Monaural Sound Localization

Anna Katharina Fuchs, Christian Feldbauer, Michael Stark

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

Abstract

The principles of human sound localization imply binaural (interaural level and time difference) as well as monaural cues. The latter are captured by the head-related transfer functions (HRTFs), which describe the direction-dependent, spectral shaping of the incident sound wave, and can be exploited to determine the direction.

In this paper an accurate talker localization strategy in the horizontal plane using the signal of only one microphone is presented. The sound localization method is developed based on a set of HRTF measurements taken from a dummy head and a statistical model of speech. High-dimensional spectral features (STFT coefficients) are taken and the direction of the sound source is evaluated with Gaussian mixture models (GMMs) using a maximum likelihood (ML) framework. An evaluation of the developed method in a synthetic test environment yields excellent localization results and leads to a promising approach which can be further investigated in future research.

Index Terms: sound localization, HRTF, single-channel, GMM, monaural localization, maximum likelihood

1. Introduction

From the human auditory system it is known that the binaural cues, interaural time and level difference (ITD, ILD) and monaural spectral cues, presented by the head-related transfer function (HRTF), are used to localize a sound position. Localization algorithms usually exploit the same cues. Typically sound localization algorithms are based on processing signals received by multiple, spatially separated sensors, e.g., microphone arrays [1].

Only few authors tried to localize sound using a single microphone.

In [2] a system is introduced which localizes sound in the vertical plane with a single microphone. A so-called neuromorphic microphone is used which copies the structure of the outer ear. The recorded sound consists of the direct sound and an echo. Successful localization is enabled by analyzing the echo time. A multilayer perceptron neural network is trained to reach final localization results. Localization within a range of 8 ° is possible.

Also in [3] localization with a single microphone is carried out. First a GMM is trained with clean speech. Afterwards, the acoustic transfer function is estimated by maximizing the likelihood with the clean speech GMM and test material uttered from each position. Afterwards a GMM for each position is trained with the estimated acoustic transfer function. The last step is to find the GMM with the maximum likelihood among the estimated GMMs corresponding to each position. As features cepstral coefficients are taken because they are assumed to effectively represent clean speech information. Experiments are taken with synthesized, reverberant speech. One, five and ten training sentences which are uttered from nine different directions, are used to train the acoustic transfer function GMMs. The evaluation is carried out on three directions. With five sentences to estimate the acoustic transfer function and in the task where three different directions have to be estimated, the direction accuracy is almost 100 %. Additionally, tests are carried out with GMMs trained with the observed speech instead of GMMs trained with the acoustic transfer functions. The accuracy of the estimation decreases.

In the follow-up paper [4] the same approach is evaluated in (a) a simulated reverberant environment, (b) a simulated noisy reverberant environment using a speaker-independent speech model and (c) in a real environment using a speaker-dependent speech model. The localization accuracy increases when the number of mixtures, the amount of training data as well as the test segment length is increased. In (a) 3, 5, 7 and 9 positions are evaluated with a localization accuracy of 51.3 % for 9 positions. In (c) 5 positions are used for training and 2 for the test which results in a localization accuracy of 94.8 % for the first position and 79 % for the second.

In [5] sound localization is carried out with a single microphone and an "artificial pinna". In this paper the prior distribution of sound as well as the direction-dependent transfer function of the pinna are modeled. A signal, recorded by one microphone, is given as the convolution of the sound source and the direction-dependent transfer function plus additive white Gaussian noise. Then the source signal is modeled using a Hidden Markov Model (HMM) with training material of different sounds (speech, animal sounds and natural sounds). The transfer functions are estimated with standard noise excitation methods. The most likely value for θ is estimated by a ML framework.

Four different "artificial pinnas" are constructed to record transfer functions that depend strongly on the direction of the sound. Then the transfer functions for angles between 0 ° and 345 ° in steps of 15 ° are estimated and the values are interpolated for finer angles. The best results for mixed human speech is delivered by a pinna where a plastic-cast, that has smooth grooves build on it in various directions, is used. In this case an average error of 7.7 ° occurs.

In this paper a localization strategy with a single microphone is developed. Based on a set of measured head-related transfer functions (HRTFs) from a dummy head [6] and a statistical model of speech, an estimation of the sound direction has been carried out. The proposed method requires the training of only a single GMM of speech. The estimation is performed in the spectral domain by use of a maximum likelihood (ML) based approach. Compared to previous works, a much higher resolution of the localization positions in the horizontal plane can be achieved.

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

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

This paper is organized as follows: In *section 2* the method for single-channel location estimation is described. *Section 3* describes the experimental evaluation. Finally, *section 4* concludes the work presented here.

2. Methods

2.1. Maximum Likelihood Classifier

In this paper classification based on the maximum likelihood principle is used [7]. This is a probabilistic approach where a pattern is assigned to the class with the maximum posterior probability $P(c_i|\mathbf{x}_l)$. $P(c_i|\mathbf{x}_l)$ indicates the probability that a given observation vector $\mathbf{x}_l$ belongs to a class c_i. The observation vector is assigned to the class c_i with the highest posterior probability, i.e. to the class the observation vector most likely belongs to. $P(c_i|\mathbf{x}_l)$ cannot be calculated directly, instead Bayes's theorem

$$P(c_i|\mathbf{x}_l) = \frac{P(\mathbf{x}_l|c_i)P(c_i)}{P(\mathbf{x}_l)} \tag{1}$$

is used. In this paper, the conditional probability $P(\mathbf{x}_l|c_i)$ is modeled for each class with a Gaussian mixture model (GMM).

Maximum likelihood is a widely used principle in which the parameter (or set of parameters) λ is set to the value that maximizes the likelihood function $P(\mathbf{X}|\lambda)$ where $\mathbf{X}$ is a sequence that contains the observation vectors $\mathbf{x}_l$. $P(\mathbf{x}_l|\lambda)$ depends on the parameters specified by λ. Then the likelihood function can be written as

$$P(\mathbf{X}|\lambda) = \prod_{l=1}^{T} P(\mathbf{x}_l|\lambda) =: \mathcal{L}(\lambda|\mathbf{X}). \tag{2}$$

2.2. Gaussian Mixture Models

Mixture models are probabilistic models. They are useful to model arbitrary density distributions. A Gaussian mixture model (GMM) consists of a linear combination of K multivariate Gaussian probability density functions given by

$$P(\mathbf{x}_l) = \frac{1}{\sqrt{\det(\mathbf{\Sigma})}(2\pi)^{D/2}} e^{\left(-\frac{1}{2}(\mathbf{x}_l-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}_l-\boldsymbol{\mu})\right)}. \tag{3}$$

The mathematical representation is given by

$$P(\mathbf{x}_l|\lambda) = \sum_{m=1}^{K} b_m P_m(\mathbf{x}_l). \tag{4}$$

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

$$\sum_{m=1}^{K} b_m = 1 \tag{5}$$

holds. With $\lambda = \{(b_m, \boldsymbol{\mu}_m, \mathbf{\Sigma}_m); m = 1, 2, \ldots, K\}$, the whole mixture model is described. The parameters λ of a model are estimated using the expectation-maximization (EM) algorithm [7].

2.3. Adaption of Speaker Model

For one of the investigated approaches, a single GMM is trained with speech material which is direction-independent. Therefore, the GMM is direction-independent. To create direction-dependent GMMs from the direction-independent GMM, the statistical models are adapted with a HRTF database. The adaption process is described by an observation model. In this model the used features are inherent in a log-spectral domain. Therefore, the direction information is incorporated by shifting the vector $\boldsymbol{\mu}$ of each mixture component of the statistical model with the HRTFs. To be more specific, the feature coefficients of the HRTFs are added to the $\boldsymbol{\mu}$'s of each component of the GMM.

With little computational effort GMMs *with* direction information can be calculated as shown in figure 1.

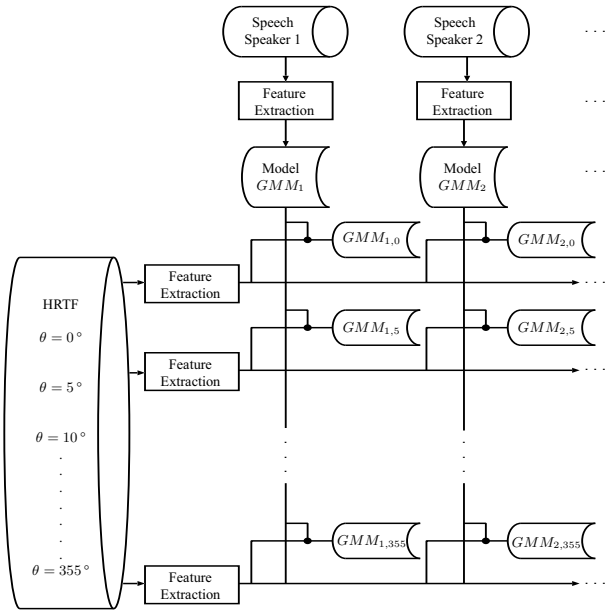

Figure 1: Adaption of the single, direction-independent GMM (Model – GMM_s for each speaker s) to yield direction-dependent GMMs ($GMM_{s,\theta}$ for each speaker s for direction θ)

Systematic Mismatch due to Adaption

Although the simple statistical model adaption is an advantage, it also brings along some problems. Due to the calculation of the Fourier transform with the DFT, the adaption of the direction-dependent model is consistent with a *circular* convolution, whereas the synthesized test utterances (as well as real-world) correspond to a *linear* convolution. As a consequence of such a *mismatched* case, two different approaches are investigated:

1. Restriction to a limited area (*LimArea*):

   Due to the adaption of the single GMM, the absolute localization error will increases. As a constraint only a half plane between 90 ° – 270 ° is taken into account (see also figure 3).

2. Direction-dependent GMMs (*GMM+HRTF*):

   Instead of training of a single GMM and the adaption afterwards, 72 direction-dependent GMMs are trained di-

rectly. The GMMs are trained from speech uttered from a certain direction.

Additionally speaker-dependent as well as speaker-independent models of both approaches are investigated. Furthermore, experiments are carried out with a varying length of the input segment.

3. Experimental Evaluation

3.1. Experimental Setup

The experiments are carried out using Matlab. The basic scheme is depicted in figure 2.

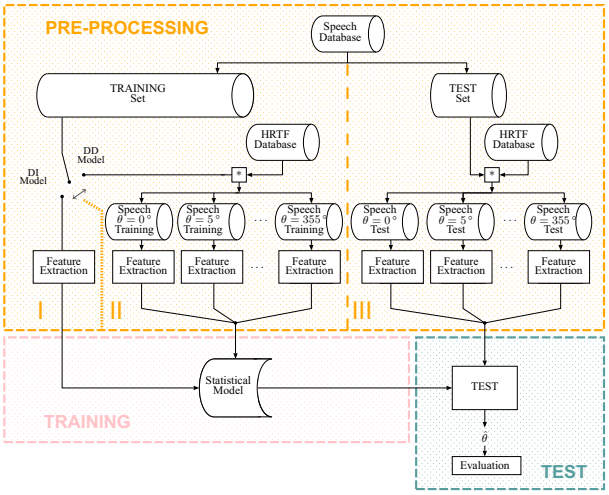

Figure 2: Block diagram of experimental setup: pre-processing with division of the speech database into training set and test set (HRTF database is the same for both), synthesizing of direction-dependent speech utterances and feature extraction; (|) for direction-independent (DI) statistical model, (||) for direction-dependent (DD) statistical model and (|||) always valid; training of either direction-independent or direction-dependent statistical model and test (classification) and evaluation of the results

There is a pre-processing step where speech material from eight different speakers (4 female, 4 male) for training and test is synthesized and features are extracted. After the pre-processing the training is carried out, where speech material from the training set is taken to train a statistical model. The statistical model can be speaker-dependent or speaker-independent. The main difference is that different input material is taken to train the statistical models. Then the results are evaluated in the test scenario.

STFT coefficients are calculated using a rectangular window with the block length of 1024 samples and a 1024-point FFT is used. Afterwards the logarithm of the absolute values of the coefficients is taken. The dimension D of the feature vector is $D = 513$. The HRTF database has been downloaded from the Internet [6]. The recordings were made in an anechoic room with the KEMAR-mannequin from Bill Gardner and Keith Martin at MIT Media Lab. The database consists of 72 HRTFs (horizontal plane – 0 °-355 °, in steps of 5 °) where the sampling frequency was reduced to $f_s = 16000$ Hz.

In this paper mixture weights for estimating the parameters of the GMMs are initialized by $b_m = \frac{1}{K}$, $\boldsymbol{\mu}_m$'s are initialized randomly and $\boldsymbol{\Sigma}_m$'s are initialized with the variance of the input training data for each dimension and is equal for each component.

A GMM is trained using the EM algorithm. As a statistical model, the parameters of the GMM are saved. In the first step in the classification process the classifier is trained to represent a class. The second step is the test scenario where a classifier is used to estimate results. Classification is done by calculating the maximum likelihood based on the posterior probability of the observed data $\mathbf{x}_l$ given the densities of the different classes. A class is represented by a Gaussian mixture model (GMM).

3.2. Results

Speech utterances from each direction are synthesized and tested in order to estimate the direction. For the evaluation the mean value $\mu_{|\epsilon|}$ and standard deviation $\sigma_{|\epsilon|}$ of the absolute error angle $|\epsilon|$ is calculated. This is done by averaging over results of all test utterances from each direction of all 8 speakers.

3.2.1. Results of Speaker-Dependent Models

In the speaker-dependent case statistical models are trained depending on the speaker, i.e. speech material from a speaker is taken to train a model for this speaker. Also in the test scenario, speech material is taken from the corresponding speaker.

In figure 3 the absolute error angle as a function of the angle to estimate can be seen. On the right-hand side the increased error due to the adaption of the model around 30 ° can be seen. To avoid the critical area, only *LimArea* is investigated.

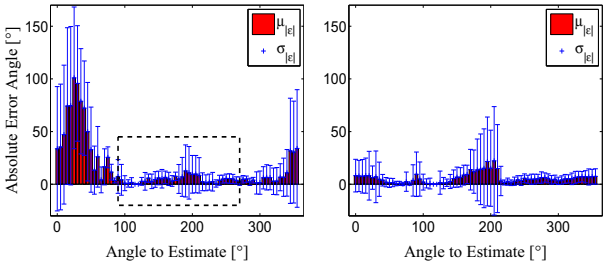

Figure 3: Mean value of absolute error angle $\mu_{|\epsilon|}$ and standard deviation $\sigma_{|\epsilon|}$ as a function of the angle to estimate; *Mismatched Case* (left) – including the *LimArea* (dashed rectangle) – and *GMM+HRTF* (right)

The evaluation of *LimArea* results in a mean absolute error of $\mu_{|\epsilon|} = 4.03\,°$ and $\sigma_{|\epsilon|} = 7.52\,°$. The mean absolute error for *GMM+HRTF* has $\mu_{|\epsilon|} = 5.86\,°$ and $\sigma_{|\epsilon|} = 15.92\,°$.

3.2.2. Results of Speaker-Independent Models

It is also important to improve the speaker-dependent case to a more general speaker-independent case. Now, the speech material from five speakers (male as well as female) is taken to train one speaker-independent model. Then tests are carried out on the remaining three speakers to assure that the training and the test is not realized based on the same speech material.

For *LimArea* the mean absolute error angle has $\mu_{|\epsilon|} = 9.04\,°$ and $\sigma_{|\epsilon|} = 14.73\,°$ whereas it has $\mu_{|\epsilon|} = 17.51\,°$ and $\sigma_{|\epsilon|} = 37.31\,°$ for *GMM+HRTF* (speaker-independent). In the speaker-independent case, which is a more general approach, the localization accuracy will decrease. Results for the two approaches and speaker-dependent and speaker-independent models are summarized in table 1.

Table 1: Comparison between the two different approaches – *GMM+HRTF* and *LimArea* – speaker-dependent (SD) and speaker-independent (SI) case

	$\mu_{\lvert\epsilon\rvert}$	$\sigma_{\lvert\epsilon\rvert}$
SD (*LimArea*)	4.03 °	7.52 °
SD (*GMM+HRTF*)	5.86 °	15.92 °
SI (*LimArea*)	9.04 °	14.73 °
SI (*GMM+HRTF*)	17.51 °	37.31 °

3.2.3. *Influence of the Segment Length*

Besides the amount of the training and test speech material and the choice and training of the statistical models, a crucial parameter for the performance is the segment length of the input utterance before a classification result is obtained. Especially for real-time applications it is important to shorten the input segments. On the one hand, the estimation of the direction of the source signal should be as accurate as possible. On the other hand, the update of the estimations should be as fast as possible. Although these two demands contradict each other, an optimal trade off has to be found.

To examine the influence of the length of the input signal on the localization accuracy, the input speech utterances are cut into pieces. Then test and evaluation are carried out as explained above. The performance is now evaluated as a function of the input segment length. So far, test utterances from each speaker arriving from each angle are taken as input. Now, all data is taken and cut into equally long segments of 1 second, 2.5 seconds, 5 seconds, 7 seconds and finally 10 seconds. Results are evaluated for *LimArea* and *GMM+HRTF* and for the speaker-dependent as well as for the speaker-independent case (see figure 4).

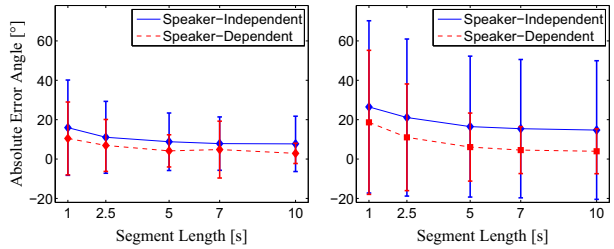

Figure 4: Comparison between speaker-independent and speaker-dependent case for *LimArea* (left) and *GMM+HRTF* (right)

The error decreases when the segment length increases. At a segment length of 2.5 seconds the mean error is 6.92 ° for the *LimArea* case which is acceptable. For the *GMM+HRTF* the error is a bit higher with a value of 11.05 °. The error increases further in the speaker-independent case to 11.07 ° (*LimArea*) and 21.09 ° (*GMM+HRTF*).

4. Conclusions

In this paper a method to estimate the direction of an incident sound wave using the signal of a single microphone is investigated. As features the logarithmic absolute STFT coefficients are taken which turned out to perform well. An incident sound wave from a certain direction and a GMM is taken to calculate the likelihood in the feature domain. A decision obtained by the values of the likelihood calculation gives the estimation of the sound source. The GMMs are first trained with direction-independent speech. Then with a simple mathematical addition in the feature domain, the direction-dependent GMMs are calculated from the direction-independent GMMs.

The negative consequence of this approach is that circular effects are introduced which destroy the localization accuracy. To get rid of these side effects, two conclusions are drawn. First, the area of investigation is restricted. Problematic areas are skipped and the localization accuracy increases, but with the drawback of a limited coverage. This is acceptable because it depends on the application whether the whole horizontal plane is needed or not.

Second, as an alternative approach, the adaption is discarded and replaced by a direct training of direction-dependent GMMs, because the reason for the circular effects is the adaption of the direction-independent GMM to produce direction-dependent GMMs. As a result localization should work in the whole range between 0 ° and 355 °.

LimArea and *GMM+HRTF* are two different approaches to estimate the angle of incident sound. Both models are promising depending on the type of application. Furthermore, the influence of the input segment length is investigated for both models as well as for speaker-dependent and speaker-independent models. Excellent localization results with an accuracy in the order of 4 ° for *LimArea* or 6 ° for *GMM+HRTF* can be presented for a speaker-independent scenario.

For further investigations in future research, the localization strategies have to be verified in a real test environment. Reverberant rooms with echoes and noise complicate the task of localization. Additionally, a device with a single microphone and an "artificial pinna" has to be built. This pinna needs to be shaped irregularly and asymmetrically. The transfer function must strongly depend on the direction of the sound source. In [5] four designs are introduced. Maybe a suitable shape can be found based on the results presented in this paper. Nevertheless, the presented localization strategy outlines a promising approach.

5. References

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed. Springer, Jun. 2001.

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

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

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

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

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

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, Oct. 2007.