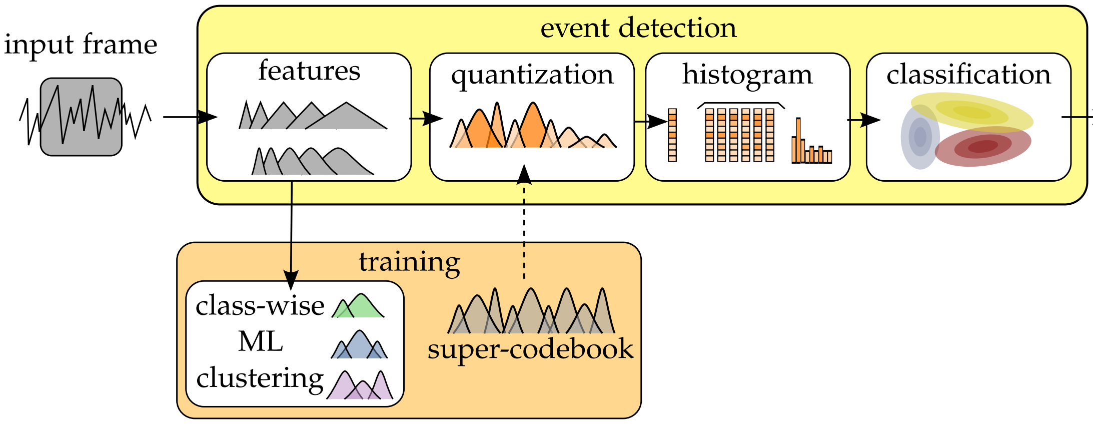

Figure 3.2: Bag of superfeatures acoustic event detection method. Both MFCC and GFCC features are calculated on a single channel input signal. In training, codebooks for each class are computed using maximum likelihood clustering. These are then concatenated into a super-codebook. When classifying an input signal, this codebook is used to compute probabilities for each Gaussian density. These probabilities are accumulated over a time window to form a histogram. The histogram is then classified using a Bayesian maximum likelihood classifier.

3.2 PROPOSED EVENT DETECTION METHOD

A refined method was devised that applies the BoF approach based on soft quantization with GMMs [PGF14]. One of the key contributions is class-wise training of codebook entries. Figure 3.2 shows the processing pipeline. First features are computed over a sliding window on the input. These features are then softly quantized by a GMM and classified by a maximum likelihood classifier. Rather than using a prior classification step to eliminate silence and background noise, as done in several systems (cf. [TMZ+07]), the rejection class Ω_0 is trained with recordings where no event occurred.

3.2.1 *Features*

A single microphone or beamformed signal is processed in short time windows of 0.6 s every 0.05 s. Within this window, short time Fourier transform (STFT) frames of $K = 1024$ samples are computed with a hop size of 512. For each frame k in this window, a feature vector $\mathbf{z}_k$ is calculated. The features for all time frames in the nth time window make up the matrix $\mathbf{Z}_n$.

Each vector $\mathbf{z}_k$ contains three types of features. First, the MFCCs coefficients are computed using the common mel filterbank of 40 filters. Second, GFCCs are computed by replacing the mel filterbank by equivalent rectangular bandwidth (ERB)-spaced linear phase Gammatone filters according to equation (2.29), as described in Section 2.2.2 on page 17, before computing the DCT on the log of the magnitudes of the filterbank output. Third, the perceptual loudness is computed on the input frame by applying an A-weighting filter[1] and subsequently computing the energy of the frame. Figure 3.3 illustrates the feature calculation process.

[1] defined in the international standard IEC 61672:2003

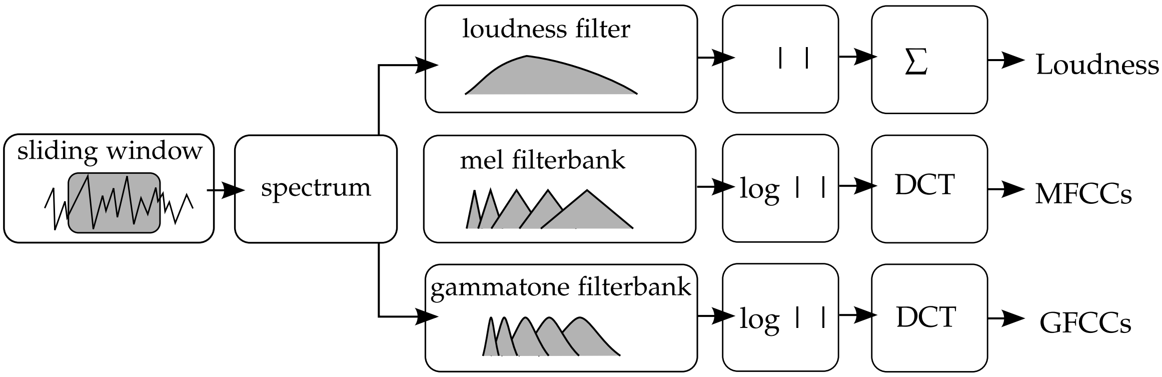

Figure 3.3: Computation of the features for acoustic event detection. The microphone signal y_i is windowed and transformed in the spectral domain with a STFT. From the magnitude spectrum, three types of features are computed. The overall perceptual loudness (top), MFCCs (middle) and GFCCs by application of a filterbank before computing the DCT of the log energies.

3.2.2 *Bag of super features*

A BoF approach is used for building a codebook of *"acoustic words"* from the training set. Unlike the common approach to estimate the codebook unsupervised, i.e., discarding class labels, the training is done in a supervised manner. Codebooks are estimated separately for all classes Ω_c and then concatenated into a large super-codebook. The expectation-maximization (EM) algorithm is applied to all feature vectors $\mathbf{z}_k$ for each class Ω_c in order to estimate I means and deviations $\mu_{i,c}, \sigma_{i,c}$, c.f. Section 2.3 on pages 20–23. The mixture weights η_c are not used, as an assignment is also estimated in training the maximum likelihood (ML) classifier. All means and deviations are concatenated into the super-codebook

$$\mathbf{v}_{l=i+cI} = (\boldsymbol{\mu}_{i,c}, \boldsymbol{\sigma}_{i,c}) \tag{3.2}$$

over all classes, therefore containing $L = I \cdot C$ means and deviations. A soft quantization of a feature vector $\mathbf{z}_k$ is computed as the soft assignment to the individual codebook entries

$$q_{k,l}(\mathbf{z}_k, \mathbf{v}_l) = \frac{\mathcal{N}(\mathbf{z}_k|\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l)}{\sum_{l'} \mathcal{N}(\mathbf{z}_k|\boldsymbol{\mu}_{l'}, \boldsymbol{\sigma}_{l'})} . \tag{3.3}$$

Then, a pseudo-histogram $\mathbf{b}$ of pseudo-frequencies can be computed over the features $\mathbf{Z}_k$ computed for all K frames of the input window by

$$b_l(\mathbf{Z}_n, \mathbf{v}_l) = \frac{1}{K} \sum_k q_{k,l}(\mathbf{z}_k, \mathbf{v}_l) . \tag{3.4}$$

This method was introduced as *"bag of super features"* [PGF14] in analogy to the supervector construct used in speaker identification (cf. [TCHJH12]).

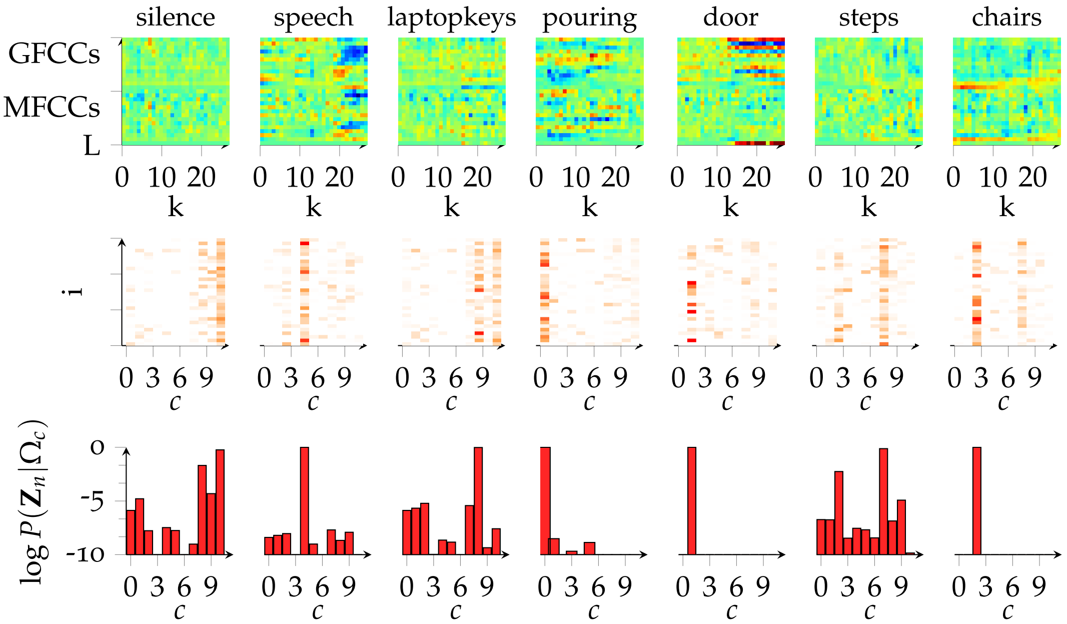

Figure 3.4: Features, GMM probabilities and ML probability (top to bottom) for several different event classes (left to right). In the top row, the color coded feature values of the MFCCs, GFCCs and the Loudness feature are shown for 30 frames. The middle row shows the GMM scores over all training classes (c, abscissa) and 30 Gaussian densities for each (i, ordinate). In the bottom row, the log scores of the Bayesian classifier for all classes are shown.

3.2.3 *Classification*

The probability of an acoustic word for a given class $P(v_l|\Omega_c)$ is estimated using a set of training features $\mathbf{Z}'_n \in \Omega_c$ for each class c. In order to handle zero-valued entries, Lidstone smoothing is used with a factor $\alpha = 0.5$:

$$P(\mathbf{v}_l|\Omega_c) = \frac{\alpha + \sum_{\mathbf{Z}'_n \in \Omega_c} b_l(\mathbf{Z}'_n, \mathbf{v}_l)}{\alpha L + \sum_{m=1}^{L} \sum_{\mathbf{Z}'_n \in \Omega_c} b_m(\mathbf{Z}'_n, \mathbf{v}_m)} \tag{3.5}$$

Since all classes are assumed to be equally likely and have the same prior, a maximum likelihood classifier is used for classification. The posterior is estimated using the relative pseudo-frequency $b_l(\mathbf{Z}_n, \mathbf{v}_l)$ of all acoustic words $\mathbf{v}_l$. By computing the product of the trained $P(\mathbf{v}_l|\Omega_c)$ likelihoods taken to the power of the pseudo-frequency $b_l(Z_n, v_l)$ for the feature vectors $\mathbf{Z}_n$ in the input window, the overall likelihood of the class is computed following a multinominal Bayesian distribution:

$$P(\mathbf{Z}_n|\Omega_c) = \prod_{l=1}^{L} P(v_l|\Omega_c)^{b_l(\mathbf{Z}_n, \mathbf{v}_l)} \ . \tag{3.6}$$

In Figure 3.4, the different stages output is shown for example input signals on seven exemplary event classes.

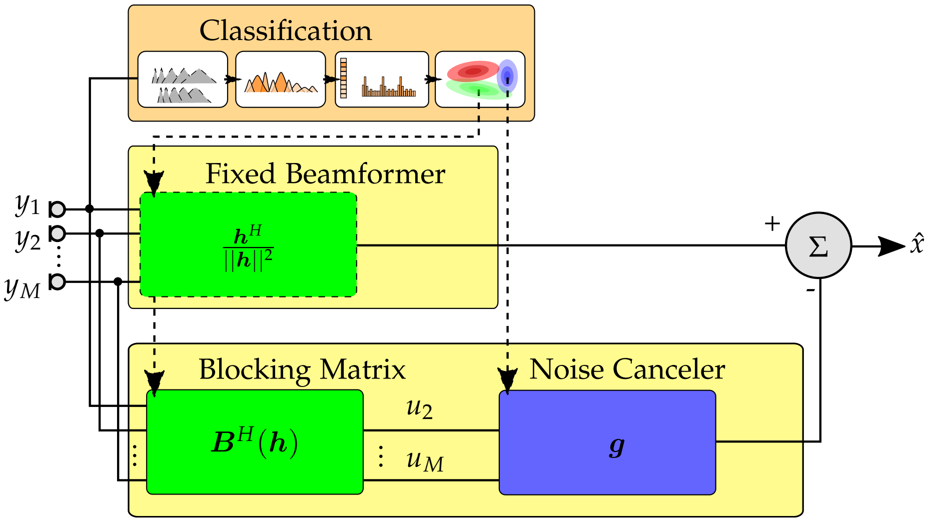

Figure 3.5: Proposed blind speech enhancement method: A classifier identifies speech or suitable noise segments that respectively update a FBF and BM or the ANC in order to estimate the clean speech signal $\hat{x}$. Illustration based on [PG16], © 2016 IEEE.

3.3 PROPOSED SPEECH DETECTION METHOD

Noise is present in many everyday situations where a smart audio device is used. Speech enhancement techniques can mitigate this noise using multiple microphones (see section 2.4 on pages 24–26). A novel method for blind speech enhancement is proposed that employs a classification based control mechanism to a beamformer.

The beamformer used here is exploiting the non-stationarity of speech in the filter estimation [GBW01]. It can provide good speech enhancement in the presence of stationary noise. It makes use of the full acoustic transfer functions (ATFs), not only the direct path, by estimating room impulse responses (RIRs) with respect to the first microphone for the filters. It is not suited to mitigate highly non-stationary noise. For this, more complex techniques have to be used [TCG11].

The classification is based on the BoF system described in the previous section. It provides control information by identifying speech and different noise types. Situations where the noise is too non-stationary for the chosen filter estimation are detected automatically. A dedicated training strategy and integration of the classification results was developed to guide the beamformers filter estimation. Figure 3.5 shows the structure of the novel blind speech enhancement system. The BoF classifier is used to classify time segments based on the single channel signal y_1 from the first microphone. This is used to estimate the components of an minimum variance distortionless response (MVDR) beamformer (see section 2.4 on pages 24–26) implemented in a generalized sidelobe canceler (GSC) structure [GBW01]. From speech segments, the relative transfer function (RTF) $\boldsymbol{h}$ is estimated as ratios of the ATFs $h_i(t, f) = a_i(t, f)/a_1(t, f)$, (cp. section 2.1.1 on pages 10–11). From this estimate, the parameters of the FBF and BM blocks are computed. Stationary noise segments are used to adapt the ANC coefficients $\boldsymbol{g}$. When non-stationary noise is detected, neither are updated [PG16].

3.3.1 *Features*

The MFCC and GFCC features are used as before. No loudness is used as this is deemed counterproductive given the unknown level of the interfering signal. The first temporal derivatives of the features are computed in addition in order to better capture the temporal characteristics hinting at the stationarity.

3.3.2 *Training*

The basic training approach of using separate isolated recordings as described before did not yield sufficient detection of speech in complex noise conditions. A dedicated training strategy was devised in order to provide the required quality in detection.

To provide a structured data basis, different types of noises are recorded. It is distinguished between types of noise based on their nonstationarity. Good results were achieved by using four different levels: The first class Ω_1 are very stationary noises such as white noise or fan sounds, the second class Ω_2 are mechanical noises, the third Ω_3 is speech-like babble noise. The fourth and final class Ω_4 is comprised of nonstationary noise like keyboard typing, for which the estimation procedure [GBW01] fails. Each of the four noise classes is trained using individual examples. Additionally, for each of them a mixture class $\Omega_1' \ldots \Omega_4'$ is trained by mixing noise types of the same level of stationarity or lower. Ω_1 is trained using mixtures of different examples for that class. Ω_2' is trained by mixing with different noise types from the categories Ω_1 and Ω_2, and so on.

In order to estimate a good representation for speech, the speech samples are mixed with samples of each of the different noise types at a high SNR of 18 dB to train the speech class Ω_0. Using lower SNRs resulted in more false alarms where noise is detected as speech.

3.3.3 *Speech detection*

The classifier is applied to a single channel input signal. Time segments classified as speech are used to estimate the FBF and BM, noise segments are used to update the ANC. In order to find the best way of applying the control, the effect of different types of classification errors have to be taken into account. The main error type to be expected is the underestimation of speech existence, especially in the transitions between speech and noise and at low SNRs.

The updating of the ANC in speech would lead to a serious deterioration of the performance, as speech would be distorted from the ANC canceling part of it. The updating of the FBF and BM is much more tolerant, as the estimation is done off-line using all segments classified as speech. Some errors in the form of a few noise time segments classified as speech will hardly change the estimation result.

In order to provide the best possible integration, a guard boundary of $d_S = 0.5\,\mathrm{s}$ around the time segments classified as speech is introduced. The ANC is only updated in noise segments that are d_S before or after the speech segments as shown in Figure 3.6.

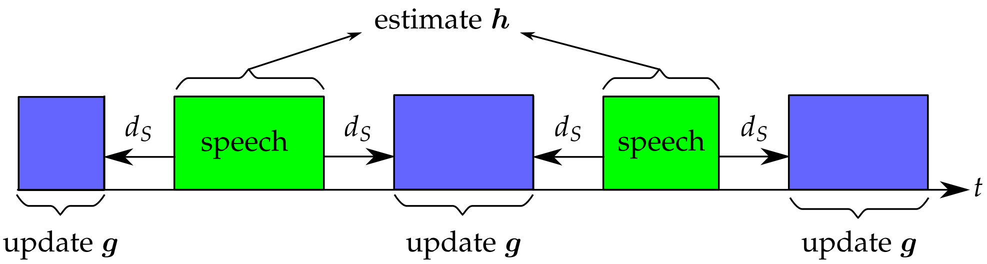

Figure 3.6: Updating strategy from classifier output. The frames classified as speech are used together to compute the fixed beamformer coefficients $\boldsymbol{h}$. The frames classified as noise farther away than a guard margin d_S from speech are used to continuously update the ANC coefficients $\boldsymbol{g}$. Illustration first used in [PG16], © 2016 IEEE.

3.4 SUMMARY

For sound classification, a variety of features and classifiers are used. Table 3.1 gives an overview of the methods described in this chapter. The classical approach of using MFCCs with competing GMM models is still used. The implementation is well established and comparatively simple. The approach shows good performance and generalization ability. For example, the approach by Vuegen et al. [VBK+13] achieved good results in the 2013 D-CASE challenge with a two stage foreground-background GMM method based on MFCC features.

Features

Besides MFCCs, auditory features have been shown recently to improve the performance. The GFCCs showed better discrimination in speaker identification [SSW07]. For event detection, they improve the performance when combined with the MFCC features [LM14]. Both SAI and SIF show promising results in conjunction with DNNs [MZX+15]. However, the computational effort of computing and handling these features is comparatively large. Given enough training data, the DNN can also learn features from raw time or spectral data [ODZ+16, HPM16].

In order to gain performance in another direction, some systems use a collection of different features. This is often combined with a classifier that is able to internally select features, e.g., a random forest. The perceptual feature set introduced by Temko et al. [TN06] is popular among these.

Classifiers

The basic GMM approach was developed for speaker identification. Nowadays, more complex approaches like UBMs and HMMs are often used.

In off-line applications for speech recognition and event detection, HMMs are widely used. While successful in speech recognition, using DNNs for output modeling does not always improve the event detection performance. This especially apparent when the amount of training data is limited [SAG16].

DNN are applicable online for detection with a sliding window. It is possible for the network to learn the feature representation by using spectral or spectrogram data as input. Detection of concurrent events is achieved by ways of a thresholded output layer or

features	quantization	classifier	online	concurrent	task	
MFCC	–	GMM	✓	–	event detection	[VBK+13]
MFCC	–	UBM	–	–	speaker identification	[TP11]
GFCC	–	GMM	–	–	speaker identification	[SSW07]
SIF	DNN	DNN	✓	–	event detection	[MZX+15]
mel energies	DNN	DNN	✓	✓	event detection	[KSWP16]
perceptual [TN06]	superframe	RF	✓	✓	event detection	[PMMM14]
MFCC, GFCC	PCA	SVM	✓	–	scream detection	[LM14]
MFCC	hard	SVM	✓	–	event detection	[PA12]
MFCC, GFCC	soft, super.	ML	✓	–	event detection	[PGF14]
MFCC, GFCC	soft, super.	ML	✓	✓	speech detection	[PG16]

Table 3.1: Overview of state-of-the-art sound classification methods with different features and classifiers

binary attribute encoding. However, they still require comparably large computational efforts and large amounts of input data, resulting in a long time for training. When the amount of training data is limited, the feature learning is not performing well. In such cases, the DNNs is outperformed by the classic MFCC GMM approach, as this is able to generalize better [KSWP16].

The BoF approach provides a practical way of implementing a two-stage classification system. It is popular due to its simplicity and good generalization ability. The first applications used hard quantization with SVM classification of the histograms [PA12]. The related superframe approach makes use of a random forest classifier. This allowed to extend the method for improved on- and offset detection [PMMM14].

Proposed method

The proposed method is a hybrid between a classical BoF approach and the GMM method, since it applies supervised training of individual GMMs for the classes before using all densities for quantization. It provides robust recognition for both acoustic events and speech in noise while being computationally efficient and requiring only a short time for training. The use of both MFCC and GFCC features improves the classification accuracy.

By a dedicated training strategy based on a hierarchy of stationarity, the detection of speech in mixtures with noise was realized. This makes the method robust against severe noises levels corrupting the speech signal. Thus it is possible to provide control information to a beamformer in order to realize blind speech enhancement.

Given the computational efficiency and robustness of the proposed method, it is possible to use the event detection as pre-filter for speaker localization and geometry calibration. It is also possible to add different classes for different speakers in order to identify them while tracking.

The difficulties that are involved in the scene analysis processes in audition often escape our notice. This example can make them more obvious: . . . your friend digs two narrow channels from the side of a lake. Each is a few feet long and a few inches wide and they are spaced a few feet apart. Halfway up each one, your friend stretches a handkerchief and fastens it to the side of the channel. As waves reach the side of the lake they travel up the channels and cause the two handkerchiefs to go into motion. You are allowed to look only at the handkerchiefs and from their motions to answer a series of questions: How many boats are there on the lake and where are they? Which is the most powerful one? Which one is closer? Is the wind blowing? Has any large object been dropped suddenly into the lake?

Albert S. Bregman: Auditory Scene Analysis (1990)

4 ACOUSTIC PERSON LOCALIZATION

The localization and tracking of speakers is one of the key applications for microphone arrays and acoustic sensor networks (ASNs). It is used for dedicated speech enhancement, camera control, and meeting annotation. In practice, all these applications require both exact speech detection of concurrent speakers and good location accuracy. The problem is important in two aspects for the geometry calibration approach described in chapter 5. First, the single node localization and speech detection is required as input. Second, the achieved geometric accuracy has a direct influence on the performance of Euclidean tracking in ASNs.

In this chapter, first an overview of related state-of-the-art methods is given. It is concluded by a description of the basic localization approach for a single microphone array, which was developed in the author's diploma thesis. Then the new method of speaker localization with a single sensor node is described. Thereafter, the novel method for speaker tracking using a distributed acoustic sensor network is presented. A summary of the methods discussed concludes this chapter.

4.1 STATE-OF-THE-ART

The problem of speaker localization and tracking is well researched, and there is a large variety of existing methods. While several modern approaches utilize complex mathematical constructs such as particle filters [PHHF11, EMN16], these were deemed beyond the scope of this work. Three basic families of approaches will be discussed that are closely related to the proposed methods and their goals.

First, methods based on cross-correlation are presented. As many methods of acoustic localization are based on the generalized cross-correlation (GCC) or steered response power with phase transform (SRP-PHAT) approach, this is an important baseline. Additional effort is required to determine speech activity and the number of concurrent speakers. Assuming perfect synchronization, the method can be applied for distant microphones for Euclidean localization and tracking in ASNs. Often the time difference of arrivals (TDoAs) are integrated over all frequencies and the spectral characteristics of the sources are not considered.

Second, approaches based on source separation will be discussed. These are exploiting the fact that not only the location, but also the spectra of the speakers are different. Based on the assumption that individual time-frequency bins are dominated by a single source, clustering and histogram techniques grouping them by direction of arrival (DoA) have been developed. Euclidean localization is possible by triangulation using the DoAs. This is better suited for ASNs as the synchronization has only to be good enough with respect to the speaker's movement.

Thirdly, approaches based on computational auditory scene analysis (CASA) will be introduced, as the proposed method also uses insights and models derived from the theory of auditory scene analysis (ASA). The human ability to localize speakers in adverse conditions has inspired a large number of computational models. Targeted at understanding the neuro-biological process, these usually use only two microphones embedded in an artificial head. A hybrid method employing a model of human hearing in conjunction with microphone arrays was introduced by the author [PHF10].

4.1.1 *Correlation based approaches*

As the correlation of microphone signals is a strong direct indicator for a sound source, several approaches rely on correlation alone. The SRP-PHAT, which evaluates the correlation in all possible directions, is very common. In order to improve the robustness and handle concurrent speakers, several extension have been proposed, including direct multiplicative combination and evaluation on intervals of dominance.

Steered response power

The SRP-PHAT-approach [BW01, pp. 157-180] uses a delay-and-sum beamformer (cf. section 2.4 on pages 24–26) that is steered into the direction where its output is maximal. This can be seen as an extension of the generalized cross-correlation with phase transform (GCC-PHAT) since the SRP-Equation (4.1) is identical to the GCC-PHAT summed over all microphone pairs (2.20) [MHA08, pp. 149–150].

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmax}} \frac{1}{F} \sum_{f=-F/2}^{F/2} \sum_{(i,j)} \frac{\Phi_{y_i y_i}(k,f)}{\left|\Phi_{y_i y_j}(k,f)\right|} e^{\mathrm{j}2\pi f/F \tau_{m,n}(\mathbf{s})} \tag{4.1}$$

$$= \underset{\mathbf{s}}{\operatorname{argmax}} \sum_{(i,j)} \underbrace{\frac{1}{F} \sum_{f=-F/2}^{F/2} \frac{\Phi_{y_i y_j}(k,f)}{\left|\Phi_{y_i y_j}(k,f)\right|} e^{\mathrm{j}2\pi f/F \tau_{i,j}(\mathbf{s})}}_{R^{PHAT}(\tau_{(i,j)}(\mathbf{s}))} . \tag{4.2}$$

When applied with time windows of sufficient length, this method is robust against noise and reverberation [ZFZ08]. A main problems with SRP-based speaker tracking is that the correlation maxima are not directly related to source presence. A peak in the correlation may be caused by reverberation or non-speech sources. To exclude the latter, a voice activity detection (VAD) may be added. The common approach to exclude the former is to apply large time averaging. This is working because the peaks due to reverberation and noise show a higher variance in TDoA.

An example of SRP-PHAT based multi-speaker tracking is the work of Lathoud et al. [LO07]. Here SRP-estimates are grouped into short-term clusters by angular and temporal neighborhood. The clusters correspond to an optimum partitioning of all detections over the time sequence. Maximum likelihood (ML) clustering is done with integrating close azimuths over short time frames, allowing for gaps due to omissions and short speech pauses. The method achieves about 2° error and good recall on multi-speaker recordings of the AV16.3 corpus.

The framework for SRP-based speaker tracking introduced by Madhu et al. [MM08] also uses maximum likelihood estimates computed via the expectation-maximization (EM) algorithm. A small variance is estimated for true source positions while noise is modeled by a Gaussian with a large variance fixed at 90°. The DoA estimates for each time frame are modeled as a mixture of Gaussians (MoG). The tracking is computed online by repeating the ML estimation for each time frame after initializing it by the result for the previous frame. This method defines a time to live (TTL) that past estimates are allowed to be associated with current ones to cover speech gaps and detection omissions. After an estimate is older than that time, the source is discarded.

Multiplicative combination

The additive combination of the SRP-PHAT leads to imaginary sources, so called 'ghosts'. This effect is already apparent if only small arrays with far field source are used. However, the highest peaks in the spatial likelihood are often from correct positions, as the TDoAs correspond to the same DoAs (cf. section 2.1.2, p. 11). If pairs with larger distances or from multiple arrays in an ASN are combined, the TDoAs correspond to hyperboloids rather than angles. Here the 'ghosts' are a more severe problem. In principle, a source should be apparent in all pairs, and the correlation estimates r^{PHAT} of the individual microphone pairs can be multiplied to find the true source positions without 'ghosts':

$$\hat{\mathbf{s}}_k = \underset{\mathbf{s}}{\operatorname{argmax}} \prod_{(i,j)\in P} r^{PHAT}\left(k, \tau_{(i,j)}(\mathbf{s})\right). \tag{4.3}$$

The product can degrade, as small values will dominate and a single zero in one of the factors will lead to the missing of a source. Thus the product itself is not robust for practical application, since errors caused by incorrect geometry calibration or noisy signals may lead to erroneously small values for one or more microphone pairs. Therefore, the family of Hamacher-t-norms was introduced as an alternative by Pertilä et al. [PKV08]. The Hamacher t-norm is defined as:

$$h_\gamma(z_1, z_2) = \frac{z_1 z_2}{\gamma + (1-\gamma)(z_1 + z_2 - z_1 z_2)} =: z_1 \odot z_2 \tag{4.4}$$

Note that a fuzzy t-norm is defined only for values z_i in the interval $[0, 1]$. By iterative application of h_γ,

$$\bigodot_{i\in I} z_i := (((z_1 \odot z_2) \odot \ldots) \odot z_n), \tag{4.5}$$

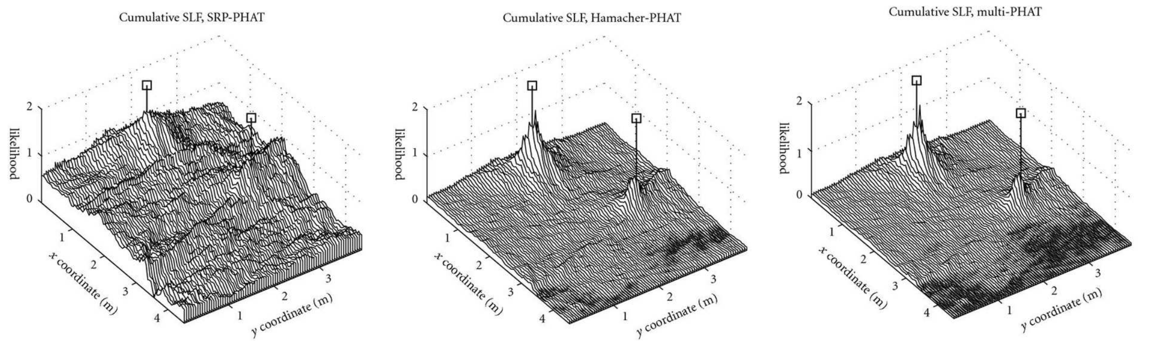

Figure 4.1: Cumulative spatial likelihood for two speakers in a living room recording using different combination methods for the microphone pairs: Additive SRP-PHAT (left), Hamacher-PHAT (middle) and Product PHAT (right). Images taken from [PKV08], Springer Open Access, Creative Commons Attribution License, © 2016 BioMed Central Ltd, © 2008 Pasi Pertilä et al..

the source position is estimated in a robust fashion:

$$\hat{\mathbf{s}}_k = \underset{\mathbf{s}}{\operatorname{argmax}} \bigodot_{(i,j)\in P} R^{PHAT}\left(k, \tau_{(i,j)}(\mathbf{s})\right). \tag{4.6}$$

Experiments with three microphone arrays in a living room show the superiority of multiplicative combination. Two concurrent speakers are reflected by clearly defined maxima at their locations for either the multiplicative and the Hamacher-PHAT. The speakers are localized within a 25 cm radius in 92% and 93% of time frames, respectively. The SRP-PHAT achieved 81%. Figure 4.1 illustrates the mitigation of ghost artifacts by use of the Hamacher norm or multiplicative combination of microphone pairs in this experiment.

Intervals of dominance

An efficient way of combining the pairwise GCCs was introduced by Oualil et al. [OFK13a]. For each GCC, intervals of dominance are computed based on the location of the peaks. Then, these intervals are mapped from TDoAs to source positions. Only source positions that are backed by a dominant peak in the correlation of all microphone pairs are considered. This is in effect a multiplicative combination. The benefit of this method is a drastic reduction of the search space as illustrated in Figure 4.2. Instead of calculating the SRP-PHAT for the whole search space (left), intervals of dominance are computed for each pair (middle) and used to define regions of interest (right).

After finding all possible source positions in this way, the SRP is computed cumulatively over all TDoAs and microphone pairs, hence the method is called cumulative steered response power (C-SRP). At the positions yielding the maximum values, the source is localized by the classical SRP-PHAT in a fine grid.

After the localization, a Bayesian classifier is used to distinguish between speech and noise. A Gaussian mixture model (GMM) is fitted to the cumulative function. Exploiting the fact that true speech sources produce sharp peaks in the spatial likelihood, while noise sources tend to have a very flat distribution, the Gaussian fitted to the C-SRP can be classified as speech or noise [OFK13b].

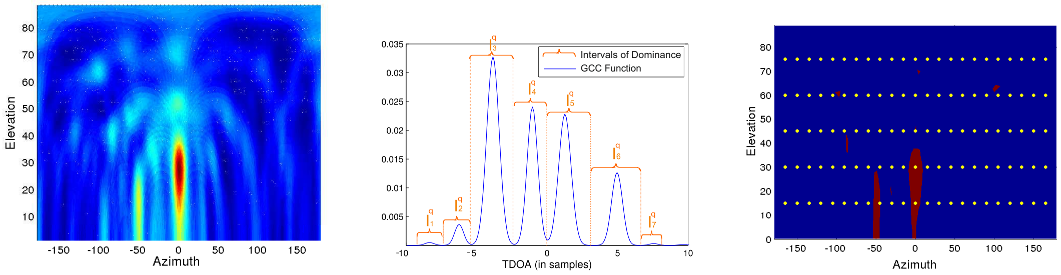

Figure 4.2: CSRP-PHAT localization [OFK13a]: Unmodified SRP-PHAT (left), intervals of dominance for one microphone pair (middle) and regions of interest (right). Images published in [OFK13a], available from EURASIP Open Library, © 2013 EURASIP.

The method achieves realtime performance and is able to track multiple concurrent speakers. Evaluation on the AV16.3 corpus shows around 50% recall with 80% precision and an angular error around 2° [OFK13a].

4.1.2 *Time-frequency localization*

Given the spectral sparsity of speech, it may be assumed that no or only few values originating from different speech sources collide in frequency. This is known as the "sparsity assumption" employed in blind source separation (BSS) [PLKP08]. This technique is used for both localization and speech enhancement, cf. Section 2.4.1 on page 25. When the direct path is dominant, strong independent components can be found for each source in the short time Fourier transform (STFT) domain. The speakers can be separated by different techniques.

A basic approach is the multiple signal classification (MUSIC) method that is based on eigenvalue decomposition (EVD), cf. [MHA08, pp. 151–154]. Here the number of sources that can be identified is bounded by the number of microphones. A mathematically more complex approach is the independent component analysis (ICA), it assumes statistical independence and employs higher order statistics, cf. [NK11]. An in-depth discussion of these methods is beyond the scope of this work. Instead, the less restricted and more practical approach of clustering will be investigated on two examples, the reference-based clustering of the directions and the identification of "single-source zones" for histogram based DoAs estimation. For both methods, extension to Euclidean tracking in ASNs have been proposed, which will be described below.

Directional clustering

The BSS based clustering method proposed by Araki et al. [ASMM06] identifies DoAs of concurrent speakers towards a microphone array. All time-frequency bins of the STFT are clustered. In order to form a common representation, one microphone is used as reference sensor and all pairs with said sensor are evaluated. The phase difference is normalized with respect to frequency, so that each time-frequency bin of each pair yields the same unit vector when pointing at the same far-field source. Experiments with three microphones in a mildly reverberant room ($T_{60} = 0.12\,\mathrm{s}$) show the methods ability to localize concurrent speakers. It is able to localize four speakers with just three microphones. Two sources as close as 22° are found with around 4° error.

An application of the clustering based DoA estimation in ASNs was introduced by Taseska et al. [TH13]. The DoAs of the two nodes with the strongest signal are used to determine the Euclidean position of the speaker. The triangulation is a plain line intersection with just two nodes. This is done independently for all time-frequency bins, so that concurrent speakers are handled by spectral association. The goal of this work is to apply an minimum variance distortionless response (MVDR) beamformer that enhances the speech of the person present at the given position or spot, hence the method is called 'spotforming'. The position based filter constraints are estimated by a minimum Bayes risk detector based on the triangulation. The spot radius around the position and the cost values of the Bayes detector are set so that the missed detection (MD) and false alarm (FA) rates have the best trade-off. While in the initial approach [TH13], the prior probability for each room position was set to a symmetric Gaussian distribution, in the later version it was refined by estimating said probability in a training stage based on room simulation [TH16]. The method is shown to provide speech enhancement in various simulated scenarios with up to four speakers and moderate to strong reverberation $(T_{60} = 0.7\,\mathrm{s})$.

Single source zones

A different BSS inspired approach was presented by Pavlidi et al. [PGP13]. Under the sparsity assumption, time-frequency bins that are dominated by a single speaker are used to estimate the DoA with a circular microphone array. After applying an STFT to the input signal for a given time frame, adjacent frequency bins dominated by one source are identified. This is done by computing the cross-correlation over adjacent pairs of microphones in each frequency bin and finding those with a correlation close to 1, called "single-source zones". By using up to eight adjacent frequency bins for each single-source zone, the DoA of that zone is computed. All DoAs determined in that fashion are cumulated in a histogram over time. In said histogram, peaks are extracted by an iterative algorithm. The largest peak is found and removed by subtracting a matched Hamming window. The process is repeated until no peak of sufficient height remains. Simulation of up to six concurrent speakers show a good performance with 3° error at $T_{60} = 0.25\,\mathrm{s}$ and 99% recall for a signal-to-noise ratio (SNR) of 20 dB that decreases to about 60% for an SNR of 0 dB. When compared with wide-band MUSIC and others, the single-source-zone method provides higher recall and slightly better precision. All three approaches solve the problem of concurrent speakers. At $T_{60} = 0.4\,\mathrm{s}$ an RMS estimation error of below 10° is achieved. The method is real-time capable and verified with real recordings.

An extension of the single source zones to ASNs was devised by Griffin et al. [GM13]. First the angular localization described above is applied in order to obtain one or more DoA estimates from a number of sensor nodes. Given the node locations, pairwise intersections of rays in the DoA direction give Euclidean positions. In order to incorporate the fact that almost parallel rays will lead to very bad localizations, pairs with a low angular difference are discarded. The remaining intersections are combined to give one or two speaker positions. If there is only one DoA per node, it is assumed that only one speaker is active and the mean of all intersections is computed as his or her position. If there are multiple DoAs for at least some nodes, the search space is divided by half-planes between them. When n is the number of nodes with two DoA detections, all 2^n possible partitions created by the half-planes are compared, and the one containing the highest number of intersection points is chosen. The mean of these intersection points is the first speaker localization. The mean of intersections on the other side of each of the chosen half-planes is the localization of the second speaker.

4.1.3 *Bio-inspired methods*

The impressive ability of humans to process and localize speech in adverse conditions has been an interesting research subject for decades. Many computational models were designed in order to imitate and thereby understand these abilities [Bla96]. Over the years, several CASA models inspired by the ASA theory [Bre90] (see section 2.2.5 on page 19) were implemented [WB06]. These CASA models of both monaural and binaural human hearing use similarities of multiple cues such as location, spectrum, and pitch for grouping and separation of speakers [CB10, HW12]. In relation to this work, binaural models for localization are of the most interest. An example of the basic model and a recent extension will be described before introducing the hybrid approach for microphone arrays.

CASA Model

A straightforward binaural localization model based on CASA principles was proposed by Roman et al. [RW08]. Two microphones mounted on an artificial head are used to localize and track multiple concurrent speakers. Their signals are filtered by a 128 channel Gammatone filterbank with center frequencies equidistantly spaced on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 50 kHz. The hair cell transduction is simulated by halfway rectification and square-root compression. The cross-correlation of the signals from left and right ear is computed in each band on 20 ms rectangular windows with a shift of 10 ms. All peaks that are defined as values higher than their left and right neighbors are considered interaural time difference (ITD) estimates. The ratio of energy between frames of left and right channel is used as interaural intensity difference (IID) estimate in each band. Reference values for ITD and IID are obtained using white noise signals applied in the azimuth range of -90° to 90°. The deviation from these reference values is modeled using a Laplacian distribution for both ITD and IID. In order to handle an unknown number of concurrent speakers, separate models are trained for one to three concurrent speakers. A small Markov model is used to model all eight possible activity patterns of three speakers as different states. The idea of a simple state model for speaker activity is also used in [OK14]. Both the ITD and IID distributions as well as the transition probabilities were estimated using samples from the TIMIT database. The model was shown to be able to track up to three overlapping speakers on various trajectories in an anechoic simulation with around 5° accuracy. Unfortunately the performance deteriorates significantly in reverberation as mild as $T_{60} = 0.05\,\text{s}$.

CASA GMM Model

A notably improved binaural CASA model was introduced by May et al. [MvK11]. The signals from an artificial human head are processed with a smaller Gammatone filterbank comprised of 32 ERB-spaced bands between 80 Hz and 5 kHz. The spike model is implemented with halfway rectification followed by square-root compression. ITD is computed as cross-correlation within the bands, followed by exponential interpolation. Interaural level difference (ILD) is computed as energy ratio, both are using 20 ms windows with a shift of 10 ms. Training was done with speech data from the TIMIT database. Both interfering speakers and a reverberation of $T_{60} = 0.5\,\text{s}$ were incorporated in

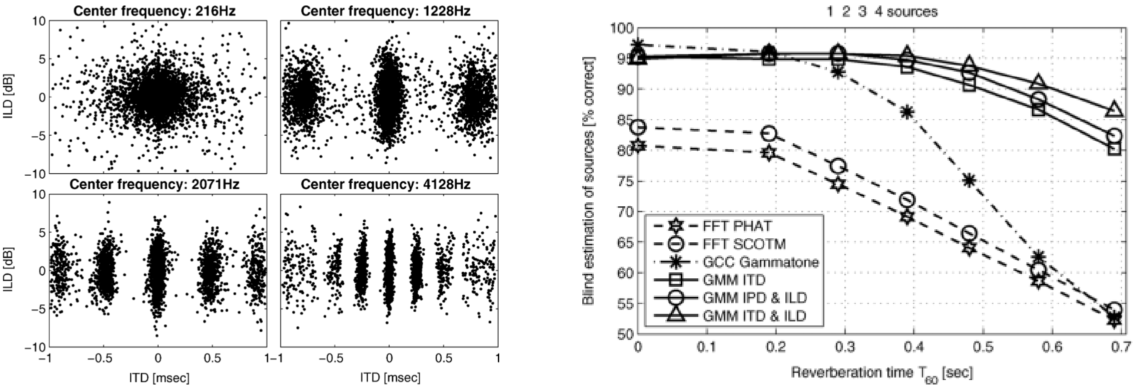

Figure 4.3: CASA GMM Localization. ITD-ILD distributions in different frequency bands (left) and localization accuracy compared to other methods (right). Images from [MvK11] © 2011 IEEE.

the training step. The ITD and ILD estimates are combined in a two dimensional feature space. Their distribution is modeled by individual GMMs for each source direction. The number of mixture components is estimated by either visual inspection or a minimum descriptor length (MDL) criterion. The GMMs are trained by EM-estimation. The localization is done by choosing the GMMs with the maximum sum log-likelihood over a number of time frames.

The localization performance degrades with both the source distance and reverberation level. This was shown in a comparative evaluation on simulated data. In a 5×7 m room, the amount of erroneous detections increases to about 15% and 45% at high reverberation of $T_{60} = 0.7\,\mathrm{s}$ for a distance of 1 and 3 m, respectively [MvK11].

The number of active speakers can be estimated by thresholding a histogram of detections over the DoAs. In the aforementioned simulation setup, up to three sources are reliably detected with 95% accuracy up to reverberation times of 0.5 s. Beyond that, the accuracy drops to about 70% at 0.7 s. A GCC-PHAT based baseline algorithm shows worse performance with 80% accuracy dropping to 50% [MP12].

4.1.4 *Hybrid approach*

Fixing the number of sensors at two — aimed at strict imitation of the human prototype — is an unnecessary constraint for a technical system. The use of multiple sensors facilitates robust localization in noisy and reverberant environments, by exploiting the redundancy among all channels. Recently, hybrid approaches applying acoustic signal processing in combination with biologically inspired neural processing to subband or circular microphone arrays were proposed [SGT07]. One of the first successful true hybrid methods was implemented by the author [PHF10].

Since the first parts of that model are also used in the proposed methods in this thesis here, the method is described in this section. Figure 4.4 shows the two common stages of processing and the third heuristic step that was later replaced. First, a cochlear model is used to compute a band-wise peak over average position (PoAP) representation of the microphones' signals. Second, a midbrain model computes a band-wise spatial likelihood with spherical coordinates. Third, an azimuth-only representation is computed and peaks are extracted yielding the speaker DoAs.

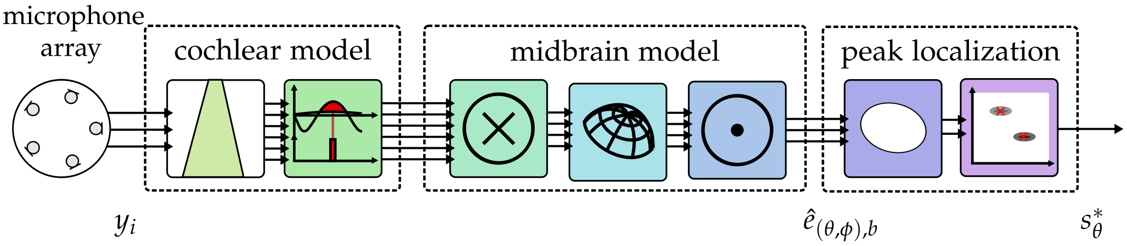

Figure 4.4: Neurologically inspired speaker localization. Cochlear and midbrain model followed by heuristic peak localization.

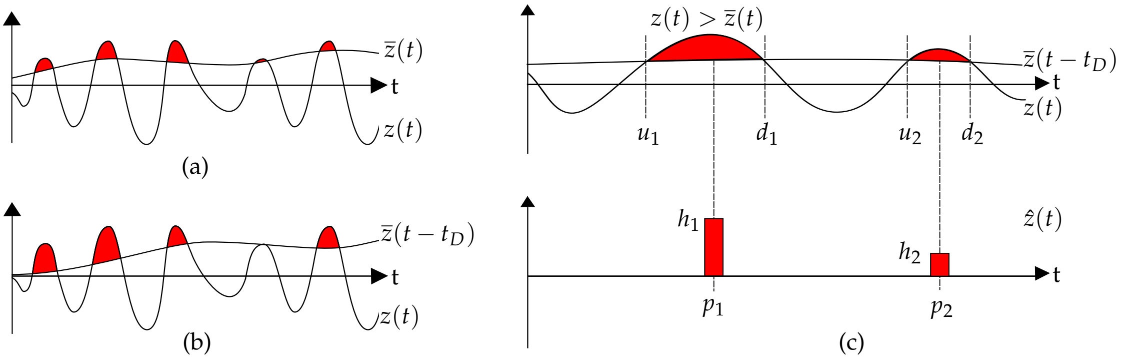

Figure 4.5: Onset dominance (left) and Peak-over-Average-Position (PoAP) spike generation (right). Band and microphone index omitted for simplicity. The difference between the signal z and its average $\bar{z}$ is shown in red (a), the first waves are enhanced if the average is shifted in time (b). The energy of the difference is encoded as spikes at the maximum position of each peak-over-average interval (c).

Cochlear model

The cochlear model is composed of a filterbank modeling the frequency response of the basilar membrane and a spike generation step modeling the cochlear nucleus. The first step is done in the STFT domain on overlapping time windows of $K = 1204$ samples with a hop size of 512. A fast Fourier transform (FFT) filter-bank is used with $B = 16$ filters defined in the spectral domain using a Gammatone approximation [UA99] defined in Equation (2.29). The center frequencies are equally distributed on the ERB scale between 300 and 3,000 Hz. For each microphone signal y_i, band filtered signals $z_{i,b}$ are generated by applying these filters in the spectral domain and back-transformation in the time domain (see section 2.2.2 on page 17).

For the phase encoding, rectangular pulses are generated phase-locked to signal maxima while intensities are coded relative to the overall amplitude using a single nonlinear time domain step, the PoAP spike generation method illustrated in Figure 4.5. It is tailored to facilitate localization in reverberant and noisy environments by three aspects: First, echo suppression is achieved by modeling neural saturation. Second, phase-locked spikes are generated based on maxima for the TDoA estimation. Third, only highly modulated parts of the signal are used.

The input signal $z_{i,b}(t)$ is compared to the 30 ms average of its halfway rectification $\bar{z}_{i,b}(t - t_D)$. By shifting the average in time, neural saturation is emulated by comparing signal to an average shifted in time [PHF10, PBW04]. This mimics the neural saturation of monaural echo suppression with minimal computational load (cf. section 2.2 on pages 16–18). As can be seen in Figure 4.5 (left), the shifted average masks the later waves but not the first one, mimicking the precedence effect.

Modulated intervals $[u_{i,b,n}, d_{i,b,n}]$ are detected as periods where the peak over average condition $z_{>}\overline{z}$ holds. In each such interval, the maximum position is determined.

$$p_{i,b,n} = \underset{u_{i,b,n} \leq t \leq d_{i,b,n}}{\operatorname{argmax}} \left(z_{i,b}(t) - \overline{z}_{i,b}(t - t_D)\right) \tag{4.7}$$

For each peak, a spike is generated as illustrated in Figure 4.5 (right). In the human brain, spikes occur in spike trains where the number of spikes per train encode the intensity of the signal (see section 2.2.1 on pages 16–17). Here, it is encoded as one number only to compress the information into a sparse signal. This is calculated as the sum of the peak-over-average amplitudes in the interval. Square root compression is used as basic model of the sensitivity:

$$h_{i,b,n} = \sum_{t=u_{i,b,n}}^{d_{i,b,n}} \sqrt{z_{i,b}(t) - \overline{z}_{i,b}(t - t_D)} \tag{4.8}$$

The output $\hat{z}_{i,b}(t)$ can be modeled as a sparse vector sequence $(p_{i,b,n}, h_{i,b,n})$ for each microphone signal with index i in each frequency band with index b. This way, a sparse representation of the spike trains encoding both the amplitude and the phase of the signals is derived. It can be efficiently stored and processed in a sparse data structure.

By accepting only peaks more than a threshold $t_g = 6\,\mathrm{dB}$ above the average, only high peaks or "glimpses" with high SNR are used as reliable witnesses for speech,

$$20 \log_{10} \left(z \left[p_{i,b,n}\right]\right) - 20 \log_{10} \left(\overline{z} \left[p_{i,b,n}\right]\right) \geq t_g. \tag{4.9}$$

Mid-brain model

As described in Section 2.2 on pages 16–18, ITD estimation between the ears in the medial superior olive (MSO) can be modeled via a cross-correlation of the two signals. As the rectangular spikes themselves are correlated, this leads to a sharp correlation figure for all frequencies, unlike the halfway rectification used in other CASA models. To reduce harmonic errors, a band and pair dependent correlation frame size $K_{b,(i,j)}$ is computed for each microphone pair (i, j). In order to capture TDoAs with the smallest possible time window, the length is computed as the sum of three values. First, the time it takes the sound to travel the microphone pair distance is used as this reflects the range of physically possible TDoAs. Second, a maximum pitch period of 12 ms is chosen, as in voiced speech the onset will occur in sync with pitch. Third, two wavelengths to the lower band edge frequency $f_b{}'$ are added to allow for deviations in the signals phases.

$$K_{b,(i,j)} = \left(\left\|\boldsymbol{m}_i - \boldsymbol{m}_j\right\| / c + 12 \text{ ms} + 2 / f_b{}'\right) f_s \tag{4.10}$$

The cross-correlations $r_{(i,j),b,\tau}$, with τ denoting time delay, are calculated using the signal in these time windows multiplied by a Hamming window function. In order to capture all possible correlations for voiced speech, this is performed in 6 ms steps, i.e. a common hop that is smaller than half the window length for all bands and microphone pairs.

Due to the time-domain sparsity of the spikes, an optimized matching algorithm is about ten times faster than performing a correlation in the spectral domain [PHF10].

This algorithm finds matching spikes $R_{b,(i,j)}(k)$ in the signals of two microphones by iterating the nonzero values in a sparse data structure used for the spikes. For each match, a small triangular function scaled by the product h of the amplitudes is added to the output at the time difference d between the spikes:

$$r_{b,(i,j),\tau}(k) = \sum_{(s,t)\in R_{b,(i,j)}(k)} (\triangle(\tau - d)h + \triangle(d-\tau)h) \quad \text{where} \tag{4.11}$$

$$R_{b,(i,j)}(k) = \{(h,d)|p_{i,b,n} - p_{j,b,n} = d \wedge h_{i,b,n}h_{j,b,n} = h \wedge kK_S \leq p_{i,b,n} < p_{j,b,n} < kK_S + K_{b,(i,j)}\}. \tag{4.12}$$

The correlations are back-projected to spherical source positions $s(\theta,\phi)$. This is implemented with a lookup table mapping the positions to time lags and linear interpolation.

$$\tau_{(i,j)}(\theta,\phi) = ||\boldsymbol{m}_j - \mathbf{s}(\theta,\phi)|| - ||\boldsymbol{m}_i - \mathbf{s}(\theta,\phi)||f_s/c \tag{4.13}$$

$$d_{(i,j),b,(\theta,\phi)}(k) \approx r_{b,(i,j),\tau'}(k) \quad \text{with} \quad \tau' = \tau_{(i,j)}(\theta,\phi) \tag{4.14}$$

Finally, all microphone pairs are combined by iterative application of the Hamacher fuzzy t-norm, as defined in Equation (4.4).

$$e_{b,(\theta,\phi)}(k) = \bigodot_{(i,j)} d_{(i,j),b,(\theta,\phi)}(k) \tag{4.15}$$

The resulting spatial likelihood is three-dimensional as it is a function of time, DoA and frequency band. For an example sequence of two, partially concurrent speakers, the three dimensional space is shown as sum projections to each pair of two dimensions in Figure 4.6.

Heuristic peak localization

In order to find speakers using the so computed spatial likelihood $e_{b,(\theta,\phi)}$, a peak detection approach is used. While the parameters of this process are naturally heuristic, consistently good results were achieved with small circular arrays in reverberant conference rooms with the values that will be given in this section.

A first reduction of the search space is done by searching only in the two spatial dimensions parallel to the floor. This can be justified as in the tabletop placement speakers can be separated by azimuth, and only small positive elevation angles are in the region of interest. Furthermore, planar arrays exhibit a bad resolution for elevation. So, as discrimination by elevation is not desired, the maximum value over a set of elevations ($\phi = 0, 5, \dots 45$) is chosen.

$$\tilde{e}_{b,\theta}(k) = \max_{\phi} e_{b,(\theta,\phi)}(k) \tag{4.16}$$

Next, the average over a longer time segment is computed. The moving average $\bar{e}_\theta$ over $L = f_s \cdot 0.5\,\text{s}$ is calculated over all data points with a shift of $L \cdot 1/4$ samples.

$$\bar{e}_{b,\theta}(k') = \sum_{k=k'-L/2}^{k'+L/2} \tilde{e}_{b,\theta}(k) \tag{4.17}$$

The spectral spread is investigated and spectral information is discarded. Speech is mostly producing a signal spread over the spectrum while several noise types are not.

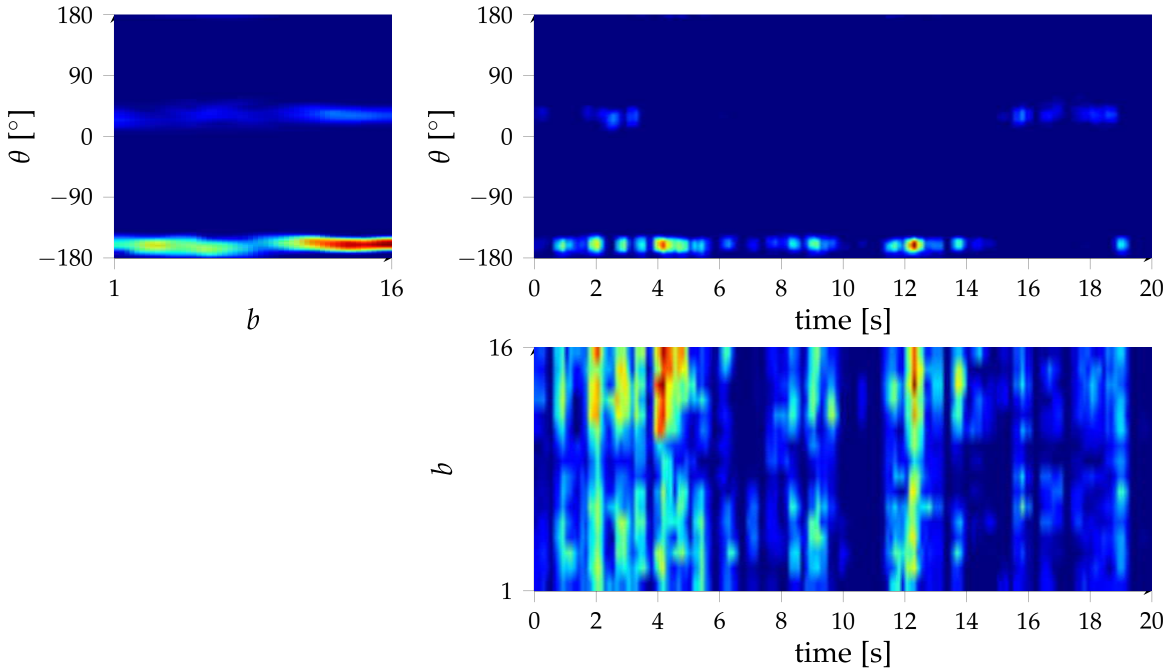

Figure 4.6: Projections of the spatial likelihood computed by the neuro-biologically inspired method. A snapshot of sequence #9 recorded in the FINCA, where two speakers have a normal conversation is shown. The tree dimensional space is shown as sum projections to each pair of two dimensions: Angle and frequency band (top left), angle and time frame (top right), and frequency band and time frame (bottom right). The colors represent the intensity from -40 dB in blue to the maximum in red. The speakers have different powers due to dissimilar distance to the array and loudness of speech.

Narrow band noise from fans and machines is typically only found in one of the frequency bands. Time domain aliasing in the correlation occurs frequency-dependent and therefore produces erroneous peaks at different source locations in different frequency bands. So by filtering out positions with low spectral spread, noise and aliasing errors can be suppressed. In practical applications of the method, speech yields nonzero values in about half of the Gammatone bands. Thus, a basic filtering is implemented by counting the bands with energy peaks, and discarding detections occurring in less than a third of the frequency bands.

$$\hat{e}_\theta(k') = \begin{cases} \sum_b \tilde{e}_{b,\theta}(k') & \#\{\bar{e}_{b,\theta}(k') > 0\} > B/3 \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

When considering larger time segments, the correlation results can be modeled as “true” peaks plus noise [LO07]. To get rid of noise induced peaks in the spatial likelihood, a processing inspired by the difference of Gaussians process found in may parts of the human sensory processing was used in [PHF10]. To incorporate the typical variations, a

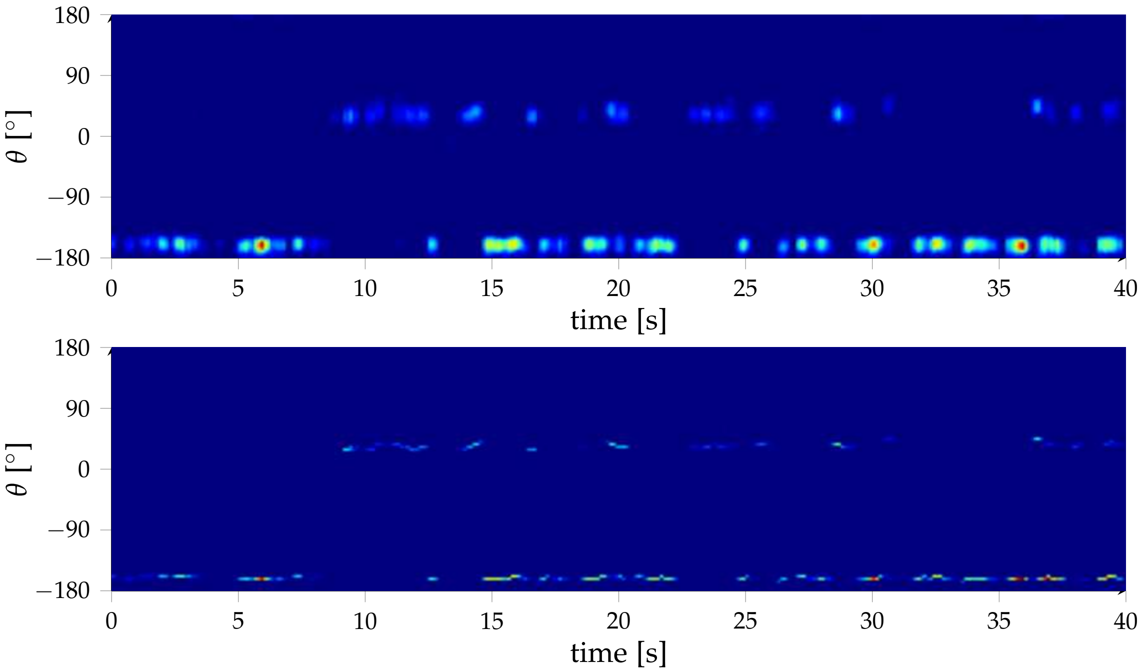

Figure 4.7: The peak localization: Spatial likelihood $\hat{e}_\theta(k')$, summed over all bands for sequence #9 (top) and the PoAP peak localization result $s^*_\theta(k')$ (bottom). Colors represent the intensity from -40 dB in blue to the maximum in red.

45° average, spanning the reverberation induced artifacts, is subtracted from a 5° average representing the signal to yield a filtered spatial likelihood:

$$f_\theta(k') = \frac{1}{5}\sum_{d=-2}^{2} \hat{e}_{(\theta+d)}(k') - \frac{1}{45}\sum_{d=-22}^{22} \hat{e}_{(\theta+d)}(k') \tag{4.19}$$

Finally, the peak-over-average algorithm is run along the azimuth to extract positions of modulated peaks. An average $\overline{f}$ is computed along the circle to find intervals of high modulation where the value is higher than the average. The maxima within these intervals are extracted as detected peaks. Given the angular width of a peak caused by a speaker in the spatial likelihood, a length of 15° for the average was chosen.

$$s^*_\theta(k') = \max_{u_n \le t \le d_n} f_\theta(k') - \overline{f}_\theta(k') \tag{4.20}$$

In Figure 4.7, this process is illustrated for a snapshot of sequence #9, where two speakers have a normal conversation. It can be seen that natural overlap of speech activity occurs. The angular PoAP operation reduced the wide likelihood areas around the speaker positions to a single value for each time frame and active speaker.

Figure 4.4 shows the whole pipeline of the PoAP speaker localization method for a single node. After computing the spike representation in the cochlear model for each microphone channel, a spatial likelihood is computed in the mid-brain model through correlation, back projection and combination of all microphone pairs. In the peak localization, the position is reduced to azimuths at the elevation with the maximum value and angular peaks are extracted.

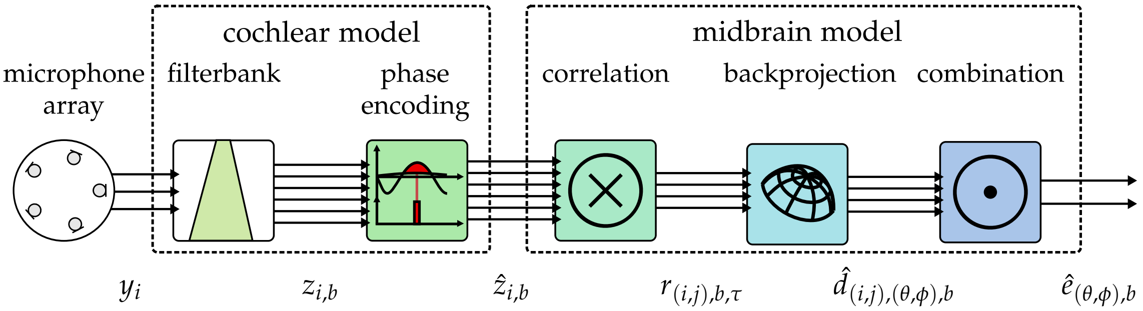

Figure 4.8: Common neurologically inspired processing used for localization.

4.2 Proposed method for single node speaker localization

The hybrid method described above calculates angular localizations for a single array [PHF10]. For the application within ASNs, it was augmented in two aspects.
The first improvement is an adaptive input gain for the signals. This eliminates the need of manually presetting a gain for each recording. More importantly, the Hamacher combination requires the signals to be clipped, as it is only defined in the interval $[0, 1]$. As the power of the input signal is unknown, a suitable gain has to be applied before clipping. By adding a method to automatically set this gain, it is no longer required to adjust this beforehand.
The second improvement is the replacement of the basic peak localization. It required adjusting the thresholds involved and is not able to separate close speakers clearly. Therefore, it was replaced by the CASA approach, where simultaneous grouping (see section 2.2.5 on page 19) is modeled by clustering. According to the ASA theory, multiple cues are used in this process. Given that the spatial likelihood is computed in a spatial and a spectral dimension, it is possible to cluster considering both important cues. This multi-criterion clustering was first used successfully in [PHF12] with the density based DBScan algorithm [EKSX96]. Here, the later approach with the EM algorithm will be described, cf. Section 2.3 on pages 20–23. The use of ML clustering allows for better modeling and removes heuristic thresholds. The output is probabilistic and spectral information is retained for each source. This is important to associate the speakers across different nodes in the tracking approach described in the next section.

4.2.1 *Cochlear and midbrain model*

The cochlear and midbrain model described on pages 49–51 is used, cf. Figure 4.8. While the first steps of the system are linear, the fuzzy combination is not. The correct gain has to be applied to the signals in order to keep the spike amplitudes in the range [0,1]. This is also important if the method is implemented with small floating point or integer precision for faster computation. The model was therefore extended by a method of automatic gain estimation. The gain is set automatically dependent on a running estimate of the current input energy. A histogram of H spike amplitudes is used to calculate the energy level. For each channel i and band b, the histogram $I_{i,b,h}$ is set to the number of spikes $h_{i,b,n}$ falling within the hth bin, i.e., with amplitudes ranging from h/H to $(h+1)/H$:

$$I_{i,b,h} = |\{h_{i,b,n}|h/H < h_{i,b,n} \leq (h+1)/H\}| \,. \tag{4.21}$$

This level $\varrho_{i,b}$ is estimated as the level at $p = 0.95$ of the spike level distribution, i.e.,

$$P(h_{i,b,n} \geq \varrho_{i,b}) \approx 0.95\,. \tag{4.22}$$

Practical tests showed small variance over the microphones. This is plausible as the nodes are of small size and thus very little difference in the speaker volume is to be expected. Similarly, the adaptation of the loudness for individual frequency bands did not prove advantageous. Thus, the level is estimated over all channels and bands:

$$\varrho^* = \frac{1}{BI} \sum_{b=1}^{B} \sum_{i=1}^{I} \varrho_{i,b} \tag{4.23}$$

A series of experiments was made in order to determine the gain required. As the analysis showed little variance over the recordings of natural speakers in different reverberant settings, the optimal gain can be approximated with a fixed emphasis function based on the frequencies. It approximates the experimentally derived gains by a simple stepwise linear function:

$$\beta(f_b) = \begin{cases} 18 f_b/860 & f_b \leq 860 \\ 18 + 12(f_b - 860)/1540 & 860 < f_b \leq 2400 \\ 30 & 2400 < f_b \end{cases}\,. \tag{4.24}$$

For performance reasons, the gain is applied to the sparse spike structure before computing the correlation and subsequent back projection and combination.

$$h'_{i,b,n} = h_{i,b,n} 10^{(\beta(f_b) - \varrho^*)/20} \tag{4.25}$$

Again, the elevation with the highest value after a short moving average $\bar{e}_{b,(\theta,\phi)}$ over L samples, e.g., $L = f_s \cdot 0.5\,\mathrm{s}$ is calculated over all data points with a shift of $L/4$ samples and detections with less than $B/3$ spectral components are excluded as non-speech sounds. The values are collected over all bands as

$$\boldsymbol{g}_{k,\theta} = \left[\bar{e}_{1,(\theta,\phi)}, \bar{e}_{2,(\theta,\phi)}, \ldots, \bar{e}_{B,(\theta,\phi)} \right]^T\,. \tag{4.26}$$

The energy values comprise a set of azimuth-spectrum tuples $G_k = \{(\theta, \boldsymbol{g}_{k,\theta})\}$ for each time frame k. In order to filter low energy noise, these are only considered speech energy detections where the sum over all bands exceeds a threshold of $t_e = -40\,\mathrm{dB}$. This sum energy over all frequency bands is interpreted as likelihood for a source ν at the given angle, as it reflects the correlation and signal strength:

$$l(\nu = (\theta, \boldsymbol{g}_{k,\theta})) = \frac{1}{B} \sum_{b} g_{b,k,\theta}\,. \tag{4.27}$$

4.2.2 *Simultaneous Grouping*

According to the ASA theory, location as well as spectral cues are used for grouping the auditory information coming from a certain source, cf. Section 2.2.5 on page 19. In order to emulate this process, ML clustering over spatial and spectral similarity is used. The probabilistic modeling of the speakers according to this concept will be described in the following before the details of the EM implementation. Then it will be explained how the number of sources is estimated in each iteration and when the algorithm is terminated.

Model

The distribution of repeated measurements of DoAs peaks produced by reverberant speech is often successfully modeled by a Gaussian distribution, cf. [LO07]. Therefore, the spatial likelihood is modeled as a MoG as in [MM08]. The probability density for a detection $\nu = (\theta, \boldsymbol{g}) \in G_k$ can be calculated with the average angle Θ and standard deviation σ:

$$p_a(\nu|\Theta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-0.5\frac{d(\theta, \Theta)^2}{\sigma^2}\right). \tag{4.28}$$

Here an angular distance function d is used. As the distance of two angles can be either clock- or counterclockwise, the direction around the circle with the shorter absolute distance is chosen:

$$d(\alpha, \beta) = \min\{360 - |\alpha - \beta|, |\alpha - \beta|\}. \tag{4.29}$$

The second clustering criterion is spectral similarity. The spectra from different speech sources are dissimilar with a high probability in most practical scenarios and noise and time domain aliasing artifacts are assumed independent across frequency. The spectral similarity of a detection $\nu = (\theta, \boldsymbol{g})$ to a model spectrum $\boldsymbol{g}'$ is calculated as normalized scalar product

$$p_s(\nu|\boldsymbol{g}') = \left\langle \frac{\boldsymbol{g}}{||\boldsymbol{g}||}, \frac{\boldsymbol{g}'}{||\boldsymbol{g}'||} \right\rangle = \frac{\sum_b g_b g'_b}{\sqrt{\sum_b g_b^2 \sum_b g'^2_b}} . \tag{4.30}$$

The pseudo-probability of ν to originate from $\Psi_c = (\Theta_c, \sigma_c, \boldsymbol{g}'_c)$ with average angle Θ_c, standard deviation σ_c and spectrum $\boldsymbol{g}'_c$ is defined as

$$p(\nu|\Psi_c) = p_s(\nu|\boldsymbol{g}'_c) p_a(\nu|\theta_c, \sigma_c) . \tag{4.31}$$

Estimation step

Sources Ψ_c are estimated over all N detections in the current and adjacent time frames $\nu \in G_{k-1} \cup G_k \cup G_{k+1}$ by maximum likelihood estimation. In the estimation step, the pseudo-probabilities are computed over the MoG using mixture weights η_c and the equations above:

$$p\left(\Psi_c|\nu\right) = \frac{\eta_c p(\nu|\Psi_c)}{\sum_{c'} \eta_{c'} p\left(\nu|\Psi_{c'}\right)}. \tag{4.32}$$

Maximization step

For the maximization step, first partial weights are computed that reflect the relative contribution of a given detection ν for a given mixture component index c:

$$\rho_c(\nu) = \frac{p(\Psi_c|\nu) l(\nu)}{\sum_{\nu'} p(\Psi_c|\nu') l(\nu')} \tag{4.33}$$

The weighting $\rho_c(\nu)$ takes the spatial sum likelihood into account. In relation to the original unweighted EM-implementation, $l(\nu = (\theta, \boldsymbol{g}))$ can be interpreted as the number of measurements for ν, so that the maximization step equals the original one for a discrete number of measurements. With these weights, the distribution parameters recomputed in order to maximize the likelihood:

$$\hat{\Theta}_c = \sum_{\nu=(\theta,\boldsymbol{g})} \rho_c(\nu)\theta \tag{4.34}$$

$$\hat{\sigma_c}^2 = \sum_{\nu=(\theta,\boldsymbol{g})} \rho_c(\nu) d(\theta, \hat{\Theta}_c)^2 \tag{4.35}$$

$$\hat{\boldsymbol{g}}_c = \sum_{\nu=(\theta,\boldsymbol{g})} \rho_c(\nu)\boldsymbol{g} \tag{4.36}$$

$$\hat{\rho_c} = \frac{1}{N}\sum_{\nu} p\left(\Psi_c|\nu\right), \tag{4.37}$$

The weighted average of angles in (4.34) has to be calculated on the circle. While the mean of a Gaussian on a circle is not defined, in practice the values are often concentrated within a small interval of angles. Thus it is sufficient to compute the average within this interval.

Number of sources

The number of sources can be estimated by observing the typical variance of speaker localizations, as in [MM08]. To allow for the number of speakers to decrease or increase from the previous time frame k, a join and split rule are evaluated. If two estimates get closer than a threshold $d(\theta_c, \theta_{c'}) < \Gamma_{\text{join}}$, the sources i, j are merged. If $\sigma_c > \Gamma_{\text{split}}$, the source i is split into two sources with $\theta_{c,c'} = \theta_c \pm \sigma_i$ as illustrated in Figure 4.9. As the angular parameters for the previously described peak extraction method, these threshold values are chosen heuristically based on the standard deviation of $\sigma = 5° - 15°$

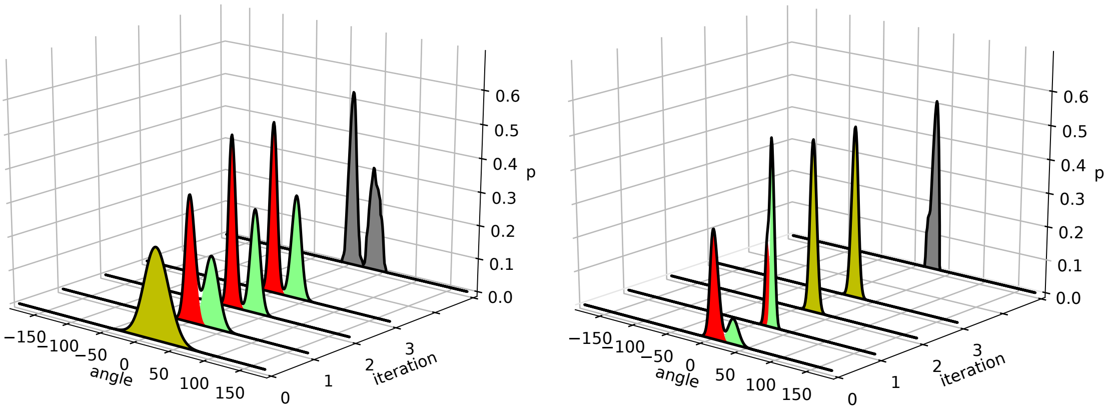

Figure 4.9: Split (left) and join (right) within the EM estimation. Mixture components are plotted in color and the mixture as contour line for each iteration. The spatial likelihood histogram $\sum_{\nu=(\theta,\boldsymbol{g})} l(\nu)$ for the angles θ to be estimated is shown in gray at the back. Iteration 0 shows the estimate from the previous time frame. Illustration first published in [PF13], available from EURASIP Open Library, © 2013 EURASIP.

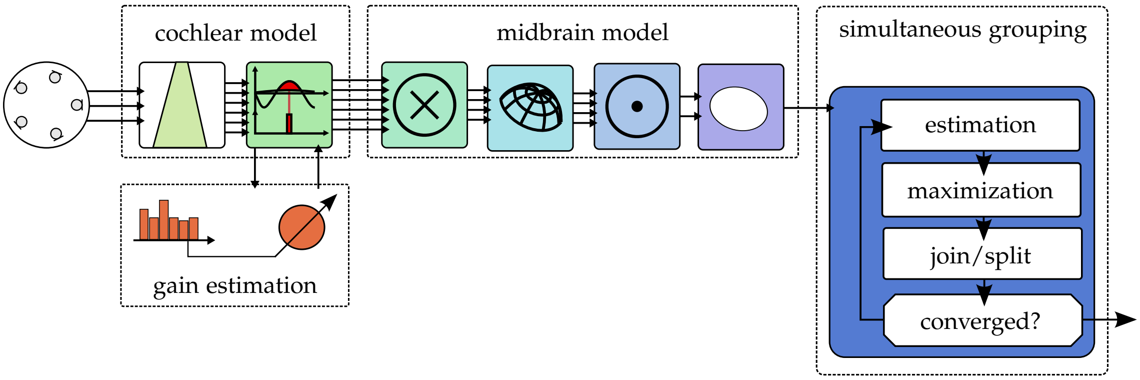

Figure 4.10: PoAP EM speaker localization. The cochlear model is used with automatic gain estimation. The midbrain model performs the correlation, back projection, and combination of microphone pairs. The EM algorithm is used to determine the probabilistic positions and spectra of an unknown number of concurrent speakers.

observed for a known speaker in the spatial likelihood in reverberant enclosures. Good results are achieved by setting $\Gamma_{\text{join}} = 11^\circ$ and $\Gamma_{\text{split}} = 22^\circ$.

Termination

The estimation loop is terminated when the likelihood does no longer change significantly. This typically happens after two to ten iterations, allowing for real-time calculation. After this step, there are clustered source estimates $\{\Psi_i\}$ for each time frame. The overall method is illustrated in Figure 4.10.

4.2.3 *Speech segment identification*

For the geometry calibration approach, static speaker positions have to be identified. This can be achieved by post-processing the PoAP EM method.

Spatial proximity is used to associate the localizations into tracks over time. Given that the PoAP EM DoA localization produces small angular errors, only DoAs localizations with an overall angular deviation of less than 5° are grouped in the same speech segment. Values with a large angular deviation are mostly caused by speaker movement. As the calibration requires static positions, these can not be used here.

Given that gaps in speech occur naturally, several approaches use a TTL, cf. [MM08]. Here, consecutive DoA localizations within a TTL of 1 s or less are grouped in time to identify speaker activity. This bridges typical small speech pauses.

4.3 Proposed Method for Speaker Tracking

In an ASN, the individual DoA localizations can be combined by triangulation in order to localize the speakers in Euclidean coordinates. This requires the knowledge of the nodes' relative geometry, which can be computed automatically as described in chapter 5 on pages 65–86. The method proposed in [PF14a] integrates the nodes' PoAP EM localizations in three steps. First, the estimates are associated using their spectra, second the Euclidean speaker coordinates are computed by triangulation and thirdly the localizations are integrated into tracks over time.

4.3.1 *Association*

The problem of ambiguity of multiple concurrent localizations by multiple nodes is illustrated in Figure 4.11. When two or more speakers are active, this results in multiple DoA localization from each node. Without additional information, it is not clear which DoA corresponds to which speaker. While some associations may be excluded geometrically, e.g., by the rays not intersecting within the room, often different associations are possible.

To associate the estimates from different nodes, their spectra are correlated using Equation (4.30), and the pairs with the strongest correlation above a minimum threshold t_s are computed. By thereafter combining all pairs with common angles, sets of angular estimates over all nodes are derived.

4.3.2 *Triangulation*

The Euclidean position of the source can be derived by triangulation using the so associated sets of DoAs and the ASN geometry. By calculating the intersection of the rays originating at two nodes with the cluster angles Θ_i and Θ_j the 2D position $\hat{\boldsymbol{s}}_{(i,j)}$ is derived as explained in 2.1.3 on pages 14–15.

Given two angles α, β the expected accuracy q of the localization by intersection may be expressed as the sine of the intersection angle to reflect the fact that an angular difference of 90° yields the highest precision and an angular difference near 0° or 180° the worst:

$$q(\alpha, \beta) = |\sin(\alpha - \beta)| \,. \tag{4.38}$$

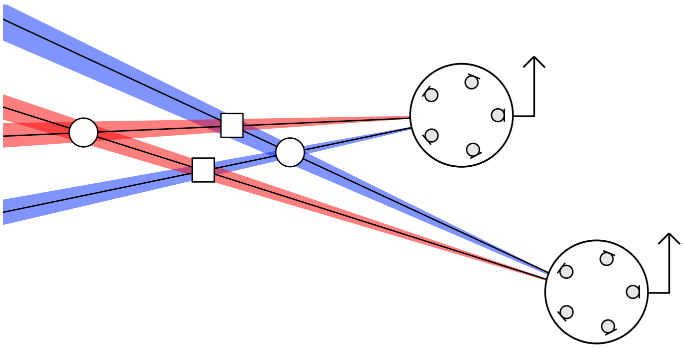

Figure 4.11: The problem of ambiguity of multiple concurrent estimates: Without additional information, all four intersections are possible source positions. Using the spectral similarity, the correct intersections (circles) are chosen and the others (squares) are discarded. Illustration first published in [PF14a], (c) IEEE 2014.

This function is a good approximation of the expected empirical error. It was derived from simulations given in the evaluation chapter. In order to calculate one point from multiple intersections, the weighted sum is used:

$$\hat{\boldsymbol{s}} = \frac{\sum_{(i,j)} q\left(\Theta_i, \Theta_j\right) \hat{\boldsymbol{s}}_{(i,j)}}{\sum_{(i,j)} q\left(\Theta_i, \Theta_j\right)} \tag{4.39}$$

As is illustrated in Figure 4.12, this leads to an improved position estimate.

4.3.3 *Tracking*

A combined tracking state $\Omega_{c,k} = (\Psi^{(m)}_{c,k}, \ldots, \Psi^{(n)}_{c,k}, \hat{\boldsymbol{s}}_{c,k})^T$ represents the states of the track with label c for time step k. It does not necessarily contain estimates from all nodes. The probability of a new detection $\Psi_{*,k+1}$ to belong to a track Ψ_c given the cluster angles is calculated for each node:

$$p_a\left(\Psi_{*,k+1} | \Psi_{c,k}\right) = p_a(\Theta_{j,k} | \Theta_{*,k+1}, (\sigma_{*,k+1} + \sigma_{c,k})/2). \tag{4.40}$$

These probabilities are then multiplied over all nodes to compute the consensus:

$$p(\Psi^{(m)}_{*,k+1}, \ldots, \Psi^{(n)}_{*,k+1} | \Omega_{j,k}) = \prod_o p_a\left(\Psi^{(o)}_{*,k+1} | \Psi^{(o)}_{j,k}\right). \tag{4.41}$$

For each set of new estimates, the track with the highest likelihood above a threshold ϵ_a is chosen from all tracks not older than a t_{TTL} (e.g. 5 s). The time-to-live covers gaps caused by speech pauses or detection or transmission failure. If no such track exists, a new one is started. Figure 4.13 illustrates the overall tracking procedure.

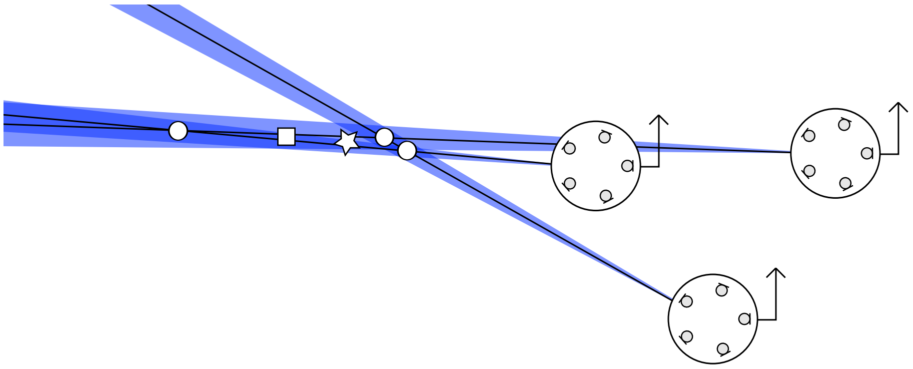

Figure 4.12: Incorporating angular intersection quality into the triangulation: Using all pairwise intersections (circles) equally, the center (square) is computed as localization. When weighting by intersection quality, the steeper angles get favored and a more precise localization is achieved (star). Illustration first published in [PF14a], (c) IEEE 2014.

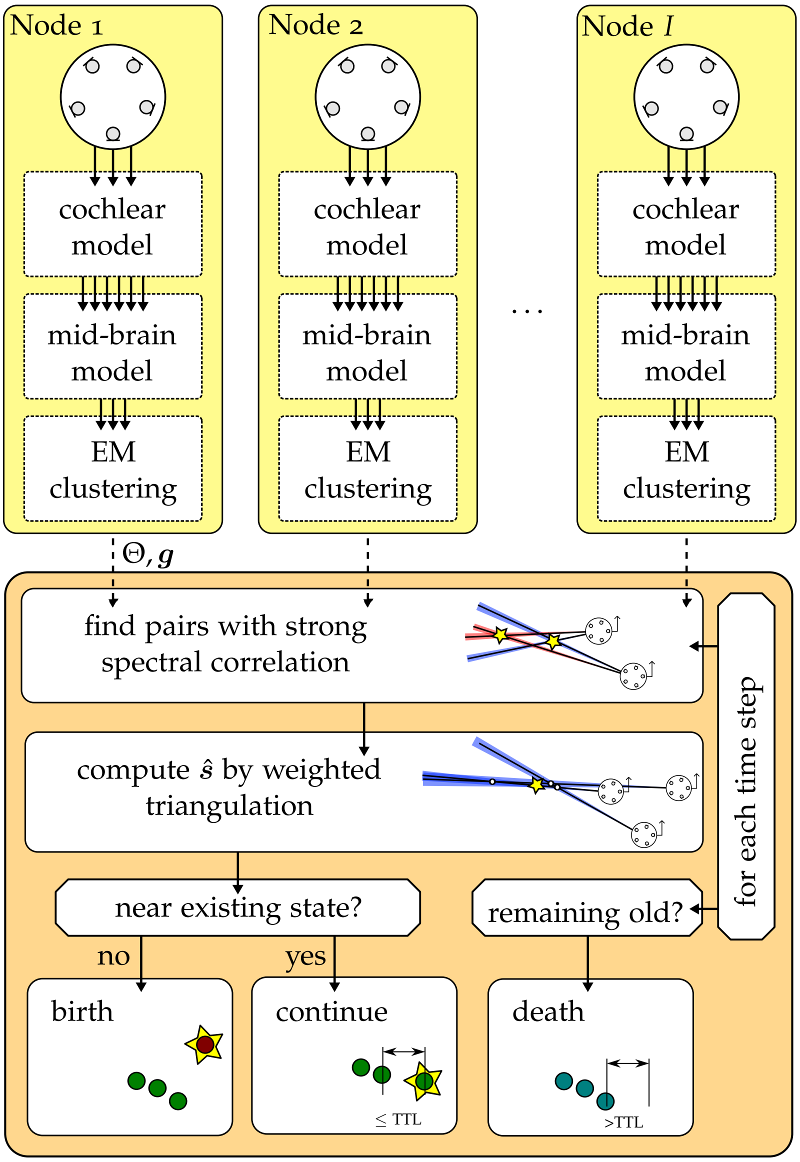

Figure 4.13: ASN tracking method: Each node computes dominant DoAs with spectral activity patterns for concurrent speakers. For each time step, all pairs of such estimates from two nodes are considered tracking candidates. If they are close to an existing state from a previous time step not longer ago than a TTL, that track is continued. Otherwise a new track is started. If there is no new detection near a track older than the TTL, that track is discontinued.

4.4 summary

There is a vast variety of speaker localization and tracking algorithms. Selected examples have been discussed in the state-of-the-art section. Based on the extended localization method for a single node, a tracking solution for ASNs was proposed. In this section, first the properties of the different localization approaches for a single node will be summarized in a comparative fashion. Then the extensions for ASN application will be described.

Single node speaker localization

Several approaches for the DoA localization with a single node were introduced. Table 4.1 on this page lists the key methods described in this chapter. One of the most important properties for application in real scenarios is the ability to handle concurrent speakers. The different approaches of the methods are summarized in the table and will be briefly discussed in comparison. Another distinguishing aspect listed in the table is the fusion of the TDoA estimates from the microphone pairs in the array. The SRP-PHAT method is still often used because of its simplicity and good performance [ZFZ08]. In its basic form, it will identify the position of any source with the maximum spatial coherence. It is not able to handle concurrent speakers, which requires additional modeling as in, e.g., [MM08, LO07].

For application in practical scenarios, VAD or speech detection should to be added, as in [OK14]. The problem of 'ghost' localizations is mitigated by using multiplicative combination of the microphone pairs.

As the TDoAs are distributed around the true value over time and frequency due to reverberation and other errors, most methods employ clustering to model this. As the distribution of correlation peaks over time is very similar to Gaussian, the use of MoG modeling and the EM algorithm is quite common.

The BSS employs the sparsity assumption by identifying time-frequency bins dominated by a single source. Even though this is less applicable in higher reverberation, both the directional clustering and single source zone approach are employed successfully in indoor environments.

method	speaker counting	mic. array
SRP-PHAT [BW01]	–	sum
SRP-PHAT EM [MM08]	EM, join	sum
C-SRP [OK14]	dominant pos.	prod.-like
Direction clustering [ASMM06]	t-f bins	direction
Single Source Zones [PGP13]	peak detection	histogram
CASA [RW08]	threshold	–
CASA GMM [MP12]	threshold	–
PoAP [PHF10]	peak detection	t-norm
PoAP EM [PF13]	EM, join / split	t-norm

Table 4.1: Single node localization methods, listing the speaker counting approach for handling concurrent speakers and the integration of microphone pairs for the handling of arrays.

CASA based methods are benefiting from insights into human perception, which makes them robust to reverberation and noise. Aimed at replication of the biological example, an artificial human head with just two channels is used. With models of the human integration and back projection, good localization of concurrent speakers is possible [MP12].
The proposed PoAP EM method for speaker localization with a single node combines the benefits and insights of the methods described above. It employs a neuro-biologically inspired cochlear and mid-brain model that makes it robust against reverberation and implicitly handles concurrent speakers [PHF10]. MoG modeling according to CASA principles was added to improve the performance and provide probabilistic estimates.

ASN speaker localization

When using an ASN comprised of nodes with small microphone arrays, Euclidean tracking based on triangulation is possible. Table 4.2 on the current page lists the methods discussed in this chapter. The Hamacher-PHAT [PKV08] uses all microphone pairs jointly for correlation. This requires a perfect synchronization and transmission of the full sample signals to a central node. It is therefore not well suited for ASN application. The other methods combine only local microphone pairs of small compact arrays in order to estimate DoAs. This relaxes the synchronization requirement and reduces the required transmission bandwidth.
A practical problem for triangulation are the bad estimates derived from near parallel DoAs. In [GM13], these are excluded when below a threshold. The proposed approach introduced an weighting based on intersection angles to resolve this issue.
Another problem is the correct association of multiple DoA localizations for concurrent speakers. In the half-plane search method [GM13], this is solved by counting the agreeing nodes for a certain intersection. Given that the sources are inside the convex hull of the sensors in their setup, this seems to work. In the clustering based approach [TH13], the association is handled by assigning time-frequency bins to the dominant DoA. The proposed method follows a similar idea by computing spectral similarity based on the reduced band information. It employs the neuro-biologically inspired PoAP EM method in the nodes in order to improve the robustness in reverberant enclosures. The algorithm also localizes concurrent speakers. The bandwidth required is minimal since only a single DoA and a coarse spectrum is transmitted for each active speaker.

method	association	triangulation	data
Hamacher-PHAT [PKV08]	position	hyperboloid-inters.	full signal
Half-spaces [GM13]	counting	counting	DoA only
Strongest pair [TH13]	spectra	strongest two	DoA / spectra
PoAP EM [PF14a]	spectra	weighted	DoA + spectra

Table 4.2: ASN tracking methods with the approach to solve the association and triangulation and the data to be transmitted between the nodes.

Let no one ignorant of geometry enter
Inscription on the entrance of Plato's academy

5 ACOUSTIC SENSOR LOCALIZATION

This chapter discusses the task of acoustic sensor localization. The goal is to compute the relative or absolute geometric arrangement of the sensors, therefore it is often also referred to as "acoustic geometry calibration". The task is the reverse of the source localization task, as in this case one or more actors (i.e. sound sources) are used to determine the positions of the sensors. However, in the source localization scenario, the sensor geometry is known, whereas in sensor localization, the position of the sources is often also unknown. In the broadest case, the problem requires the estimation of the geometry of sensors, sources and the timing. In order to be able solve the underlying equations, the rank of the measurements has to be equal or higher than the rank of the unknowns. If certain constraints are imposed on the geometry or the signals used, the search space becomes more restricted, allowing for easier estimation or even direct computation of geometry parameters.

The knowledge of the spatial arrangement of acoustic sensor nodes is required for both localization and some speech enhancement algorithms. Automated methods are required since manual measurement is cumbersome and impractical in ad hoc scenarios. It is also favorable that the calibration can be performed solely by acoustic signals received by the nodes, thus eliminating the need for additional sensors, speakers in the nodes or external calibration devices. As the application of wireless acoustic sensor networks (WASNs) has become increasingly popular, an increasing number of methods has been published in recent years. These methods progress at reducing the number of practical constraints imposed for the calibration.

In this chapter, first criteria to distinguish the vast variety of acoustic geometry calibration methods will be introduced. Thereafter, several state-of-the-art methods will be outlined. Finally, the novel methods for multimodal and acoustic calibration of microphone array configurations will be described in detail.

5.1 TAXONOMIES

To categorize the variety of methods, several criteria can be employed. Here the taxonomy introduced in [PJHUF16] will be used: As main criterion, the scenario defining the quantities to be estimated is chosen. Within each scenario, additional constraints used can be identified to distinguish the methods. An important distinction is the type of signal and synchronization required, yielding different measured quantities. From these, mathematical procedures can be derived in order to estimate the geometry. These criteria to distinguish the different approaches will be outlined before discussing selected examples of the different methods.

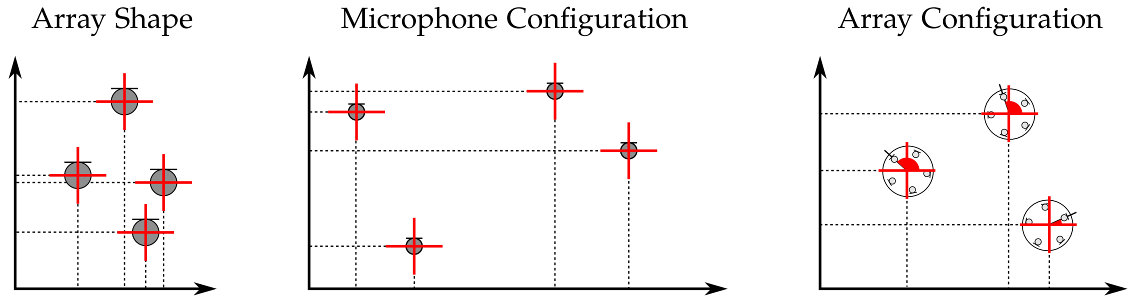

Figure 5.1: Scenarios of sensor geometry calibration. From left to right: The calibration of the shape of a small compact array, the calibration of the position of distributed microphones, and the calibration of distributed microphone arrays. Illustration after an idea used in [PJHUF16].

5.1.1 *Scenarios*

The first distinction is the actual geometry of the microphones to be estimated. Considering the extent of the geometric configuration, three types can be distinguished, cf. Figure 5.1:

ARRAY SHAPE The first scenario deals with a small compact microphone array. This will be referred to as array shape calibration. The relative geometric positioning of the microphones is to be estimated. The scenario is characterized by the limited physical extent of the array. Here, the small inter-microphone distance allows for the application of otherwise unusable or impractical techniques such as manual distance measurement or exploiting the diffuse noise coherence.

MICROPHONE CONFIGURATION The second scenario addresses individual distributed microphones. This is the task commonly referred to as microphone geometry calibration. In distinction to the previous scenario, the microphones can be distributed with large inter-microphone distances all over a room or, e.g., a conference table.

ARRAY CONFIGURATION The third scenario is specifically addressing distributed microphone arrays. The intra-array geometry is assumed to be known, either by measurement or application of one of the previous methods. Here not only the position, but also the orientation of the arrays is to be estimated.

5.1.2 *Measurement type*

The second distinction is the measurement type, which is closely related to the imposed constraints on the calibration process. Four types of acoustic measurements can be distinguished, cf. Section 2.1.2 on pages 11–12.

PAIRWISE DISTANCE This case is the measurement of all pairwise distances between the microphones or nodes [Bir03]. This can be achieved by using active nodes, each of which emits a sound [RD04] or by using diffuse noise in the case of small apertures like the array shape case [ML08]. The resulting measurement is of high rank and allows for direct computation of the geometry.

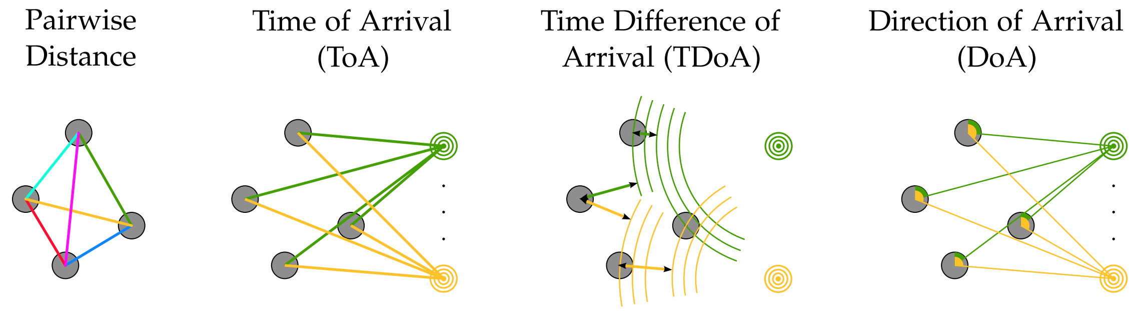

Figure 5.2: Measurement types used in geometry calibration. From left to right: Pairwise distances as measured with active devices of diffuse noise, the time of arrival of sound from a synchronized source, the time difference of arrival measured at pairs of microphones, and the direction of arrival. Illustration first used in [PJHUF16] (c) IEEE 2016.

TOA The first alleviation is to use fewer sound sources that are not geometrically linked to the nodes. When the emission time is known by synchronizing the nodes, absolute time of arrival (ToA) measurements can be used. When the source positions are unknown, they have to be estimated as well. This requires a larger number of sound emissions.

TDOA Without such synchronization, only time difference of arrivals (TDoAs) between pairs of microphones is directly available. The emission times can estimated as well, allowing to reduce the problem to the previous one of ToA based calibration. This is often done by imposing additional restrictions such as using a previously known sequence of sounds in order to reduce the number of measurements required. A single but notable exception uses the maximum TDoAs to derive the pairwise distances.

DOA In the array configuration scenario, the direction of arrival (DoA) can be measured as well and be used to estimate the nodes' orientation. This can be done separately. Either the geometry is estimated from the DoAs first and the scaling is estimated thereafter by using the TDoAs [SJHU+11, JSHU12]. Or the nodes' positions are estimated first and the orientation is estimated in a second step [PMH11].

5.1.3 *Mathematical approaches*

The mathematical procedure of estimating the geometry is closely linked to the type of measurement and the number of known quantities. It is common to successively reduce the number of unknowns by separate minimization steps.

A large portion of the state-of-the-art approaches exploits the fact that the rank of the problem is limited by the physical dimensions and uses either eigenvalue decomposition (EVD) or singular value decomposition (SVD).

When the pairwise distances between some or all microphones or nodes have been derived from the measurement, the $p = 1, 2$, or 3 eigenvectors corresponding to the p largest eigenvalues of the centered square distance matrix will yield a solution for the $1, 2$, or 3D geometry [Bir03]. This is exploiting the fact that the variance in the measurements created by the physical distance is larger compared to the variance produced by measurement errors. The EVD algorithm is so simple and robust that it is applied even in cases where underlying assumptions are not fully fulfilled. It is applied to derive an initial solution in active device calibration without considering the individual distances between speaker and microphones at each of the nodes [RKL05]. The solution is then refined by least squares (LS) or maximum likelihood (ML) minimization that takes the exact positioning into account.

It was shown that the ToA measurement of distances between unknown source and microphone positions can be decomposed in a similar manner, using either EVD [BS05] or SVD [CDM12]. The latter approach was subsequently extended to TDoA measurements by first estimating emission times [GKH13]. It was then further enhanced to use measurements from unsynchronized devices by previously estimating the individual recording offsets [GKH14].

5.2 STATE-OF-THE-ART

As both the unknowns and the complexity of the estimation procedure increase with the scenarios ordered as introduced here, the description of the individual methods will be ordered by the scenarios in the following. Examples of active and passive calibration methods will be described. As the EVD based approach is fundamental to many methods, the description will start with its original application to array shape calibration.

5.2.1 *Array shape calibration*

One of the first methods for the calibration of microphone arrays was the use of tape measured pairwise distances. Given all pairwise distances, the geometrical shape can be computed analytically using multidimensional scaling (MDS). The approach was extended to base-point classical multidimensional scaling (BCMDS). Here, rather than using all distances between the microphones, the ToA from a set of base points to each microphone is measured. The base-points are either a-priori known or estimated within the calibration process. Since the geometry estimation problem is of low rank, a few base-points are sufficient to compute a basis that can be used to apply multidimensional scaling, afterwards. It was shown that using $p+1$ base-points, the measurements of their relative distance and the distance between each base-point and microphone is sufficient to derive such a basis [BS05].

Multidimensional scaling

Applying MDS can be related to the principal component analysis (PCA). In both methods the eigenvectors corresponding to large eigenvalues yield the desired representation. Here a matrix consisting of the squared distances of the sensors is used. The eigenvectors of the decomposed inner product matrix are an estimate of the relative sensor coordinates.

First the square distance matrix $\hat{\boldsymbol{D}}_{ij} = \hat{d}^2_{ij}$ is composed from all pairwise distance estimates. Using the double centering matrix $\boldsymbol{H}$, the mean is removed and the approximate inner product matrix $\hat{\boldsymbol{B}}$ is computed as

$$\hat{\boldsymbol{B}} = -\frac{1}{2}\boldsymbol{H}\hat{\boldsymbol{D}}\boldsymbol{H} \quad \text{where} \quad \boldsymbol{H} = \boldsymbol{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T . \tag{5.1}$$

The inner product matrix can be factorized to determine the $N \times p$ position matrix $\boldsymbol{M} = [\boldsymbol{m}_1 \ldots \boldsymbol{m}_N]^T$ containing the positions of the N microphones since

$$\boldsymbol{B} \approx \boldsymbol{M}\boldsymbol{M}^T. \tag{5.2}$$

$\hat{\boldsymbol{B}}$ is positive semidefinite, therefore it can be factorized using EVD into the diagonal of eigenvalues $\boldsymbol{\Lambda}$ and the matrix composed of the corresponding unit eigenvectors $\boldsymbol{V}$:

$$\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T = \text{evd}(\hat{\boldsymbol{B}}). \tag{5.3}$$

We can observe the relation to the SVD:

$$\boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{W}^T = \text{svd}(\hat{\boldsymbol{B}}) \tag{5.4}$$

Exploiting the fact that the p largest eigenvalues correspond to the largest variance in $\hat{B}$, which span the $p = 2$ or 3 dimensional room containing the microphones, we can estimate the geometry using the p largest eigenvalues. The geometry estimate is computed as the product of the square-root of the diagonal matrix containing the two largest eigenvalues $\boldsymbol{\Lambda}_p$ multiplied with the $N \times p$ sub-matrix $\boldsymbol{V}_p$ containing the corresponding eigenvectors:

$$\hat{\boldsymbol{M}} = \boldsymbol{V}_p\boldsymbol{\Lambda}_p^{\frac{1}{2}}. \tag{5.5}$$

The estimate has an arbitrary rotation and translation in relation to the true positions.

Diffuse noise coherence

In the case of array shape calibration, the distance between the microphones is so small that the received signals will be spatially coherent. When the microphones pick up spatially diffuse noise, the pairwise correlation has a characteristic form dependent on the microphone distance. Thus it is possible to estimate the microphones' pairwise distance by matching the pairwise correlation of the microphone signals to the coherence function for a diffuse noise field. This was used by McCowan et al. [ML08] to estimate pairwise distances and subsequently compute the geometry using MDS as described above.
The complex coherence between two microphone signals y_i and y_j is computed as:

$$\Gamma_{y_j,y_i}(f) = \frac{\Phi_{y_j,y_i}(f)}{\sqrt{\Phi_{y_j,y_j}(f)\Phi_{y_i,y_i}(f)}}, \tag{5.6}$$

where $\Phi_{y_j,y_i}(f)$ and $\Phi_{y_j,y_j}(f)$ denote cross- and auto-spectral densities at signal frequency f, respectively.
Assuming a diffuse noise field and omni-directional microphones in distance d, the theoretical coherence function (cf. [BW01, chap. 4], [HG07]) is given as

$$\Gamma^{\text{diffuse}}(f, d) = \text{sinc}\left(\frac{2\pi f d}{c}\right) = \begin{cases} 1 & f = 0 \\ \frac{\sin(2\pi f d c^{-1})}{2\pi f d c^{-1}} & f \neq 0 \end{cases}. \tag{5.7}$$

Note that the noise field model can also be derived for directional microphones. Obtaining distances $d_{m,n}$ from coherence measurements is formulated as a non-linear least-squares model fitting problem:

$$\hat{d}_{m,n} = \underset{d}{\text{argmin}} \sum_{f=0}^{f_s/2} \|\Gamma^{\text{diffuse}}(f, d) - \Gamma_{y_j,y_i}(f)\| \tag{5.8}$$

where $f_s/2$ is the Nyquist frequency. This optimization problem is solved for all pairs by minimizing Equation (5.8), e.g., via the Levenberg-Marquardt algorithm.

Applications

McCowan et al. [ML08] used this technique to estimate the two-dimensional geometry of linear and circular microphone arrays. The microphones' positions were estimated with around 1 cm accuracy. The error converged to this value after using about ten seconds of noise produced by electronic devices' fans.

Different tests were performed in order to assess the influence on practical applications. First, the beampattern for a super-directive beamformer was investigated. When using the estimated positions in case of a linear array, the steering error was about 10-15°. The circular array proved more robust and showed little difference in the beampattern using the estimated microphone positions. In a different set of experiments, the microphone calibration technique was used to estimate the shape of two circular arrays for speech recognition in a scenario with two concurrent speakers [HML08]. After calibrating the arrays, the steered response power with phase transform (SRP-PHAT) was used for speaker localization. The speaker position estimates were used to steer a superdirective beamformer. A basic postfilter was added that set the time-frequency bins to zero in case the competing speaker showed higher activity. When using the calibrated positions, the localization and speech recognition performance was close to the one achieved with the measured positions. When using the localization, the speech recognition performance was actually slightly better compared to using approximate ground truth speaker positions [ML08].

Hennecke et al. applied the ambient noise and MDS approach in a smart conference room setting. Circular and T-shaped arrays were calibrated with about 1 cm accuracy [HPFHU09]. In one case, the error increased up to 4 cm, according to Hennecke et al. most likely due to the placement of the array near an non-orthogonal room corner. This hints at the fact that spatial diffuseness can be lower in certain physical configurations and thereby limit the accuracy of this method.

5.2.2 *Estimation of distributed microphones' geometry*

The estimation of the geometry of distributed microphones has been a popular research topic in recent years. This is most likely due to the advent of WASNs in the form of ad hoc assemblies of smart phones and laptops. Three types of measurement acquisition can be identified in order to sort the various approaches: First, the use of active devices equipped with a speaker in addition to the microphones. Second, the use of separate sound sources for the calibration sequence. Third, the use of speech events only, requiring no additional equipment.

Several methods employ speakers in the devices, each one emitting a calibration sound. This is then used to compute the pairwise distances from the ToA for MDS. Such methods are employed with laptops [RKL05] and smartphones [HPFHU09].

Alternatively a sequence of chirps or hand claps emitted from unknown positions in the room is used [GKH13]. Assuming the direct path is the shortest to all microphones, the TDoA can then be estimated using onset detection [KA13]. In order to simplify the setup, a single moving source continuously emitting short chirps can be used [GKH14].

The most interesting approach is the use of distributed speech events. This allows to calibrate without additional devices or a dedicated calibration sequence. The method introduced by Pertilä et al. uses the TDoA of speech events to jointly estimate the synchronization and geometry [PMH12].

Active devices

When active devices are used, the fact that the distance of speaker and microphone is small compared to the inter-device distance allows for a two-step estimation approach: First, the speaker configuration is estimated by base-point MDS (see section 5.2.1 on pages 68–69) neglecting the distance of speaker and microphone in the device. Second, the geometry is estimated by maximum likelihood (ML) estimation. The method was tested with $p+1$ of the microphones paired with a speaker. In an experiment, 32 microphones in a three-dimensional convex hull arrangement could be localized after a speaker was playing a chirp on five positions close to a microphone [RD04]. The same approach was later used to calibrate distributed laptops placed on a table in close proximity. The position error was 3.8 cm with synchronization and 6.2 cm without it [RKL05]. A similar approach was employed for the calibration of distributed microphones in the form of smartphones on a table by Hennecke et al. [HF11]. Here, the geometry is estimated from sound events created by the phones themselves. A short logarithmic sweep chirp is played and the ToA is computed by each phone via correlation with the known signal. From the ToAs, pairwise TDoAs are computed for all pairs of phones, using the two events where each is once playing the sound. Again, in the first step, the distance of the phones own speaker to its microphone is neglected. The pairwise distance is estimated from the difference of the TDoA between two devices when they alternately play a sound. As the sound takes the direct path to device j when device i plays a sound, the TDoA $\tau_{i,(i,j)}$ is directly proportional to their distance. By using the devices alternately, any constant time offset $\delta_{(i,j)}$ cancels out.

$$d_{i,j} = \|\boldsymbol{m}_j - \boldsymbol{m}_i\| \approx \frac{c}{2f_s}\left(\tau_{i,(i,j)} + \delta_{(i,j)} - \tau_{j,(i,j)} - \delta_{(i,j)}\right) = \frac{c}{2f_s}\left(\tau_{i,(i,j)} - \tau_{j,(i,j)}\right) \tag{5.9}$$

Using the so obtained distance estimates, the geometry of the microphones is estimated by MDS. This is again used as initialization for a general minimization-based estimation in a second step. The known microphone to loudspeaker distances and the global orientation of the phones as provided by their sensors is incorporated. Real word experiments were done with three ad hoc configurations of six identical smartphones in about 40 cm distance, cf. Figure 5.3 on the following page. Over ten recordings of each configuration, the geometry was estimated with around 7 cm accuracy. An experiment with increasing pairwise distance of two phones showed a steep increase of the localization error with the microphone distance. Problems when using smartphones like this are the unknown effects of the microphones and speakers directivity, the sound propagation of loudspeakers lying on a table, and the uncontrolled effects of the phones audio software including noise canceling and gain control.

Calibration sounds

Many methods estimating the geometry using a set of sounds emitted from unknown positions are again based on the idea of aligning the geometry by decomposition of the measurement matrix. As shown by Crocco et al. [CDM12], a geometry estimate can be obtained by SVD, again using the p biggest singular values and the corresponding

Figure 5.3: Experimental setup for smart phone geometry calibration. Several phones placed on a conference table in close proximity. Image taken form [HF11] (c) IEEE 2011.

vectors. The product of the receiver and source position matrices is proportional to the distance matrix obtained from the ToA measurements.

$$-2\boldsymbol{M}\boldsymbol{S}^T \approx \tilde{\boldsymbol{D}} \tag{5.10}$$

This allows for the estimate to be computed from the truncated SVD using a $p \times p$ mixing matrix $\boldsymbol{C}$:

$$\boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{W}^T = \text{svd}(\tilde{\boldsymbol{D}}) \tag{5.11}$$

$$\hat{\boldsymbol{M}} \approx \boldsymbol{C}\boldsymbol{V}_p \quad \text{and} \quad -2\hat{\boldsymbol{S}}^T \approx \boldsymbol{C}^{-1}\boldsymbol{\Lambda}_p^{\frac{1}{2}}\boldsymbol{W}_p^T \,. \tag{5.12}$$

A simple-to-use method for calibrating a set of devices with a single microphone each using a mobile phone emitting a calibration sequence was proposed by Gaubitch et al. [GKH14]. The phone is playing a sequence of chirps with known time delays, while being moved around the microphones to be calibrated. As the chirps are short relative to the movement velocity, the emissions can be treated as originating from static unknown positions. A matched filter is used to extract peaks corresponding to the emission time in each device. Given that the delays between the chirps are known, the internal offset δ_i of each device i can be computed by aligning the sequence to the device's clock. This is done jointly for all devices by composing a matrix from all measured TDoAs to estimate the ToAs. Using the offsets and ToAs, the relative geometry is estimated by minimizing the difference to the measurement as

$$\hat{\boldsymbol{M}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{\delta}} = \underset{\boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\delta}}{\text{argmin}} \sum_{t=1}^{T}\sum_{i=1}^{I} \|\boldsymbol{m}_i - \boldsymbol{s}_t\| - ct_{t,i} - \delta_i \,. \tag{5.13}$$

Given that the minimization can end up in local minima far from the correct solution, initialization of the gradient descent is crucial. Therefore, the offsets are initialized by the previous estimation and the positions using an SVD estimate of the geometry. Both simulations and a single experiment with microphones distributed on a table showed an accuracy of 3 cm for this method.

Maximum TDoA of speech

A recent approach uses maximum time difference of arrival (mTDoA) values of speech events to estimate pairwise distances and classical MDS to calibrate the geometry [PMH12]. When a source is in the endfire position $\Theta_{t,(i,j)} = 0$ for a pair of microphones (see section 2.1.2, p. 12), the TDoA (2.11) reaches its maximum value:

$$\begin{aligned} \tau_{i,j}^{\max} &= \max_t \left\{ \tau_{t,(i,j)} \right\} && (5.14) \\ &\approx \max \left\{ \cos\left(\Theta_{t,(i,j)}\right) \|\boldsymbol{m}_i - \boldsymbol{m}_j\| f_s / c \right\} = \|\boldsymbol{m}_i - \boldsymbol{m}_j\| f_s / c. && (5.15) \end{aligned}$$

Equally, for the other endfire position $\Theta_{t,(i,j)} = 180°$, the minimum value is reached as

$$\tau_{i,j}^{\min} = -\|\boldsymbol{m}_i - \boldsymbol{m}_j\| f_s / c \; . \tag{5.16}$$

In the case of asynchronous devices, each device has an unknown time offset, resulting in unknown pairwise time offsets $\delta_{i,j}$. Thus the measured TDoA is

$$\begin{aligned} \tilde{\tau}_{i,j}^{\max} &= \|\boldsymbol{m}_i - \boldsymbol{m}_j\| f_s / c + \delta_{i,j} \quad \text{and} && (5.17) \\ \tilde{\tau}_{i,j}^{\min} &= -\|\boldsymbol{m}_i - \boldsymbol{m}_j\| f_s / c + \delta_{i,j} = \|\boldsymbol{m}_j - \boldsymbol{m}_i\| f_s / c + \delta_{i,j} && (5.18) \end{aligned}$$

Thus it follows that the offset can be computed as [PHM13]

$$\delta_{i,j} = \frac{1}{2} \left(\tilde{\tau}_{i,j}^{\max} + \tilde{\tau}_{i,j}^{\min} \right) \tag{5.19}$$

and the distance as [PMH12]

$$d_{i,j} = \frac{c}{2 f_s} \left(\tilde{\tau}_{i,j}^{\max} - \tilde{\tau}_{i,j}^{\min} \right) \; . \tag{5.20}$$

In order to find valid mTDoA measurements, first an voice activity detection (VAD) is used to detect speech. The TDoAs themselves are obtained using the generalized cross-correlation with phase transform (GCC-PHAT) (see section 2.1.2, p. 13). A histogram based filtering is employed to exclude outliers. Using the so obtained distance estimates, the geometry of the microphones is estimated by MDS as described in Section 5.2.1 on pages 68–69. The position of distributed smartphones and laptops with a single microphone was successfully calibrated with around 10 cm accuracy. In meeting room experiments wirelessly coupled devices were calibrated with an accuracy of 7–15 cm after synchronization [PPH14].

5.2.3 *Estimation of microphone array configurations*

Since there is a growing number of devices with more than one microphone, this scenario has been of increased interest in recent years. This is of high relevance in ad hoc scenarios. Both the case of active devices and ambient sound are investigated.

For most applications, the correct estimation of the orientations of the devices is of vital importance. Most types of spatial processing using several small arrays are in effect using triangulation, and therefore even a small orientation error will lead to large errors in the estimation of source positions.

In order to achieve a good orientation estimate, the methods for calibrations of distributed microphones described in the previous section are not sufficient. The error in the position estimate is often large compared to the device dimensions, so the implicitly estimated rotation is not of sufficient precision. Better results are achieved when the known intra-array microphone geometry is used to derive DoAs measurements in the calibration and include these in the geometry estimation process.

Active devices

Using mobile lab prototypes with four microphones and one speaker in the center, Pertilä et al. [PMH11] devised an active calibration method. Each device emits a maximum length sequence (MLS) signal. By using the TDoA between pairs of microphones from different devices, the pairwise device distances are estimated. From that, the inter-array translation is estimated using MDS. In a second step, the rotation and reflection of each array is estimated by aligning the SVD-based rotation estimate to the measured DoAs. In a series of experiments with four devices placed on a table in a mildly reverberant room ($T_{60} \approx 0.26\,\text{s}$), their geometry was calibrated with around 1 cm and 6° accuracy.

Speech-based RANSAC methods

A hierarchical approach by Hennecke et. al [HPFHU09] starts by ambient noise coherence based detection and calibration of the individual arrays as described in section 5.2.1 on pages 68–70. Subsequently, a five minute random walk of a speaker is used to derive a large set of TDoA measurements. Given that the arrays were located at the ceiling of a conference room, they are approximately positioned in a plane parallel to the moving speaker. Therefore, a constant height offset to the source can be estimated. The source is localized relative to each array by SRP-PHAT. Then SVD-based data set matching (DSM) is used to derive the relative geometry of the arrays. Over 100 Monte Carlo trials with a random sampling consensus (RANSAC) optimization, two circular arrays placed in the ceiling were calibrated with about 25 cm accuracy using natural speech. In a comparable experiment with white noise, 10 cm accuracy was achieved.

Using wall-mounted arrays consisting of only two microphones, Schmalenstroeer et al. measured the DoA of a moving speaker by beamforming [SJHU+11]. The relative position and the orientation of microphone arrays was estimated in two dimensions by an angular matching and the RANSAC approach. The problem of scale indeterminacy inherent to DoA-only observations is solved by estimating the scale in a second step using the TDoA. Tests in a mildly reverberant room ($T_{60} = 0.15\,\text{s}$) showed a translation error of about 25 cm and a rotation error of 2°. Extensions of the method features better cost functions [JSHU12, JSHU13]. Extensive simulation showed that the RANSAC procedure is required to remove outliers for even mild reverberation above $T_{60} = 0.05\,\text{s}$. For medium reverberation around $T_{60} = 0.4\,\text{s}$ the improved circular cost function decreased the position error from around 70 to 30 cm.

5.3 Proposed multimodal method for array configuration calibration

In a smart room setting, often several cameras are mounted. As they are mostly fixed on the walls or ceiling, they can be calibrated once. Sometimes microphone arrays are also mounted at these fixed positions. The disadvantage of this positioning is that they mostly receive indirect sound after reflections, as the speakers barely face the walls. Thus it is practical to put the microphones inside the room, e.g., on a conference table. This means that the microphone array positions are not fixed. Even if they are mounted on the table, the table itself may be moved. Therefore, the proposed method aims at automatically calibrating the geometric microphone array configuration using known camera positions. A human speaker will say a few sentences while moving through the room. He is localized visually by the cameras while the microphone arrays gather directional measurements.

The goal of the method is to find the absolute geometry $\hat{\gamma}_i$ of each microphone array using these two measurements. Thus, both the position $\boldsymbol{r}_i = [r_{1i}, r_{2i}]^T$ and the orientation o_i have to be estimated, cf. Section 2.1.3 on page 14. The possible range for the position is given by the space between the cameras, o_i is in the range $[-180°, 180°)$.

$$\hat{\gamma}_i = [\hat{r}_{1i}, \hat{r}_{2i}, \hat{o}_i]^T \tag{5.21}$$

The method first proposed in [PF14b] uses a speaker talking at static positions in the room to estimate the microphone arrays' geometry in the following way: Suitable time periods for a number of positions are identified from the acoustic recording. The Euclidean positions of a speaker are estimated by visual detection and triangulation. The DoA of the utterances at each microphone array and position are estimated. Both Euclidean position and DoA are computed for the projection to the ground floor. Using sets of matched visual 2D localizations and acoustic DoAs, an estimate of the absolute position and orientation of the microphone arrays is computed. By computing a consensus over several such estimates, a reliable estimation is derived.

5.3.1 *Acoustic speaker localization*

The robust bio-inspired speaker localization described in Section 4.2 on pages 54–58 is used since it is robust against reverberation and provides an implicit speech/non-speech decision. Time periods where a speaker is static and robustly localized are identified as periods with a large number of similar estimates as described in Section 4.2.3 on page 58. For each person position with index n and microphone array with index i, the median DoA $\Theta_{n,i}$ with respect to the ground plane is computed.

5.3.2 *Visual person localization*

A visual localization estimates positions with respect to the ground plane. Given a conference setting where the person may be sitting, upper body detections from the camera images are computed with histograms of oriented gradients (HOGs) [DT05]. Especially if visual clutter is present, background subtraction can be used to restrict the visual search area [KB01].

The method from [Bri13] is used here for upper body detection. It uses background subtraction followed by HoG computation and a support vector machine (SVM) detector for each camera image. Thereafter, triangulation is used to determine the person location. Only areas that differ form the background camera image are searched, as the person

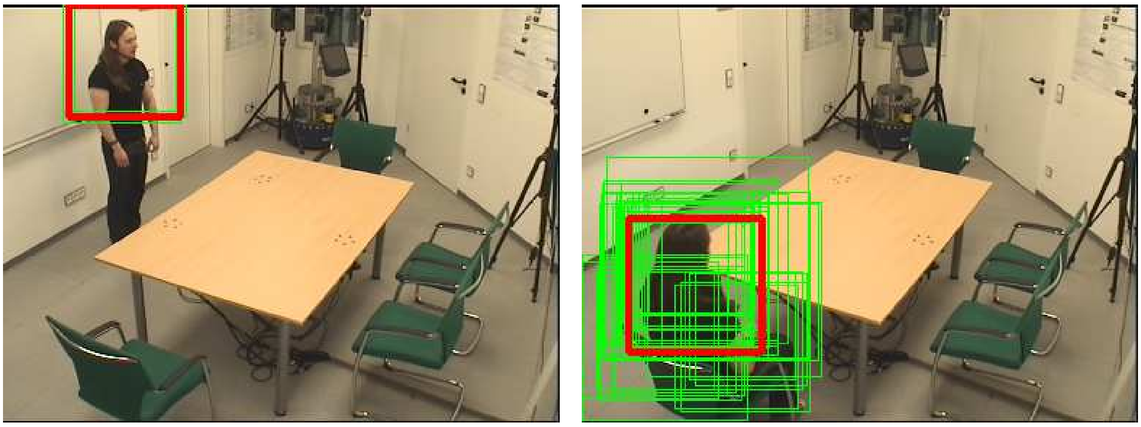

Figure 5.4: Example of upper body detection in the FINCA (cf. on page 91). Individual detections in areas preselected by background subtraction (green) and merged detections (red). Computed using models from [Bri13].

appearing will be different from the background image of the empty room. Within these areas, a sliding window search is done on multiple scales. An SVM classifier is used to detect an upper body in each. After this, overlapping detections are merged into a single one with a simple center of gravity method. Examples for detections and the merging are given in Figure 5.4. The so found detections in the camera images are back projected into the room using the know position and orientation of the cameras. If a person is seen by more than one camera, their Euclidean position is computed by weighted triangulation. The weighting function was derived for the acoustic speaker tracking method developed in this thesis, see page 60. If the person is detected by only one camera, the distance of the person is estimated by assuming the detection window width corresponds to an average shoulder width of 0.5 m. Thus for each position with index n, an absolute two-dimensional localization $\boldsymbol{s}_n$ with respect to the ground plane is estimated.

5.3.3 *Geometry Estimation*

In order to find the geometry automatically, a target function is used that expresses the error of the geometry estimate. Thus the function will reach its minimum output value for the correct geometry as input. It uses the above mentioned measurements and the geometric relation to the speaker position.

For each person position and microphone array, the vector from the source s to the receiver r can be expressed by the unit vector in the DoA as defined in equation Equation (2.13) on page 12 and the distance $k_{n,i}$ as illustrated in Figure 5.5 on the next page:

$$\hat{\boldsymbol{r}}_i = \tilde{\boldsymbol{s}}_n - k_{n,i}\, \boldsymbol{\alpha}(\hat{o}_i + \tilde{\Theta}_{n,i})\ . \tag{5.22}$$

This equation holds when there is no error in the localization data or geometry estimate. We can reformulate these equations to describe the Euclidean error of the geometry estimate $\gamma_i = [\hat{r}_{i1}, \hat{r}_{i2}, \hat{o}_i]$. This error reflects both errors in the position and angular localization as well as the geometry estimate.

$$\varepsilon^{\mathrm{AV}}_{n,i} = \hat{\boldsymbol{r}}_i - \tilde{\boldsymbol{s}}_n + k_{n,i}\boldsymbol{\alpha}\left(\hat{o}_i + \tilde{\Theta}_{n,i}\right) \tag{5.23}$$

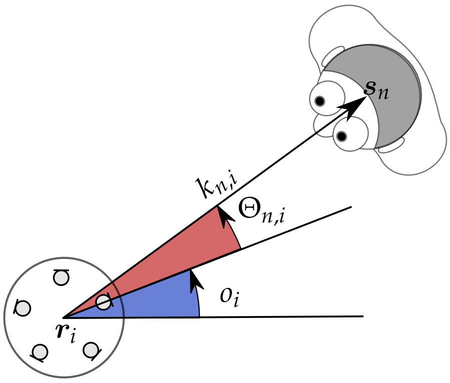

Figure 5.5: Geometric relations between microphone array and speaker. The vector $\boldsymbol{s}_n - \boldsymbol{r}_i$ from the node to the speaker position corresponds to a vector of length $k_{n,i}$ pointing in the direction of the DoA relative to the nodes' absolute orientation $\boldsymbol{\alpha}(o_i + \Theta_{n,i})$. Illustration from [PF14b], available from EURASIP Open Library, © 2013 EURASIP.

It can be seen that $\|\varepsilon_{n,i}\|$ is a convex function with a minimum at the correct geometry. When $\hat{\boldsymbol{r}}_i$ moves away from the true $\boldsymbol{r}_i$, Equation (5.22) no longer holds and the value is strictly increasing with the distance to the true position. Likewise, an wrong value of o_i will rotate the speaker away and thus the function is also strictly increasing for the orientation.

As this equation is under-determined, a set of speaker positions $S \in \mathcal{P}(\{1, 2, \ldots, N\})$ with a fixed number of source positions $J = |S|$ is employed in order to estimate the geometry. We derive a minimization problem stating that the squared error should be minimal for the correct estimates given the measurements for these positions.

$$\mathcal{J}_i^{\mathrm{AV}}(S) = \sqrt{\sum_{n \in S} \left(\varepsilon_{n,i}^{\mathrm{AV}}\right)^2} \tag{5.24}$$

The positions $\tilde{\boldsymbol{s}}_n$ and DoAs $\tilde{\Theta}_{n,i}$ are given by the acoustic and visual localization. The offsets o_i and positions, $\boldsymbol{r}_i$, and the distances $\boldsymbol{k}_i = k_{n,i} | n \in S$ have to be estimated. Hence (5.24) estimates $3 + J$ unknowns with $2J$ equations, and is determined for $J \geq 3$. An estimate for $\boldsymbol{\gamma}_i = (r_{i1}, r_{i2}, o_i)^T$ is computed by minimizing (5.24).

$$\hat{\gamma}_i^{(S)} = \left[\hat{r}_{1i}^{(S)}, \hat{r}_{2i}^{(S)}, \hat{o}_i^{(S)}\right]^T \quad \text{with} \quad \hat{\mathbf{r}}_i^{(S)}, \hat{o}_i^{(S)}, \hat{\mathbf{k}}_i^{(S)} = \operatorname*{argmin}_{\mathbf{r}_i, \hat{o}_i, \hat{\mathbf{k}}} \mathcal{J}_i^{\mathrm{AV}}(S) \tag{5.25}$$

The function is visualized in Figure 5.6. The search space is bounded in the possible orientations $o_i \in [-180°, 180°)$. It can be restricted for the possible positions by the maximum array extension $|r_{ij}| \leq r_{\max}$, and for the speaker distances by the room size $0 < k_{i,t} < k_{\max}$. The function is convex for perfect measurements, as the Euclidean distance will be minimal for the same true position $\mathbf{r}_i$ and orientation o_i. In practice with measurement errors, the different $\varepsilon_{t,i}$ might have minima for different γ_i, and their sum is no longer necessarily convex. However, it is still practical to use the minimum, as this is where the most measurements agree. The common Broyden–Fletcher–Goldfarb–Shanno algorithm for bounded gradient descent is employed to find the geometry with the minimum target function value [BLNZ95].

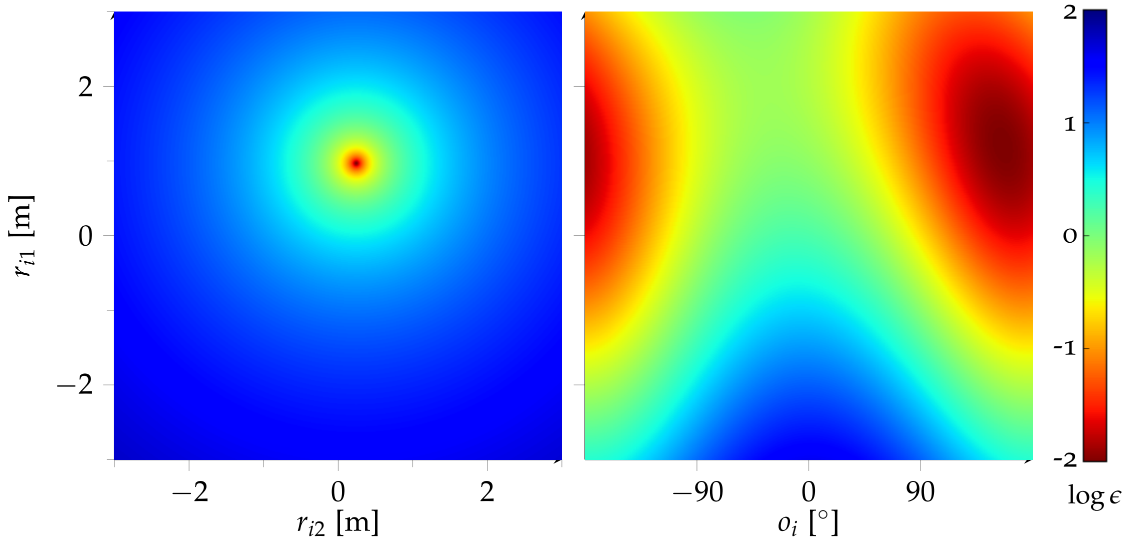

Figure 5.6: Cuts through the seven-dimensional target function for one microphone array and four source positions in a smartroom recording. Values shown in rainbow colors (blue = highest, red = lowest) as parameter of position and orientation, using the best estimated values for the other dimensions.

5.3.4 *Consensus*

The estimates may contain outliers due to errors in the localization or the positions being close to co-linear. Q random sets $\mathcal{S}_J = S_1 \ldots S_S$ of a fixed number of $J \geq 3$ positions are chosen and corresponding geometry estimates are computed for each set.
Over the Q geometry estimates, a refined position estimate is computed as the median two-dimensional position over all individual estimates.

$$\hat{\boldsymbol{r}}_i^{m} = \text{median}\{\hat{\boldsymbol{r}}_i{}^{(S)} | S \in \mathcal{S}\} \tag{5.26}$$

The set $\mathcal{S}'$ of the $Q' = Q/3$ estimates with the smallest Euclidean distance to the median is used to compute an improved estimate. The mean position and orientation of these estimates is the final "consensus" estimate.

$$\hat{\gamma}_i^* = \frac{1}{Q'} \sum_{S \in \mathcal{S}'} \hat{\gamma}_i^{(S)} \tag{5.27}$$

Figure 5.7 summarizes the overall procedure. Both the acoustic and visual recording is done in the calibration sequence where the speaker talks from several static positions. Then, first the acoustic localization identifies the time segments with speech. The speaker is localized for each such segment in the camera images and the position is computed by triangulation. Using both types of measurements, each node can be localized. Several subsets of speaker positions are chosen and a geometry calibration solution is computed for each. Then the median position is computed and the final mean estimate is computed for the consensus set of close estimates.

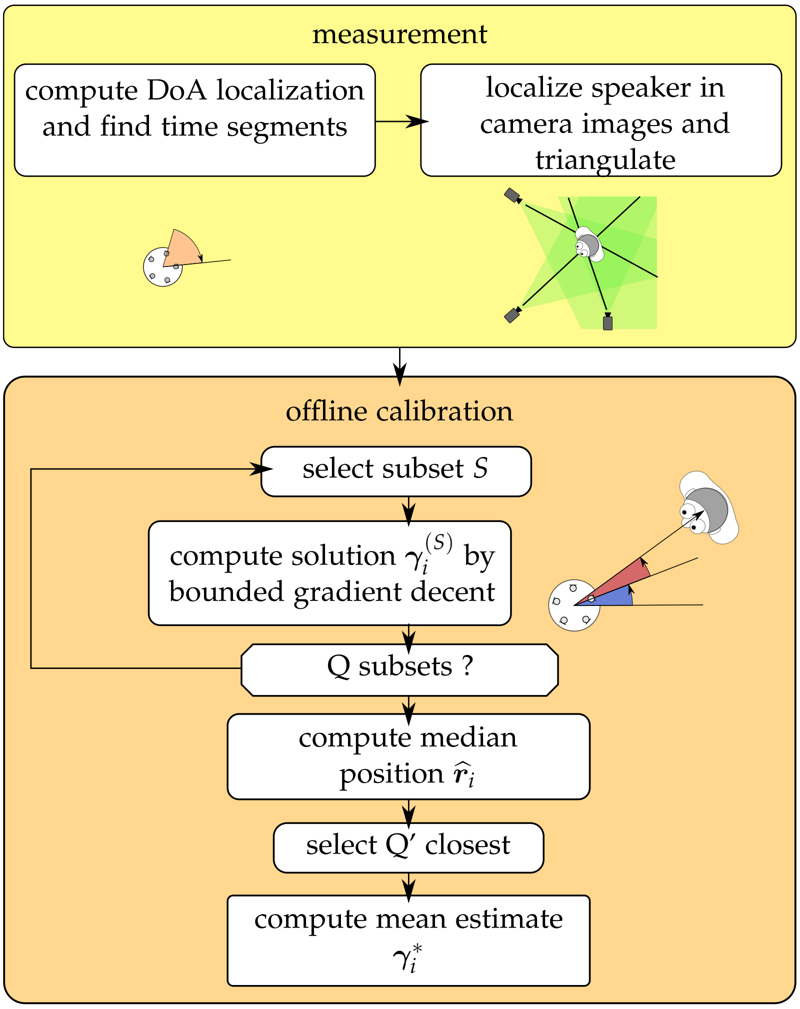
measurement
compute DoA localization and find time segments
localize speaker in camera images and triangulate
offline calibration
select subset S
compute solution $\gamma_i^{(S)}$ by bounded gradient decent
Q subsets ?
compute median position $\hat{\boldsymbol{r}}_i$
select Q′ closest
compute mean estimate γ_i^*

Figure 5.7: Multimodal geometry calibration method to localize a microphone array from angular and visual speaker localizations.

5.4 Proposed Acoustic Methods for Array Configuration Calibration

In order to calibrate the geometry of microphone arrays from sound only, the source positions have to be estimated as well. To solve this problem, additional information regarding the distance between the microphone arrays is used. Additionally to the DoA, the TDoA between the arrays is measured. The sampling of the nodes has to be synchronized by a suitable method, e.g., [MGGC12, PHM13, CG14, SJHU14].

The joint DoA-TDoA optimization was first proposed in [PF14c] and later refined in [PFG17]. Different methods for the measurement and computation of an estimate were proposed. While in the first approach, the TDoA estimation was done using the GCC-PHAT, the methods were later unified by employing the correlation of the peak over average position (PoAP) spikes already computed for the DoA localization. Since the search space is highly non-continuous, exhaustive search was used at first. The computation was quite slow, so this was later replaced by an evolutionary optimization scheme. This enabled real-time application.

As defined in Section 2.1.3 on page 14, R sensor nodes with planar nonlinear microphone arrays are placed at unknown positions r_i and orientations o_i. As there is no anchoring information in this case, only the relative geometry can be estimated. $\boldsymbol{r}_0$ and o_0 are fixed to an arbitrary value and only the others $\boldsymbol{r}_i, o_i$ for $i > 0$ are estimated.

5.4.1 *DoA and TDoA Measurements*

Sounds played or spoken at a set of fixed unknown source positions s_n and received by all microphone arrays. First, DoAs are estimated and the time segments corresponding to the speaker positions are identified. Second, an intra-array TDoA estimate $\tau_{n,(i,j)}$ is computed for each sound event.

As in the multimodal approach, the localization method described in Section 4.2 on pages 54–58 is used that allows to isolate the events and computes the DoA $\Theta_{n,i}$ with respect to the ground plane for each of the sound events with index n and microphone array with index i [PF13]. The events themselves are identified automatically as time segments with low DoA variance as described in Section 4.2.3 on page 58.

The common approach is to find the maximum in the correlation using PHAT weighting (2.20). This was used in [PF14c]. The computation requires large time segments in order to be robust against reverberation [ZFZ08], which leads to a relatively large bandwidth requirement for exchanging the information in a WASN.

An alternative method was introduced in [PFG17]. The sparse spike representation computed in the cochlear model is used for correlation. The band-wise spikes used for the DoA measurement are summed up over the bands and correlated with the data from other nodes. This has two advantages. First, the amount of data to be exchanged is significantly reduced. Second, the measurement is more robust to reverberation.

For symmetrical microphone arrays, the TDoA between two microphone arrays i, j for sound event n can be computed as the average over all microphone pairs between the arrays. When the TDoA estimated from the correlation has a standard deviation exceeding a threshold of, e.g., 100 cm, it is considered to be unreliable. These pairs are not used and their estimates are discarded before computing the mean TDoA. If the number of unreliable pairs is more than half of all the pairs, the speaker position is discarded.

5.4.2 *Estimation*

For each pair i, j of arrays, the source position can be computed by triangulation as explained in 2.1.3 on pages 14–15. When both distances $k_{n,i}$ and $k_{n,j}$ are positive, the

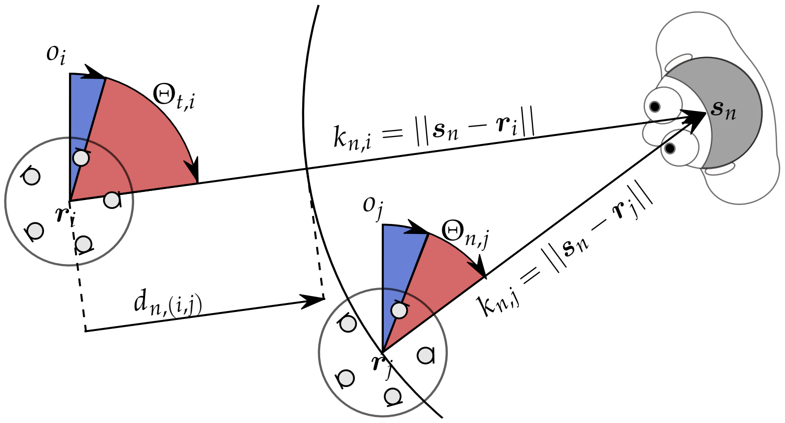

Figure 5.8: Geometric relations between two microphone arrays and a single speaker. The vectors $\boldsymbol{s}_n - \boldsymbol{r}_i$ and $\boldsymbol{s}_n - \boldsymbol{r}_j$ between the nodes i, j and the speakers' position t again correspond to the vectors oriented in the DoAs relative to the nodes positions. The distance difference between the length of the two vectors corresponds to the TDoA $\tau_{n,(i,j)}$ multiplied by the speed of sound c. Illustration first used in [PF14c] ©2014 IEEE.

rays starting at $\hat{\boldsymbol{r}}_i$ in direction $\hat{o}_i + \tilde{\Theta}_{n,i}$ and $\hat{\boldsymbol{r}}_j$ in direction $\hat{o}_j + \tilde{\Theta}_{n,j}$ have an intersection at $\hat{\boldsymbol{s}}_{n,(i,j)}$ computed by equation (2.21), cf. Figure 5.8. In order to get the best possible estimate of the speaker position, a joint estimate over all pairs is computed as the mean of the individual intersections

$$\hat{s}_n = \frac{2}{R(R-1)} \sum_{i<j} \hat{s}_{n,(i,j)} \,. \tag{5.28}$$

As illustrated in Figure 5.8, the difference of the two array to speaker distances corresponds to the projected distance between the microphone arrays. Therefore, the distance inferred by the measured TDoA $\tilde{\tau}_{n,(i,j)}$ should be identical:

$$k_{n,i} - k_{n,j} \approx \tilde{d}_{n,(i,j)} = \tilde{\tau}_{n,(i,j)} c / f_s \,. \tag{5.29}$$

An error of the estimates with respect to the TDoA estimates is computed as difference between the TDoA and triangulation based estimate of the relative microphone array distance in speaker direction. Cases where the two rays do not intersect are penalized by a constant ε_{NI} chosen clearly larger than the possible intra-node distance, e.g., 10 m.

$$\varepsilon_n^{\mathrm{A}\,2} = \sum_{i<j} \begin{cases} \left(||\hat{s}_n - \hat{r}_i|| - ||\hat{s}_n - \hat{r}_j|| - \tilde{d}_{n,(i,j)} \right)^2 & \text{when } k_{n,i} > 0 \wedge k_{n,j} > 0 \\ \varepsilon_{NI}{}^2 & \text{otherwise} \end{cases} \tag{5.30}$$

As in the multimodal case, this equation is under-determined, so again a set of speaker positions $S \in \mathcal{P}(\{1, 2, \ldots, N\})$ with a fixed number of source positions $J = |S|$ is employed in order to estimate the geometry. The target function is therefore again a function of a set of measurements.

$$\mathcal{J}^{\mathrm{A}}(S) = \sum_{n \in S} \varepsilon_n^{\mathrm{A}\,2} \tag{5.31}$$

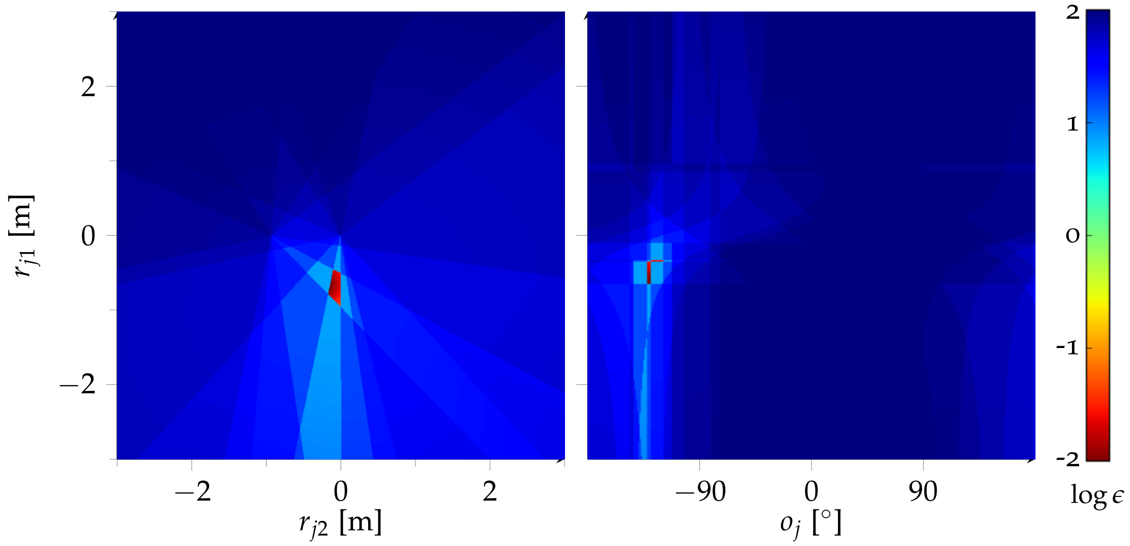

Figure 5.9: Target function values (blue = highest, red = lowest) for one microphone array in the full search space using the best estimated values for the others. The plateaus are due to non-intersections penalized by ϵ_{NI}.

By minimizing (5.30) for a set S of source positions the joint estimate $\hat{\gamma}^{(S)}$ for the whole multi-array configuration is computed:

$$\hat{\gamma}^{(S)} = (\hat{r}_{11}^{(S)}, \hat{r}_{12}^{(S)}, \dots, \hat{r}_{R2}^{(S)}, \hat{o}_1^{(S)}, \dots, \hat{o}_R^{(S)}) \quad \text{with} \quad \hat{\boldsymbol{o}}^{(S)}, \hat{\boldsymbol{r}}^{(S)} = \underset{\boldsymbol{o},\boldsymbol{r}}{\operatorname{argmin}}\, \mathcal{J}^{\mathrm{A}}(S) \tag{5.32}$$

Figure 5.9 shows two cuts through the search space. It is non-convex and non-continuous because of the penalties. Therefore, direct gradient descent is not applicable. Two different optimization strategies were implemented, hierarchical search and evolutionary optimization as described in the next sections.

As the estimation may be biased because of an individual error in measurement or the source positions being close to co-linear, subsets of positions are used as in the multimodal approach. In order to get a better estimate, multiple sets $S \in \mathcal{P}(\{1 \dots T\})$ of a fixed number of positions $J = |S|$ are used. Their average is computed weighted by the reciprocal of the error:

$$\hat{\gamma}^{*} = \left(\sum_{S} \frac{1}{\mathcal{J}^{\mathrm{A}}(S)} \right)^{-1} \left(\sum_{S} \frac{\hat{\gamma}^{(S)},}{\mathcal{J}^{\mathrm{A}}(S)} \right) \tag{5.33}$$

5.4.3 *Hierarchical search*

The method proposed in [PF14c] applies exhaustive search in order to find the geometry best fitting a selected subset. The search space is bounded in the possible orientations $o_i \in [-180, 180)$ and can be restricted for the possible positions by the maximum array extension $|r_{ij}| \leq r_{\max}$. This space is divided into an equidistant grid of, e.g., 10 cm and 2°. As this is computationally heavy, a hierarchical approach was devised. First, a solution for pairs $(0, i)$ of microphone arrays is found by exhaustive grid search. The solution is refined by bounded gradient descent [BLNZ95]. The so found solutions for all arrays are used as a starting point for a second gradient descent optimizing a joint solution

for all arrays. Given the non-convex property of the search space, the so found solution might still lie outside the convex target area. In order to exclude these cases, a solution is discarded when the target function value is too large, e.g. $\epsilon_S > 10\,\mathrm{cm}$.
This is repeated for several subsets until a predefined number N of subset solutions is present. Then the overall solution is computed using Equation (5.33), cf. Figure 5.10.

5.4.4 *Evolutionary optimization*

The strategy was later replaced by an online method [PFG17]. By the use of a genetic algorithm, it is possible to compute a solution for the full configuration in real time. A differential evolutionary algorithm with a binomial distribution mutation of the best member of each generation is used [SP97].
A population of, e.g. 25, candidate solutions $\gamma^{(i)}$ is iteratively optimized over several generations. They are initialized with random values and evolutionary optimized by iterating a mutation step followed by a selection step over several generations. In the mutation step, trial candidates ρ are generated by mutating the member γ^* with the best fitness value ε^*. The individual values are mutated using two members j, k of the current generation that are chosen by a binomially distributed random variable Z. The difference is weighted by a mutation factor ν and added to the value of γ^*. The individual values of γ^i are replaced if the value of Z exceeds a given crossover threshold ϑ_{CR}.

$$\gamma_d = \begin{cases} \gamma^* + \nu(\gamma_d^{(j)} - \gamma_d^{(k)}) & \text{if } Z > \vartheta_{CR} \\ \gamma_d^{(i)} & \text{otherwise} \end{cases} \tag{5.34}$$

In the selection step, the trial candidate replaces the current one if its fitness value is better. If it is better than the best member, it also replaces that one.

$$\gamma^{(i)} \leftarrow \boldsymbol{\rho} \quad \text{if } \varepsilon(\boldsymbol{\rho}) < \varepsilon(\gamma^{(i)}) \tag{5.35}$$

$$\gamma^* \leftarrow \boldsymbol{\rho} \quad \text{if } \varepsilon(\boldsymbol{\rho}) < \varepsilon^* \tag{5.36}$$

Once the population converges to a set with low variance, the optimization terminates. This method converges fast to an accurate solution for all nodes. This allows for distributed online computation of the geometry as illustrated in Figure 5.11.
Whenever there is a speech event, the DoA is computed by the node. The result is broadcast with the corresponding spike segment to all other nodes. Upon receiving this, each node computes the TDoA to all other nodes. The result is subsequently broadcast to all other nodes.
While there is no speech, each node selects subsets of the speech events received so far and computes a geometry estimate using the differential evolution algorithm. The result is broadcast to the other nodes. The subset estimates computed by all nodes are then used to update the weighted mean using Equation (5.33), providing the current overall estimate.

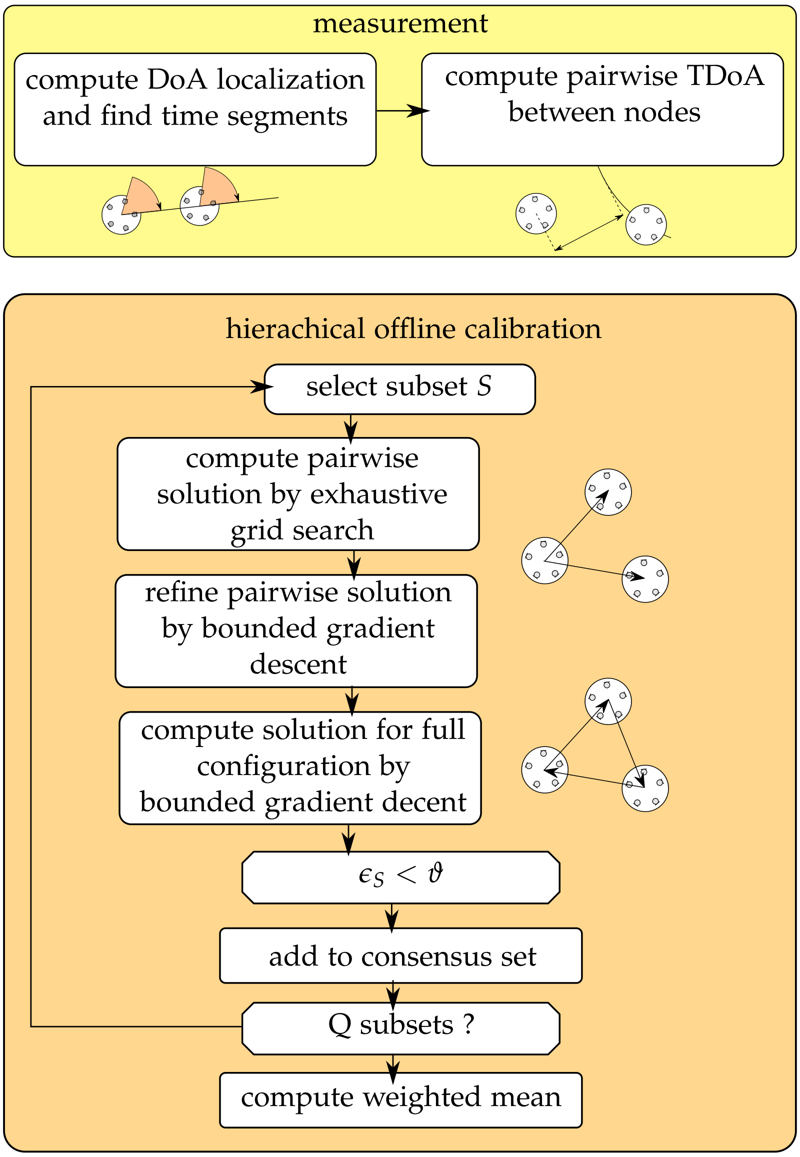
measurement
compute DoA localization and find time segments
compute pairwise TDoA between nodes
hierachical offline calibration
select subset S
compute pairwise solution by exhaustive grid search
refine pairwise solution by bounded gradient descent
compute solution for full configuration by bounded gradient decent
$\epsilon_S < \vartheta$
add to consensus set
Q subsets ?
compute weighted mean

Figure 5.10: Acoustic offline calibration method using exhaustive search. First a sequence is recorded where a speaker talks from a number of static positions. In the recording, speech segments corresponding to the individual positions are identified by the DoA localization. For each position, the TDoA between the nodes is measured by correlation of the signals from pairs of nodes. With these two measurements, the calibration is computed. Random subsets of positions are chosen. For each an individual solution is computed in three steps: For pairs of microphone arrays, a solution is found by exhaustive search in a fixed grid. These are then refined by gradient descent. Then a joint estimate over all nodes is computed by bounded gradient descent. When estimates for Q subsets are found, the overall estimate is computed as weighted mean.

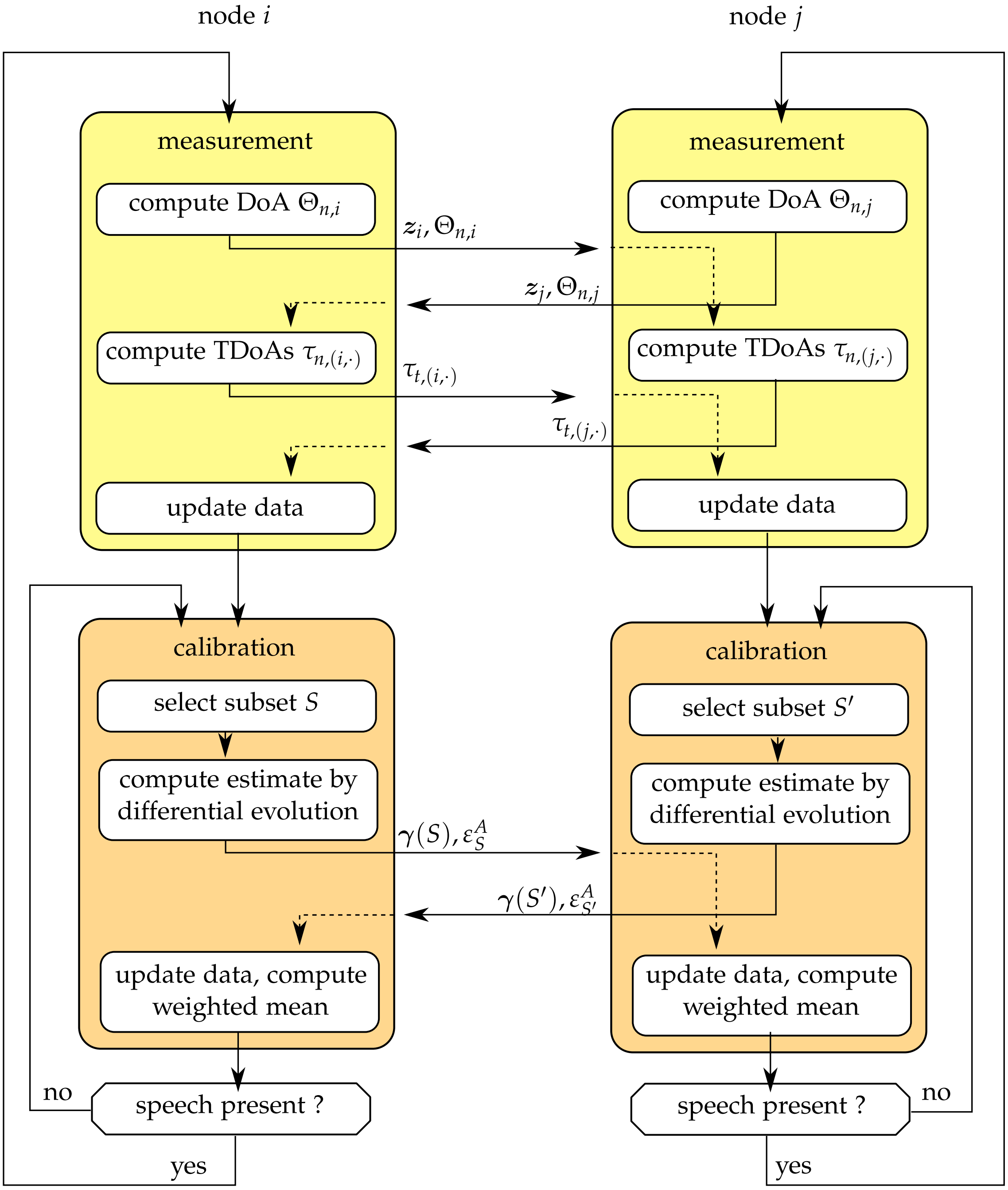
node i
node j
measurement
compute DoA $\Theta_{n,i}$
compute TDoAs $\tau_{n,(i,\cdot)}$
update data
measurement
compute DoA $\Theta_{n,j}$
compute TDoAs $\tau_{n,(j,\cdot)}$
update data
$z_i, \Theta_{n,i}$
$z_j, \Theta_{n,j}$
$\tau_{t,(i,\cdot)}$
$\tau_{t,(j,\cdot)}$
calibration
select subset S
compute estimate by differential evolution
update data, compute weighted mean
calibration
select subset S'
compute estimate by differential evolution
update data, compute weighted mean
$\gamma(S), \varepsilon_S^A$
$\gamma(S'), \varepsilon_{S'}^A$
speech present ?
no
yes
speech present ?
no
yes

Figure 5.11: Acoustic online calibration method using a genetic algorithm and distributed computation. When speech occurs, the DoA is computed in each node. The spike representation of the microphone signals is shared with the others in order to compute TDoAs. After that, at each node several random subsets of positions are used to compute overall estimates by differential evolution. These can again be shared in order to compute a better weighted mean estimate. Illustration based on [PFG17] © 2017 IEEE

5.5 SUMMARY

Three types of scenarios to be calibrated were introduced to organize the state-of-the-art: array shape, distributed microphones and distributed microphone configurations. Different methods were presented for each. They were grouped by the constraints they impose on the scenario. It was distinguished between the use of active devices, dedicated sound sources and ambient sounds such as speech. Table 5.1 lists the main methods according to these two criteria.

For the array shape calibration, only the diffuse noise based MDS estimation is used. It achieves around 1 cm accuracy if the spatial diffuseness is good enough.

For the geometry calibration of distributed microphones, a multitude of methods has been introduced. Active devices that each play a chirp can be calibrated by MDS with subsequent optimization based on the device geometry. The devices do not need to be synchronized. About 7 cm accuracy was achieved in practical experiments with laptops and smartphones. With calibrations sounds, an SVD-based approach can be employed. By playback of a known sequence, the need for synchronization was eliminated. About 3 cm accuracy was achieved by playing a chirp sequence from a moving smartphone. The use of the mTDoA allows to compute the geometry from ambient speech sources. This approach requires the speakers to be near the endfire positions at some point. About 10 cm accuracy was achieved with smartphones and laptops.

The newer scenario of distributed microphone arrays requires the estimation of the nodes' position as well as the orientation. Active devices can be calibrated using MDS and a subsequent SVD orientation alignment. Around 1 cm and 6° accuracy were achieved with laboratory mockup devices. For passive calibration, a single speaker walking on a random trajectory for about 5 minutes was used. With RANSAC estimation, about 25 cm were achieved.

The proposed methods allow for automated calibration from speech events form a number of static positions. The multi-modal approach uses cameras at known positions for person detection and subsequently aligns the nodes. The acoustic methods employs both DoA and TDoA measurements computed by a neuro-biologically inspired model. The two measurements are combined in a joint objective function. By using evolutionary optimization, this function can be minimized online. As will be shown in the evaluation on pages 123–135, the accuracy achieved by the proposed methods is better than the state-of-the-art.

	active	sound sources	speech
array shape	–	diffuse noise → EVD • MDS [Bir03, ML08] • base-point MDS [BS05]	–
mic. config	MDS + ML • laptops [RD04] • smartphones [HF11]	ToA → SVD • known emission [CDM12] • known sequence [GKH14]	max TDoA • introduction [PMH12] • + data assoc. [PPH14]
array config	MDS + SVD orientation • mockups [PMH11]		SRP-PHAT • trajectory [HPFHU09] DoA + RANSAC • trajectory [SJHU+11] • +cost. [JSHU12, JSHU13] Proposed • off-line [PF14c] • online [PFG17]

Table 5.1: State-of-the-art acoustic geometry calibration approaches ordered by scenario and measurement method.

Nature is relentless and unchangeable, and it is indifferent as to whether its hidden reasons and actions are understandable to man or not.

Galileo Galilei

6 EVALUATION

The methods were evaluated in simulations and many real life scenarios. In the beginning, an overview of the simulations in as well as the recorded datasets used in will be given in Section 6.1 and Section 6.2, followed by description of the methodology in Section 6.3. Then, the different tasks will be investigated. We begin with the acoustic event detection, as it can be used as pre-filter for the other methods in Section 6.4. Second, the blind speech enhancement based on it in Section 6.5. Third, the single node speech localization will be focused on in Section 6.6. It is the basis for the sensor network methods. Fourth, the different proposed methods for geometry calibration of sensor nodes with multiple microphones will be thoroughly evaluated in simulation and with real recordings in Section 6.7. Fifth, the Euclidean speaker tracking in the so-calibrated network will be tested in Section. 6.8. Finally, event detection, calibration and tracking methods will be used in a combined experiment on recordings of natural speakers in a reverberant conference room in Section 6.9.

6.1 SIMULATIONS

Two types of simulations were performed in order to assess the properties of the proposed methods. In order to work on defined artificial microphone signals, acoustic room simulations with synthetic room impulse responses (RIRs) were performed. To systematically investigate the effect of defined measurement errors, randomly generated errors were added to the ground truth.

6.1.1 *Acoustic room simulation*

Simulations of the acoustic wave propagations were done using the image source model (ISM) [AB79]. The room is approximated by a "shoebox" of six walls enclosing the source and receivers. The wave propagation is approximated by assuming a spherical waves radiating from the source. When the signal reaches a wall, it is reflected and attenuated by multiplication with a reflection coefficient.

The simulation was done using MATALB code released into the public domain by Eric Lehmann [LJN07].[1] The code provides a "fast" version that approximates the tail of the room impulse response. This option did not provide comparable results with the actual recordings. Therefore, the "non-fast" version of the simulation was used, where the tail is computed fully.

[1] `http://www.eric-lehmann.com/ism_code.html`

6.1.2 *Measurement errors*

Especially for the sensor localization, the question of the influence of measurement errors was investigated. For a given root mean square (RMS) ϵ, individual errors were generated by a zero mean Gaussian distribution. Theses were then added to the ground truth values in order to generate the erroneous measurement.

$$\tilde{\theta}_i = \theta_i + e_i \quad \text{with} \quad e_i \sim \mathcal{N}(0, \epsilon) \quad \text{for} \quad i = 1, \ldots, I \tag{6.1}$$

In the case of two dimensional measurements, pairs of values drawn from a normal distribution were used:

$$\tilde{\boldsymbol{s}}_n = \boldsymbol{s}_n + [e_{2n-1}, e_{2n}]^T \quad \text{with} \quad e_i \sim \mathcal{N}\left(0, \frac{\epsilon}{\sqrt{2}}\right) \quad \text{for} \quad i = 1, \ldots, 2N \tag{6.2}$$

Multiple Monte Carlo simulations were done with this method. In all cases, the average RMS of 100 Monte Carlo runs was close to the intended ϵ, with an error of below 30%.

6.2 RECORDINGS

For the evaluation of both sensor geometry calibration and speaker tracking, several recordings with one or more natural speakers were made in the FINCA smartroom. To test the acoustic event classification and detection performance, a dedicated set of recordings was done there as well. During a research visit to Israel, dedicated constructed smart phone mockups were used for recordings. In order to compare to published results from other researchers, the AV16.3 corpus and the publicly available D-CASE dataset from the IEEE AASP challenge in 2013 was used [GSB+13].

6.2.1 *FINCA Dortmund, Germany*

The experiments were in large part done in a conference room setup in a laboratory at the robotics research institute at TU Dortmund university, called the "FINCA", before it was dismantled in September 2014. The room is roughly 3.5 × 6.5 × 2.4 meters in size and has a clipped edge near the door, see Figure 6.1. It was equipped with microphones and cameras in the following configuration:

The installation featured three m-Audio Delta1010 soundcards. They were synchronized by wired clock connection, recordings of coherent white noise showed a remaining jitter of 22 μs between the sound cards. Behringer ECM8000 microphones were placed in a table as three uniform circular microphone arrays with five microphones each. The microphones from each microphone array were connected to an individual soundcard, capturing the signals at $f_s = 48\,\text{kHz}$. A reverberation time of $670 \pm 89\,\text{ms}$ over the microphone signals was calculated using an estimation algorithm [LYJV10].

Three circular microphone arrays with five microphones in a circle with 5 cm radius were embedded in a table as shown in Figure. 6.1. They were placed in a non-symmetrical triangle with about 1m edges.

Up to five Sony EVI-D70P cameras were used in the experiments. Their field of view (FOV) is 48°. They were connected to a PAL framegrabber that delivered 388x284 images with up to 30 fps. In the installation, four cameras were mounted at the ceiling as shown in Figure 6.1 (right). Since the coverage of the room is limited, an additional camera was set on a tripod in the room as shown in Figure. 6.1 (right) in gray. With this configuration,

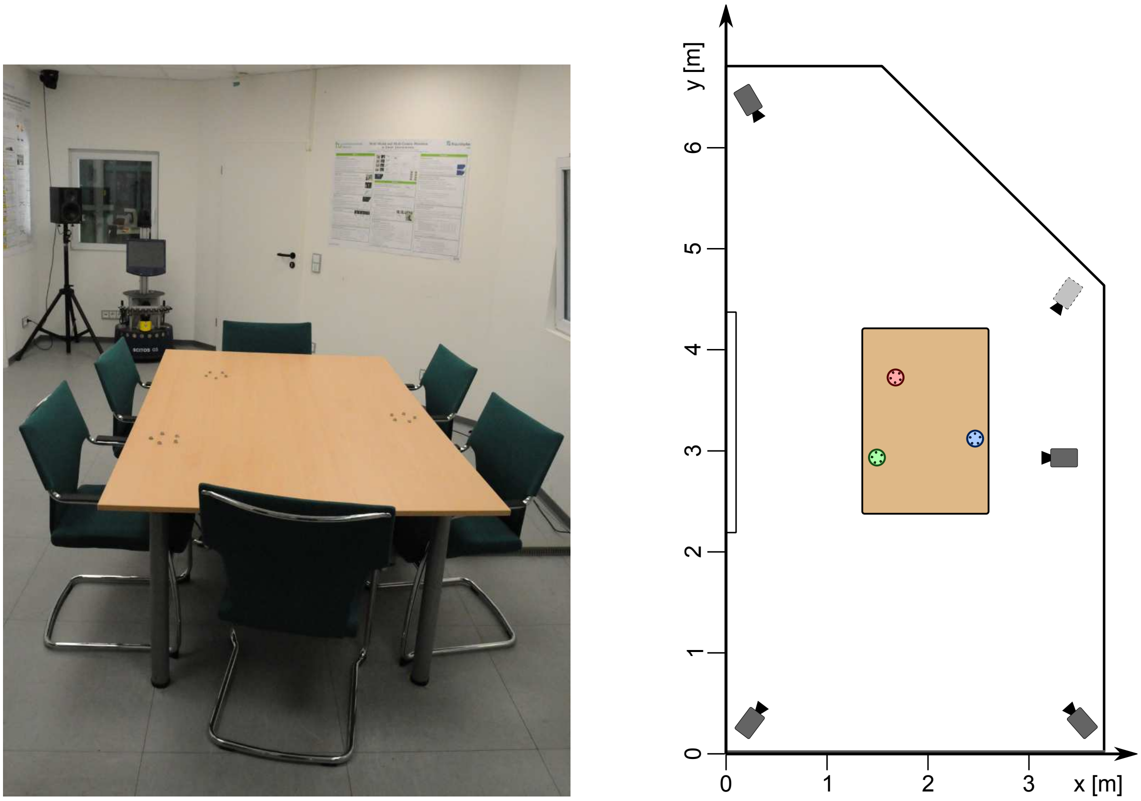

Figure 6.1: Table with three embedded circular microphone arrays in the FINCA (left) Camera and microphone array positions (right).

#	nodes	cams	description
#1	3	–	A mobile phone playing white noise at table height, using the same positions as in #2.
#2	3	5	One speaker speaking at 10 static positions sitting and standing around the table.
#3	3	5	One speaker speaking at 15 static positions sitting and standing around the table.
#4	3	5	One speaker speaking at 19 static positions sitting and standing around the table.
#5	3	4	Two speakers taking up six positions and talking alternately.
#6	3	4	Two speakers taking up six positions and talking alternately.
#7	3	4	Four speakers standing and sitting in various positions while discussing research topics.
#8	3	4	Three concurrent speakers.
#9	3	4	Two speakers discussing.

Table 6.1: Recordings of natural speakers in the FINCA used for the evaluation of tracking, speaker identification and geometry calibration.

the capture of a person around the table by at least two cameras is ensured for most positions.

Table 6.1 lists the recordings made. For off-line calibration and testing of the localization, several sequences were recorded where one or more speakers were taking up a number of fixed positions in the room and uttering a few sentences at each. To test the limits towards concurrent speakers, three speakers were doing their best to talk simultaneously in sequence #8. Additionally, some unconstrained recordings were made. #9 is a natural discussion between two speakers to show the natural overlap. In #7, several persons were discussing and changing positions at will.

6.2.2 *AV16.3 Dataset*

For comparison with the literature, the AV16.3 dataset was used [LOGP05]. Here two uniform circular microphone arrays with eight microphones and a diameter of 20 cm are placed on a table in a mildly reverberant conference room ($T_{60} \approx 0.5\,\mathrm{s}$). The sequences were recorded in parallel with three cameras. Speakers were wearing colored balls on their heads in order to get position data by visual triangulation. The recordings were done at 16 kHz sampling rate and 16 bit resolution.

6.2.3 *FINCA AED Dataset*

The FINCA acoustic event detection dataset was recorded in the smartroom using one microphone embedded in the table. The dataset was published along with the paper detailing the bag-of-features (BoF) approach to acoustic event detection [PGF14]. The following eleven sound event classes used are listed in Table 6.2. For creating independent training and test sets, each sound class was produced twice by a different person on a different day. Each recording was longer than 60s. Additionally, two scripted recordings containing a large portion of the sound events were created.

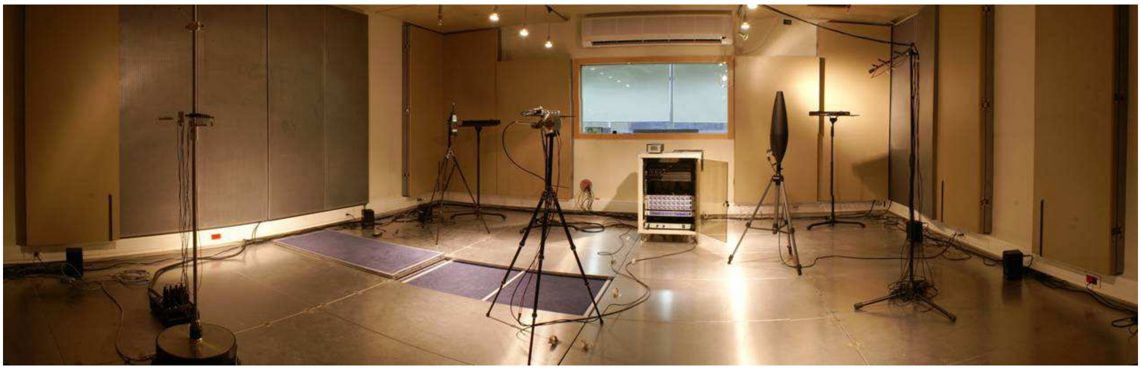

Figure 6.2: Acoustic laboratory at Bar Ilan University.

6.2.4 *D-CASE AED dataset*

The D-CASE 2013 challenge was a public competition in acoustic event and scene recognition [GSB+13]. The acoustic event dataset is comprised of the 16 event classes listed in table 6.3 on the following page.

For each class, 20 training examples of varying length are included. The public development dataset included three 'skript' sequences of events occurring in an office environment. Two sets of annotations were provided for all recordings.

The results of the challenge were published at the WASPAA conference in October 2013 [SGB+15]. The proposed methods were not taking part in the challenge, but the results on the D-CASE development dataset were published in the ICASSP conference in 2014 [PGF14].

6.2.5 *BIU Ramat Gan, Israel*

Another set of recordings was made at Bar Ilan University in the speech and signal processing laboratories at the faculty of engineering [HHVG14]. The lab is about 6 × 6 × 2.5 m in size. It is acoustically isolated from the environment. The lab is equipped with acoustic panels on the walls and ceiling that allow to change the reverberation properties of the room, c.f. Figure 6.2. The panels are reflective on one side and absorbent on the other and can be flipped so that either side faces the room. Up to 64 microphone signals can be recorded simultaneously with synchronized sampling equipment. At the same time, multiple signals can be played from loudspeakers.

Speech enhancement dataset

For experiments in speech enhancement, the room's T_{60} was adjusted to 320 ms by opening and closing specific panels. Special smartphone mockups were constructed by the university workshop to simulate future generation smartphones. The mockup consists of a plastic body and four microphones mounted near the edges. They can be mounted in an 12x8 cm rectangular pattern as shown in Figure 6.3. The mockup was placed in the center of the room. Three loudspeakers were placed around it as shown in Figure 6.4. The desired speaker was placed at 1.2 m distance at $-30°$ in position s_1. Two interfering speakers were placed at 2.0 m distance at $45°$ and $135°$, at position s_2 and s_3 respectively. The recordings were executed with 48 kHz sampling rate and 24 bit resolution.

Several sound samples were played back from the speakers. Noise samples from the NOISEX-92 database were used [VS93]. In order to add some more realistic office

sound	description
silence	No event happening but background noise from the room and electrical noise from the recording equipment.
door	Opening the door of the smart room.
steps	A person walking around.
chairs	Pulling one of the chairs.
rolling	Moving a rolling chair mounted on small wheels.
paper	Turning pages of a multi-page printout or knocking a stack of papers on the table.
keyboard	Typing on a keyboard on the table with heavy pressure keys.
laptopkeys	Typing on a laptop keyboard with soft keys.
speech	A single person talking in the room.
cups	Moving around cups on the table or lifting them up and setting them down again.
pouring	Pouring a liquid from a can in the cups.

Table 6.2: Sounds in the FINCA AED dataset

sound	description
alert	a short alert (beep) sound.
clearthroat	someone clearing his throat.
cough	a person coughing.
doorslam	a door slammed shut.
drawer	the opening of a desk drawer.
keyboard	keyboard clicks.
keys	keys put on a table.
knock	someone knocking on a door.
laughter	a person laughing.
mouse	a computer mouse click.
pageturn	turning a page in a book or printout.
pendrop	a pen, pencil, or marker touching table surfaces.
phone	a phone ringing
printer	an office printer working.
speech	a person speaking.
switch	a very soft and short click made by a lightswitch.

Table 6.3: Sounds in the D-CASE AED dataset

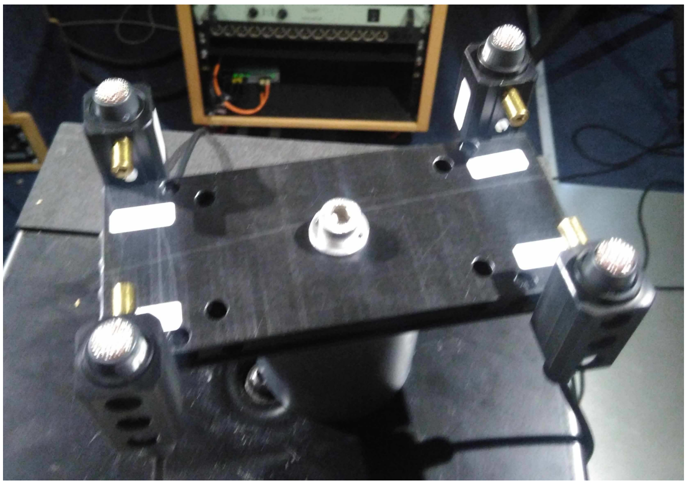

Figure 6.3: Smartphone mockup comprised of four microphone mounts attachable to a plastic body.

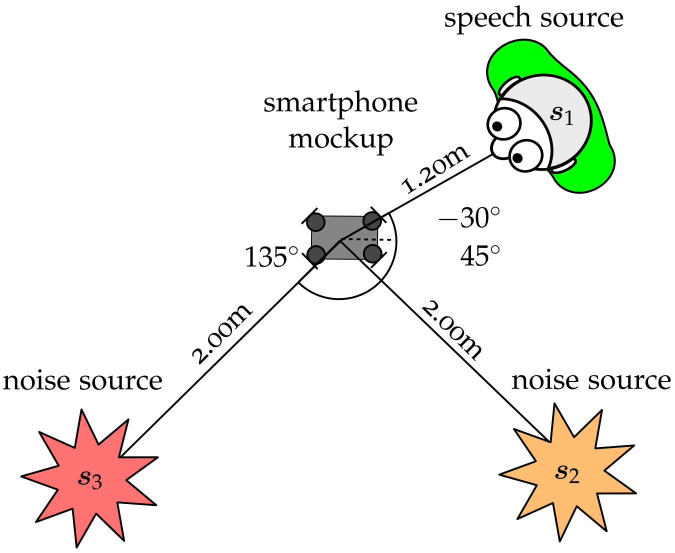

Figure 6.4: Recording setup for speech enhancement with one smartphone mockup at BIU.

and work noises, additional samples were extracted from the freesound database[2]. The sound classes were the following:

Ω_0 SPEECH Playback of anechoic speech recordings

Ω_1 STATIONARY NOISE 'white' and 'pink' noise from NOISEX-92, as well as 'roaring fan', a rather loud humming fan, and 'ventilation' air conditioning noise from the freesound database

Ω_2 MECHANICAL NOISE 'factory1' and 'factory2' from the NOISEX-92 database

Ω_3 BABBLE NOISE from the NOISEX-92 database

Ω_4 NONSTATIONARY NOISE constant keyboard typing from freesound

To generate the test data, speech played from s_1 was mixed with a single noise played from s_2 or s_3 and with two different noises played from s_2 and s_3 simultaneously. Two different anechoic speech sequences from the same speaker at s_1 were played. In each sequence, there are four speech segments of 2-4 s. Overall, they were 18.5 s and 16.5 s long, where speech is present half of the total time. For a single noise test, each noise was played individually from each of the noise speaker positions $s_{2,3}$ and added to each speech sequence at signal-to-noise ratios (SNRs) of 0,6, and 12 dB. For mixed noise testing, sequences with two different noises were generated. Two different noise samples played from the speaker at s_2 and s_3 were added simultaneously to the speech.
The classifier was trained with data from a different recording session using the same mockup, the speakers were placed at slightly different positions. A 45 s long anechoic speech sequence was used. The different noise samples were played for up to 120 s.

6.3 METRICS AND REPRESENTATION

Throughout the evaluation, the main goal is not only to list quantitative results, but to present them in an adequate way and to apply the appropriate statistical tests. One major concern is that many of the quantities under investigation are not justifiably modeled by a predefined statistical distribution. To portray or test them as such would therefore be inadequate. The charts and statistical tests employed were therefore chosen to avoid introducing any unjustified model assumptions. They will be described in this section.

6.3.1 *Classification metrics*

Every correctly detected event is counted as true positive (TP). A detection of wrong class as false positive (FP), a missed event as false negative (FN). Then precision P and recall R can be defined along with the f-score F in the usual way:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F = \frac{2PR}{P + R}. \tag{6.3}$$

For the event detection performance, the non-event class Ω_0 is excluded in the counts. The metrics are evaluated frame-based and class-based, for the latter all classes are evaluated individually and the average is computed.

[2] http://www.freesound.org

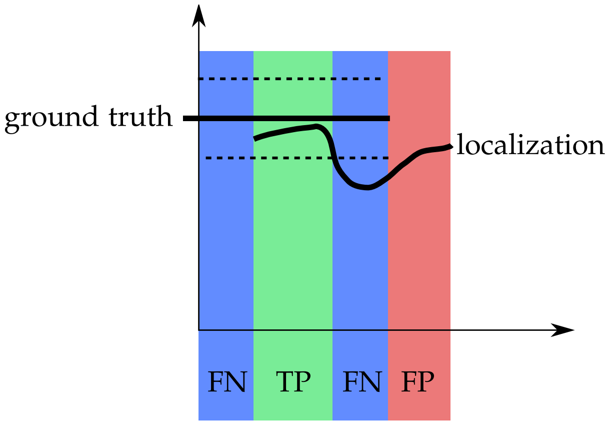

Figure 6.5: Evaluation of tracking results

6.3.2 *Localization and tracking evaluation*

The most direct measure for the error of a location estimate is the mean absolute difference between ground truth and measurement. In the case of vectorial values, such as two dimensional coordinates, the mean of the Euclidean distance for each measurement is used. In the case of multiple detections and/or ground truth values, the pairs that minimize the metric are used as in common tracking measures like the Wasserstein or OSPA distance [RVCV11]. This is computed by testing all permutations of the detections against the ground truth and choosing the one that minimizes the error.

The evaluation of the tracking as a detection task was performed as illustrated in Figure 6.5. A practical maximum distance between the ground truth and the localization was set as a threshold, this is 15° or 0.5 m. These values were chosen with respect to practical applications such as camera control, where larger values would lead to steering the camera to a position where the speakers' face is not visible. Whenever there is a localization within that margin, it is counted as true detection or TP. If it is further away, the frame is counted as FN or missed detection. The same is done in the case that there is no detection at all. When there is a detection but no ground truth value, the frame is counted as FP or false alarm. Thereafter, precision and recall are calculated as in Equation (6.3).

6.3.3 *Geometry estimate evaluation*

The geometry estimates from array configuration calibration methods can exhibit an arbitrary mirroring, rotation and translation to the reference geometry. In order to relate the estimate $\hat{\boldsymbol{M}}$ to the true geometry, the translation and rotation is estimated using singular value decomposition (svd). First, we subtract the center $\overline{\boldsymbol{m}} = \frac{1}{N}\sum_i \boldsymbol{m}_i$ from the microphone coordinates, to use the matrix $\hat{\boldsymbol{M}}'$ of centered positions $\hat{\boldsymbol{m}}_i' = \hat{\boldsymbol{m}}_i - \overline{\boldsymbol{m}}$ to compute the dispersion matrix $\boldsymbol{O}$. Its singular value decomposition yields the rotation and mirroring matrix $\boldsymbol{R}$. Using this, the translation $\boldsymbol{v}$ with respect to the true positions is computed, which allows to compute the aligned sensor positions $\hat{\mathbf{m}}_i''$.

$$\boldsymbol{O} = \frac{1}{N}\boldsymbol{M}'^{T}\hat{\boldsymbol{M}}' \tag{6.4}$$

$$\boldsymbol{R} = \boldsymbol{J}\boldsymbol{L}^{T} \quad \text{with} \quad \boldsymbol{J}\boldsymbol{K}\boldsymbol{L} = \mathrm{svd}(\boldsymbol{O}) \tag{6.5}$$

$$\boldsymbol{v} = \boldsymbol{M}' - \boldsymbol{R}\hat{\boldsymbol{M}}' \tag{6.6}$$

$$\hat{\boldsymbol{m}}_i'' = \mathbf{R}\hat{\boldsymbol{m}}_i + \boldsymbol{v} \; . \tag{6.7}$$

The error ϵ_p is then computed as Euclidean distance of the estimated positions $\hat{\mathbf{m}}_i''$ to the true positions $\hat{\mathbf{m}}_i$.
The position error e_r for inter-array calibration is computed with the same procedure as Euclidean distance of the aligned array positions $\hat{\boldsymbol{r}}_i''$ to the ground truth $\boldsymbol{r}_i$ over all R nodes. The overall orientation error e_o is computed relative to the optimal rotation d_o.

$$e_o = \left\{ |o_r - (\hat{o}_r - d_o)| |\forall_g \right\} \text{ with } d_o = -180^\circ + \frac{1}{R} \sum_{r=1}^{R} (o_r - \hat{o}_r + 180^\circ) \bmod 360^\circ \tag{6.8}$$

6.3.4 *Speech enhancement evaluation*

In order to asses the improvement of speech while considering the distortion introduced by the processing method, the overall output SNR is often found not to be a good indicator [LK11]. The frequency weighted segmental SNR (fwSNRseg) [HL08] was found to be the best objective measure reflecting subjective listening quality [KDG$^+$16]. The idea of this measure is to weigh the differences of the processed signal $\hat{x}$ to the clean signal x weighted with respect to the energy in the clean signal. The signal is decomposed by a Mel-filterbank into $B = 32$ bands. The magnitude spectrum in each band is taken to the power of $\gamma = 0.2$ to produce the corresponding weight $W(b,k)$ for the band b and the current time frame k. Then the difference in SNR is computed for each time frame by summing the weighted difference per band:

$$\begin{aligned} w(x,\hat{x}) = & \frac{10}{K} \sum_{k=1}^{K} \frac{\sum_{b=1}^{B} W(b,k) \left(\log_{10} |x(b,k)|^2 - \log_{10} (|x(b,k) - \hat{x}(b,k)|)^2 \right)}{\sum_{b=1}^{B} W(b,k)} \\ & \text{with } W(b,k) = |x(b,k)|^\gamma \end{aligned} \tag{6.9}$$

The relative improvement is computed as a difference in the fwSNRseg w between the input and output. The output fwSNRseg is computed between the speech signal as received by the reference microphone used in the estimation, e.g., the first microphone x_1 to the processed output $\hat{x}$, $w(x_1, \hat{x})$. The input fwSNRseg is computed between the speech only signal on the reference microphone s_1 and the mixed input signal y_1 as $w(x_1, y_1)$. The improvement is then computed as

$$\Delta w = w\left(x_1, \hat{x}\right) - w\left(x_1, y_1\right) . \tag{6.10}$$

6.3.5 *Significance*

A significance test is used to determine how certain one can be that the results are unlikely to have occurred by chance alone. The significance level p is the probability with which this assumption, referred to as the $\mathcal{H}_0$ hypothesis, is wrong. Historically, some process with a p-value of below 0.05 (or 5%) was called "significant" by R. A. Fisher in the first applications of significance testing. This is practiced by many researchers up to date. Since allowing an 1:20 error is not uncontroversial, a threshold of $p < 0.01$ is advocated by many researchers. Likewise, the term "highly significant" is not well defined, it is used for $p < 0.01$, $p < 0.005$ or $p < 0.001$ throughout the literature. Within this text, the p value will be given alongside the word significant to be precise. The tests will aim for the highest significance level of $p < 0.001$.

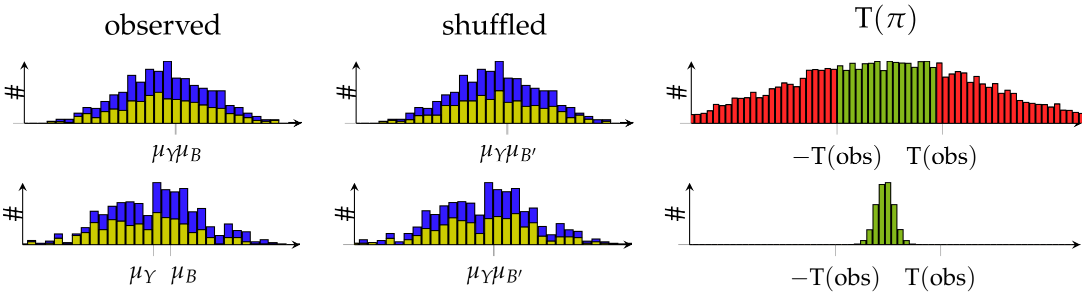

Figure 6.6: Visualization of the randomization test principle: Histogram of two data series with 500 elements (left), random shuffling of labels (middle) and 10,000 random permutations' differences of means (right). In the histograms, permutations will only change the relative part of the two quantities yellow and blue of the distribution in each histogram bin while not changing the overall distribution or the total amount of blue and yellow. It is unlikely to move the means farther apart than they are initially in the bottom case, while it is simple in the top case. The relative number of differences of means larger than the initially observed $\mathrm{T}(\mathrm{obs}) = \mu_Y - \mu_B$ shown in red approximates the significance level p. If it is large (top, $p = 0.74$), Y and B are not significantly different, if it is small (bottom, $p < 0.001$), they are.

The significance testing will aim at determining whether one method or set of parameters is superior to another, or if the difference in results could be attributed to chance. For many of the metrics compared, it would be unreasonable to assume a given statistical distribution. There is a straightforward technique of testing on significance that does not require any model assumption. This is the class of randomization or permutation testing, which will be described in the following subsection.

6.3.6 *Randomization test*

The randomization test can determine the significance level on any two sets of numbers by shuffling them and comparing the difference of means (i.e. expected values) before and after the shuffling [OG10]. Imagine the numbers being some performance measurement results from two different algorithms on the same input data. This can be anything like, e.g., recognition rate, error rate, or execution time. Consider the numbers as a set with the labels "algorithm A" and "algorithm B". If we mix up the labels on the numbers, the means for A and B should change if the algorithms perform differently. In particular, the chance of producing a larger difference of means by shuffling the two is very slim if the two sets of numbers were generated by different processes. If the numbers are different for the two processes, mixing them will lead to more similar means. So the chances of producing means with a larger difference are better if they were generated by the same process. In this case shuffling obviously makes little difference. The principle is visualized in Figure 6.6.

Formally, given two sets of numbers $x_{1..n}$ and $x_{n+1..n+m}$, we compute the mean difference

$$\mathrm{T}(\mathrm{obs}) = \left| \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{m} \sum_{j=n+1}^{n+m} x_j \right| \tag{6.11}$$

then we create a random permutation $\pi \in \mathcal{P}(n+m, n)$ of n out of $n+m$ values and distribute the values into two sets with the original sizes, i.e., using the index sets π and $\overline{\pi} = \{1 \leq i \leq n+m \wedge i \notin \pi\}$ and compute the difference of means

$$\mathrm{T}(\pi) = \left| \frac{1}{n} \sum_{i=\pi_1}^{\pi_n} x_i - \frac{1}{m} \sum_{j=\overline{\pi}_1}^{\overline{\pi}_m} x_j \right| = \left| \frac{1}{n} \sum_{i=\pi_1}^{\pi_n} x_i - \frac{1}{m} \left(\sum_{i=1}^{n+m} x_i - \sum_{i=\pi_1}^{\pi_n} x_i \right) \right| \tag{6.12}$$

and count the number of times that the difference of means of the permutation is larger $\mathrm{T}(\pi) > \mathrm{T}(\mathrm{obs})$. If we divide this number by the count of subsets, we get the exact two-sided p-value, therefore the permutation test is also called an exact test.

$$p_{\text{exact}} = \frac{|\{\mathrm{T}(\pi) > \mathrm{T}(\mathrm{obs}) \mid \pi \in \mathcal{P}(n+m, n)\}|}{|\mathcal{P}(n+m, n)|} \tag{6.13}$$

There are $\binom{n+m}{n}$ subsets of length n out of $n+m$ values. In practice, this number is often way to large for the exact test to be feasible. However, it can be observed that when using only a small number of permutations, the result of the permutation test converges to the true value very quickly [OG10].

In order to find out how many iterations are necessary, we can employ statistical theory. For any Monte Carlo approximation with k iterations and a true p-value of p_T, we can compute the standard deviation of the p-value as described in [Goo00]:

$$\delta p = \sqrt{\frac{p_\mathrm{T}(1 - p_\mathrm{T})}{k}} \tag{6.14}$$

Since p_T is notoriously unknown, it is practical to compute the standard deviation of the p-value [OG10] the significance level α

$$\delta p \approx \sqrt{\frac{\alpha(1-\alpha)}{k}} \tag{6.15}$$

or the upper bound

$$\delta p \leq \frac{1}{2\sqrt{k}}. \tag{6.16}$$

Assuming we want $\delta p \leq 0.001$ for $\alpha = 0.01$, we require $k \geq 10,000$ iterations according to Equation (6.15) or $k \geq 250,000$ when using the upper bound Equation (6.16).

Much is to say for the permutation test method since it requires no model assumption while producing accurate results. As Schmucker et al. said: *"Before the era of cheap computer power, the randomization test was impractical for all but the smallest experiments. As such, statisticians created significance tests that replaced the actual score differences with the ranks of scores"* [SAC07, p. 625]. Which means, in return, that given today's computing power, there is little need for anything but the randomization test to determine the significance. Given that one can rarely assume the system in question to produce normally distributed errors, this is the method of choice for comparison of algorithm performance.

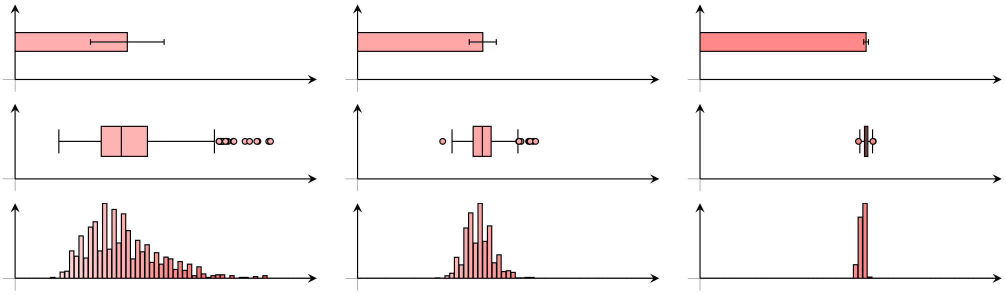

Figure 6.7: Three distributions (left to right) displayed as bar, box and histogram plot (top to bottom).

6.3.7 *Plots*

Within this thesis, sets of numbers will be displayed graphically mostly in two ways: Bar charts and box plots, see Figure 6.7 for examples.

When the number of data points is small, bar charts are displayed. The bar length corresponds to the mean value. The standard deviation is shown by symmetrical error bars. When the number of data points in the set is large, box plots will be used to provide a concise representation of the distributions. The box represents the interquartile range between the 1st and 3rd quartile. The line in the middle marks the median. Values farther away than 1,5 times the interquartile range from the box are considered outliers and plotted as individual points. The whiskers are therefore drawn to the maximum value above the 3rd quartile that is still below the 3rd quartile plus 1,5 times the interquartile range, and the minimum value no more than 1.5 times interquartile range below the 1st quartile.

6.4 ACOUSTIC EVENT CLASSIFICATION

Several experiments on acoustic event detection and classification were performed. Several different approaches were implemented. They will be outlined in section 6.4.1. For the recordings made in the FINCA (see section 6.2.1 on pages 90–92), a detailed comparison of systems and the performance for different classes and features will be described in this section. Then the proposed method will be evaluated on the D-CASE dataset in order to compare it with state-of-the-art results from the literature.

6.4.1 *Systems*

Several event classification and detection systems were implemented for comparison. Each system was trained on C input classes, including the non-event class. For all systems, features z_k were computed on windows of 1024 samples at 48,000 Hz sampling rate, i.e., 21.3 ms. A total of 0.6 s, i.e., $K = 27$ consecutive feature vectors were used for classification. The following ways of codebook estimation were used:

HQ-U Hard vector quantization. One codebook with a fixed number of $I \cdot C$ centroids was estimated from the training data using Lloyd's algorithm [LBG80, Llo82].

SQ-U Soft vector quantization. One codebook with a fixed number of $I \cdot C$ Gaussian densities was estimated from the training using the expectation-maximization (EM) algorithm, cf. Section 2.3 on pages 20–23.

SQ-S Supervised soft vector quantization. For each of the C classes, a fixed number of I Gaussian densities was estimated using the EM algorithm, cf. section 3.2.2 on page 35. All Gaussians were concatenated to a super-codebook with $I \cdot C$ densities.

Using the estimated codebooks, the bag of features (BoF) method was used with different classifiers:

SVM$_{\text{LIN}}$ A multi-class support vector machine (SVM) classifier with a linear kernel. The slack parameter was determined by a grid search on the training data.

SVM$_{\text{RBF}}$ A multi-class SVM classifier with a radial basis function (RBF) kernel. The slack and γ parameter were determined by a grid search.

SVM$_{\text{HI}}$ A multi-class SVM with a histogram intersection kernel, like the bag of features event detection approach proposed by Pancoast et al. [PA12] as described in Section 3.1.2 on page 33.

ML A maximum likelihood Bayesian classifier as described in Section 3.2.3 on page 36.

Additionally, the following classifiers were used for comparison:

GMM The Gaussian mixture model (GMM) implementation follows the standard "bag of frames" approach [GSB^{+}13]. For each class, an individual Gaussian mixture model with a fixed number of I densities was trained. The means were computed using Lloyd's algorithm as initialization of the EM algorithm.

For each 0.6 s detection window, the posterior probability (score) for each individual GMM model was computed as sum of the log-likelihoods. The class belonging to the model with the highest score is chosen as the detection.

HMM A straightforward hidden Markov model (HMM) approach using a Gaussian mixture model for semi-continuous modeling of the observation. $L = I \cdot C$ Gaussian densities with diagonal covariance were estimated from the training data, either supervised (SQ-S) or unsupervised (SQ-U). The ESMERALDA toolkit was used [FP08], which estimated the initial means by the k-means algorithm [Mac67] before the densities were estimated using the EM-algorithm.

For each of the C classes, a linear state structure with the same number of $S = 15$ fixed states was used. Each sub-model was trained individually with the Baum-Welch algorithm [Fin14, pp. 92–96]. All class sub-models were connected by pseudo-states for the Viterbi decoding [Fin14, pp. 85–87], allowing the hypothesis to provide an alignment using any sequence of the classes sub-models.

In off-line mode, the whole data file was decoded once using the Viterbi algorithm and the class corresponding to the sub-model where the optimal path was at each frame was used as the detection result. The most often occurring class in each 0.6 s time window was used as final decision. In online mode, up to 10 s of previous frames before the current one were decoded.

DNN The deep neural network (DNN) baseline system published with the 2016 D-CASE challenge was applied for comparison. The sliding windows were used for training and test. Three fully connected rectified linear unit (ReLU) layers of 500 neurons with 10% dropout are followed by a one-of-k coding with sigmoid output. The background class is not used in the training. An event is considered detected if the output is larger than 0.5, otherwise the window is considered background.

For all classifiers, 100 different training iterations were computed. The estimation of the Gaussian densities was randomized by shuffling the input data.

6.4.2 *Classification on FINCA dataset*

The classification was tested using the data for each class in the training set for training and using the test set to evaluate the performance, (see section 6.2.3 on page 92). For a comparison of the different classifiers, a codebook size of $I = 30$ densities per class was chosen for all supervisedly trained codebooks. For the unsupervised trained codebook, a corresponding size of $I \cdot C = 330$ was chosen, except the BoF SVM_{HI} approach. For the latter, the codebook size of 1,000 densities was chosen as in [PA12]. Figure 6.8 show the overall performance of the different classifiers using the loudness, mel frequency cepstral coefficient (MFCC) and Gammatone frequency cepstral coefficient (GFCC) features.

In the offline case, one run of the Viterbi algorithm was done by the HMMs over the full sequence. The supervised codebook (SQ-S) is significantly better in all measures $(p \leq 0.001)$ than the unsupervised one.
When using the HMM online, the performance deteriorates as expected. However, the HMM using the supervised codebook performs best. The bag of super features [PGF14] (ML BoF SQ-S) is the next best classifier with 92.0% f-score, showing significantly better precision than the rest. It can also be seen that its performance is much more consistent over the 100 randomized runs. This shows its greater abstraction ability. Overall, the average performance of the GMM is similar, leading to a close f-score value of 91.7%. The HMM with the unsupervised codebook comes close, but with a higher deviation due to the random codebooks.

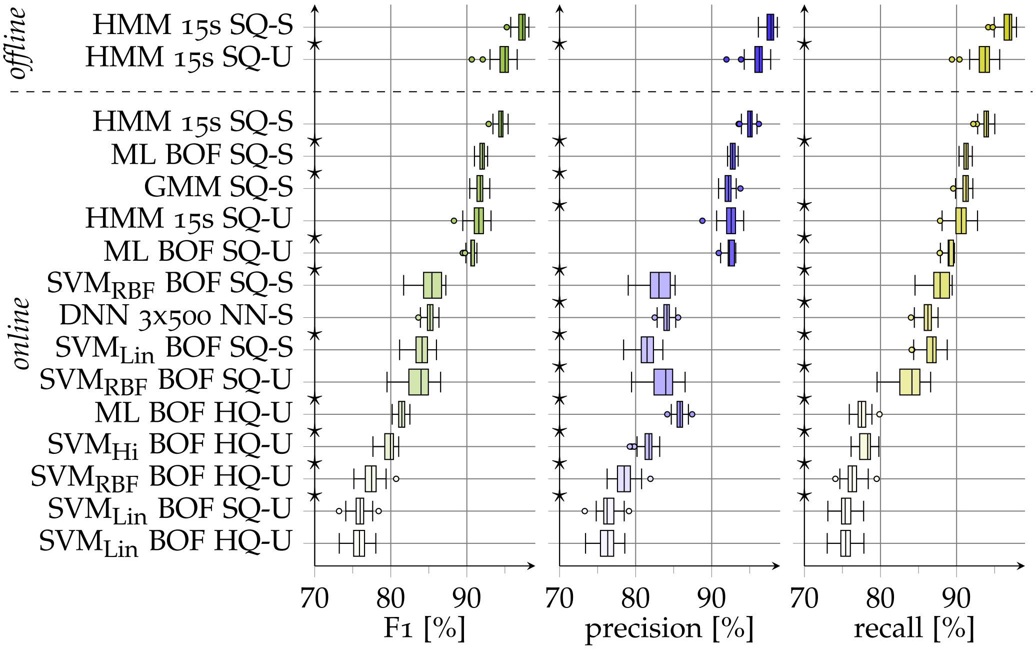

Figure 6.8: Classwise classification results on the FINCA dataset for different classifiers using the loudness, MFCC and GFCC features. Box plots of the results of 100 runs. A star marks that the performance is significantly different than the next best one below according to a randomization test ($N = 10^5$).

The DNN achieved 84.4% f-score with the feature combination of loudness, MFCCs and GFCCs. The silence detection is controlled by the threshold applied to the output neurons. While the f-scores are similar, the precision and recall of 91.7% and 78.1% for a threshold close to 1.0 change to 82.1% and 85.7% for the proposed threshold of 0.5. The DNN performs worse than the GMM but better than plain BoF methods. As the DNN is in principle able to learn and optimize its own feature representation, it was also applied with spectra and mel band energies as input. 83.7% and 76.1% f-score were achieved, respectively. This shows that the proposed feature combination provides advantageous information compared to the standard mel band energy approach. The better performance using the plain amplitude spectrum shows the DNNs ability to infer better features. Still, even with this it is not close to outperforming the proposed method or the GMM.

Within the rest, a clear ordering among codebook and classifier is visible. The supervised codebook is always better than the unsupervised one, and soft quantization is always better than hard. The maximum likelihood (ML) classifier clearly outperforms the SVM. The RBF kernel is always working better than the linear one. The SVM with histogram intersection kernel (BOF SVM-HI) and hard quantization [PA12] shows better performance than the RBF or linear ones.

Codebook estimation and size

Given the clear superiority of the supervised codebooks of the same size, it was interesting to see if larger unsupervised codebooks converge towards the performance of the supervised ones. Figure 6.9 shows the classwise performance for the HMM and bag of features maximum likelihood (ML) Bayes classifier for different codebook sizes and

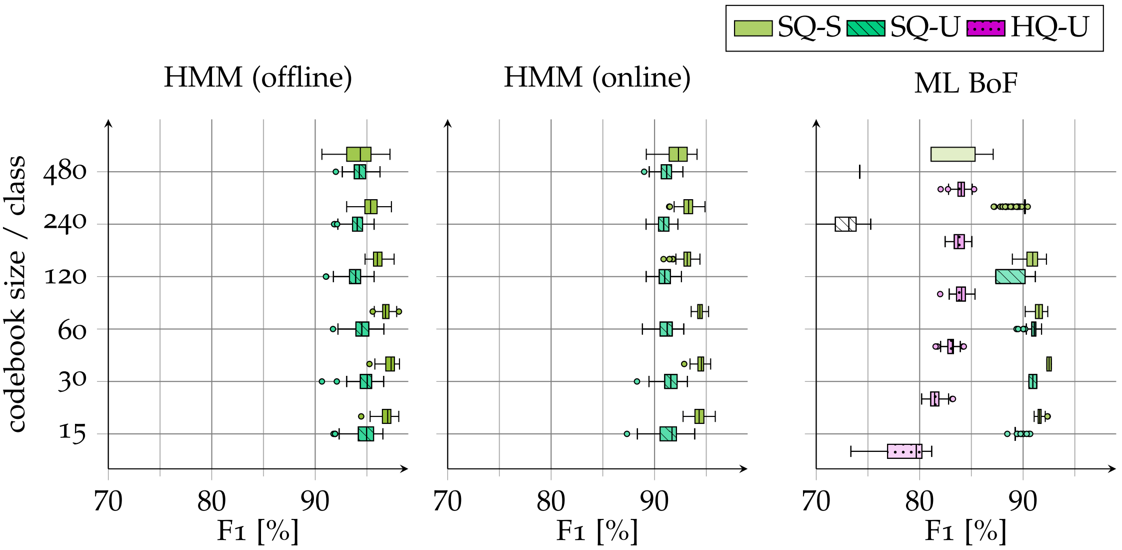

Figure 6.9: Classwise classification results on the FINCA dataset using supervised and unsupervised codebooks of different size. Box plot over 100 runs with randomized training.

supervised and unsupervised clustering. Exponentially increasing codebook sizes of 15, 30, 60, 120, 240, and 480 Gaussian densities per class with diagonal covariance were used, i.e., 165 to 5,280 densities in total.

The unsupervised case is almost similar for all codebook sizes, the performance becomes slightly more consistent for larger codebooks. All results using a supervised trained codebook are better on average than using an unsupervisedly trained one the same size. Pairwise randomization test ($N = 10^5$) showed that the results of an HMM using supervised codebooks are significantly different ($p < 0.001$) from an HMM using unsupervised codebooks independent of codebook size, i.e. all HMMs using supervised codebooks produce better results than any HMM using an unsupervised codebook.

For the bag of super features classifier, the supervised case performs slightly worse for larger codebooks than 30, which might be a sign of overfitting. Beyond 120 centroids, the performance deteriorates. In the unsupervised case, the performance becomes both better and more consistent with increasing codebook size up to 120. After that, the performance decreases, probably since the data is insufficient to support more centroids. However, it never comes close to the supervised case. The results for supervised and unsupervised training are disjoint up to as size of 120, i.e. the best result for unsupervised training is still worse than the worst result for supervised training. As for the HMM, all classifiers with unsupervised trained codebooks show significantly ($p < 0.001$) different results than any supervised one according to a pairwise randomization test ($N = 10^5$).

The performance using hard quantization does not come close to the soft quantization, even for much larger codebooks. Its median f-score improves form about 80% to 85% when increasing the codebook size from 165 up to 1,320 centroids, while the soft quantization is in the range between 90% and 95%.

Event classes and features

Figure 6.10 shows the results for the eleven different classes using an HMM, GMM and the bag of super features [PGF14] classifier. For all classifiers, “speech” is the best per-

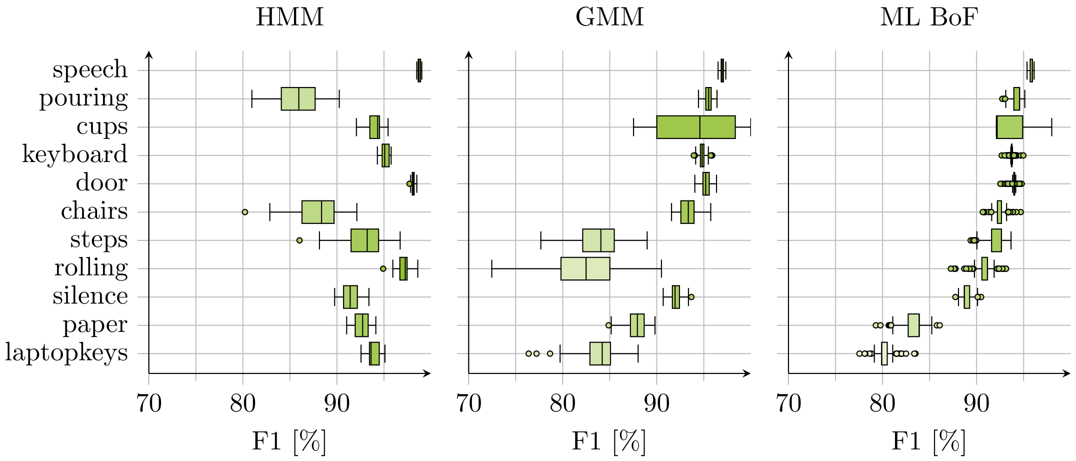

Figure 6.10: F-score for the different classes using loudness, MFCC and GFCC features on the FINCA dataset for different sound event classes using the HMM (online), GMM and BOF (ML) classifiers. Box plots over 100 runs.

forming class. The HMM performs slightly worse for the classes "pouring" and "chairs", this may be due to some unjustified generalization from the training data. In contrast, it outperforms the other methods clearly for "paper" and "laptopkeys", both of which contain a lot of transients. Here, the improved temporal modeling seems to be working better. The ML classifier works better and more consistent than the GMM, especially for "steps", "rolling" and "door". This is most likely to attribute to its better generalization ability and utilizing the Gaussians of the other classes.

Another aspect to look at are the features used. As the ML BoF approach showed the most consistent results, different feature sets were used with this classifier in order to see their influence on the classification performance. The combination of MFCC and GFCC features works best. The use of the MFCCs alone produces results close to the combination with GFCCs. This is understandable as the GFCCs themselves perform significantly worse than the other feature sets.

In order to find out where the benefit of the GFCCs lies, it is necessary to look at the individual classes. Table 6.4 lists the results for each of the classes and the different feature sets. For "speech", all feature sets perform well, which may be due to its wide spectral spread. A slight advantage of adding the GFCC features can be seen with the "paper", "steps", and "door" class, even though the GFCCs alone only perform badly on "door" by themselves. The combination achieves similar results to the MFCCs alone for the other classes. These features are also compared with the perceptual feature representation from [TN06], as it is used in numerous methods and encompasses a large selection of promising features. These work slightly better on "pouring" and "speech", but worse over all.

6.4.3 *Event detection example*

The results for the proposed method for the skript recording in the FINCA are visualized in Figure 6.11. The ML BoF classifier with loudness, MFCC and GFCC features was used. As expected, the speech events are detected quite precisely. The worst confusion is that of *"chairs"* and *"steps"*. This is not only the result of them being very similar in sound, but also due to the rough annotation of what is in practice a mixture of both events. The

features	chairs	cups	door	keyboard	laptopkeys	paper	pouring	rolling	silence	speech	steps	mean
L MFCC GFCC	**93.6**	93.3	**96.2**	**94.7**	**78.4**	86.1	94.5	**93.9**	**87.5**	95.6	**92.4**	**91.5**
MFCC GFCC	89.2	89.0	**95.4**	92.5	76.6	81.4	92.1	88.9	86.4	94.9	90.6	88.8
L MFCC (40)	**93.2**	**94.7**	94.0	93.3	77.0	80.4	94.6	92.6	83.0	**96.5**	91.3	90.1
MFCC (40)	**92.6**	**95.7**	93.9	92.7	**78.7**	79.8	94.3	**93.2**	85.4	**96.7**	91.0	90.3
L MFCC	**92.8**	93.7	94.0	**94.9**	76.8	81.9	93.8	91.5	**86.9**	95.4	90.0	90.1
MFCC	90.0	90.2	92.0	92.7	75.2	77.9	89.5	86.6	86.0	94.3	89.0	87.6
L GFCC	89.5	92.2	67.3	91.3	55.4	86.5	94.3	92.2	79.4	94.7	86.9	84.5
GFCC	79.1	86.3	50.0	91.9	37.9	**89.6**	91.5	87.1	70.6	94.6	81.6	78.2
perceptual [TN06]	90.1	93.0	86.8	91.7	62.2	80.1	**96.6**	88.9	27.9	**96.6**	87.8	82.0

Table 6.4: Classwise f-score [%] for different feature sets for the event classes using the ML BoF classifier on the FINCA event classification task. Colored between 100% 🟩 and 75% □. Scores close to 1% of the best for the class are set in bold.

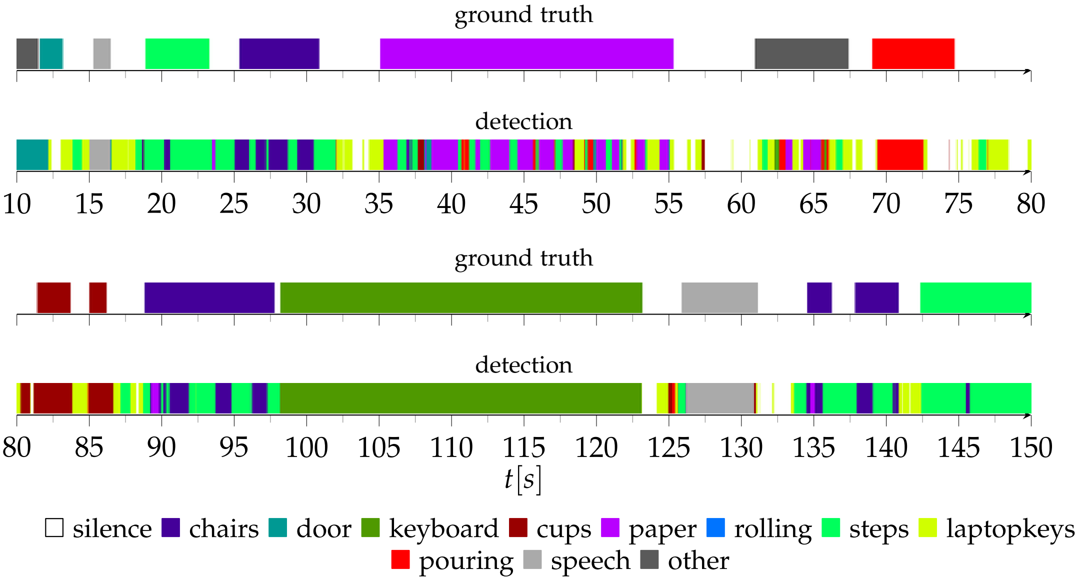

Figure 6.11: Event detection results for FINCA skript.

detection of "paper" is not very clear, as it is often confused with other classes. Partially, the background is classified as "laptopkeys", not surprising as this was the class with the worst classification performance.

6.4.4 *Event detection on D-CASE dataset*

In order to compare to literature values, the development set of the D-CASE challenge [GSB+13] was used also to evaluate the event detection performance. The dataset is similar to the FINCA. It was recorded in an office and contains training snippets and skript recordings, see Section 6.2.4 on page 93 for a description of the dataset. Several contenders published their performance on the development set.[3] The systems used were the following:

[SMS+13] An HMM with features computed by a Gabor filterbank as extension of the MFCCs and their temporal derivatives. A noise-reduction algorithm was used as preprocessing. The Gabor filters are used as two dimensional feature modeling the modulation along time and frequency in the spectrogram. The event detection is done by a two-layer HMM. The first layer is a fully connected HMM with each state corresponding to an event class. The observations of this layer are sub-HMMs, which model the filterbank responses. This entry performed best in the challenge with a 61.52% framewise f-score on the test set.

[NVM13] An HMM with meta classification. From the training data, a variety of technical features was calculated. These were used to train a random forest classifier using random subsets for each class. The detection is performed by a two-layer HMM. The top layer contains a state for each event class plus extra finishing states to explicitly model class transitions. Each class state is connected to a sub event cluster of states modeling the observations in the second layer. This entry was second with an 45.50% framewise f-score.

[VBK+13] An extension of the classical MFCC GMM approach. Based on a threshold criterion, either a shared background GMM or foreground GMM for the given class is trained. The output of both is combined and classified by another GMM per class. This entry performed close to second best with 43.42% framewise f-score.

[GVK+13] A HMM with non-negative matrix factorization (NMF) spectra as features. A dictionary of magnitude spectra was calculated by NMF on the training data. Additionally to the given classes, a background or non-event class was trained on the not annotated parts of the training files. For the event detection, the magnitude spectra were mapped to the event classes by means of a matrix mapping the estimated dictionary entries to the classes. The mapping was converted into posterior probabilities by scaling and normalization. A basic HMM with a single state per class was used for classification. The transition probabilities were set to be equal for all classes using fixed self transition, event to event, and event to background probabilities. This entry achieved an 31.94% framewise f-score on the test set.

[DHV13] A basic MFCC HMM approach. They did well on the development set, but achieved only 26.0% framewise f-score on the test set.

[GSB+13] The baseline system published with the challenge. It utilizes NMF spectra classified by a multiclass SVM.

[KSWP16] The DNN baseline system published with the 2016 D-CASE challenge was used again to compare. The proposed 40 mel band energies over the sliding window are used as features. After the input layer, three fully connected ReLU layers of 500 neurons with 10% dropout are followed by a one-of-k coding with sigmoid output. An event is considered detected if the output is close to 1, otherwise the window is considered silence.

[3] Results and papers are availiable at www.elec.qmul.ac.uk/digitalmusic/sceneseventschallenge

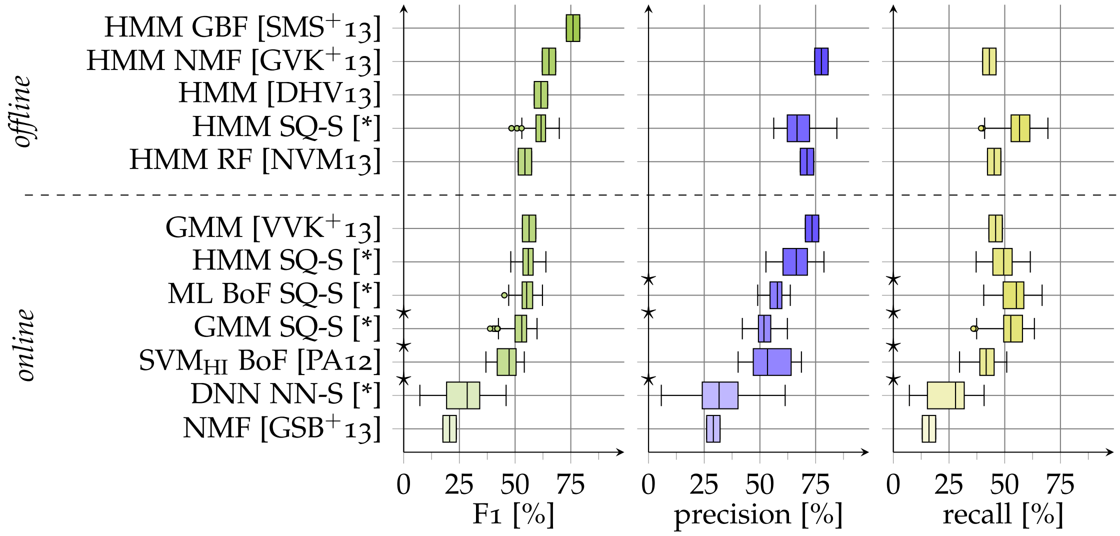

Figure 6.12: Acoustic event detection results on D-CASE development set. Framewise f-score, precision, and recall over all three scripts and both annotations computed for the methods reimplemented [*], the bag of audio words approach [PA12] and values from the literature.

Again, several of the implemented methods tested on the FINCA dataset were evaluated. The Loudness, MFCC, and GFCC features were used. As the codebook creation using supervised soft clustering (SQ-S) was clearly superior, only this method is used. The codebook size was set to $n_g = 30$, resulting in a total of 510 Gaussians for the 17 classes. The HMM was used both offline decoding the whole sequence and online decoding up to each time frame. The bag of super features method [PGF14] (ML BOF) and the basic GMM approach were used as described in the previous section. The bag of words approach was applied with MFCC-Delta and energy as features and hard vector quantization with a codebook size of 1,000 as in [PA12].

Classifiers

The classifiers were run on all three skripts, fifty times each to capture the effect of random initialization. The background parts of the other two scripts were taken for training of the non-event class. Both sets of annotations were used separately, thus a total of 300 runs was computed for each classifier. For the implemented methods, the Loudness, MFCC, and GFCC features were used together with supervised clustering. In order to replicate the "bag of words" approach [PA12], MFCC-Delta and energy was used with hard vector quantization. Figure 6.12 on the current page shows the results on the D-CASE development set.

The off-line HMM performs close to the other off-line methods with 61.3% f-score, 56.7% precision, and 67.7% recall. [DHV13] report an f-score of 61.6% for their MFCC HMM method on one of the annotations. The NMF based HMM achieved 65.2% f-score according to [GVK+13]. The meta recogition HMM performed worst with 54.4% according to [NVM13].

In the online event detection, the foreground-background GMM achieved an f-score of 56.3% both the HMM and ML BoF approach achieve a similar f-score of 55.9% and 55.4%, respectively. The difference in f-score is not significant. The precision and recall differ significantly according to a permutation test $(p < 0.001, N = 10^5)$ The HMM

features	alert	clearthroat	cough	doorslam	drawer	keyboard	keys	knock	laughter	mouse	pageturn	pendrop	phone	printer	silence	speech	switch	mean
L MFCC GFCC	12.1	53.7	46.3	34.2	**35.6**	61.5	44.3	**69.7**	17.9	5.2	46.5	**29.0**	30.8	40.6	**92.7**	65.1	0.0	40.3
MFCC GFCC	18.4	59.5	31.7	3.3	31.8	56.9	42.2	**68.9**	16.5	**6.0**	50.3	**28.2**	27.9	**44.3**	**92.6**	61.0	0.0	37.6
L MFCC	1.5	48.9	47.9	38.7	30.7	**63.3**	43.5	60.4	15.7	**6.5**	48.6	10.7	29.4	41.8	**92.5**	44.9	0.0	36.8
MFCC	5.6	52.8	25.1	2.1	28.0	58.2	43.3	58.3	15.6	**6.9**	48.1	14.3	30.1	41.9	**92.3**	45.6	0.0	33.4
L GFCC	27.9	65.7	**54.0**	**44.5**	**35.8**	41.8	32.4	68.5	16.3	0.1	32.9	1.8	28.9	23.1	**92.7**	69.9	0.0	37.4
GFCC	**33.3**	**74.1**	46.1	0.0	23.4	40.7	30.3	65.5	16.3	0.4	46.9	1.6	27.9	24.5	**92.6**	**80.1**	0.0	35.5
perceptual [TN06]	1.1	49.2	43.6	31.5	31.4	53.5	**48.8**	50.6	20.2	4.0	39.5	8.8	32.0	**44.8**	**92.4**	23.2	0.0	33.8

Table 6.5: Classwise F score [%] for different feature sets for the event classes using the ML BoF classifier on the D-CASE dataset. Colored between 100% 🟩 and 0% □. Scores close to 1% of the best for the class are set in bold.

achieves this by lower precision and higher recall, it seems to handle the background class slightly better. The GMM performs slightly worse, with significantly smaller f-score, precision, and recall according to the permutation test $(p < 0.001, N = 10^5)$. The bag of audio words approach [PA12] performs worse than even the GMM, likely due to the unsupervised hard quantization. The DNN baseline system with mel band energy features [KSWP16] shows a large variance over different training runs. It is unable to beat any of the GMM approaches, which is consistent with its performance in the D-CASE 2016 challenge on Task 2. The training material seems to be insufficient.

Features and Classes

Again, it is interesting to see if the use of the GFCC features is helping. Table 6.5 shows the results per event class and feature set for the ML BoF SQ-S classifier.

Averaged over all classes, the combination of loudness, MFCCs, and GFCCs performs well. The combination is significantly better than loudness and MFCC or GFCC alone according to a permutation test over the mean f-score over all classes for the 100 runs $(p < 0.001, N = 10^5)$. Adding the loudness feature also leads to significantly better results $(p < 0.001, N = 10^5)$ for MFCC, GFCC, and the combination of both.

The performance of the features sets varies largely depending on the event class. The f-score for speech is increased from 45% to 70% with the GFCC. The combination of loudness, MFCC, and GFCC is also close to 70%, so the trade-off of incorporating the other non-speech oriented features is acceptable. The loudness feature actually decreases the performance for speech.

Detection example

Figure 6.13 shows the detections of the BoF ML classifier for one run on the first skript. There are a few misclassifications of the rather soft "pendrop" and "phone" classes, the "pageturn" is sometimes missed. The "alert" and "laughter" classes are also misclassified. Speech is basically detected, although it is confused with "laughter" and "clearthroat" at the on or offset. The "printer" is detected before it is audible according to the ground truth annotations.

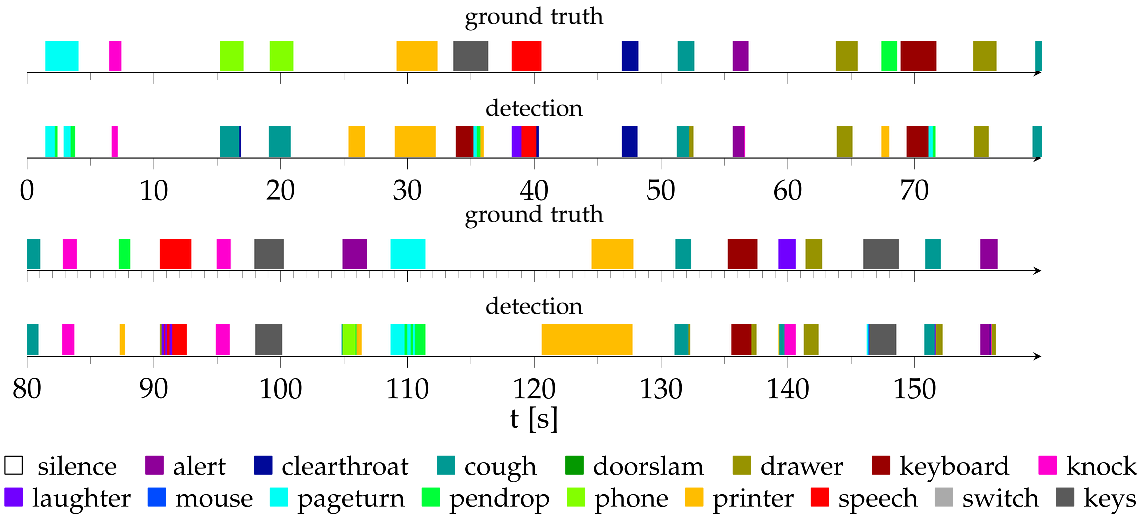

Figure 6.13: Detections on D-CASE skript using the BoF ML classifier and L,MFCC and GFCC features.

6.4.5 *Summary*

The proposed method for detection of acoustic events was thoroughly evaluated on data recorded in the FINCA smartroom and data from the D-CASE challenge. The addition of the GFCC to the MFCC features increases the overall performance, in the case of the smartroom recordings more so than using more MFCCs. For both the smartroom and the D-CASE dataset, the proposed feature combination outperforms the common perceptual features [TN06].

A systematic comparison showed the superiority of the super-codebook estimation. Estimating separate codebooks for each of the event classes with separate applications of the EM algorithm and concatenating them works better for HMM, GMM, and BoF classification. The HMM is performing best in either off-line or online application. The DNN baseline system used for comparison is outperformed by the proposed approach as well as the GMM on both datasets. This is most likely due to the training data being insufficient for the approach, cf. [GPF17].

The proposed BoF method achieves similar results to the online HMM. It clearly outperforms both the GMM and standard BoF approach. Further enhancement may be possible through the use of multiple channels as shown in [KGPF16].

6.5 BLIND SPEECH ENHANCEMENT

Speech enhancement is one of the key applications for acoustic sensor nodes. This can be achieved by dedicated beamforming methods (see section 2.4 on pages 24–26). To be applicable in practical scenarios, they need some type of control information. One fundamental type is the detection of speech activity over time. As the proposed BoF classifier provides robust results in real time, it was applied for this task. The speech detection was used to control a minimum variance distortionless response (MVDR) beamformer realized as generalized sidelobe canceler (GSC) as described in Section 3.3 on pages 37–38. The filter estimation assumes stationary noise and exploits the nonstationarity of the speech signal in order to estimate it [GBW01]. The relative transfer function (RTF) of this signal is estimated, thus the direction of arrivals (DoAs) of the speech and noise signals do not have to be known. Dedicated recordings with a single smartphone mockup were done in the acoustic lab at Bar-Ilan university, cf. Section 6.2.5 on pages 93–96. Up to two concurrent noises from different directions were added to speech signals. A novel training strategy was devised to handle difficult scenarios with real noise types. It was evaluated in comparison to simpler alternatives. Different scenarios with different noise types and SNRs were used. The speech classification and its influence on the enhancement performance were evaluated.

6.5.1 *Features and Training*

On the features level, only the addition of deltas is different from the event detection method (see section 3.2 on pages 34–36). No higher level temporal augmentation was found to improve the performance, neither using a spatial pyramid [PGF14] nor temporal feature augmentation [GPF15].

A dedicated training strategy was devised. As described in Section 3.3.2 on page 38, the training data is divided into levels according to nonstationarity to form a hierarchy of noises. The training data for each level is trained as individual classes. Additionally, mixtures up to a certain level are used to train mixture classes. The introduction of hierarchical mixing was found necessary in order to handle mixed noise cases well. Using only high SNR mixtures of speech with noise was found important in order to ensure high recall.

Each change was retracted individually and the resulting classifier was evaluated on the test set. Figure 6.14 shows the speech classification results for different training strategies

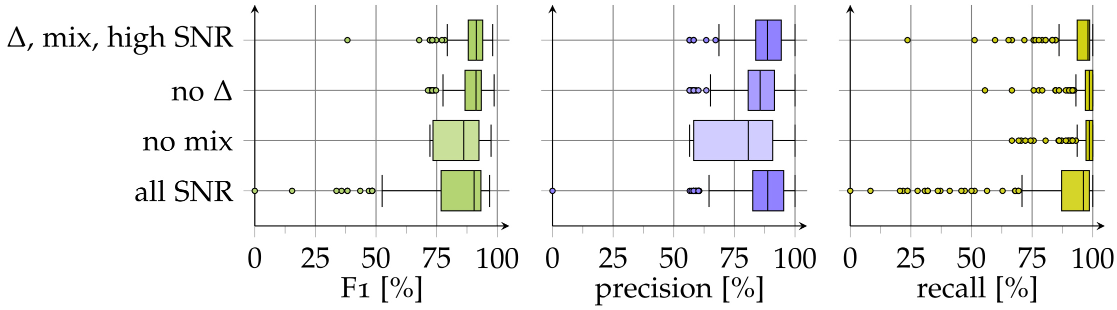

Figure 6.14: Classification of speech using different training sets and features: the proposed method using deltas, mixtures and only high SNR speech samples, the same without using deltas, training without noise mixtures, and using speech mixed in all SNRs (from top to bottom).

detection		Ω_0	Ω_1	Ω_2'	Ω_2	Ω_3'	Ω_3	Ω_4'	Ω_4
speech	Ω_0	92.3%	23.5%	23.6%	15.0%	44.8%	13.4%		
stationary	Ω_1	0.5%	55.2%						
	Ω_2'	1.1%	18.9%	30.3%	1.1%	0.6%	1.3%		
mechanical	Ω_2	1.0%		1.2%	54.9%	10.1%	28.6%		
	Ω_3'	0.7%	0.4%	44.7%	26.6%	30.1%			
babble	Ω_3	1.0%			1.5%	14.3%	56.7%		
	Ω_4'	0.1%	2.0%	0.2%	1.0%				
nonstationary	Ω_4	3.1%							100%

Table 6.6: Confusion matrix for the different speech and noise classes. Detection results over the full test set. Class type of the data is enumerated in columns, detections in rows. The cell color is scaled linearly between 100% 🟩 and 0% □.

and features. In order to assess the significance of the changes, a permutation test was performed between the proposed method and each of the variants.

The proposed methods achieves a 89.6% precision and 93.8% recall in the mean over all sequences, resulting in a mean f-score of 87.4% . When using no delta-features, the performance is slightly worse. This effect was found to be slightly significant ($p < 0.02, N = 10^5$). Both training set changes resulted in a significantly worse f-score with a mean value of 76.5% for no mixing ($p < 0.001, N = 10^5$) from a reduced precision to a mean value of 76.5% ($p < 0.001, N = 10^5$). This is caused by more noise as speech classifications. On the contrary, the use of training samples for speech mixed with noise in lower SNRs significantly reduces the recall to a mean value of 85.4% ($p < 0.001, N = 10^5$). This is caused by more speech as noise classifications.

Table 6.6 shows the class-wise confusion for the chosen strategy and features. There is little detection of speech (Ω_0) as noise, the worst is non-stationary noise with 3% of the speech data in the test sequences with keyboard noise (Ω_4). All non-stationary noise frames are classified correctly. This is important as the beamforming algorithm can not be applied in this case, and has to be switched off. Some noises, especially babble noise mixed with others, are wrongly classified as speech. This mostly occurs in the transitions before or after the speech segments. The pure noise classes are classified correctly more than half of the time. Especially in the mixed cases, there is some confusion within the different stationary noise classes ($\Omega_1 - \Omega_3'$).

6.5.2 *Scenarios*

Figure 6.15 shows the speech classification results for the different scenarios using the chosen training strategy and features. The results are worse for lower SNRs, as is to be expected. Interestingly, the variation of the precision is slightly higher for the cases with only one noise source. It seems to be slightly easier for the classifier to distinguish speech from a mixture of noises. The all important recall is very close to 100% in all cases except 0 dB SNR. Here it is still rather good with a mean of 87.2% and 91.0% for one and two interfering noises, respectively. The precision is slightly worse for the high SNR cases, meaning that more confusion of noise with speech takes place in these cases. The mean precision degrades from 95.6% and 88.6% to 87.6% and 83.2% for one and two interfering noises, respectively. This is understandable as high noise levels make the

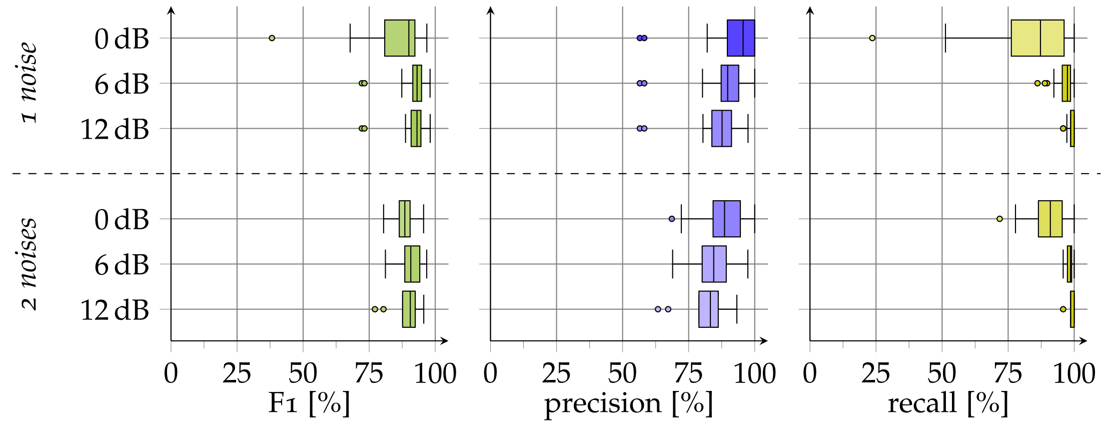

Figure 6.15: Classification of speech using the proposed method for different SNRs and number of noise sources.

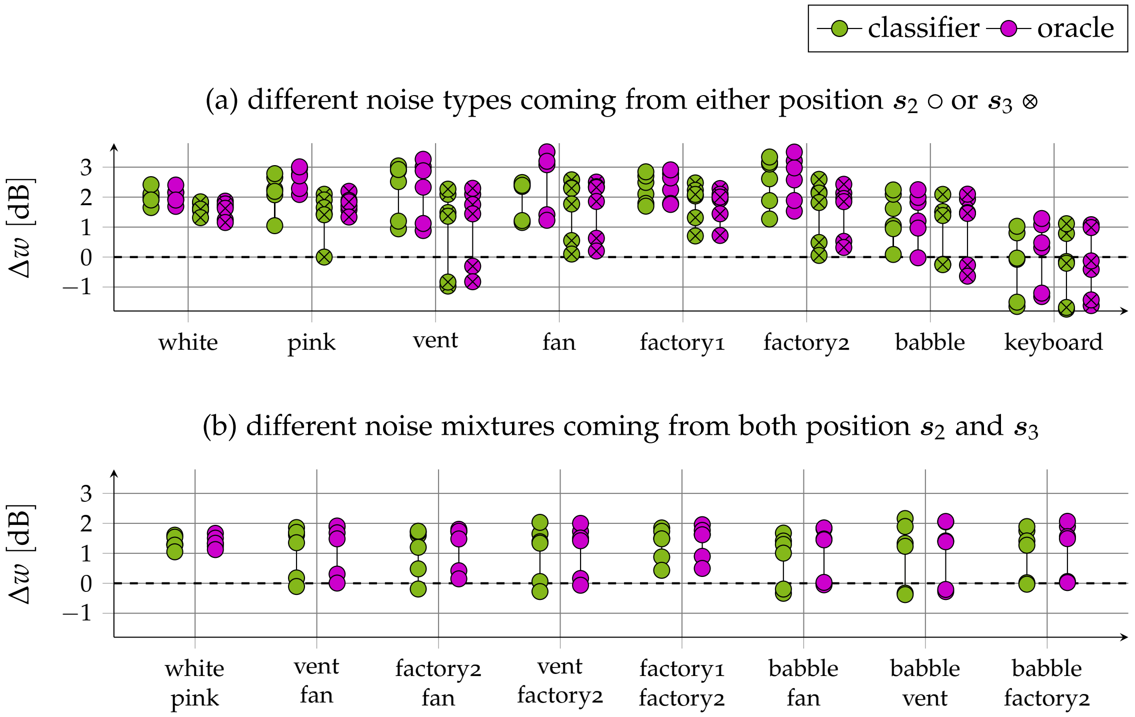

Figure 6.16: Improvement measured by mean and standard deviation difference in fwSNRseg for different noise scenarios computed over the three SNRs (0,6,12) and both speech sequences, totaling six data points. The classifier is compared with using oracle annotations of speech.

noise more prominent, even in the transitions. Together with the bad recall in the lower SNR, this results in a mean f-score of 90.0% and 88.5% for the 0 dB case that goes up to 93.1% and 90.6% for one and two interfering noises, respectively.

6.5.3 *Enhancement*

Figure 6.16 shows the quantitative improvement in fwSNRseg for both a single noise and noise mixtures recorded with the smartphone mockup (see section 6.2.5 on pages 93–96). For a single noise source, the mean improvement in fwSNRseg Δw over all noise types, excluding the keyboard, is 1.75 ± 0.89 dB. When using an oracle in the form of ground truth annotations instead of the classifier, it is only slightly better with 1.87 ± 0.92 dB. When looking at the individual results plotted in Figure 6.16a, the classification only clearly mitigates the result in the case of pink noise. Only six out of the 84 test sequences have a negative result, i.e., the fwSNRseg is decreased after the processing. This happens in a few cases of ventilation and babble noise, both only when they are coming from position s_3. This is consistent with the fact that the results are slightly worse for this position over all. In the case of 'keyboard' noise, the proposed method is not able to consistently improve the fwSNRseg; In half of the cases, the fwSNRseg decreases. As the presence of this kind of noise is detected, the algorithm can be switched off.

When two different noises are coming from different directions, the task is more difficult as the adaptive noise canceler (ANC) has to cancel them both. The mean improvement in fwSNRseg is 1.09 ± 0.76 dB compared to 1.21 ± 0.74 dB with the oracle, cf. Figure 6.16b. There is a clear improvement over all cases. In seven out of the 48 test sequences, the fwSNRseg is slightly worse after the processing.

The proposed method clearly suppresses the noise while very little distortion is introduced to the speech signal.[4] Figure 6.17 shows an application of the method on one of the test signals. The classification and output signal for both oracle and classifier are shown. As the speech detection is very close to the ground truth, there is little difference in the output.

6.5.4 *Summary*

A fully blind system for speech enhancement with multiple microphones was proposed. The BoF classifier is used to provide the control information for a beamformer.

The classifier performs very well in most cases, as speech and non-stationary noise are classified with high accuracy. The training strategy of using classes of mixtures is able to generalize well enough. There is some confusion between the mixed classes and their counterparts, which is not relevant for the application. Overall the idea of using mixtures of different noises in the training is required to handle the overlapping noises in practice. By training only speech with high SNR in mixtures with the various noises, the misclassification of speech as noise was minimized. This is vital as this would lead to cancellation of speech. An additional guard margin around the speech segments was introduced to avoid this. As the speech enhancement quality is very close to using the ground truth instead of the classifier, successful automation was achieved, making the system truly blind. As the system already uses a sensor node with multiple microphones, the performance may be enhanced using multiple microphone information [GPKM14, PMH$^+$15, KGPF16].

There is a solid improvement achieved by the proposed method for a single noise source and for two noise signals from different directions even in 0 dB SNR. The speech enhancement quality is good, as a gain in fwSNRseg of 1-3 dB is achieved. This is en par with state-of-the-art results; Cuachi et al. [CKR$^+$15] showed a similar improvement for an MVDR beamformer using eight microphones on the evaluation set of the reverb challenge.

In the case of highly non-stationary noise, there is little improvement by the proposed method. This is expected, as the filter estimation assumes the noise to be stationary. Since the classifier detects this situation, the ANC adaptation can be switched off.

[4] Audio samples available at `www.eng.biu.ac.il/gannot/speech-enhancement/sam16`

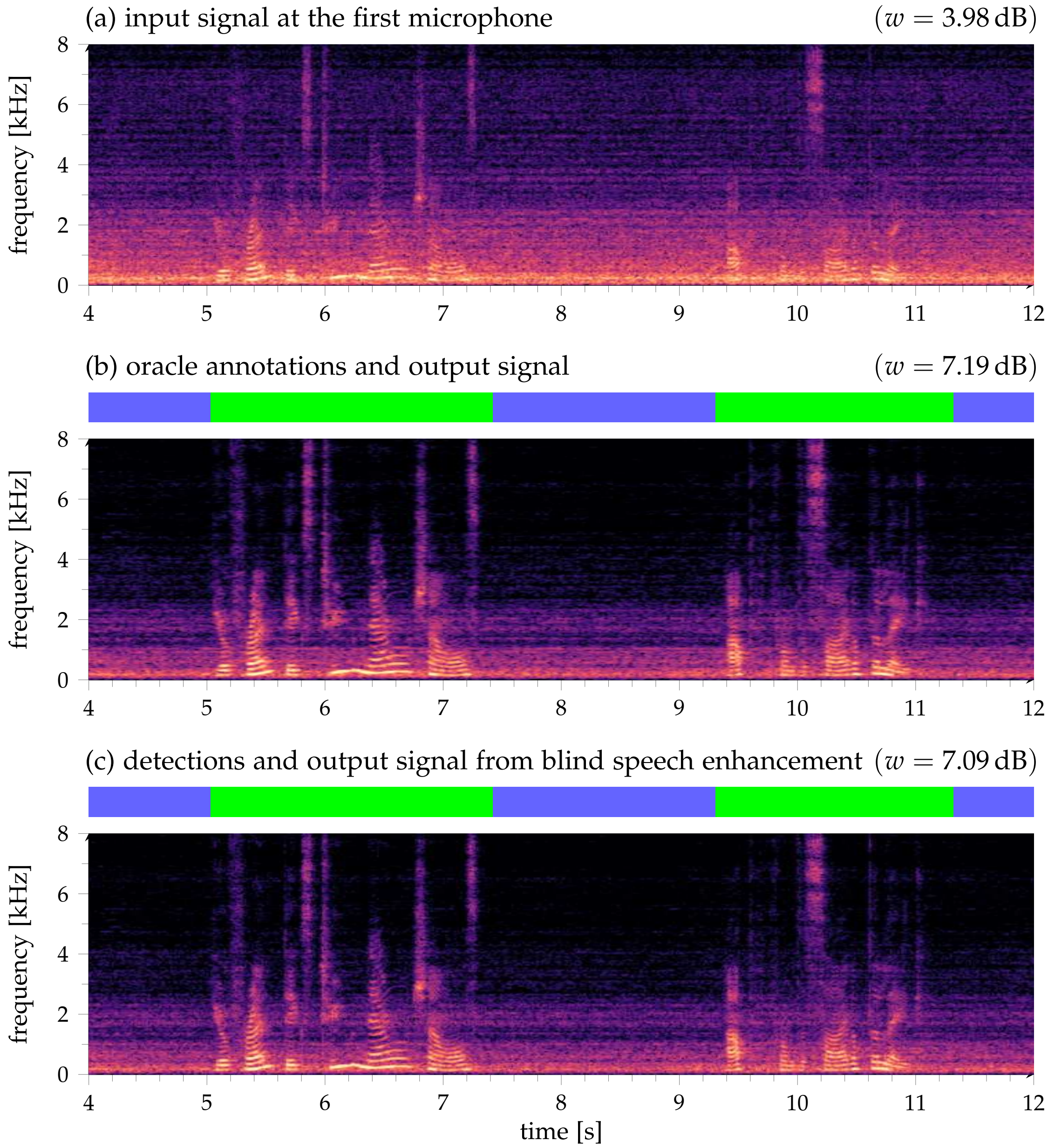

Figure 6.17: Spectrograms of speech from s_1 distorted by 'factory' noise from s_3. Input signal (a), output using oracle ground truth annotations (b), and output of the proposed method (c). Classification is shown on top of the output signals, speech in green (🟩) and noise in blue (🟦).

6.6 SINGLE NODE SPEAKER LOCALIZATION

The DoA localization method [PHF10] was extended with probabilistic clustering via the EM algorithm according to computational auditory scene analysis (CASA) principles [PF13]. The method described in Section 4.2 on pages 54–58 is used for two tasks within this thesis. First, to provide detection of speech segments with DoA measurements for the geometry calibration. Second, to provide running DoA and spectral estimates for the subsequent tracking with the acoustic sensor network (ASN).

In this section, the method is evaluated towards these goals. First, an encompassing simulation is used to investigate the robustness against reverberation and concurrent speakers. Second, it is applied to different recordings in smart rooms to show its ability to handle moving and concurrent speakers and detect valid speech segments in real situations.

6.6.1 *Simulation*

For systematic evaluation, a single node was simulated using the ISM method with a shoe-box model as described in 6.1.1 (p. 89). An uniform circular array with eight microphones and a diameter of 10 cm was placed in the middle of a $5 \times 6 \times 2.5$ m room. Speakers were placed at seven positions in 1-2 m distance at $-170°$, $-120°$, $-70°$, $-20°$, $30°$, $80°$, and $130°$ as shown in Figure 6.18. Snippets of eight seconds from three different anechoic recordings were used as speech data. The first two are of male and female speaker reading a text in normal voice, the third one is a theatrical performance with high volume modulation.[5] The reverberation time T_{60} was varied between 0 and 2 s.

DoA localization

The individual speaker signals were used to compare the proposed method with its predecessor and the steered response power with phase transform (SRP-PHAT). The

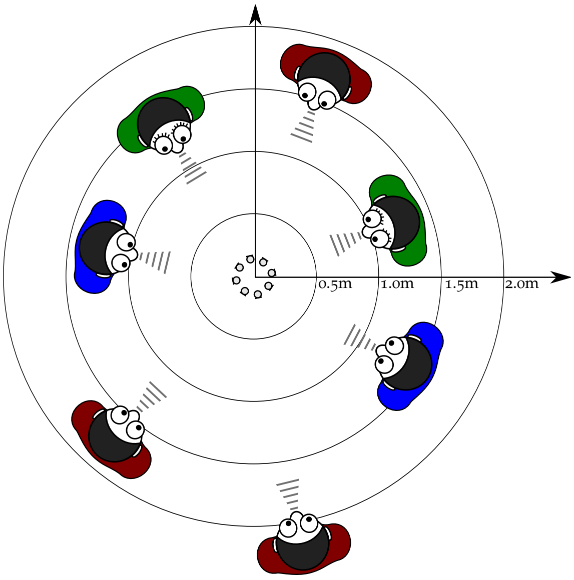

Figure 6.18: Simulated speakers around a circular array.

[5] Courtesy of the university of North Carolina school of arts, sound stage test recordings http://faculty.uncsa.edu/dandp/romneyj/testrecordings/ 12.09.2016

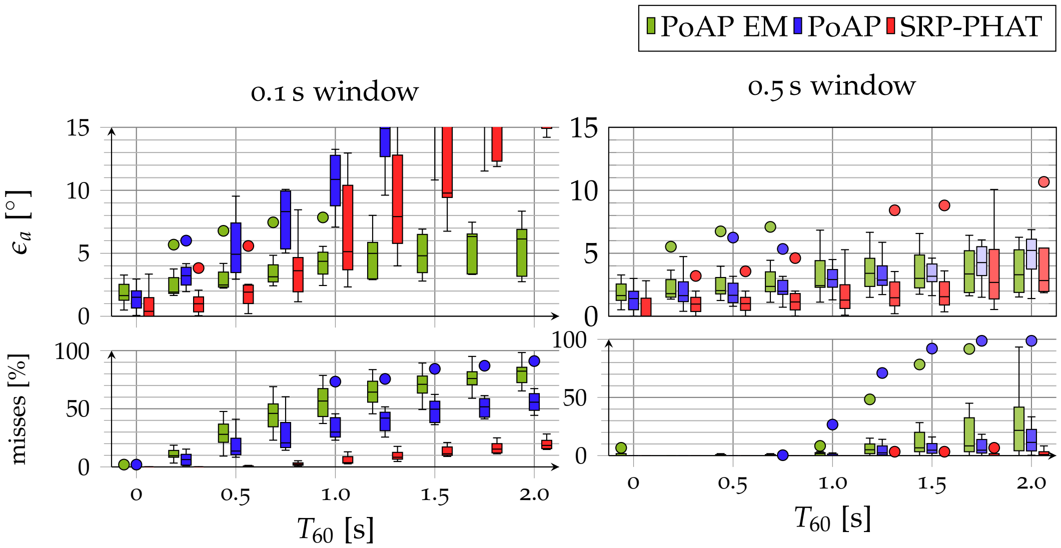

Figure 6.19: Localization error of a single speaker using different localization methods for a single node at varying reverberation times.

default parameters were used ($\gamma = 0.5, t_g = 6\,\text{dB}, t_e = -40\,\text{dB}$). Given that the proposed method is able to work with comparably small time windows, both a 0.1 s and a 0.5 s sliding window were used for all methods. Figure 6.19 shows the resulting angular localization error ϵ_a and the number of missed detections.

Using the smaller time windows, only the proposed method is able to localize the speakers consistently with below 10° error up to very heavy reverberation of $T_{60} = 2\,\text{s}$. This is achieved by rejecting about half of the time windows as non-reliable. Both the non-probabilistic peak over average position (PoAP) method and the SRP-PHAT accept more time windows, which results in higher angular errors. It may be possible to improve their performance by adaptive thresholding. This is not necessary for the proposed method. The gain estimation and probabilistic clustering handle all scenarios well.

When the window size is increased to half a second, all methods perform reasonable. As in these simple scenarios the speaker is always the strongest signal, choosing the maximum correlation is sufficient.

Speaker counting

To assess the ability to detect concurrent speakers, ten mixtures with up to five simultaneously active sources were generated from the simulated data. For the case of a single speaker, the seven positions were used individually. The PoAP EM method for DoA localization was applied to all mixtures with 0.5 s time windows and the default parameters. Table 6.7 on the next page lists the number of speakers found.

A single speaker is found in all cases. Up to three concurrent speakers are found in moderate reverberation. Four or five speakers are only found consequently in mildly reverberant conditions. This may be due to the fact that the spatial likelihood saturates with artifacts. As the reverberation increases, there are more missed detections due to reverberation artifacts until speakers are not found in the whole 8 s sequence. Figure 6.20 illustrates the effect of increased reverberation on a choice of three concurrent speakers.

speakers \ T_{60}	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
1	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
2	100.0%	100.0%	100.0%	97.5%	92.5%	82.5%	77.5%	71.1%	65.8%
3	100.0%	100.0%	100.0%	93.3%	86.7%	77.2%	70.2%	63.2%	50.9%
4	100.0%	100.0%	95.8%	76.4%	72.2%	57.4%	48.4%	41.7%	34.6%
5	97.5%	100.0%	76.2%	56.2%	40.0%	31.3%	27.5%	25.3%	25.5%

Table 6.7: Single node detection of concurrent speakers. Percentage of detected speakers for different reverberation times and number of speakers.

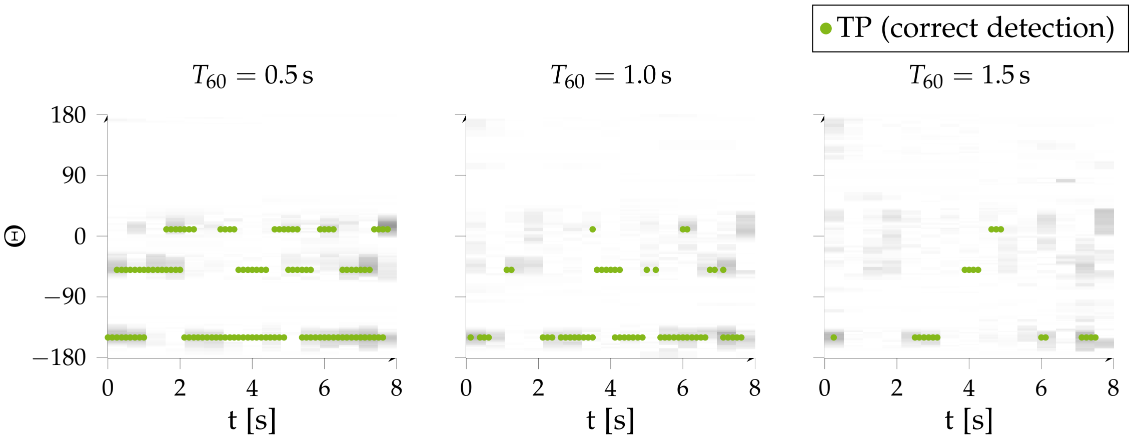

Figure 6.20: Localization of concurrent static speakers in simulation with the PoAP EM method. The spatial likelihood before EM clustering is shown in gray in the background.

6.6.2 *Smart room recordings*

The method was tested with smart room recordings for the main objective of detection and localization of speech activity. It is compared to the SRP-PHAT approach. Additionally, recordings from the AV16.3 dataset (see section 6.2.2 on page 92) were used to compare to the cumulative steered response power (C-SRP) method. The data is not ideal for the PoAP model for two reasons: First, it is sampled only at 16 kHz giving poorer phase resolution and second, the microphone array has a larger diameter of 20 cm which induces more spatial aliasing.

Speech detection and localization

In order to test the ability of the proposed method to find speech segments for the geometry calibration, it was run on sequence #2 and #4 where a single speaker talks from different static position towards the table with the microphone arrays. The PoAP EM was applied with 0.5 s time windows and the default parameters. For comparison, the basic SRP-PHAT approach was used as well with a window size of 0.5 s. As the SRP-PHAT requires additional speech detection, the best possible performance was approximated using an oracle threshold, below which localizations were discarded as non-speech. The threshold was selected so that it minimizes the harmonic mean of misses and false alarms on the given sequence. The overall DoA error was computed as well as the precison and recall as described in Section 6.3.2 on page 97. Table 6.8 lists the results.

sequence	speakers	method	ε_a	precision	recall
FINCA #2	1 static	SRP-PHAT	2.57± 8.42	71.3%	98.3%
		SRP-PHAT, oracle th.	2.54± 8.41	74.7%	98.0%
		PoAP EM	1.77± 1.46	92.6%	92.1%
FINCA #4	1 static	SRP-PHAT	3.59± 5.18	70.9%	93.6%
		SRP-PHAT, oracle th.	3.59± 5.18	72.6%	93.6%
		PoAP EM	2.47± 2.08	86.5%	91.6%
AV16.3 #01	1 static	SRP-PHAT	7.35±27.28	53.3%	94.3%
		SRP-PHAT, oracle th.	3.97±17.68	58.7%	94.3%
		C-SRP [OFK13a]	1.64± 1.23	96.0%	64.3%
		PoAP EM	1.98± 1.64	91.4%	95.3%
AV16.3 #11	1 moving	SRP-PHAT	3.08± 4.41	58.3%	96.1%
		SRP-PHAT, oracle th.	3.08± 4.41	58.7%	96.1%
		C-SRP [OFK13a]	3.46± 2.46	97.3%	94.8%
		PoAP EM	3.10± 2.91	78.7%	96.1%
AV16.3 #18	2 moving	C-SRP [OFK13a]	1.75± 1.68	100.0%	46.9%
		PoAP EM	3.09± 2.76	100.0%	44.0%

Table 6.8: Single node localization of a speakers talking from different positions towards the node in conference rooms.

The proposed method achieves best DoA accuracy in both sequences recorded in the FINCA, about 1° less error than the SRP-PHAT even with oracle information. The SRP-PHAT produces more false alarms, resulting in lower precision.

In order to allow for comparison with the literature, the methods were run on sequence #1, #11 and #18 of the AV16.3 corpus [LOGP05]. For this sequences, the output of the C-SRP method was provided by the authors [OFK13a].

In the fist sequence, a single speaker again utters two sentences at different static positions. In Figure 6.21 the localization results from the different methods is visualized. It can be seen that the SRP-PHAT also localizes a laptop fan present at around 100° in the speech pauses. Both the C-SRP and the proposed method do not suffer from this problem, as they include speech detection. The C-SRP has more misses, which might also be an effect of the speech classification step in that approach.

In sequence # 11, a single speaker moves his head very rapidly while talking continuously. Here we can see the ability of the methods to follow a moving source. All methods perform good, as the speed is not too fast nor is the movement too irregular for the 0.5 s time window, cf. [TR16].

In sequence # 18, two speakers move their heads together till touching two times while trying to talk continuously. As the ground truth assumes constant speech activity that is not always discernible, misses are expected. The straightforward SRP-PHAT approach could not be applied as it can only localize a single speaker at a time. The proposed approach shows a slightly higher angular error compared to the C-SRP as it does not follow the head movements as closely. Overall, the performance is quite similar with just a few more missed detections.

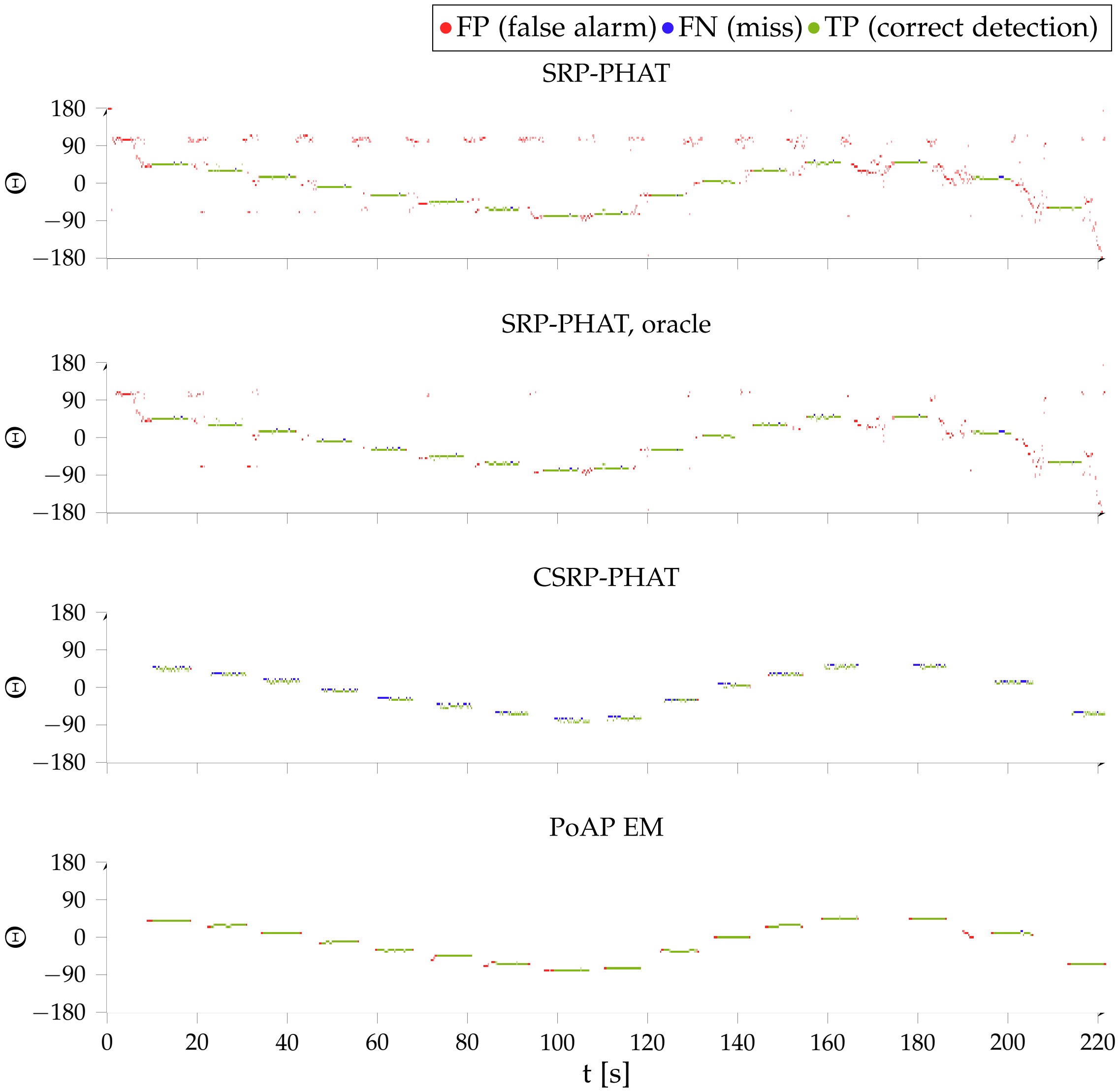

Figure 6.21: Localization of static speaker in AV16.3 sequence # 1

Speech segment identification

After the frame-wise EM, speech segments were automatically identified as consecutive detections over more than 2 s with an angular distance of less than 5° and a time to live (TTL) of 1 s. Table 6.9 lists the results for all three nodes in FINCA recordings with static speech from various positions. All positions were found with no false alarms. The smaller precision in #4 may be the result of head movement not captured by the ground truth. Figure 6.22 shows the segments found.

6.6.3 *Summary*

The PoAP EM DoA localization method is designed to automatically detect speech activity from an unknown number of sources. It was shown to be robust against reverberation in simulation with reverberation times up to 2 s.
When localizing a single speaker, the proposed method can localize correctly with about 5° accuracy even in short time windows. Both the gain estimation and probabilistic clustering make it more robust than its predecessor or the SRP-PHAT.

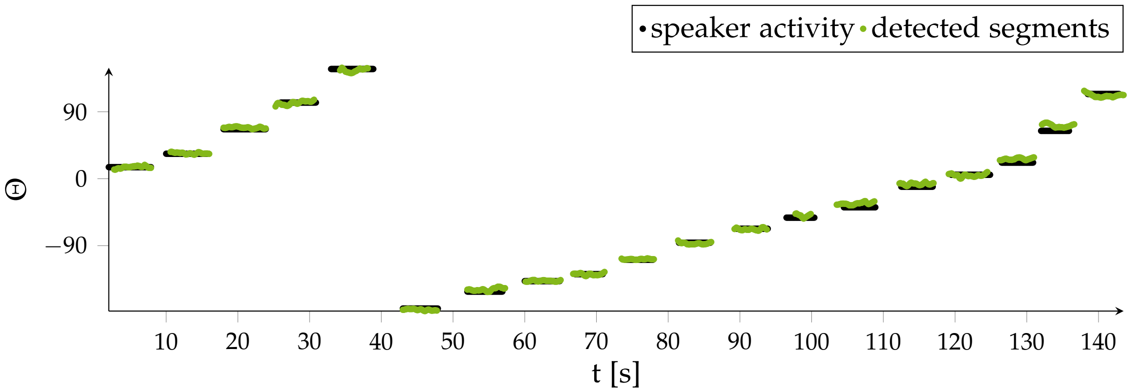

Figure 6.22: Speech segment detections of a static speaker in the smart room in sequence #4 with the PoAP EM method using a single node.

sequence	precision	recall	segments found
FINCA #2	95.4%, 94.4%, 90.9%	94.9%, 95.8%, 95.8%	100.00%, 100.00%, 100.00%
FINCA #3	95.9%, 96.7%, 94.3%	81.4%, 86.0%, 84.9%	100.00%, 100.00%, 100.00%
FINCA #4	88.7%, 87.4%, 86.0%	93.2%, 91.9%, 93.6%	100.00%, 100.00%, 100.00%
FINCA #5	97.8%, 99.6%, 98.2%	82.4%, 80.4%, 80.4%	100.00%, 100.00%, 100.00%
FINCA #6	98.2%, 98.2%, 99.2%	89.5%, 88.5%, 83.4%	100.00%, 100.00%, 100.00%

Table 6.9: Speech segment identification results for smart room recordings in the FINCA for all three arrays.

Unlike the SRP-PHAT, the proposed method is able to automatically detect speech from one or more concurrent speakers due to its neuro-inspired pre-processing and CASA based integration. Up to three speakers are detected in low to moderate reverberations. This is likely sufficient for many realistic scenarios, as overlap between more than three close talking speakers is rather unusual.

The detection and localization ability is also shown in application to real recordings from both the FINCA and the AV16.3 corpus. Fewer false alarms and comparable misses are achieved than when using the SRP-PHAT, even with an oracle correlation threshold as voice activity detection (VAD). This method would require a more reliable VAD, while the proposed method works as is. The results for the C-SRP on the AV16.3 data are similar to the proposed method. The speech classification included in that approach works sufficiently as well [OFK13a].

Most importantly, the PoAP EM approach allows to identify speech segments from static positions with good accuracy in real recordings. There are few false detections as localizations caused by footfall noise and other noises are discarded when they occur outside a constant speech segment. These segments are the basis for the geometry calibration method evaluated in the next section.

6.7 SENSOR NODE LOCALIZATION

The automated geometry calibration of ASNs is the central task discussed in this thesis. Multiple methods were developed to solve it in different scenarios. They all employ the PoAP EM DoA localization evaluated in the last section.

First, a multimodal scenario, where video cameras in fixed locations are used to localize the speaker (see section 5.3 on pages 75–78). It was evaluated in a number of simulations and smart room recordings, as described in section 6.7.1 and 6.7.2.

Second, an acoustic scenario where the speaker positions are unknown is addressed (see section 5.4 on pages 80–83). Here not only the DoA measurements, but also the time difference of arrivals (TDoAs) between the nodes are employed. To make sure that the assumption of air conduction and the resulting proportionality of time and distance are correct, other types of sounds should be rejected by the event or speech detection method. The off-line methods for the acoustic calculation will be evaluated with systematic simulations and real recordings in section 6.7.3 and 6.7.4. The online method for acoustic calibration will be evaluated using a recording from a meeting in section 6.7.5.

The recordings made in the FINCA allow to use the same recordings for both scenarios. Using this, the methods developed will be compared to each other and an online method from the literature in section 6.7.6.

6.7.1 *Simulation of the multimodal approach*

In order to test the viability of the method, several simulations were performed. All simulations used the microphone configuration in the smart room and the 10 person positions located around them in recording #2. The positions were chosen with regard to a conference scenario, either sitting at the table, standing near the table or near the whiteboard. Localization errors were simulated using Gaussian distributed random offsets with zero mean and around a predefined RMS value. All simulated errors were generated as described in Section 6.1.2 on page 90. For each configuration investigated, 100 such errors were generated for each of the three nodes, resulting in 300 runs. Over these, the actual RMS was deviated below 30% from the given value.

The method estimates the absolute geometry of each node individually. While $J = 3$ speaker positions are sufficient for estimation of the two dimensional position and orientation, it has been observed that using more positions increases the robustness. Rather than using all positions at once in one big set, multiple randomly chosen subsets can be even more beneficial. This aspect will be first investigated. The number of position sets and positions per set were varied for a simulation with an RMS of $\epsilon_l = 20\,\mathrm{cm}$ and $\epsilon_a = 4°$. These values correspond to the error found in practice by the localization algorithms used. Then the measurement errors themselves were varied to investigate its effect on the calibration accuracy.

Number of position sets

Assuming a zero-mean distributed error of the localizations, it is clear that the average error for multiple position sets will decrease with the number of sets used. Figure 6.23 on the next page shows the geometry estimation errors for position ϵ_r and orientation ϵ_o. The mean of different numbers of sets N for estimations with $J = 6$ positions was computed over 128 choices.

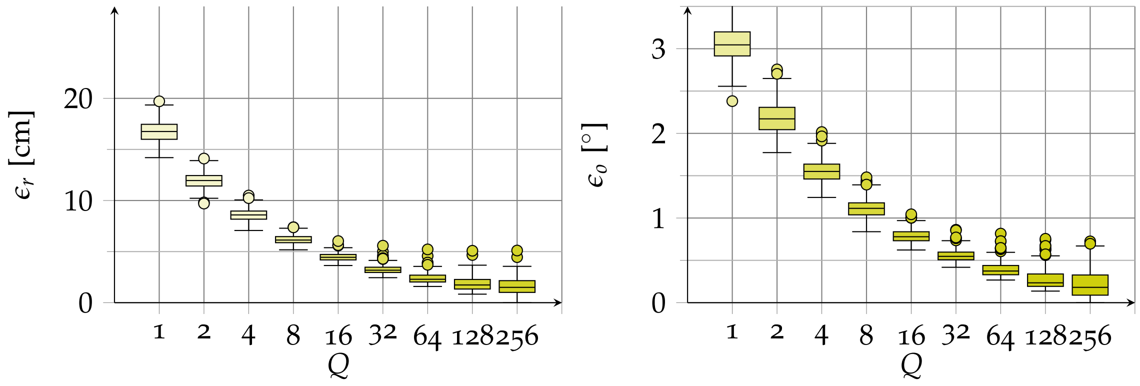

Figure 6.23: Estimation error and its standard deviation for the mean from a different numbers of position sets Q of $J = 6$ positions. Based on 100 simulations with a localization error of $\epsilon_l = 20\,\text{cm}$ and $\epsilon_a = 4°$ for each of the three nodes.

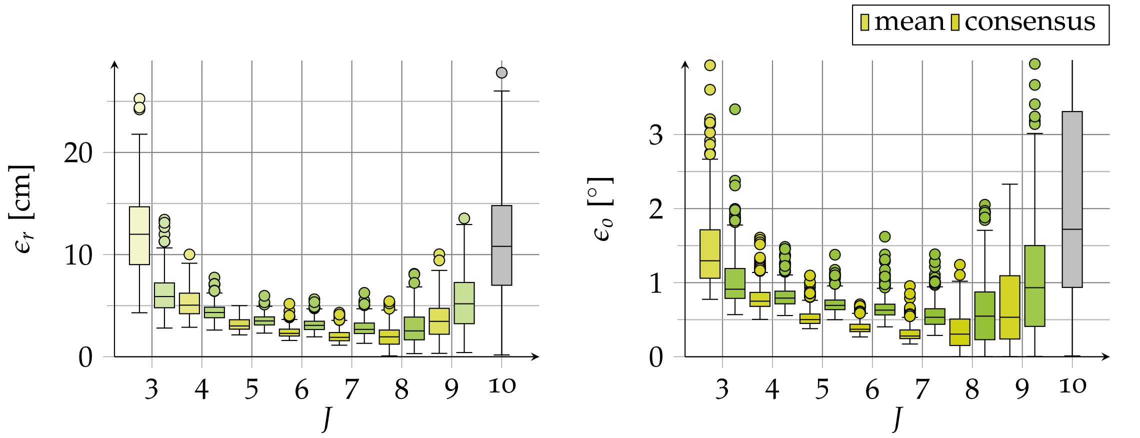

Figure 6.24: Estimation error for different numbers of positions J in the sets using the mean of $Q = 64$ choices and a $Q' = 16$ consensus. Results for 3x100 simulations with a measurement error of $\epsilon_l = 20\,\text{cm}$ and $\epsilon_a = 4°$.

As the problem is not necessarily convex for erroneous measurements, the choice of the Broyden–Fletcher–Goldfarb–Shanno algorithm could be suboptimal. In order to investigate this, the differential evolution was used on the same simulations in place of the bounded gradient descent. The results were almost numerically identical and not significantly different for any configuration ($p > 0.2, N = 250,000$).
The estimation error decreases with the number of position sets used. Each increase in the number of sets lead to significantly better results according to a permutation test ($p < 0.001, N = 250,000$). It is also lower than the error of the initial measurement, as the different errors compensate each other.

Number of positions

The number of positions J used in the set was varied from three to all ten in the next experiment. $Q = 64$ sets were chosen, and the $Q' = 16$ estimates closest to the median were used as consensus to compute an improved estimate. Figure 6.24 shows the resulting errors using either the consensus or the mean. Again, the evolutionary optimization was run in comparison on the same data. The results were again almost identical and showed no significant difference according to permutation tests ($p > 0.7, N = 250,000$).

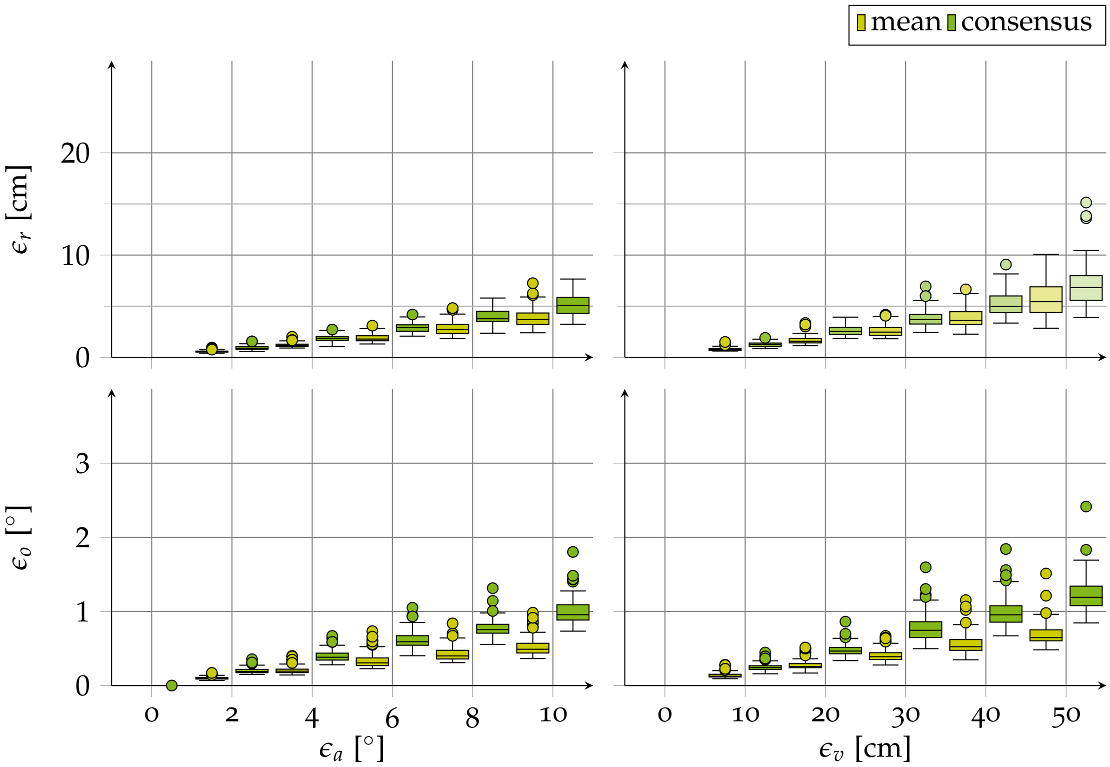

Figure 6.25: Estimation error of orientation ϵ_o and absolute position ϵ_r and its standard deviation for different audio and video localization errors ϵ_a, ϵ_v. The mean and standard deviation over 3x32 simulations is plotted for both the mean and consensus method for sets of size $J = 6$.

The error decreases up to $J = 7, 8$ out of the ten positions, then slowly rise again. Using all positions jointly (shown in gray) performed worse than $J = 5 - 9$. The differences were found significant except between $J = 7$ and $J = 8$. This shows that using subsets increases the robustness of the approach.

The consensus has a s smaller position error than the mean for small sets $J = 3, 4$ while for bigger sets the mean is better as the zero mean errors cancel each other out. This does not necessarily hold for other error distributions.

Measurement error

In order to investigate the influence of measurement errors, different audio localization errors $\epsilon_a = 0, 2, \ldots, 10°$ and video localization errors $\epsilon_l = 0, 10, \ldots 50\,\text{cm}$ were simulated. $N = 64$ sets of positions and a $N' = 16$ consensus were used with $J = 6$. The resulting geometry estimation errors are shown in Figure 6.25. For all simulations, the orientation error ϵ_o is lower than $3°$, which is beneficial for localization target applications, since the triangulation quality decreases rapidly with angular errors. The visual localization error ϵ_v translates to a position estimation error. As the result is averaged over several positions, the individual localization errors cancel themselves out partially. The resulting position error is well below half the visual localization error. The influence on the estimated orientation is similar but even less strong, with about one degree orientation error for $\epsilon_v = 30\,\text{cm}$.

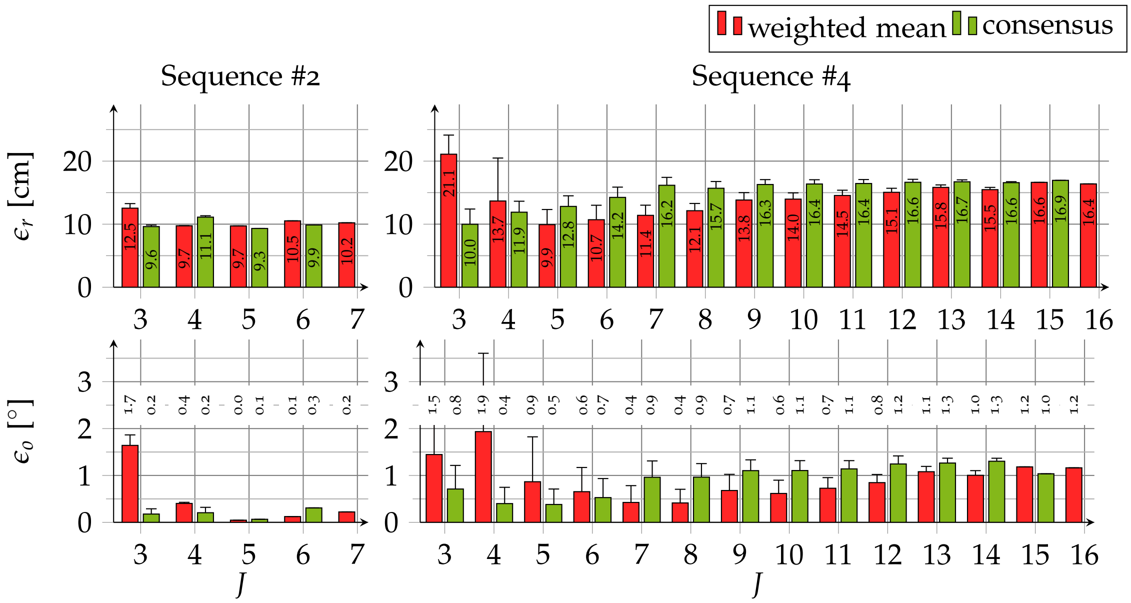

Figure 6.26: Calibration results for the second array in the multimodal approach for different consensus strategies. Mean and standard deviation are plotted as the number of runs is limited.

6.7.2 *Multimodal smart room recordings*

Two recordings with both cameras and microphone arrays were used to test the multimodal calibration approach. Sequence #2 with the speaker taking up ten positions and sequence #4 with the speaker at a total of 19 positions. The DoA localization method (see section 4.2 on pages 54–58) was applied in order to determine speech segments and the DoA of the speaker at the arrays. Time segments with low angular deviation in all the arrays were grouped as described in Section 4.2.3 on page 58. The mean DoA estimates of the positions had an error of $\epsilon_a = 3.00°$ and $3.58°$ over all arrays, respectively. For each of these time segments, the visual localization using histograms of oriented gradients (HoG) descriptors and background subtraction filtering was applied to the camera images [Bri13]. As the speaker was not found in the camera image at all positions, only 7 and 16 positions could be used for the calibration algorithm, respectively. The Euclidean localizations had a mean RMS error of 16.7 and 21.9 cm, respectively.

Figure 6.26 shows the results using different positions set counts J for both sequences using either the weighted mean or consensus. Both the position and orientation error are plotted over random choices of position sets.

Due to some big errors in the visual localizations for sequence #4, the mean performs badly up to a position set size of 6. The consensus is much more robust against this kind of error. Here good results are already achieved at a set size of 4.

The results for the more concise recording #2 are generally better. This reflects the fact that a well organized calibration sequence with close and distributed positions is not only sufficient but favorable for the calibration task. Below 10 cm location error and around 1° orientation error are readily achieved. In contrast, the longer sequence #4 leads to slightly worse calibration accuracy. Around 15 cm and 1.5° are achieved. Even if the outliers are removed by the consensus method, the overall worse measurement accuracy has a negative effect that is not fully compensated by the use of more positions.

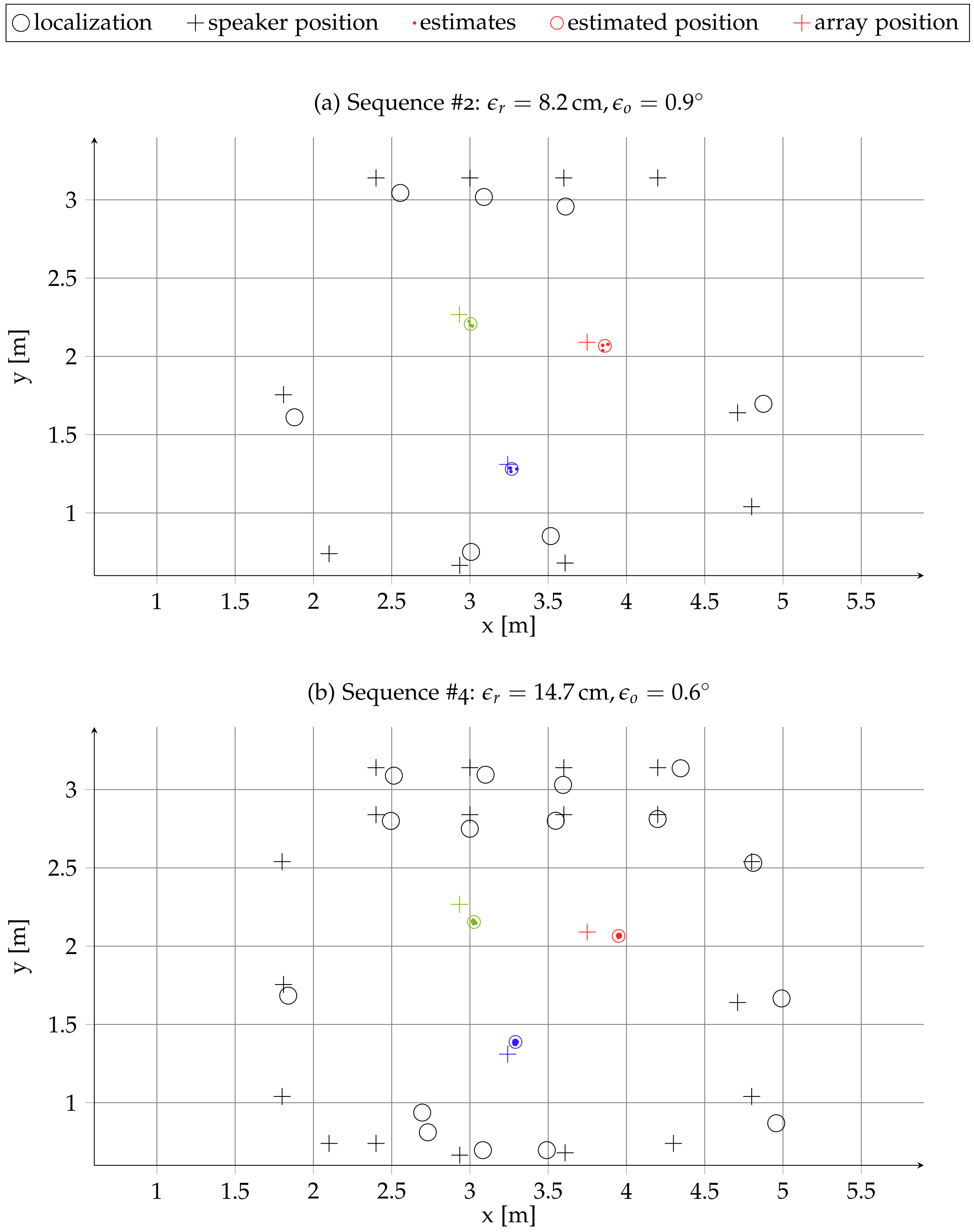

Figure 6.27: Multimodal calibration of sequence #2 (top) and #4 (bottom). The 10 and 19 speaker positions and the 7 and 16 visual localizations are shown along with the array localization in the FINCA. The consensus of $J = 5$ positions sets is used, all estimates used in the consensus are plotted together with the final estimates for each of the three arrays in red, green, and blue, respectively.

Figure 6.27 shows the result of a single calibration run for both sequences along with the visual localizations. The outlier in the visual localizations for sequence #4 is clearly visible. The array positions results are more strongly biased which may be the result of the bias introduced by errors in the visual localization.

6.7.3 *Simulation of the acoustic approach*

Several simulations were performed to investigate the properties of the acoustic geometry calibration approach (see section 5.4 on pages 80–83). The simulations were all based on the ground truth data of sequence #1. First, the different optimization methods were applied in comparison. Second, measurement errors were simulated to determine the influence on the accuracy of the method.

Optimization methods

Two different optimization strategies were proposed, hierarchical grid search [PF14c] and differential evolution optimization [PFG17]. As the optimization is different from the multimodal approach, it is again interesting to see the influence of the position set size. So, all possible set sizes $J = 3 \dots 10$ were used in the simulation. The discarding of estimates with a high error function value was turned on and off for both methods. Several runs with a normally distributed measurement error with an RMS of 4° and 5 cm around the three-dimensional ground truth positions were computed. Up to 64 position sets were used to compute the weighted mean. The results are shown in Figure 6.28.
It can be seen that the position error is very similar for all strategies. The thresholding criterion improves the result when using all positions, which means that for several runs of the grid search a worse overall estimate was found.
The orientation error is almost consistently better for the differential evolution optimization. The thresholding has no visible effect, meaning that the target function values are typically already lower than 10 cm. In the grid search case, the results improve though the thresholding, showing that some bad solutions are found.
The computation times were measured using the python implementation on a core i7 processor. Four subsets were computed simultaneously on four cores. The differential evolution optimization is faster to compute than the grid search. The latter is about one order of magnitude slower. It can be seen that the subset size has a small influence on the computation time necessary. This is the result of two effects working against each other. While the computation of a larger set takes longer, there are less than the maximum 64 sets when choosing sets of size $J \geq 8$ out of the ten positions. The single choice of all positions was computed in around 10 s and 300 s compared to 500 s and 8,000 s for $J = 7$ with the differential evolution and the grid search, respectively.
Overall, there is a slightly better performance for the differential evolution optimization. This method was chosen for the rest of the evaluation, as it provides more consistent results with less outliers and is faster to compute. There is no clear choice for the number of positions J in the sets.

Measurement errors

As with the multimodal approach, different measurement error values for both types of measurement were simulated, cf. Section 6.7.1 on page 125. As the method does assume a planar geometry, additional distance estimate errors are induced by the height of the speaker relative to the table. Therefore, the simulation was done with the source placed

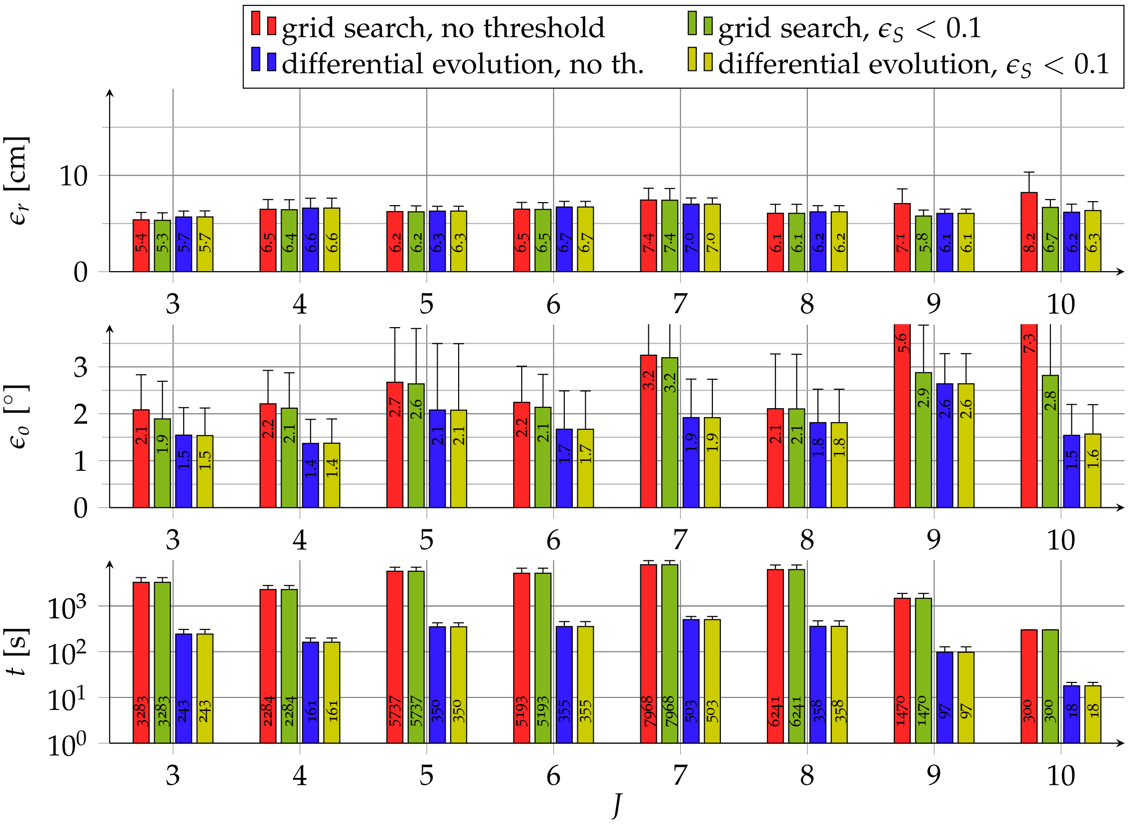

Figure 6.28: Estimation error of absolute position ϵ_r and orientation ϵ_o as well as computation times, means and standard deviations for different optimization strategies. The acoustic geometry calibration applied to a simulation of scenario #2 with random distributed errors of $\epsilon_a = 4°$ and $\epsilon_\tau = 5$ cm.

at table height as in sequence #1 and the speaker standing or sitting as in sequence #2. The results are plotted in Figure 6.29. It can be seen that the method is more sensitive to DoA measurement errors than the multimodal approach, cf. Figure 6.25 on page 125.

The planar assumption induces an additional position error of around 6 cm as can be seen in the simulation of angular measurement errors. This reflects the shortening of the TDoAs due to speaker elevation. There is no observable effect on the estimated orientations.

Errors in the TDoA measurement do not affect the orientation estimation much, the error increases up to around 2° for an distance measurement error of 20 cm and does not increase much for higher errors. They again induce position errors, as the measurement error in the co-planar case. The effect is rather small as the error slowly increases to around 10 cm for distance measurement errors of 50 cm. The effect of the source elevation becomes weaker for higher errors.

Errors in the DoA measurement cause both position and orientation errors. As the geometry is deduced over all arrays, it is understandable that the faulty orientation of the arrays leads to a certain position offset. This offset increases with the DoA error to about 6 cm for an DoA error of 8°. The orientation estimate also degrades with increasing DoA errors, it is below half the measurement error. Compared to the multimodal approach, the effect of DoA errors is stronger, which is understandable as the geometry is computed over all three arrays.

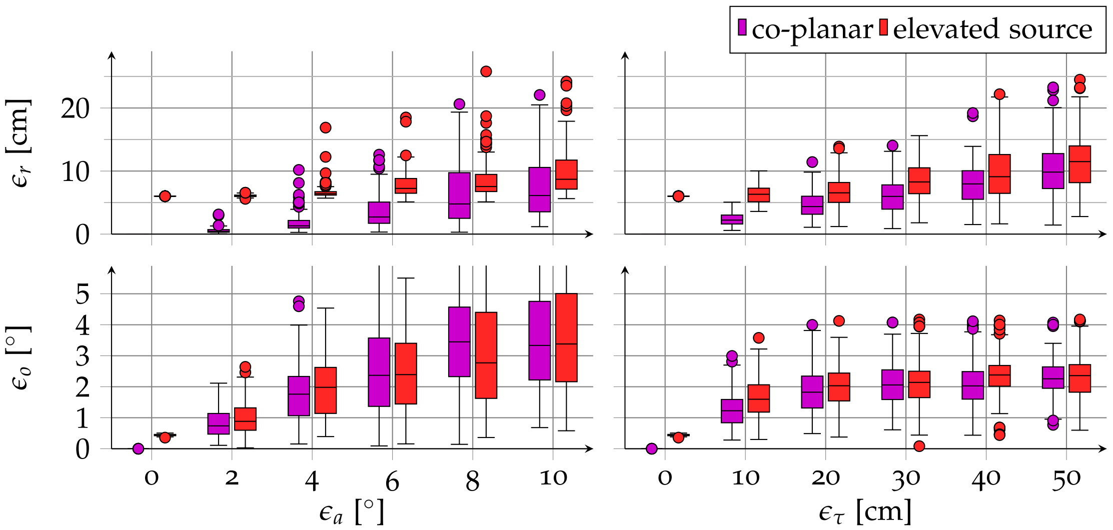

Figure 6.29: Estimation error of orientation ϵ_o and absolute position ϵ_r and its standard deviation for different angle and distance errors $\epsilon_a, \epsilon_\tau$. The mean and standard deviation of the estimate provided by the differential evolution optimization on all positions is plotted over a total of 100 simulations each. Both a co-planar source and a standing and sitting speaker that is elevated from the table plane was used for comparison.

6.7.4 *Offline ASN calibration with real recordings*

Several recordings with a single source directed towards the table in the FINCA (see section 6.2.1 on pages 90–92) were used to test the acoustic calibration method (see section 5.4 on pages 80–83). In sequence #1 a smartphone was used to play white noise at table height, in the other sequences #2, #3, and #4 a human speaker was saying a sentence from positions sitting in a chair or standing up in the room. In sequence #1 and sequence #2, the same ten positions were used. In #3 and #4 15 and 19 positions were used, respectively. The additional positions were situated at the whiteboard further away from the table. Table 6.10 summarizes the accuracy of the measurements and the resulting geometry calibration error using differential evolution optimization on all positions.

Again, the DoA estimation was used to automatically determine the time segments corresponding to the different positions (see section 4.2 on pages 54–58). The angular RMS error was 4.4°, 3.9°, 7.3°, and 5.4° with respect to the ground truth positions for sequence #1 to #4, respectively. The slightly higher error for the latter sequences is probably caused by the speaker being further away and the signals having higher reverberation.

The TDoA measurements based on SRP-PHAT had an RMS error of 3.4, 8.9, 19.5 cm and 12.1 with respect to the two-dimensional ground truth positions for sequence #1 to #4, respectively. The higher deviation in the human speaker case stem most likely from his elevation. When computing the TDoA error with respect to the three-dimensional ground truth positions, the RMS error is lowered by 4-6 cm to 4.5, 15.8 cm, and 3.3 for sequence #2 to #4. When using the PoAP spike representation in order to compute the TDoA measurements, the RMS error with respect to the two-dimensional ground truth positions is similar or better than the SRP-PHAT based estimate with 6.5, 19.5 and 10.2 cm for sequence #2 to #4. For the white noise it is slightly worse with 5.3 cm.

		measurement RMS			calibration error			
sequence	T	ϵ_a	ϵ_τ		ϵ_o		ϵ_r	
		PoAP	PHAT	PoAP	PHAT	PoAP	PHAT	PoAP
#1 noise	10	4.44°	3.34 cm	5.27 cm	2.40°	2.01°	2.84 cm	2.73 cm
#2 speech	10	3.93°	8.92 cm	6.47 cm	2.79°	1.97°	5.94 cm	3.64 cm
#3 speech	15	7.28°	12.89 cm	12.96 cm	1.43°	1.48°	9.06 cm	8.90 cm
#4 speech	19	5.45°	12.06 cm	10.17 cm	1.35°	0.69°	8.92 cm	7.06 cm

Table 6.10: Error of the measurements used for acoustic geometry estimation with calibration sequences and the resulting error of the calibration with differential evolution optimization on all positions.

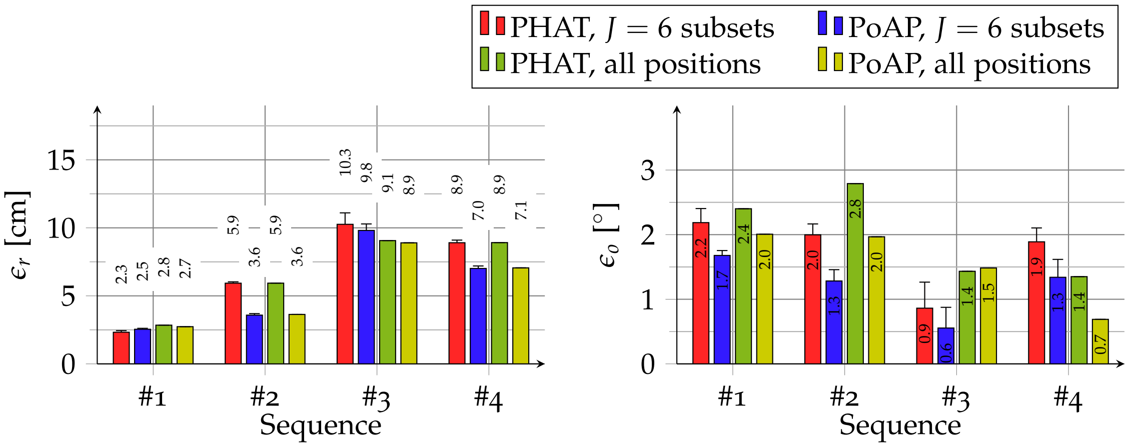

Figure 6.30: Calibration results for the acoustic approach on dedicated calibration sequences recorded in the FINCA. The differential evolution optimization was used and the weighed mean of estimates on random subsets of $J = 6$ positions or all positions were used to calculate the final geometry estimate. Either the SRP-PHAT or PoAP data was used to estimate the TDoAs. In sequence #1 smartphone was used to play white noise from 10 positions at table height, in the other sequences #2, #3, and #4 a human speaker was saying a sentence from positions sitting in a chair or standing up in the room from 10, 15 and 19 positions, respectively.

This can be understood since the spike representation is not tuned for broadband noise signals.

Figure 6.30 shows the calibration results using the differential evolution optimization. Two methodical aspects were varied: First, either the weighted mean of estimates for random subsets of $J = 6$ positions was used or the estimate was computed once over all positions. Second, either the SRP-PHAT or the correlation of the PoAP spikes was used to measure the TDoAs. The calibration error is about 2 cm and 2° for the noise recording #1. When the human speaker is used instead of the smartphone, the position error increases to about 6 cm. This is in line with the expected error resulting from the TDoA estimates being shortened by the speaker's elevation. The same effect was seen in the simulation (see section 6.7.3 on pages 128–129). Using the sequences with more speaker positions, the position error increases to around 10 cm while the orientation error decreases slightly. Overall, the accuracy provided is more than sufficient for most applications. The effect of the calibration on the speaker tracking performance is investigated in Section 6.8.2 on pages 140–141.

The use of the sparse spike representation as shared information for the calculation of the TDoAs does not reduce the achieved accuracy. In most cases the error is lower than when applying the SRP-PHAT on the time signals. This may hint at more consistent data.

The calculation time on a core i7 processor running at 3.4 GHz was 8-21 s when using all estimates. Assuming four parallel cores sharing the computation of the 64 sets of positions, it was 92-142 s for the subset method. As the subset-based mean is not clearly better in terms of calibration accuracy and it is much faster to compute an estimate over all positions, the latter seems preferable for dedicated calibration sequences.

6.7.5 *Online ASN calibration*

When using position subsets, the acoustic calibration can be performed online with a growing set of measurements. After each speech event, solutions for new sets of positions and a new weighted mean estimate is computed (see section 5.4.4 on page 83). As the amount of information to share is reduced by about two orders of magnitude when using the PoAP spikes instead of the signals for correlation, this is the feasible choice for wireless acoustic sensor networks (WASNs).

This method was applied to the meeting recording in sequence #7. Here, five people entered the FINCA, greeted each other and then sat or stood at a random position while talking to each other. The measurement of DoA and TDoA was done within the first five seconds of each utterance. The computation in the nodes was simulated by running the python implementation on a core i7 processor at 3.4 GHz. As the computation time for the geometry with the differential evolution algorithm is around 10 s per set for small sets on one core, multiple estimates can be computed between speech events. The number of positions J in each set is fixed and the weighted mean estimate is computed over random choices of such sets.

In order to investigate the influence of the number of positions in the subsets $J = |S_k|$, it was varied between the minimal value of three up to seven. Likewise the population size for the differential evolutionary optimization was varied. Table 6.11 shows the number of sets evaluated $\#S_k$, the time used to optimize one set of positions t_ϵ, the position error ϵ_r and orientation error ϵ_o over 100 runs of the algorithm. The errors were computed as average over the time of the meeting. The results are quite similar, showing the robustness of the approach. Using small population sizes U lead to slightly higher errors, even though much more individual solutions can be computed through the decreased time required for each one. Position set sizes of $4 - 6$ perform well, beyond that the accuracy begins to drop.

In order to visualize the algorithms behavior over time, in Figure 6.31 the speech events and the calibration error are plotted for one run of the algorithm with $J = 5$ and two cores. In order to provide an indication as soon as possible, the geometry calibration process is started when three speech events are available. When there are five or more, the nodes start collecting estimates for the weighted mean. In the initial phase with less than six speech events, the error already decreases as the first speech events are rather well localized. This can not be expected in general, though. When enough speech events are measured for the weighted mean, the results are more reliable. The position error has decreased to 7 cm and stays around this value for the whole meeting. In the time

$\lvert S_k \rvert$	U	$\#S_k$	t_ϵ[s]	ϵ_r [cm]	ϵ_o [°]
5	5	459±102	2.4±0.4	8.2±0.2	1.7±0.1
5	15	145±039	7.2±1.5	8.0±0.3	1.5±0.2
5	**25**	**77±021**	**11.9±2.2**	**7.9±0.4**	**1.4±0.3**
5	35	52±012	16.4±3.7	7.9±0.5	1.5±0.3
5	45	43±004	19.2±3.9	7.8±0.4	1.4±0.3
3	25	127±027	8.6±1.5	7.8±0.9	1.9±0.5
4	25	110±026	9.3±1.8	7.9±0.5	1.6±0.3
5	**25**	**77±021**	**11.9±2.2**	**7.9±0.4**	**1.4±0.3**
6	25	59±020	13.8±2.9	8.0±0.2	1.4±0.2
7	25	53±013	14.2±3.1	8.2±0.3	1.4±0.2

Table 6.11: Comparison of different parameterizations of the online calibration approach

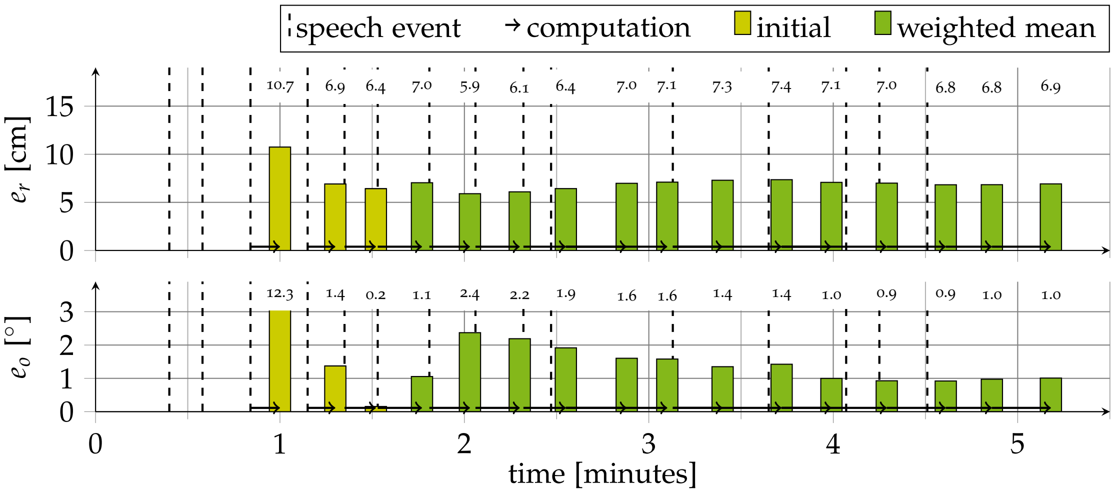

Figure 6.31: Online calibration. Realtime processing using the measurements (dotted lines) to update the geometry estimate over the course of a meeting. The resulting position errors are shown in the top graph, the orientation errors on the bottom.

period of two to four minutes, the orientation error decreases from around 2° to 1° and stays there.

6.7.6 *Summary*

Different approaches for the calibration of distributed microphone arrays were proposed. A multi-modal approach, and an acoustic approach that can use different optimization strategies. All methods use the DoA localization computed by the neuro-biologically inspired approach described in Section 4.2 on pages 54–58.

The multi-modal method requires video cameras to be mounted at known positions, therefore the applicability is limited to such conference room scenarios. It provides absolute positions that can be directly used to perform, e.g., acoustic speaker tracking for camera control. The use of a consensus over several random subsets leads to higher

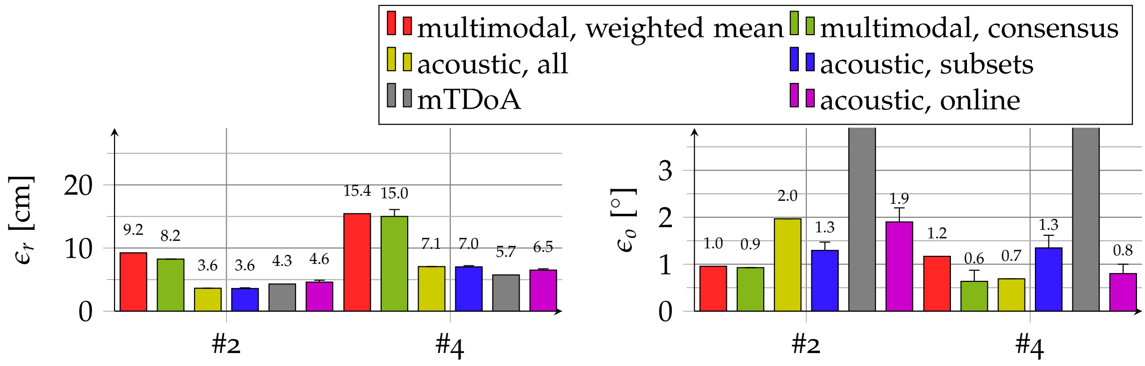

Figure 6.32: Comparison of different methods for array configuration calibration. Position (left) and orientation error (right) for two calibration sequences with 10 and 19 speaker positions. The results for the multimodal and the acoustic method with PoAP spikes and differential evolution are shown. The mTDoA method was evaluated for comparison.

accuracy and robustness. This is also reflected by the small influence of measurement errors.

The acoustic method can only provide relative positions for the sensor nodes. Since it requires only the nodes themselves, it is applicable in a range of ad hoc scenarios. The differential evolution optimization provides better and faster results than the grid based approach. The use of the PoAP spikes for correlation allows to reduce the amount of shared information without reducing the accuracy of the method. Thus, the use of the SRP-PHAT is not recommended. For calibration sequences, the use of all positions in a joint optimization is as accurate as using position subsets, therefore the former is recommended as it requires less computational effort. In the case of online calibration, this is not so. The weighted mean over random subsets ensures the robustness against outliers which are more likely in an unconstrained scenario. A continuously growing set of estimates allows for real-time application.

In Figure 6.32 the results for two calibration sequences are shown in comparison. The multi-modal approach is compared to the acoustic one. For the acoustic approach, the PoAP spike correlation was used in conjunction with the differential evolution optimization. Both the off-line and online version were run with position subsets of size $J = 6$. All proposed acoustic methods calibrate with an error of 7 cm and around 1°. From the simulation, it is clear that the position error is partly due to the speaker's elevation. Similar performance was achieved in the online application of the method during a meeting. In the short dedicated calibration sequence #2, the position error is lower with about 4 cm. This is likely due to the speaker's higher proximity to the nodes.

For comparison, the maximum time difference of arrival (mTDoA) and multidimensional scaling (MDS) approach was re-implemented [PMH12]. It achieves a comparable position accuracy. However, the orientation can not be estimated reliably from the microphone positions. The orientation error is much higher than 5° when using singular value decomposition (SVD) (as in [PMH11]) or minimum angular difference for alignment of the known geometry.

As the multi-modal approach optimizes the array's positions independently, the error in measurement translates more directly. Additionally, the visual localization had a higher error than the correlation based distance measurements. Thus, a higher position error of

15 cm and 10 cm is observed. The multimodal approach has slightly better orientation accuracy than the acoustic approach in the short sequence, the values are similar for the long sequence.

It can be remarked that both proposed methods perform well within the accuracy required for speaker tracking (see section 6.8.2 on pages 140–141) and better than the state-of-the-art methods working with speech, cf. pages 86–86.

6.8 ASN SPEAKER TRACKING

In order to evaluate the multi speaker tracking method proposed for acoustic sensor networks [PF14a], several simulations and recordings in different settings were performed. First, the effectiveness and robustness was investigated using a dedicated simulation with a single speaker. The proposed weighted triangulation was compared to other approaches. The effect of the number and position of sensor nodes was tested. The robustness against reverberation, transmission errors and jitter is shown. As the geometry calibration method is the basis for employing the tracking in an ad hoc configuration, the influence of geometry calibration errors was investigated.
Several recordings of human speakers in the smart room with three sensor nodes in the table were used. The effect of speaker movement was investigated. By using recordings with concurrent speakers, the ability to automatically detect the number of speakers and associate the DoAs across the nodes is shown.

6.8.1 *Triangulation error*

One of the novelties introduced in the speaker tracking is the weighted triangulation. The weighting function was derived from a simulation experiment. Two nodes were placed 4 m apart and DoAs with intersection angles α were set with and without angular errors ϵ_a.

$$\begin{aligned} &\boldsymbol{r}_1 = [2,0]^T, && \boldsymbol{r}_2 = [-2,0]^T, && o_1 = o_2 = 0 \\ &\theta_{n,1} = 0,1,\dots,359, && \theta_{n,2} = \theta_1 + \alpha, && \alpha = 1^\circ,2^\circ,\dots,170^\circ \\ &\hat{\theta}_{n,1} = \theta_{n,1}, && \hat{\theta}_{n,2} = \theta_{n,2} + \epsilon_a, && \epsilon_a = 1^\circ,2^\circ,\dots,10^\circ \end{aligned} \tag{6.17}$$

The source position without error $\boldsymbol{s}_{n,(1,2)}$ and with error $\hat{\boldsymbol{s}}_{n,(1,2)}$ was computed by triangulation as described in Section 2.1.3 on pages 14–15 using either $\theta_{n,1}, \theta_{n,2}$ or $\hat{\theta}_{n,1}, \hat{\theta}_{n,2}$, respectively. The Euclidean distance between these two position gives the localization error ϵ_l.

$$\epsilon_l = \left\| \hat{\boldsymbol{s}}_{n,(1,2)} - \boldsymbol{s}_{n,(1,2)} \right\| \tag{6.18}$$

Figure 6.33 shows the resulting error and the reciprocal of the weighting function. It can be seen that the sine reciprocal gives a reasonable fit, especially in the range of 0° to 90°. At the other end, angles close to 180°, the metric error is smaller as the intersection is

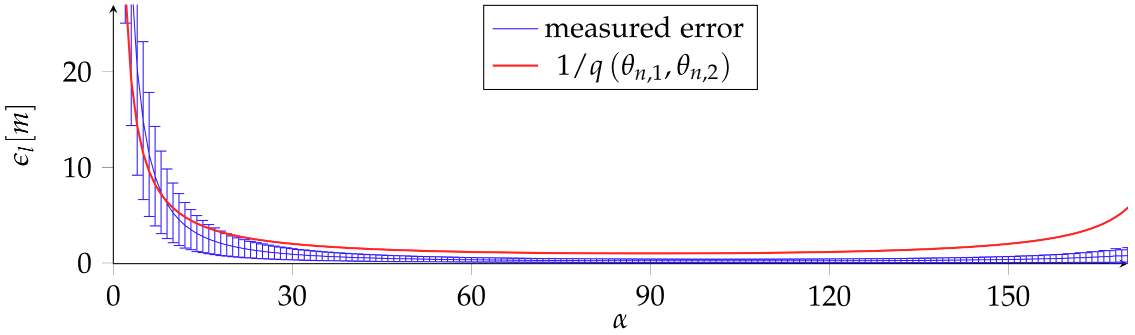

Figure 6.33: Triangulation error as function of the intersection angle from simulation and the weighting function.

closer to the nodes. While the reliability of the intersection decreases again above 90°, the absolute error decreases due to the proximity. Modeling this more exactly would require a more complex weighting function taking the node distance into account. The configurations of nodes considered in this thesis are not likely to produce angles close to 180°, since the speakers would have to move in between nodes. The issue is not further investigated here.

6.8.2 *Room simulation*

To evaluate the system's basic properties and its fitness for the task, a number of experiments were performed on signals computed by simulation. A rectangular $6.5 \times 3.5 \times 2.5\,\text{m}$ room was simulated using the ISM with a shoe-box model as described in section 6.1.1 on page 89. Five nodes were placed in the inner part of the room at table height and a single speaker speaks from 18 positions around the arrays as illustrated in Figure 6.34. The reverberation time was varied from 0 to 2.0 s in 0.25 s steps. According to the approximation formula (2.2), the critical distance r_D decreases to 0.3 m at $T_{60} = 2.0\,\text{s}$. The simulation was repeated seven times with different speech signals at the different positions in order to make the results independent of the actual speech content. Each speech event is 4 s long.

The tracking algorithm described in Section 4.3 on pages 59–60 was run on DoA localizations computed by the method described in Section 4.2 on pages 54–58 in a moving window of 0.5 s with a time step of 0.25 s. In the experiments, the Euclidean localization was not constrained by the actual room dimensions. However, positions farther than 9 m from the nearest array were rejected automatically.

Triangulation

The first thing investigated is the triangulation strategy. Different methods were tested with the multi-node simulation. The weighted triangulation introduced in this thesis is compared to an unweighted combination of the pairs' intersections. Additionally, the two nodes with the highest signal amplitude were used as in [TKH14]. The idea behind

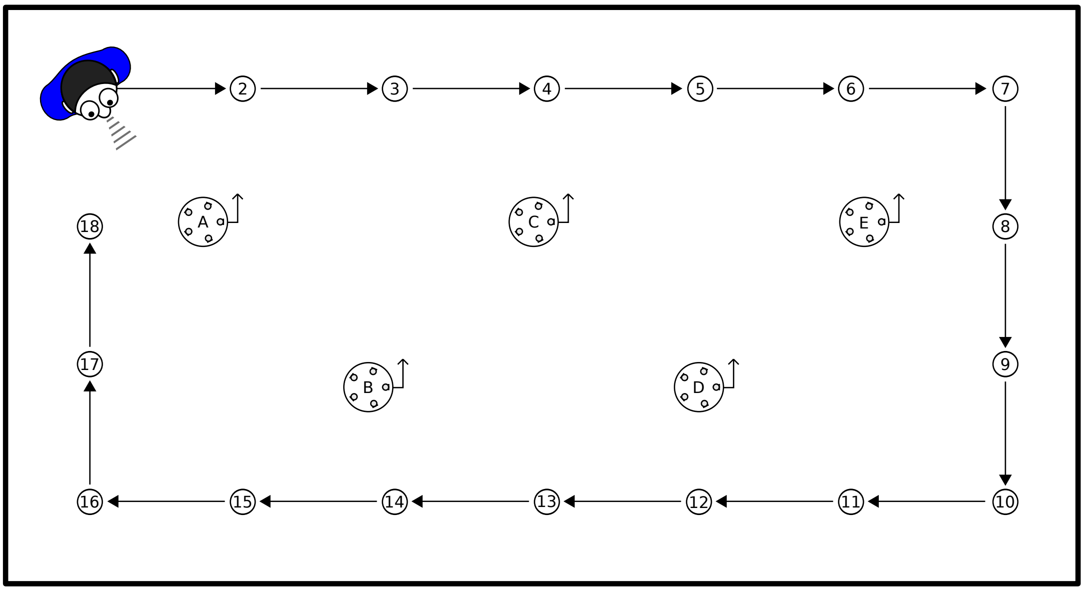

Figure 6.34: Simulated room with five sensor nodes and a single speaker at 18 positions.

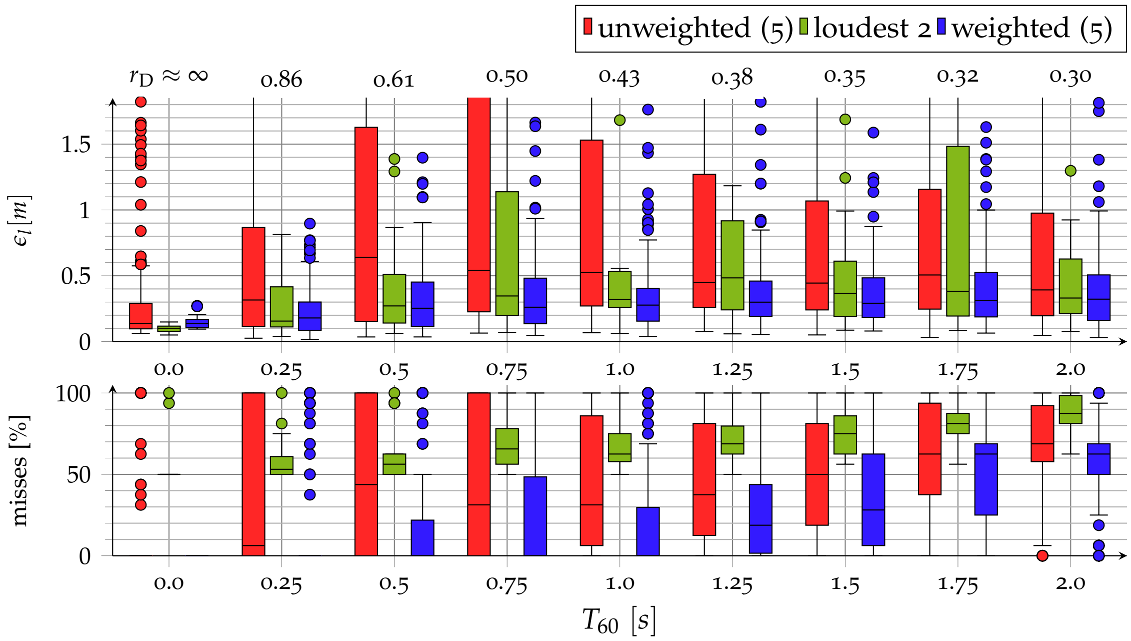

Figure 6.35: Localization error and misses for a simulation of a single speaker tracked by five sensor nodes at different reverberation times for different triangulation strategies. Distribution over all speaker positions and simulation runs with different utterances.

this is to use the closest nodes that most likely have the best angular estimates. Using only two angles, triangulation is straightforward line intersection.

Figure 6.35 shows the two-dimensional localization error for different reverberation times when using these three different modes for triangulation. In the bottom, the number of missed frames is plotted. A frame is counted as missed if there is no localization or the speaker is further than 0.5 m away from the nearest localization.

The weighted triangulation has the lowest localization error, increasing from about 20 cm at $T_{60} = 0.25$ s to about 50 cm at $T_{60} = 2.0$ s. The median of the number of misses is 0% below $T_{60} = 1.0$ s and then steadily increases to about 60% at $T_{60} = 2.0$ s. The unweighted triangulation already degrades to 1 m at $T_{60} = 0.5$ s on average and is significantly worse in most scenarios ($p < 0.001, N = 250,000$). The use of the two nodes with the highest amplitudes is performing better, close in localization error to the weighted triangulation for low reverberation levels. However, there are more missed frames. In both respects it is significantly worse than the weighted triangulation between $T_{60} = 0.75$ and 1.75 ($p < 0.01, N = 250,000$).

Node count and positioning

Not only the number but also the positioning of the nodes used has direct influence on the result. Figure 6.36 shows the localization error for different subsets of the simulated sensor nodes.

When only two nodes can be used, the accuracy is worse because the triangulation is using close angles in some cases. Interestingly, the two nodes B,D that are close together in the middle of the room perform far worse than the use of the nodes A,E that are further apart. This is likely related to the larger intra-node distance. The fact that the speaker's trajectory is closer to A,E in a few more instances does not seem to be the major

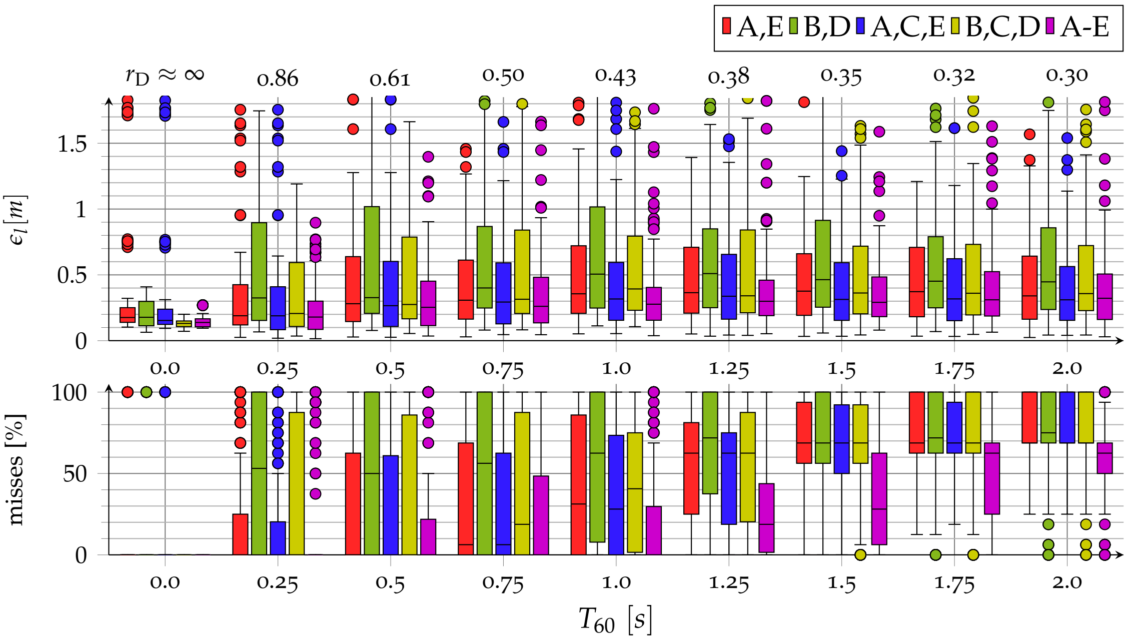

Figure 6.36: Localization error (top) and number of missed frames (bottom) of different selections of nodes at different reverberation times over all simulations and speaker positions.

influence since the co-linear nodes A,C,E spanning 5 m also achieve better localizations than the three central nodes B,C,D with a maximum inter-node distance of 2 m. The advantage of the bigger aperture is also reflected in the number of missed detections which is clearly higher for the smaller choices. As the variation over the positions and simulation runs is rather large, a permutation test only showed a significantly better localization for the use of all nodes (A-E) versus the other configurations ($p < 0.01, N = 250,000$) for T_{60} below 1.5 s.

Transmission failure of some nodes may temporarily change the number of used sensor nodes. Depending on the conditions, not all nodes may be able to send their localization at a given time step. In order to investigate the influence of the node count and the robustness of the system, a fixed number of nodes were selected randomly for each time step. Figure 6.37 on the following page shows the localization error ϵ_l for different counts and reverberation times as well as the number of missed detections.

Again, using only two nodes performs significantly worse and the accuracy increases with the number of nodes. Starting with three nodes, the median localization error is below 0.5 m. The gain from using five over using four nodes is minimal. The miss rate increases with the reverberation.

Drift and Jitter

Regardless whether the signals from the nodes are transmitted over wireless or wired connections, by exchanging information every frame in real time and using the integration node's clock as reference, drift can be avoided. Severe jitter of up to a time step (0.25 s) may be the result of different clocks at the nodes or transmission delays. In order to simulate this, random jitter with a given RMS was added to the nodes inputs signals at $T_{60} = 0.5$ s. The results are shown in Figure 6.38. The tracking error shows no significant increase, even if the jitter reaches unrealistic values that can not be caused by

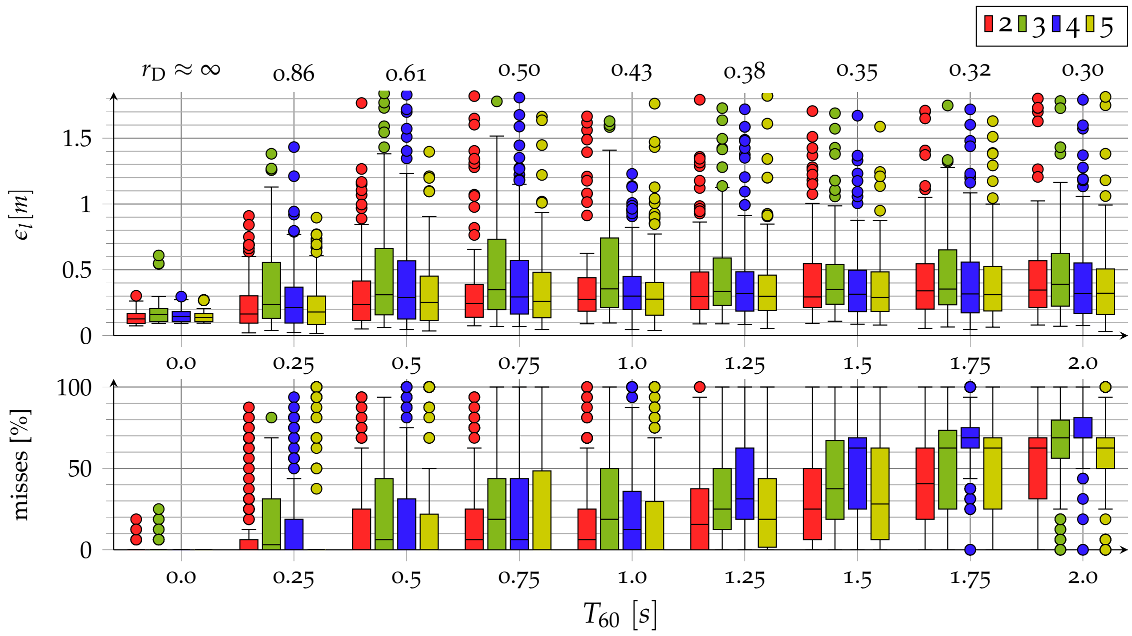

Figure 6.37: Localization error (top) and number of missed frames (bottom) of different counts of randomly selected nodes at different reverberation times over all simulations and speaker positions.

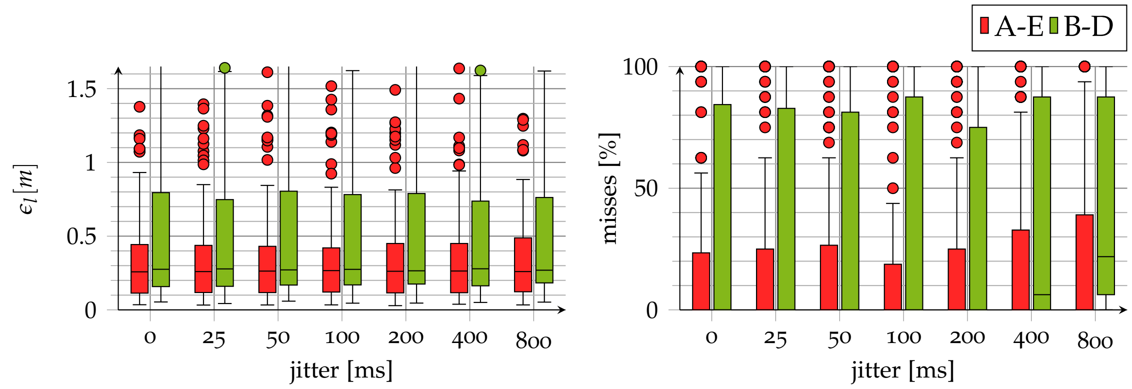

Figure 6.38: Localization error and missed detections for different jitters between the nodes at $T_{60} = 0.5$ s.

lack of synchronization alone, such as 400 or 800 ms. However, the number of missed detections shows a slight increase. This is to be expected as there are DoA detections missing when the nodes are out of sync. From a system design standpoint, we may therefore conclude that jitter can be neglected as long as the nodes communicate at a reasonable rate.

Geometry Calibration

One important practical aspect in the tracking with multiple arrays is the accuracy of the geometry calibration. In order to find out the dependency, different calibration errors were simulated using the three center nodes and all five at $T_{60} = 0.5$ s. One hundred Monte Carlo trials were performed for fixed geometry errors as described in Section 6.1.2 on page 90.

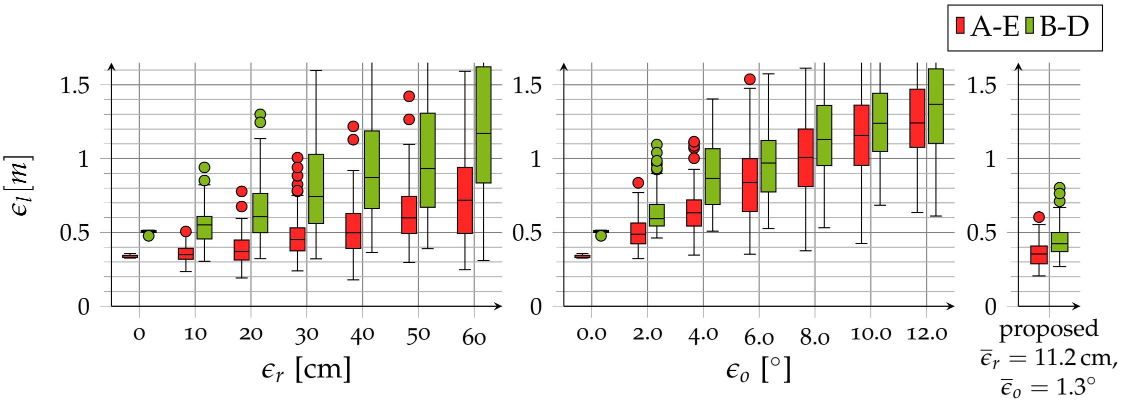

Figure 6.39: Localization error as function of Euclidean (left) and angular (middle) geometry calibration error as well as calibration by the proposed method (right) for three and five nodes at a T_{60} of 0.5 s. Box plots over 100 Monte Carlo runs.

First, a fixed random displacement of the arrays was generated. Figure 6.39 (left) shows the resulting localization quality. While a small error up to about 20 cm leads only to a minimal small decrease of accuracy, at 50 cm the results deteriorate. When using only the three inner nodes, the localization deteriorates much quicker. This may be due to the fact that the zero mean array positions errors are in effect averaged by the localization algorithm.

Secondly, the incorrect calibration of the angular position of the nodes was simulated. Due to the use of triangulation based on the angles, small angular errors already deteriorate the localization accuracy significantly as is shown in Figure 6.39 (middle). An error of $2°$ already decreases the accuracy by 20 cm and an error of $4°$ yields localization errors beyond 1 m rendering the approach unfeasible. The compensating effect of using more nodes is much smaller here.

Using the off-line calibration with evolutionary optimization (described in section 5.4 on pages 80–83), a calibration error of 11.2 ± 2.0 cm and $1.3 \pm 0.4°$ was observed over 100 subset choices ($J = 5$). The localization was run with the geometry estimate obtained from each choice. The resulting localization error is plotted in Figure 6.39 (right). The accuracy achieved by the proposed calibration method with below 0.5 m is sufficient for most tracking applications.

6.8.3 *Smart room recordings*

Several recordings were made in the FINCA with the circular arrays embedded in the table. The ground truth annotations are based on floor positions and assume linear movement, thus they do not reflect slight head movement or speed and position variations. For recall and precision calculation, a localization is considered correct if it is within a typical person's shoulder width of 0.5 m.

In sequence #8, one speaker starts talking while walking into the room. After that, three speakers are talking concurrently. This is an extreme case which allows to investigate the association by spectrum. With a strong threshold for the minimal spectral correlation ($t_s = 0.9$), no association errors occur. Figure 6.40 shows the tracking result.

The tracking was also applied to the two static speaker sequences #2 and #4. In order to test the influence of the automated calibration, the node positions and orientations

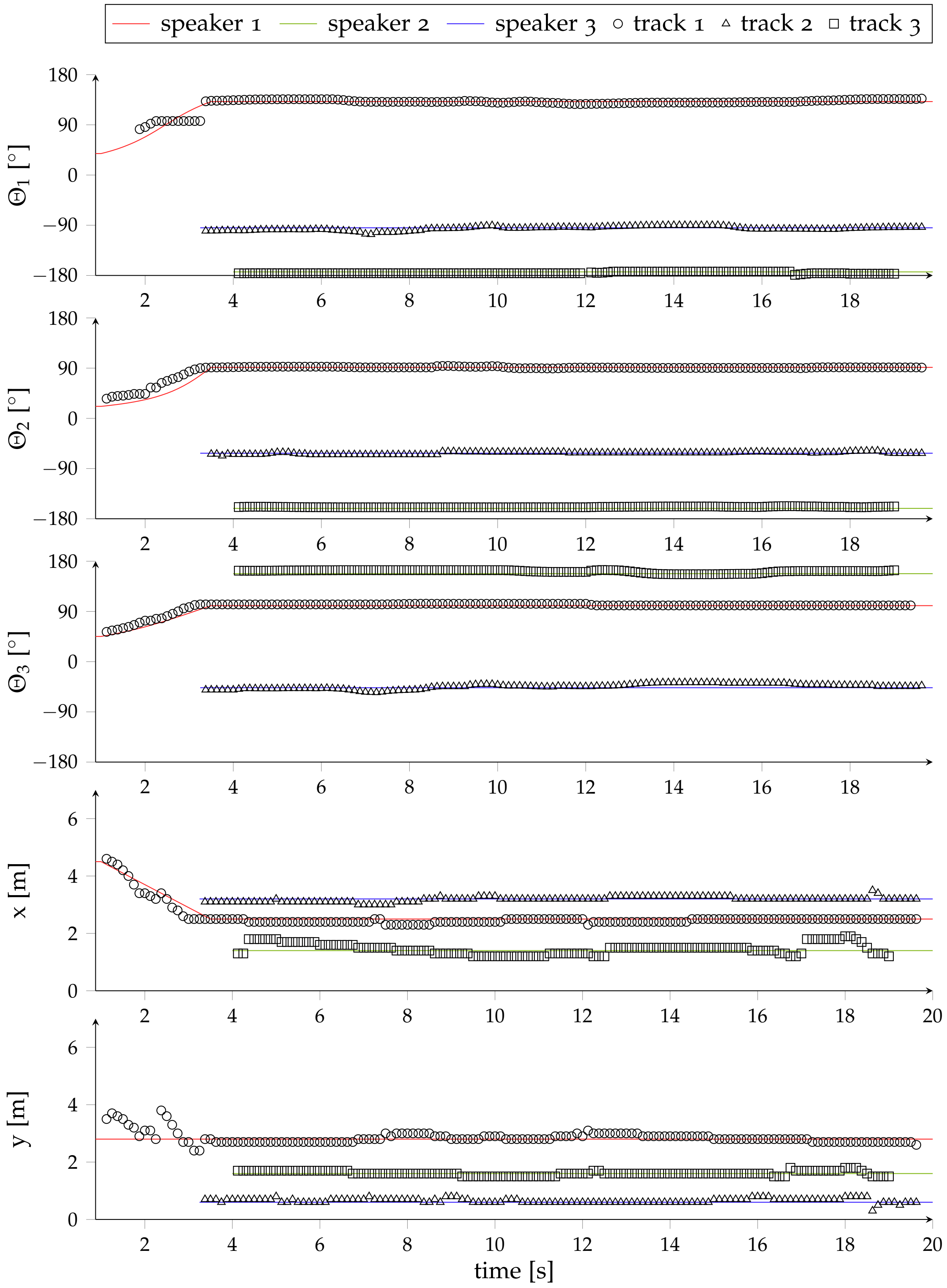

Figure 6.40: Tracking three concurrent speakers in the smart room.

sequence		ϵ_a [°]			ϵ_l [m]	pr. [%]	re. [%]
#8	measured	2.78±3.01	2.38±3.38	3.79±2.67	0.15±0.14	100.0	93.3
#2	measured	5.16±2.65	2.21±2.48	3.24±3.23	0.25±0.19	100.0	93.8
	calibrated	5.14±2.65	2.72±2.91	3.21±2.82	0.30±0.16	100.0	91.4
#4	measured	5.97±4.65	3.04±2.50	3.97±3.38	0.27±0.21	97.6	89.7
	calibrated	5.97±4.63	4.37±3.01	3.85±3.32	0.31±0.21	97.6	86.5

Table 6.12: ASN tracking results in the smart room

were derived from manual measurement as well as from the off-line calibration ($J = 6$, evolutionary optimization) applied to the same sequence. Table 6.12 lists all tracking results. The speaker is localized successfully at all positions. The automated geometry calibration has little influence, precision and recall are almost the same. The position error increases by about 5 cm.

6.8.4 *Summary*

The proposed speaker tracking approach for acoustic sensor networks was evaluated in simulation and with real recordings. In simulation it was shown to handle even severe reverberation levels of $T_{60} > 1.0$ s by maintaining 0.5 m accuracy in the majority of the cases at the cost of an increased number of missed detections. The systematic comparison showed that the proposed weighted triangulation is outperforming other methods. By randomly omitting node transmissions and introducing severe jitter, it was shown that the method is very robust against these type of errors. The accuracy achieved by the proposed calibration methods is good enough not to deteriorate the tracking performance.

With actual smart room recordings, it was shown that the method is applicable in practice. Despite considerable reverberation, the proposed method tracks the speakers well within the accuracy required for practical applications. The association by spectra can handle concurrent speakers.

6.9 combined experiment

In order to show what the methods developed for ASNs can achieve when working together, another experiment was made that employs the event detection, geometry calibration and tracking together on the basis of sequences #5 and #6. In both sequences, two speakers talk alternately while moving through the room.

Single node speech detection and localization

The neuro-inspired PoAP EM localization already includes a basic speech model by use of the spectral spread in the Gammatone filters. To safely exclude non-speech events, the acoustic event detection (see section 3.2 on pages 34–36) was computed on a single microphone of each node. Localizations in time windows where the classifier does not detect speech are removed. Table 6.13 shows the speaker localization performance in comparison. The classifier increases the precision by removing non-speech events. The identification of constant DoA segments used as preprocessing for the geometry calibration naturally has a low angular error with respect to the constant ground truth. However, the classification based approach achieves the highest precision with an only slightly increased angular variance. The recall is reduced by both the constant DoA and classification filtering. This is of little consequence for the calibration, as there are estimates for each positions.

Figure 6.41 shows the effect of filtering out non-speech events by event classification in recording #6. Localizations in a time frame classified as speech are shown in green, while other events, mostly 'steps' and 'background' are plotted in gray. It can be seen that several non-speech events are filtered out, mostly footfall noise between the utterances.

Geometry calibration

Using the speech segments detected by the single node localization (see section 4.2 on pages 54–58) on sequence #5, the geometry of the nodes was calibrated. The off-line geometry calibration (see section 5.4 on pages 80–83) was run on recording #5 using 64 subsets $(J = 5)$ and the evolutionary optimization. The resulting geometry had an error of $e_r = 9.8\,\text{cm}$ and $e_o = 0.76°$.

sequence	method	ϵ_a	precision	recall
#5	PoAP EM	5.69°	95.67%	**81.04%**
	PoAP EM + const. seg. ident.	4.41°	96.40%	77.54%
	PoAP EM + AED = speech	**4.37°**	**98.74%**	65.94%
#6	PoAP EM	4.89°	83.59%	**95.27%**
	PoAP EM + const. seg. ident.	**4.39°**	93.36%	89.70%
	PoAP EM + AED = speech	4.51°	**97.03%**	85.57%

Table 6.13: Localization results for different methods of speech detection. The basic PoAP EM method, added filtering by identifying constant segments and added filtering by event detection. Mean values over all three nodes.

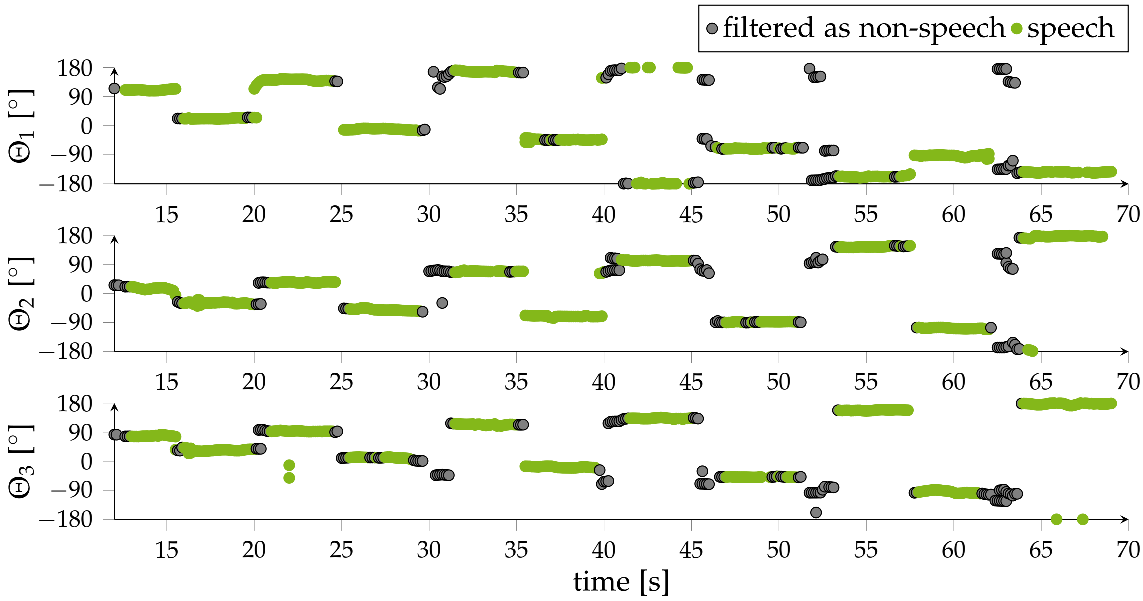

Figure 6.41: Speech localizations filtered by event detection.

Speaker tracking and identification

Using the so-calibrated ASN, the proposed ASN tracking method was used to compute tracks of the speakers in sequence #6. The speech-filtered localizations were used. The resulting tracks are shown in Figure 6.42. The position error relative to the ground truth is 0.36±0.17 m with 97.4% precision and 75.4% recall. Without the event detection it is slightly worse with 0.37±0.17 m and 96.2% precision and 76.2% recall. The position errors are not significantly different according to a permutation test. When using the measured ASN geometry instead of the automatic calibration, the position error decreases to 0.27±0.17 m and 97.6% precision and 82.5% recall are achieved. This decrease is significant $(p < 0.001, N = 250,000)$.

The ability of the event detection method to jointly identify the speakers was tested as well. The two speakers were trained as individual speech classes along with the other acoustic events. Each track computed from the tracking was classified and assigned the corresponding speaker. Both speakers were identified correctly in all cases for all nodes. This shows that it is possible to track and identify known speakers with the same method used for event classification.

Summary & Discussion

The three proposed methods were used in combination with real recordings in the reverberant smart room. The single node localization was applied to identify speech events. Then the event detection was used to filter out non-speech events. Here, superior precision compared to the heuristic identification of speech segments based on constant DoAs was achieved. The angular variance is comparable.

In order to use the calibration in unconstrained scenarios, the classification is necessary to exclude sounds not conducted by air, such as footsteps or chair movement. In the combined experiment, the calibration was applied successfully once again. A very low orientation error of below 1° was achieved. This is important in order not to induce large triangulation errors in subsequent spatial processing.

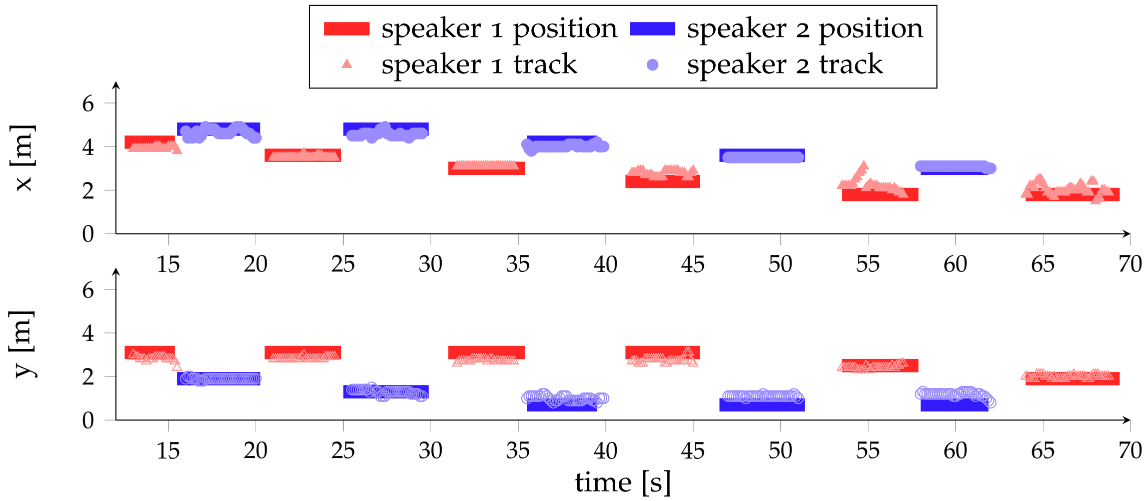

Figure 6.42: Speaker tracks computed using a combination of proposed methods. The calibration was done based on the single node PoAP EM localization. The tracking was performed with the calibration result. The speaker tracks were assigned to the different speakers based on the classification.

The tracking was applied successfully based on the calibration. Using the classification as pre-filter to exclude non-speech events sightly increased the precision. Compared to measurement of the spatial node configuration, the automated calibration led to an increased Euclidean position error. This is a direct result of the 10 cm position error of the calibration. However, the error of 36 cm is still small enough not to impact most applications.

The speakers were identified successfully by classifying the tracks. This is an important result for practical applications of the ASN, as now not only the location but the identity of the speakers is available. Whether the tracking result is used for speech enhancement or camera control, specific speakers can be processed individually.

The overall combination of the proposed methods was demonstrated successfully with recordings in a reverberant smart room. The location and identity of different speakers was inferred automatically with the ASN. Thus it was shown that the proposed methods can work together to provide information on the acoustic scene in real time.

The open secret of real success is to throw your whole personality into your problem.

George Pólya: How to Solve it (1957), p. 207

7 CONCLUSION

The number of computation devices with communication links and acoustics sensors around us is increasing. Thus, acoustic sensor networks (ASNs) are a growing platform that opens the possibility for many practical applications. ASN based speech enhancement, source localization, and event detection can be applied for teleconferencing, camera control, automation, or assisted living. For this kind of applications, the awareness of auditory objects and their spatial positioning are key properties. In order to provide these two kinds of information, novel methods have been developed in this thesis. Information on the type of auditory objects is provided by a novel real-time sound classification method. Information on the position of human speakers is provided by a novel localization and tracking method. In order to provide this kind of information within the ASN, the relative arrangement of the sensor nodes has to be known. Therefore, different novel geometry calibration methods were developed.

In the following, the individual methods and their validation in the evaluation will be summarized. Thereafter, their common properties and combined application will be addressed. A short outlook to future developments concludes this chapter.

Event detection

The proposed bag-of-features (BoF) event detection method is robust and fast while achieving state-of-the-art results. It can be easily integrated in the overall ASN processing. As the acoustic calibration relies on the propagation speed of sound, such a classification step is necessary to exclude other sound events. Since the underlying Gaussian mixture model (GMM) approach is suitable for speaker identification, the same method can be used to distinguish speakers when tracking them.

As shown in the evaluation, the novel combination of mel frequency cepstral coefficient (MFCC) and Gammatone frequency cepstral coefficient (GFCC) features leads to higher classification accuracy. The introduction of supervised codebook training into the BoF paradigm boosts the performance, so that the method clearly surpasses other BoF approaches as well as the basic GMM method. Unlike state-of-the-art deep learning methods, the proposed method can generalize well from limited training data. This is important because the amount of data available is often constrained in practical applications.

In order to detect speech overlapped by noise, a dedicated training strategy was devised. It creates a hierarchy of sound classes based on their stationarity. With this training, the BoF approach is able to provide control information for a beamformer. This way, a fully blind speech enhancement system is realized. By incorporating the full room transfer function, better speech enhancement is achieved compared to using only the direct path to the speaker.

In the evaluation, the speech enhancement performance was compared to the achievable optimum in classification. This was done by using the ground truth annotations as oracle for the beamformer control. It was shown that the proposed method provides similar performance in most cases. Even in cases with multiple directed noise sources and low signal-to-noise ratio (SNR), a significant improvement in speech intelligibility is achieved.

Speaker tracking

For speaker localization with a microphone array, the peak over average position (PoAP) expectation-maximization (EM) method for direction of arrival (DoA) localization was proposed. The neuro-biologically inspired method uses a dedicated cochlear and midbrain model, which make it robust against the reverberation found in indoor rooms. It implicitly handles the concurrency of speech activity found in natural conversations. An automated gain estimation was introduced that allows to use the method without prior manual adaptation to the scenario. EM clustering based on computational auditory scene analysis (CASA) principles was introduced that emulates the simultaneous grouping according to spatial as well as spectral cues. Based on the localizations, the speech segments required for the geometry calibration methods are detected.

The evaluation showed that the proposed PoAP EM method for DoA localization of speakers is able to handle two or more concurrent speakers. It is robust against the reverberation typically found in indoor environments.

The proposed method for speaker tracking in ASNs employs the PoAP EM localization in each node. Each node shares probabilistic DoA estimates together with an estimate of the spectral distribution with the network. As this information is relatively sparse, it can be transmitted with low bandwidth. The information from all nodes is integrated according to spectral similarity. By incorporating the intersection angle in the triangulation, the precision of the Euclidean localization is improved. Speaker tracks are computed over time.

With dedicated simulations, the tracking method was shown to be robust against transmission errors and jitter. Its ability to track concurrent speakers was shown with recordings of real persons in reverberant rooms.

Geometry calibration

The central task of geometry calibration has been solved with focus on sensor nodes equipped with multiple microphones. In addition to the off-line audio-visual and acoustic calibration methods, an online method employing a genetic algorithm with incremental measurements was introduced. By using the robust speech localization method, the calibration is computed in parallel to the tracking. As speech events can be used, this is possible without additional devices. Unlike previous methods that only infer the positioning of distributed microphones, the proposed methods incorporate the DoA and are able to calibrate the orientation of the nodes with a high accuracy. This is very important for all applications using the spatial information, as the triangulation error increases dramatically with bad orientation estimates.

The evaluation showed that both the audiovisual and audio only method can be applied for off-line calibration. The subset sampling strategy makes it robust against measurement errors. The resulting orientation accuracy was consistently 2° when the method was tested with recordings in a smart room. The position error of around 10 cm is well within the accuracy required for practical applications. When the method is applied as basis for the proposed speaker tracking, the localization error increases only slightly compared to using measured positions. The online method was shown to be able to calibrate the ASN in real time. Thus it is possible to perform the calibration in parallel to the speaker tracking.

Informed ASNs

All new methods are important building blocks for the use of ASNs. The online methods for localization and calibration both make use of the neuro-biologically inspired processing in the nodes which leads to state-of-the-art results, even in reverberant enclosures. The high robustness and reliability can be improved even more by including the event detection method in order to exclude non-speech events. When all methods are combined, both semantic information on what is happening in the acoustic scene as well as spatial information on the positioning of the speakers and sensor nodes is automatically acquired in real time. This realizes truly informed audio processing in ASNs.

The combined experiment showed that all methods work in conjunction. The PoAP EM localization provides the speakers DoA while the event detection reliably filters out non-speech events. The so found speech events are the basis for the geometry calibration, which provides the relative ASN geometry. Once the geometry is established, the localized speech events can be combined to speaker tracks with the Euclidean tracking. The tracks can be assigned to individual speakers by the BoF method.

Summary & Outlook

In this thesis, novel methods that enable employing smart devices in a collaborative way as ASNs were developed. All methods were evaluated with recordings of real persons in reverberant rooms, showing their practical applicability. By combined application, it was shown that the methods work together in order to provide both classification and spatial information to the network.

The novel methods enable a multitude of applications in ASNs. As such networks can be built from smartphones, tablets, laptops, and hearing aids in an ad hoc assembly, these can be applied in a growing number of everyday life situations.

For example, teleconferencing can be automated. In a meeting with multiple participants, the closest devices to the speaker can be used for audio pickup. Cameras can be selected and steered automatically.

Another example is the improvement of distributed speech enhancement. The speech detection and localization could be incorporated in a linearly constrained minimum variance (LCMV) approach, allowing to selectively suppress or enhance speakers based on their position and identity. Thus an ad hoc network of, e.g., smartphones and hearing aids, can be used to provide enhanced speech to a hearing impaired person.

These examples illustrate how this thesis provides an important contribution that can stay a part of practical applications. Therefore, the contribution of this thesis is not only advancing the state-of-the-art in automatically acquiring information on the acoustic scene, but also pushing the practical applicability of such methods.

ACKNOWLEDGMENTS

Writing a thesis is a special kind of journey. As with every longer journey, there are good and bad times, and it would be impossible to undertake it without both of them. What makes the good times remarkable and the bad times bearable are the people along the way. So I take this opportunity to thank all of them.

Of course my thanks go to *Gernot A. Fink* for his supervision and friendly advice throughout the work on my thesis. Thank you, *Gernot*, for staying at my side and always making time for discussions and feedback.

Another big thank you to *Rainer Martin* for agreeing to be my second reviewer. Thank you for the friendly encouragement and discussions.

Thanks to my colleagues in Dortmund, first of all *René Grzeszick* for making our office a good working environment and all the discussions and collaboration. Another thank you goes to *Leonhard Rothacker* and for countless scientific and technical discussions. I also have to thank the students *Balasubramanian Raja, Steve Brischke, Pujan Shaha* and *Julian Kürby* for their collaboration and help with the experiments.

Thanks to my project partners in Paderborn, *Reinhold Häb-Umbach* and *Florian Jacob* for the fruitful collaboration in the DFG project and the good work on the joint review article.

Great thanks to the *DAAD* and *Sharon Gannot* for enabling me to stay in Israel. Thank you, *Sharon*, for your hospitality, the constructive collaboration and valuable advice in the writing process. And "toda" to everyone in the acoustic lab: *Boaz Schwartz*, for always having an open ear, *Pini Tandeitnik* for all your help and keeping the lab working, *Shmulik Markovich-Golan, Ofer Schwartz, Yuval Dorfan*, and *Gershon Hazan* for your ongoing collaboration, *David Levin*, and *Schlomi Chazan*, for inspiring discussions, *Dani Cherkassky*, and *Yiftach Richter* for encouragement, as well as *Bracha Laufer*, and *Tamar Maromm* for volunteering in the experiments.

Another round of thanks goes to my colleagues from near and far that provided insight and encouragement. Thanks to *Simon Doclo* and *Emanuël A. P. Habets* for friendly support, *Richard F. Lyon* for discussions and encouragement, *Tobias May* and *Huy Phan* for helpful discussions and insights. A big thank you to *Yussef Ouali* for not only discussing but also sharing data and results on speaker tracking.

Last but by no way least I have to thank my friends and family for staying at my side during sometimes difficult times and proving to me that there is life outside the thesis.

LIST OF FIGURES

LIST OF TABLES

INDEX

BIBLIOGRAPHY

[AB79] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[AMM$^+$03] Shoko Araki, Ryo Mukai, Shoji Makino, Tsuyoki Nishikawa, and Hiroshi Saruwatari. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Transactions on Speech and Audio Processing*, 11(2):109–116, March 2003.

[ASMM06] Shoko Araki, Hiroshi Sawada, Ryo Mukai, and Shoji Makino. DOA estimation for multiple sparse sources with normalized observation vector clustering. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 33–36, October 2006.

[ANS10] Shoko Araki, Tomohiro Nakatani, and Hiroshi Sawada. Simultaneous clustering of mixing and spectral model parameters for blind sparse source separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5–8, 2010.

[AN11] Shoko Araki and Tomohiro Nakatani. Hybrid approach for multichannel source separation combining time-frequency mask with multi-channel wiener filter. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 225–228, May 2011.

[ADP07] Jean-Julien Aucouturier, Boris Defreville, and Francois Pachet. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, 122(2):881–891, 2007.

[BKA10] Jorg-Hendrik Bach, Birger Kollmeier, and Jorn Anemuller. Modulation-based detection of speech in real background noise: Generalization to novel background classes. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 41–44, 2010.

[BYRN99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

[BGSP15] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark D. Plumbley. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, 32(3):16–34, May 2015.

[BvH07] Moritz Bürck and J. Leo van Hemmen. Modeling the cochlear nucleus: A site for monaural echo suppression? *Journal of the Acoustical Society of America*, 122:2226–2235, 2007.

[BMC05] Jacob Benesty, Shoji Makino, and Jingdong Chen, editors. *Speech Enhancement*. Springer, Berlin, Germany, 2005.

[BSH08] Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors. *Springer Handbook of Speech Processing*. Springer, Berlin Heidelberg, 2008.

[BM12] Alexander Bertrand and Marc Moonen. Distributed node-specific LCMV beamforming in wireless sensor networks. *IEEE Transactions on Signal Processing*, 60(1):233–246, January 2012.

[Bil98] Jeff A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical report, International Computer Science Institute, 1998.

[Bir03] Stanley T. Birchfield. Geometric microphone array calibration by multidimensional scaling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong (cancelled), April 2003.

[BS05] Stanley T. Birchfield and Amar Subramanya. Microphone array position calibration by basis-point classical multidimensional scaling. *IEEE Transactions on Speech and Audio Processing*, 13(5):1025–1034, September 2005.

[Bla96] Jens Blauert. *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press, October 1996.

[BW01] Michael Brandstein and Darren Ward, editors. *Microphone Arrays*. Springer, 2001.

[Bre90] Albert S. Bregman. *Auditory Scene Analysis*. MIT Press, 1990.

[Bri13] Steve Brischke. Multikamera Personenlokalisierung in Intelligenten Umgebungen (multicamera person localization in intelligent environments). Diplomarbeit, TU Dortmund University, December 2013.

[BAK04] Herbert Buchner, Robert Aichner, and Walter Kellermann. TRINICON: A versatile framework for multichannel blind signal processing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, page iii–889, Montreal, Canda, 2004.

[BLNZ95] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16:1190–1208, 1995.

[CFP+13] Vincenzo Carletti, Pasquale Foggia, Gennaro Percannella, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Audio surveillance using a bag of aural words classifier. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 81–86, August 2013.

[CKR+15] Benjamin Cauchi, Ina Kodrasi, Robert Rehr, Stephan Gerlach, Ante Jukić, Timo Gerkmann, Simon Doclo, and Stefan Goetze. Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. *EURASIP Journal on Advances in Signal Processing*, 2015(1):61, 2015.

[CLVZ11] Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman. The devil is in the details: An evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011.

[CG14] Dani Cherkassky and Sharon Gannot. Blind synchronization in wireless sensor networks with application to speech enhancement. In *International Workshop on Acoustic Signal Enhancement*, Antibes – Juan les Pins, France, September 2014.

[CB10] Heidi Christensen and Jon Barker. Speaker turn tracking with mobile microphones: combining location and pitch information. In *European Signal Processing Conference*, pages 954–958, Aalborg, Denmark, August 2010.

[Coo06] Martin Cooke. A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America*, 119:1562–1573, 2006.

[CDM12] Marco Crocco, Alessio Del Bue, and Vittorio Murino. A bilinear approach to the position self-calibration of multiple sensors. *IEEE Transactions on Signal Processing*, 60(2):660–673, February 2012.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[DTL11] Jonathan Dennis, Huy Dat Tran, and Haizhou Li. Spectrogram image feature for sound event classification in mismatched conditions. *Signal Processing Letters, IEEE*, 18(2):130–133, Feb 2011.

[DTC13] Jonathan Dennis, Huy Dat Tran, and Eng Siong Chng. Image feature representation of the subband power distribution for robust sound event classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):367–377, Feb 2013.

[DIH$^+$09] Sasha Devore, Antje Ihlefeld, Kenneth Hancock, Barbara Shinn-Cunningham, and Bertrand Delgutte. Accurate sound localization in reverberant environments is mediated by robust encoding of spatial cues in the auditory midbrain. *Neuron*, 16;62(1):123–34, April 2009.

[DHV13] Aleksandr Diment, Toni Heittola, and Tuomas Virtanen. Sound event detection for office live and office synthetic AASP challenge. Technical report, Tampere University of Technology, 2013.

[DM03] Simon Doclo and Marc Moonen. Design of far-field and near-field broadband beamformers using eigenfilters. *Signal Processing*, 83(12):2641 – 2673, 2003.

[DSWM07] Simon Doclo, Ann Spriet, Jan Wouters, and Marc Moonen. Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction. *Speech Communication*, 49(7-8):636–656, July 2007.

[DCG15] Yuval Dorfan, Dani Cherkassky, and Sharon Gannot. Speaker localization and separation using incremental distributed expectation-maximization. In *European Signal Processing Conference*, pages 1256–1260, Nice, France, August 2015.

[DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, New York, USA, 2 edition, 2001.

[EC83] Meng Er and Antonio Cantoni. Derivative constraints for broad-band element space antenna array processors. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(6):1378–1393, December 1983.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, OR, USA, 1996.

[EMN16] Christine Evers, Alastair H. Moore, and Patrik A. Naylor. Localization of moving microphone arrays from moving sound sources for robot audition. In *European Signal Processing Conference*, pages 1008–1012, Budapest, Hungary, August 2016.

[FP08] Gernot A. Fink and Thomas Plötz. Developing pattern recognition systems based on Markov models: The ESMERALDA framework. *Pattern Recognition and Image Analysis*, 18(2):207–215, June 2008.

[Fin14] Gernot A. Fink. *Markov Models for Pattern Recognition, From Theory to Applications*. Advances in Computer Vision and Pattern Recognition. Springer, London, 2 edition, 2014.

[GVMGO17] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730, April 2017.

[GBW01] Sharon Gannot, David Burshtein, and Ehud Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614–1626, 2001.

[GKH13] Nikolay D. Gaubitch, Willem Bastiaan Kleijn, and Richard Heusdens. Auto-localization in ad-hoc microphone arrays. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[GKH14] Nikolay D. Gaubitch, Willem Bastiaan Kleijn, and Richard Heusdens. Calibration of distributed sound acquisition systems using ToA measurements from a moving source. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 7505–7509, 2014.

[GVK$^+$13] Jort F. Gemmeke, Lode Vuegen, Peter Karsmakers, Bart Vanrumste, and Hugo Van hamme. An exemplar-based nmf approach to audio event detection. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, October 2013.

[GSB$^+$13] Dimitrios Giannoulis, Dan Stowell, Emmanouil Benetos, Mathias Rossignol, and Mathieu Lagrange. A database and challenge for acoustic scene classification and event detection. In *European Signal Processing Conference*, Marrakech, Morocco, September 2013.

[GPKM14] Panagiotis Giannoulis, Gerasimos Potamianos, Athanasios Katsamanis, and Petros Maragos. Multi-microphone fusion for detection of speech and acoustic events in smart spaces. In *European Signal Processing Conference*, pages 2375–2379, Lisbon, Portugal, September 2014.

[GM90] Brian R. Glasberg and Brian C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1–2):103–138, August 1990.

[Goo00] Phillip Good. *Permutation Tests – A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer Series in Statistics. Springer, 2 edition, 2000.

[GB08] Davi Grangier and Samy Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1371–1384, Aug 2008.

[GM13] Anthony Griffin and Athanasios Mouchtaris. Localizing multiple audio sources from DoA estimates in a wireless acoustic sensor network. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[GJ82] Lloyd J. Griffiths and Charles W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, January 1982.

[Gro03] Benedikt Grothe. New roles for synaptic inhibtion in sound localisation. *Nature*, 4(7):540–550, 2003.

[GPF15] René Grzeszick, Axel Plinge, and Gernot A. Fink. Temporal acoustic words for online acoustic event detection. In Juergen Gall, Peter Gehler, and Bastian Leibe, editors, *German Conference on Pattern Recognition*, volume 9358 of *Lecture Notes in Computer Science*, Cham, 2015. Springer International Publishing.

[GPF17] René Grzeszick, Axel Plinge, and Gernot A. Fink. Bag-of-Features methods for acoustic event detection and classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1242–1252, June 2017.

[HG07] Emanuël A. P. Habets and Sharon Gannot. Generating sensor signals in isotropic noise fields. *Journal of the Acoustical Society of America*, 122(6):3464–3470, December 2007.

[HBC+10] Emanuel A. P. Habets, Jacob Benesty, Israel Cohen, Sharon Gannot, and Jacek Dmochowski. New insights into the MVDR beamformer in room acoustics. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):158–170, January 2010.

[HHVG14] Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot. Multichannel audio database in various acoustic environments. In *International Workshop on Acoustic Signal Enhancement*, pages 313–317, Juan les Pins, France, September 2014.

[Han89] Stephen Handel. *Listening*. MIT Press, Cambridge, MA, USA, 1989.

[HPFHU09] Marius H. Hennecke, Thomas Plötz, Gernot A. Fink, and Reinhold Haeb-Umbach. A hierarchical approach to unsupervised shape calibration of microphone array networks. In *IEEE Workshop on Stat. Signal Proc.*, pages 257–260, Cardiff, Wales, UK, 2009.

[HF11] Marius H. Hennecke and Gernot A. Fink. Towards acoustic self-localization of ad hoc smartphone arrays. In *Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, pages 127–132, Edinburgh, UK, 2011.

[HPM16] Lars Hertel, Huy Phan, and Alfred Mertins. Comparing time and frequency domain for audio event recognition using deep learning. In *IEEE International Joint Conference on Neural Networks*, Vancouver, Canada, 2016.

[HZH+12] Richard Heusdens, Guoqiang Zhang, Richard C. Hendriks, Yuan Zeng, and W. Bastiaan Kleijn. Distributed mvdr beamforming for (wireless) microphone networks using message passing. In *International Workshop on Acoustic Echo and Noise Control*, September 2012.

[HML08] Ivan Himawan, Iain McCowan, and Mike Lincoln. Microphone array beamforming approach to blind speech separation. In Andrei Popescu-Belis, Steve Renals, and Hervé Bourlard, editors, *Machine Learning for Multimodal Interaction*, volume 4892 of *Lecture Notes in Computer Science*, pages 295–305. Springer Berlin Heidelberg, 2008.

[HDY+12] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, and Tara Sainath. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[HW10] Guoning Hu and DeLiang Wang. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2067–2079, 2010.

[HW12] Ke Hu and DeLiang Wang. An unsupervised approach to cochannel speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21:120–129, 2012.

[HL08] Yi Hu and Philipos C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, jan 2008.

[HOS95] Jie Huang, N. Ohnishi, and N. Sugie. A biomimetic system for localization and separation of multiple sound sources. *IEEE Transactions on Instrumentation and Measurement*, 44(3):733–738, 1995.

[HAH01] Xuedong Huang, Alejandro Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall, Upper Saddle River, NJ, USA, 2001.

[JSHU12] Florian Jacob, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach. Microphone array position self-calibration from reverberant speech input. In *International Workshop on Acoustic Signal Enhancement*, September 2012.

[JSHU13] Florian Jacob, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach. Doa-based microphone array position self-calibration using circular statistisc. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 116–120, 2013.

[JF95] Ea-Ee Jan and James Flanagan. Microphone arrays for speech processing. In *URSI International Symposium on Signals, Systems, and Electronics*, pages 373–376, San Francisco, California, USA, October 1995.

[Jef48] Lloyd A. Jeffress. A place theory of sound localization. *Journal of Comparative & Physiological Psychology*, 41:35–39, 1948.

[JW09] Zhaozhang Jin and DeLiang Wang. A supervised learning approach to monaural segregation of reverberant speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):625–638, May 2009.

[KB01] P. KadewTraKuPong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *European Workshop on Advanced Video-Based Surveillance Systems*, 2001.

[KAK06] Young-Ik Kim, Sung Jun An, and Rhee Man Kil. Zero-crossing based binaural mask estimation for missing data speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, Toulouse, France, 2006.

[KL10] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12–40, 2010.

[KDG$^+$16] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A. P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, Armin Sehr, and Takuya Yoshioka. A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016(1), January 2016.

[KSJM09] Holger Klinck, Roland Stelzer, Karim Jafarmadar, and David K. Mellinger. AAS endurance: An autonomous acoustic sailboat for marine mammal research. In *Int. Robotic Sailing Conference*, Matosinhos, Portugal, July 2009.

[KC76] Charles H. Knapp and G. Clifford Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[KSWP16] Qiuqiang Kong, Iwona Sobieraj, Wenwu Wang, and Mark Plumbley. Deep neural network baseline for DCASE challenge 2016. In *Detection and Classification of Acoustic Scenes and Events Workshop*, pages 50–54, Budapest, Hungary, September 2016.

[KA13] Yubin Kuang and Kalle Aström. Stratified sensor network self-calibration from TDoA measurements. In *European Signal Processing Conference*, Marrakesh, Morocco, 2013.

[KGPF16] Julian Kürby, René Grzeszick, Axel Plinge, and Gernot A. Fink. Bag-of-features acoustic event detection for sensor networks. In *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pages 55–59, Budapest, Hungary, September 2016.

[Kut00] Heinrich Kuttruff. *Room Acoustics*. Taylor & Francis, 4 edition, 2000.

[LOGP05] Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez. Av16.3: An audio-visual corpus for speaker localization and tracking. In *International conference on Machine Learning for Multimodal Interaction*, volume 3361 of *LNCS*, pages 182–195, Martigny, Switzerland, 2005.

[LO07] Guillaume Lathoud and Jean-Marc Odobez. Short-Term Spatio-Temporal Clustering applied to Multiple Moving Speakers. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[LJN07] Eric A. Lehmann, Anders M. Johansson, and Sven Nordholm. Reverberation-time prediction method for room impulse responses simulated with the image-source model. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007.

[LM14] Baiying Lei and Man-Wai Mak. Sound-event partitioning and feature normalization for robust sound-event detection. In *IEEE International Conference on Digital Signal Processing*, pages 389–394, Hong Kong, April 2014. IEEE.

[LBG80] Yoseph Linde, Andrés Buzo, and Robert M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, January 1980.

[Llo82] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[LK11] Philipos C. Loizou and Gibak Kim. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):47–56, January 2011.

[LYJV10] Heinrich W. Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary. An improved algorithm for blind reverberation time estimation. In *International Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, August 2010.

[Lyo82] Richard F. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 1282–1285, Paris, France, 1982.

[Lyo83] Richard F. Lyon. A computational model of binaural localization and separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 1148–1151, Boston, Massachusetts, USA, 1983.

[Lyo10] Richard F. Lyon. Machine hearing – An emerging field. *IEEE Signal Processing Magazine*, 27(5):131–139, September 2010.

[LPC11] Richard F. Lyon, J. Ponte, and G. Chechik. Sparse coding of auditory features for machine hearing in interference. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5876–5879, May 2011.

[Lyo17] Richard F. Lyon. *Human and Machine Hearing: Extracting Meaning from Sound*. Cambridge University Press, June 2017.

[Mac67] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[MM08] Nilesh Madhu and Rainer Martin. A scalable framework for multiple speaker localization and tracking. In *International Workshop on Acoustic Echo and Noise Control*, Seattle, WA, USA, September 2008.

[MM11] Nilesh Madhu and Rainer Martin. A versatile framework for speaker separation using a model-based speaker localization approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1900–1912, September 2011.

[MLS07] Shoji Makino, Te-Won Lee, and Hiroshi Sawada. *Blind Speech Separation*. Springer, Berlin, Heidelberg, 2007.

[MGGC09] Shmulik Markovich-Golan, Sharon Gannot, and Israel Cohen. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1071–1086, August 2009.

[MGGC12] Shmulik Markovich-Golan, Sharon Gannot, and Israel Cohen. Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming. In *International Workshop on Acoustic Signal Enhancement*, Aachen, Germany, September 2012.

[MHA08] Rainer Martin, Ulrich Heute, and Christiane Antweiler. *Advances in Digital Speech Transmission*. Wiley, 1 edition, 2008.

[MvK11] Tobias May, Steven van de Par, and Armin Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1), 2011.

[MP12] Tobias May and Steven van de Par. Blind estimation of the number of speech sources in reverberant multisource scenarios based on binaural signals. In *International Workshop on Acoustic Signal Enhancement*, September 2012.

[MPK13] Tobias May, Steven van de Par, and Armin Kohlrausch. Binaural localization and detection of speakers in complex acoustic scenes. In Jens Blauert, editor, *The Technology of Binaural Listening*, pages 397–425. Springer, Berlin, Heidelberg, 2013.

[ML08] Iain McCowan and Mike Lincoln. Microphone array shape calibration in diffuse noise fields. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):666–670, 2008.

[MZX$^+$15] Ian McLoughlin, Haomin Zhang, Zhipeng Xie, Yan Song, and Wei Xiao. Robust sound event classification using deep neural networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(3):540–552, March 2015.

[MHEV10] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. Acoustic event detection in real-life recordings. In *European Signal Processing Conference*, pages 1267–1271, Aalborg, Denmark, 2010.

[MK16] Hendrik Meutzner and Dorothea Kolossa. A non-speech audio captcha based on acoustic event detection and classification. In *European Signal Processing Conference*, Budapest, Hungary, September 2016.

[MBE$^+$12] Xavier Miro Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Acoustic, Speech, and Language Processing*, 20(2):356–370, 2012.

[MDH11] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):14–22, 2011.

[NHG95] Climent Nadeu, Javier Hernando, and Monica Gorricho. On the decorrelation of filter-bank energies in speech recognition. In *European Signal Processing Conference*, pages 1381–1384, 1995.

[NK11] Ganesh R. Naik and Dinesh K. Kumar. An overview of independent component analysis and its applications. *Informatica: An International Journal of Computing and Informatics*, 35(1):63–81, 2011.

[NH93] Radford M. Neal and Geo E. Hinton. A new view of the EM algorithm that justifies incremental and other variants. *Learning in Graphical Models*, pages 355–368, 1993.

[NOS08] Francesco Nesta, Maurizio Omologo, and Piergiorgio Svaizer. Multiple TDoA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS. In *Machine Learning for Signal Processing*, Cancun, Mexico, 2008.

[NVM13] Maria E. Niessen, Tim L. M. Van Kasteren, and Andreas Merentitis. Hierarchical modeling using automated sub-clustering for sound recognition. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2013.

[OG10] Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *The Journal of Machine Learning Research*, 11:1833–1863, 2010.

[ODZ$^+$16] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[OFK13a] Youssef Oualil, Friedrich Faubel, and Dietrich Klakow. A fast cumulative steered response power for multiple speaker detection and localization. In *European Signal Processing Conference*, Marrakech, Morocco, 2013.

[OFK13b] Youssef Oualil, Friedrich Faubel, and Dietrich Klakow. An unsupervised Bayesian classifier for multiple speaker detection and localization. In *INTERSPEECH*, pages 2943–2947, Lyon, France, August 2013.

[OK14] Youssef Oualil and Dietrich Klakow. Multiple concurrent speaker short-term tracking using a Kalman filter bank. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1458–1462, Florence, Italy, May 2014.

[Ows85] Norman L. Owsley. Signal subspace based minimum-variance spatial array processing. In *Asilomar Conference on Circuits, Systems and Computers*, pages 94–97, November 1985.

[PBW04] Kalle J. Palomäki, G. J. Brown, and DeLiang Wang. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication*, 43(4):361–378, 2004.

[Pó88] George Pólya. *How to Solve It*. Princeton University Press, 1988.

[PA12] Stephanie Pancoast and Murat Akbacak. Bag-of-audio-words approach for multimedia event classification. In *INTERSPEECH*, Portland, OR, USA, 2012.

[PS06] Hyung-Min Park and Richard M. Stern. Spatial separation of speech signals using continuously-variable masks estimated from comparisons of zero crossings. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.

[PA02] Lucas Parra and Christopher Alvino. Geometric source separation: merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6):352–362, sep 2002.

[PS03] Lucas Parra and Paul Sajda. Blind source separation via generalized eigenvalue decomposition. *Journal of Machine Learning Research*, 4:1261–1269, December 2003.

[PPH14] Mikko Parviainen, Pasi Pertilä, and Matti S. Hämäläinen. Self-localization of wireless acoustic sensors in meeting rooms. In *Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, pages 152–156, Villers lès Nancy, France, May 2014.

[PRH$^+$92] Roy D. Patterson, Ken Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. In *Auditory physiology and perception; International Symposium on Hearing*, pages 429–446, 1992.

[PGP13] Despoina Pavlidi, Anthony Griffin, and Matthieu Puigt. Real-time multiple sound source localization and counting using a circular microphone array. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2193–2206, 2013.

[PLKP08] Michael S. Pedersen, Jan Larsen, Ulrik Kjems, and Lucas C. Parra. Convolutive blind source separation methods. In Benesty et al. [BSH08], chapter 53, pages 1065–1093.

[PKV08] Pasi Pertilä, Teemu Korhonen, and Ari Visa. Measurement combination for acoustic source localization in a room environment. *EURASIP Journal on Audio, Speech, and Music Processing*, 2008:1–14, 2008.

[PMH11] Pasi Pertilä, Mikael Mieskolainen, and Matti S. Hämäläinen. Closed-form self-localization of asynchronous microphone arrays. In *Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, pages 139–144, Edinburgh, UK, May 2011.

[PMH12] Pasi Pertilä, Mikael Mieskolainen, and Matti S. Hämäläinen. Passive self-localization of microphones using ambient sounds. In *European Signal Processing Conference*, pages 1314–1318, Bucharest, Romania, August 2012.

[PT13] Pasi Pertilä and Aki Tinakari. Time-of-arrival estimation for blind beamforming. In *International Conference on Digital Signal Processing*, Santorini, Greece, July 2013.

[PHM13] Pasi Pertilä, Matti S. Hämäläinen, and Mikael Mieskolainen. Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2393–2402, Nov 2013.

[PMMM14] Huy Phan, Marco Maasz, Radoslaw Mazur, and Alfred Mertins. Random regression forests for acoustic event detection and classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 20–31, 2014.

[PM14] Huy Phan and Alfred Mertins. Exploiting superframe coopccurence for acoustic event recognition. In *European Signal Processing Conference*, Lisbon, Portugal, September 2014.

[PMH+15] Huy Phan, Marco Maass, Lars Hertel, Radoslaw Mazur, and Alfred Mertins. A multi-channel fusion framework for audio event detection. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2015.

[Pli10] Axel Plinge. Neurobiologisch inspirierte Lokalisierung von Sprechern in realen Umgebungen. Diplomarbeit, TU Dortmund; Fakultät für Informatik in Zusammenarbeit mit dem Institut für Roboterforschung, Dortmund, Germany, May 2010.

[PHF10] Axel Plinge, Marius H. Hennecke, and Gernot A. Fink. Robust neuro-fuzzy speaker localization using a circular microphone array. In *International Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, August 2010.

[PHHF11] Axel Plinge, Daniel Hauschildt, Marius H. Hennecke, and Gernot A. Fink. Multiple speaker tracking using a microphone array by combining auditory processing and a gaussian mixture cardinalized probability hypothesis density filter. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2476–2479, Prague, Czech Republic, 2011.

[PHF12] Axel Plinge, Marius H. Hennecke, and Gernot A. Fink. Reverberation-robust online multi-speaker tracking by using a microphone array and CASA processing. In *International Workshop on Acoustic Signal Enhancement*, Aachen, Germany, September 2012.

[PF13] Axel Plinge and Gernot A. Fink. Online multi-speaker tracking using multiple microphone arrays informed by auditory scene analysis. In *European Signal Processing Conference*, Marrakesh, Morocco, September 2013.

[PGF14] Axel Plinge, René Grzeszick, and Gernot A. Fink. A Bag-of-Features approach to acoustic event detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014.

[PF14a] Axel Plinge and Gernot A. Fink. Multi-speaker tracking using multiple distributed microphone arrays. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 614–618, Florence, Italy, May 2014.

[PF14b] Axel Plinge and Gernot A. Fink. Geometry calibration of distributed microphone arrays exploiting audio-visual correspondences. In *European Signal Processing Conference*, Lisbon, Portugal, September 2014.

[PF14c] Axel Plinge and Gernot A. Fink. Geometry calibration of multiple microphone arrays in highly reverberant environments. In *International Workshop on Acoustic Signal Enhancement*, Antibes – Juan les Pins, France, September 2014.

[PJHUF16] Axel Plinge, Florian Jacob, Reinhold Haeb-Umbach, and Gernot A. Fink. Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms. *IEEE Signal Processing Magazine*, 33(4):14–29, July 2016.

[PG16] Axel Plinge and Sharon Gannot. Multi-microphone speech enhancement informed by auditory scene analysis. In *Sensor Array and Multichannel Signal Processing Workshop*, Rio de Janeiro, Brazil, July 2016.

[PFG17] Axel Plinge, Gernot A. Fink, and Sharon Gannot. Passive online geometry calibration of acoustic sensor networks. *IEEE Signal Processing Letters*, 2017(3):324–328, March 2017.

[RD04] Vikas C. Raykar and Ramani Duraiswami. Automatic position calibration of multiple microphones. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 4:iv–69–72, 2004.

[RKL05] Vikas C. Raykar, Igor V. Kozintsev, and Rainer Lienhart. Position calibration of microphones and loudspeakers in distributed computing platforms. *IEEE Transactions on Speech and Audio Processing*, 13(1):70–83, 2005.

[RLB+09] Martin Rehn, Richard F. Lyon, Samy Bengio, Thomas C. Walters, and Gal Chechik. Sound ranking using auditory sparse-code representations. In *ICML Workshop Sparse Methods for Music Audio*, Montreal, Canada, 2009.

[RMGB+13] Klaus Reindl, Shmulik Markovich-Golan, Hendrik Barfuss, Sharon Gannot, and Walter Kellermann. Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2013.

[Rey95] Douglas A. Reynolds. Speaker identification speaker models. *Speech Communication*, 17:91–108, 1995.

[RVCV11] Branko Ristic, Ba-Ngu Vo, Daniel Clark, and Ba-Tuong Vo. A metric for performance evaluation of multi-target tracking algorithms. *IEEE Transactions on Signal Processing*, 59(7):3452–3457, 2011.

[RWB03] Nicoleta Roman, DeLiang Wang, and Guy J. Brown. Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114(4):2236–2252, 2003.

[RW08] Nicoleta Roman and DeLiang Wang. Binaural tracking of multiple moving sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):728–739, 2008.

[SMAM04] Hiroshi Sawada, Ryo Mukai, Shoko Araki, and Shoji Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, 12(5):530–538, September 2004.

[SJHU+11] Joerg Schmalenstroeer, Florian Jacob, Reinhold Haeb-Umbach, Marius H. Hennecke, and Gernot A. Fink. Unsupervised geometry calibration of acoustic sensor networks using source correspondences. In *INTERSPEECH*, 2011.

[SJHU14] Joerg Schmalenstroeer, Patrick Jebramcik, and Reinhold Haeb-Umbach. A gossiping approach to sampling clock synchronization in wireless acoustic sensor networks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014.

[SMS+13] Jens Schröder, Niko Moritz, Marc Rene Schaedler, Benjamin Cauchi, Kamil Adiloglu, Joern Anemueller, Simon Doclo, Birger Kollmeier, and Stefan Goetze. On the use of spectro-temporal deatures for the ieee aasp challenge 'detection and classification of acoustic scenes and events'. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[SAG16] Jens Schröder, Jörn Anemüller, and Stefan Goetze. Performance comparison of GMM, HMM and DNN based approaches for acoustic event detection within task 3 of the DCASE 2016 challenge. In *Detection and Classification of Acoustic Scenes and Events Workshop*, pages 80–84, Budapest, Hungary, September 2016.

[SSW07] Yang Shao, Soundararajan Srinivasan, and DeLiang Wang. Incorporating auditory feature uncertainties in robust speaker identification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 277–280, 2007.

[SSKP16] Siddharth Sigtia, Adam Stark, Sacha Krstulovic, and Mark Plumbley. Automatic Environmental Sound Recognition: Performance versus Computational Cost. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11), 2016.

[SZ03] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003.

[Sla93] Malcom Slaney. An efficient implementation of the Patterson-Holdsworth auditory filter bank. Technical Report 35, Apple Computer, Inc., 1993.

[Smi99] Stephen W. Smith. *The Scientists and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, 2 edition, 1999.

[SAC07] Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 623–632, New York, NY, USA, 2007. ACM.

[SHK+14] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[SKG13] Daniel Steele, Dirkjan Krijnders, and Catherine Guastavino. The sensor city initiative: Cognitive sensors for soundscape transformations. In *GIS Ostrava 2013 – Geoinformatics for City Transformation*, Ostrava, Czech Republic, January 2013.

[SGT07] Richard M. Stern, Evandro B. Gouvêa, and Govindarajan Thattai. Polyaural array processing for automatic speech recognition in degraded environments. In *INTERSPEECH*, pages 926–929, Antwerp, Belgium, 2007.

[SP97] Rainer Storn and Kenneth Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal on Global Optimization*, 11:341 – 359, 1997.

[SGB$^+$15] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, Oct 2015.

[TCG11] Ronen Talmon, Israel Cohen, and Sharon Gannot. Transient noise reduction using nonlocal diffusion filters. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1584–1599, August 2011.

[TCHJH12] Hao Tang, Stephen M. Chu, Mark Hasegawa-Johnson, and Thomas S. Huang. Partially supervised speaker clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):959–971, 2012.

[TH13] Maja Taseska and Emanuël A. P. Habets. Spotforming using distributed microphone arrays. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2013.

[TKH14] Maja Taseska, Affan Hasan Khan, and Emanuël A. P. Habets. Speech enhancement with a low-complexity online source number estimator using distributed arrays. In *European Signal Processing Conference*, Lisbon, Portugal, September 2014.

[TH16] Maja Taseska and Emanuël A. P. Habets. Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7):1291–1304, 2016.

[TN06] Andrey Temko and Climent Nadeu. Classification of acoustic events using SVM-based clustering schemes. *Pattern Recognition*, 39(4):682–694, 2006.

[TMZ$^+$07] Andrey Temko, Robert Malkin, Christian Zieger, Dušan Macho, Climent Nadeu, and Maurizio Omologo. CLEAR evaluation of acoustic event detection and classification systems. In Rainer Stiefelhagen and John Garofolo, editors, *Multimodal Technologies for Perception of Humans*, volume 4122 of *Lecture Notes in Computer Science*, pages 311–322. Springer Berlin Heidelberg, 2007.

[TN09] Andrey Temko and Climent Nadeu. Acoustic event detection in meeting-room environments. *Pattern Recognition Letters*, 30(14):1281–1288, 2009.

[TTH14] Oliver Thiergart, Maja Taseska, and Emanuel A. P. Habets. An informed parametric spatial filter based on instantaneous direction-of-arrival estimates. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(12):2182–2196, December 2014.

[TP11] Roberto Togneri and Daniel Pullella. An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits and Systems Magazine*, 11(2):23–61, 2011.

[TR16] Vladimir Tourbabin and Boaz Rafaely. Analysis of distortion in audio signals introduced by microphone motion. In *European Signal Processing Conference*, Budapest, Hungary, August 2016.

[UA99] Masashi Unoki and Masato Akagi. A method of signal extraction from noisy signal based on auditory scene analysis. *Speech Communication*, 27(3):261–279, 1999.

[VB95] Chandra Vaidyanathan and Kevin M. Buckley. Performance analysis of the MVDR spatial spectrum estimator. *IEEE Transactions on Signal Processing*, 43(6):1427–1437, June 1995.

[VS93] Andrew Varga and Herman J. M. Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993.

[VB88] Barry D. van Veen and Kevin M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988.

[VBK$^+$13] Lode Vuegen, B. van den Broek, Peter Karsmakers, J. F. Gemmeke, Bart Vanrumste, and Hugo van Hamme. An MFCC-GMM approach for event detection and classification. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2013.

[WB06] DeLiang Wang and Guy J. Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. IEEE Press/Wiley Interscience, 2006.

[Wu83] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11(1):95–103, 03 1983.

[YS01] Stuart H. Young and Michael V. Scanlon. Robotic vehicle uses acoustic array for detection and localization in urban environments. *SPIE Proceedings, Mobile Robot Perception*, 4364:264–273, September 2001.

[ZH14] Yuan Zeng and Richard C. Hendriks. Distributed delay and sum beamformer for speech enhancement via randomized gossip. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):260–273, January 2014.

[ZSB13] Matthias Zeppelzauer, Angela S. Stöger, and Christian Breiteneder. Acoustic detection of elephant presence in noisy environments. In *ACM International Workshop on Multimedia Analysis for Ecological Data*, pages 3–8. ACM, 2013.

[ZFZ08] Cha Zhang, D. Florencio, and Zhengyou Zhang. Why does PHAT work well in lownoise, reverberative environments? In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2565–2568, 2008.

[ZSW12] Xiaojia Zhao, Yang Shao, and DeLiang Wang. CASA-based robust speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1608–1616, 2012.