Titel der Ausarbeitung

Autor A, Autor B, Autor C

***Zusammenfassung*—**

***Abstract*—**

I. Einführung

In nature the ability to hear or in other words the ability to gain informations about your environment by sound processing is an essential skill for many animals as well as for humans. Whether it is for hunting prey, for communication, or for drawing attention to potential threads, audition can help solving a variety of different tasks.

In today's world, humans use their ears in road traffic most of the time. At the same time, the technological advances in the development of autonomous cars are making significant progress. It is reasonable to assume that an auditory system could improve even more the performance of autonomous car driving. While most humans use their car horns to communicate warnings, sirens are an important tool for police, fire and rescue services that are using them officially to indicate an emergency. An auditory system can improve the decision-making of the autonomous car based on informations e.g. on the road character, the car condition, and the squealing of the tires.

While most of the development effort concentrated on robot locomotion and vision systems, an effective communication and interaction method between robots and their environment is based on auditory systems. A core component of this system is the sound source localization (SSL): The robot must be able to find the location of the voice source. In the case of autonomous car driving, the source localization contains important information for its planned driving route. ....

State of the Art

In 2003, J.-M. Valin *et al.* showed that a mobile robot can localize different types of sound sources over a range of 3 m with a precision of 3° in real time using an array of 8 microphones [4]. By 2016, they were able to localize and track simultaneous different moving sound sources over a range of 7 m using beam forming and particle filtering [5]. Liu *et al.* took a different approach with a biologically inspired spiking neural network for sound localisation in 2008 [6]. Their exprimental results showed that their model could localize a sound source from the azimuth angle, the angle of incidence, -90 to 90 degree. In 2009, Murray *et al.* presented a hybrid architecture using cross-correlation and recurrent neural networks for accoustic tracking in robots [7]. Using only two microphones, their model has shown comparable results with the capabilities of the human auditory cortex with the azimuth localisation differing by an average of $\pm 0.4°$.

Diese Arbeit wurde von Akad.Titel Vorname Name unterstützt.

Murase *et al.* used an array of 8 microphones mounted on a mobile robot in order to track multiple moving speaker. Their two key ideas were to use beamforming to locate the sound sources and to use a set of Kalman filters to track the non-linear movements of the speaker [8]. The used filters had different history lengths in order to reduce errors under noisy and echoic environments. As a result, multiple moving speakers could be tracked successfully even when speakers and the mobile robot moved non-linearly. So far, most of those systems have in common that they are built to work in closed or crowded environments to interact with people. Focusing on auditive systems for cars, we find that Fazenda *et al.* demonstrated an acoustic based safety emergency vehicle detection for intelligent transport systems in 2009 [9]. Based on a cross microphone array, they were capable of determining the incoming direction of a siren as a sound source. For their suggested array radius, their methods, which were based on time delay estimation, outperformed those, based on calculating the intensity at the microphone array.

So far, most of the mentioned system focused on sound localization, tracking and separation. Another important aspect for a hearing car is source classification. While the goal for most auditory systems for mobile robots is speech recognition, an autonomous car will more likely be confronted with environmental sounds. Performing source classification with neural networks is an active field of research. In 2012 Shen *et al.* [10] proposed a system based on pattern recognition using a Gaussian mixture models and Mel-Frequency Cepstal Coefficients features. Their system showed an average accuracy of 91.36% for offline tests with 8 different sound sources. Further evaluation in online tests yield good results as well. Piczak [11] used a convolutional neural network to classify short audio clips of environmental sounds. His model outperformed baseline implementations relying on mel-frequency cepstral coefficients and performed comparable to other state of the art approaches. This was validated with 3 public available data sets for environmental and urban recordings.

Baelde [**?**]

A. Requirements and constrains

Several requirements have to be fulfilled to enable the practical use of auditory systems for autonomous car driving. A detailled overview can be found in the thesis by Marko Durkovic [3]. The following requirements are necessary for our auditory system:

- **Robustness towards reverberation:** As sound waves propagate through space, they get reflected on surfaces in their environment. A captured sound under real conditions always means that the recording consists of the

original sound source and its reflections. The environment of a car changes constantly. Whether it be driving at urban terrain or countryside, a changing environment will yield different magnitude of reverberation which can become very intense while driving through a tunnel for instance. The auditory system for an autonomous car needs to be robust towards this changing environments without losing its accuracy.

- **Robustness towards noise:** Captured sound has usually to deal with two different kind of noise: One part is sensor noise, also know as self-noise introduced by the microphones. The other part are unwanted sound sources created by the environment. In the environment of an autonomous car, the proportion of sensor noise compared to the sound sources of your enviroment will be small. Sensor noise often becomes a problem with high pre-amplification whereas we will have to deal with high intensity noise given for instance by the airflow around the microphone array or by engine sounds around the car. Therefore, we assume that the sensor noise is negligible whereas the produced noise by our environment will be a major factor our auditory system has to deal with.
- **General applicability:** The auditory system of an autonomous car will likely be faced with many different types of sound sources. Therefore the system needs to maintain performance even if multiple sources with different signal characteristics are present.
- **Number of sources:** In a dynamic enviroment like road traffic it is not possible to determine the number of sound sources that the car will get in simultaneous contact. Therefore, the auditory system needs to be able to process observations where the amount of active sound sources is not known beforehand and possibly greater than one. The lower and upper bound of sources the system shall be able to track depends on the number of the events and their likelihood to occur. Limiting the system to identifying horns and sirens an upper bound of two different active sources can be sufficient. The more informations the system is supposed to gather, the more complex it becomes.
- **Number of dimensions:** The position of an active sound source can be described by three parameters relative to your own coordinate system. In order to track sirens or horns, the auditory system should at least be able to estimate the direction of the sound source. [That allows combined with tracking for a 2 dimensional localisation. Even though an estimation of the distance of the source would be desirable.?]
- **Source classification:** For the auditory system of an autonomous car the source classification is a mandatory feature. It has to identify sirens and preferably other desired sources with high accuracy. It needs to be robust for noise depending on the preprocessing of the sound.
- **ROS:**

II. Grundlagen

A. *Modules of an auditory system*

Subsection text.

B. *HARK and why we chose it*

Subsubsection text.

C. *Sound source localization*

The sound source localization algorithm was implemented based on the MUSIC (MUltiple SIgnal Classification) algorithm. Figure 1 displays the structure of the sound source localization algorithm [**?**]. The localization algorithm performs first by acquiring the multi-channel sound signals from the microphones and getting the multi-channel spectrum by means of the Fourier transformation. Then the cross-spectrum correlation matrix is computed to make the eigenvalue decomposition of the averaged correlation matrix over a time interval. The next stage employs the steering/position vectors and the eigenvectors of the noise subspace to calculate the MUSIC responses of each frequency bin. The final stages include the averaging of the MUSIC responses over a frequency range and identifying the direction of arrival (DOA) of the sound sources by means of peak picking. The following subsections go into the details of the MUSIC algorithm and remaining stages.

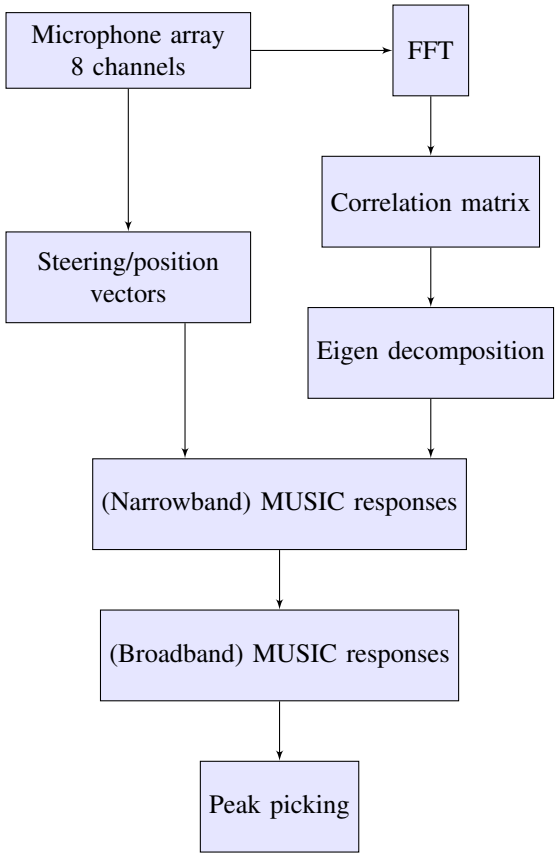

Bild 1. Block diagram of the sound source localization

D. *MUSIC algorithm*

The Multiple Signal Classification algorithm can be defined as the determination of different parameters of multiple wavefronts that enter an array of antennas or sensors [**?**].
The MUSIC algorithm provides asymptotically unbiased estimates of different parameters such as number of signals,

direction of arrival (DOA), polarization, strength and cross correlation among the directional waveforms, and many others. The M array elements receive the waveforms from the sources that are linear combinations of the D incidents wavefronts[1] and noise. This can be expressed as in equation 1.

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} = \begin{bmatrix} \mathbf{a}(\theta_1) & \mathbf{a}(\theta_2) & \dots & \mathbf{a}(\theta_D) \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_D \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_M \end{bmatrix} \tag{1}$$

$$\text{or } \mathbf{X} = \mathbf{AF} + \mathbf{W} \tag{2}$$

where the vector $\mathbf{X}$ represents the M received waveforms, the incident signals are represented by the vector $\mathbf{F}$ in phase and amplitude at some reference point, and the complex noise is represented by the vector $\mathbf{W}$.
The elements a_{ij} from the matrix $\mathbf{A}$ are functions of the angle of arrival and the position of the array elements. Therefore, each element a_{ij} depends on the position relative to the origin of the ith array element and the response of the incident signal from the direction jth. The jth column of matrix $\mathbf{A}$ is known as the mode vector of responses to the direction of arrival θ_j of the jth signal. That is, the mode vector $\mathbf{a}_j$ is equivalently to the direction of arrival θ_j.

It can be deducted from equation 1 that the vector $\mathbf{X}$ is a linear combination of the mode vectors $\mathbf{a}(\theta_j)$ in which the elements of the matrix $\mathbf{F}$ are the coefficients of this combination. The directions of arrival of multiple incident wavefronts are calculated by determining the intersections of the $\mathbf{a}(\theta)$ continuum with the range space of $\mathbf{A}$.

For determining the source localization, the correlation matrix of the input signals needs to be calculated. That is, the covariance matrix of the $\mathbf{X}$ vector is

$$\mathbf{S} = \mathbf{XX}^* = \mathbf{AFF}^*\mathbf{A}^* + \mathbf{WW}^*. \tag{3}$$

Where the * operator indicates the conjugate transpose operator. With the covariance matrix of the microphone array observations is performed a principal component analysis on this matrix to separate the disjoint signal and noise subspaces. For these means, the eigenvalue decomposition or singular decomposition of the covariance matrix $\mathbf{S}$ is carried out. The M-th space correlation matrix obtained in 3 is decomposed in the signal and noise subspaces as per

$$\mathbf{S} = \mathbf{Q\Lambda Q}^{-1} \tag{4}$$

in which $\mathbf{Q}$ is a $M \times M$ matrix with the eigenvectors $\mathbf{q}_i$ of $\mathbf{S}$ written in the column i-th, and the matrix $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues.
The singular vectors $\mathbf{q}_i$ are orthogonal to the space spanned by the columns of $\mathbf{A}$ or in other words the vectors contained in the columns of $\mathbf{A}$ are perpendicular to each other. Since the eigenvectors $\mathbf{q}_i$ have correlation to the power of the incident wavefronts, the eigen vectors can be divided into N noise eigenvectors and D incident signal mode vectors, in which the D vectors correspond to the eigenvalues with greatest value. For instance, the matrix $\mathbf{Q}$ can be written as

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_M]$$

and the split matrix gives the mode vectors $[\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_D]$ and the noise eigenvectors $[\mathbf{q}_1, \dots, \mathbf{q}_N]$.

The next step is to solve for the incident mode vectors. In order to do that the spectrum for SSL is calculated using the noise eigenvectors as per

$$P = \frac{\mid \mathbf{H}^*\mathbf{H} \mid}{\sum_{i=1}^{N} \mid \mathbf{H}^*\mathbf{q}_i \mid}. \tag{5}$$

Where $\mathbf{H}$ is the transfer function of the sound propagation. The transfer function can be determined numerically or by measurements. The transfer function is a multichannel expression that depends on the sound S_i in direction θ_i in view of the microphone array to the i-th microphone and is expressed as

$$\mathbf{H} = [H_1, \dots, H_M]. \tag{6}$$

The transfer function is to be determined with anteriority and is also called as the steering vector.

The numerator of equation 5 represents the multiplication between the steering vector (transfer function) and the noise-related eigenvector. This product is theoretically zero if the transfer function is a vector corresponding to the desired sound; making the quotient of the spectrum diverge infinitely. In reality, the denominator does not go exactly to zero due to the effects of noise but an abrupt peak can be observed.
Equation 5 represents the spectrum for every frequency a broadband SSL is performed for the desired frequencies as

$$\hat{P} = \sum_{\omega=\omega_{min}}^{\omega_{max}} W_\Lambda W_\omega P. \tag{7}$$

ω_{min} and ω_{max} determine the desired frequency bands, W_Λ is the eigenvalue weight and is the square root of the maximum eigenvalue. Finally, W_{omega} is the spectrum weight factor.

The search of the sound is performed in the frequency bands aforementioned for $\hat{P}$ where the maxima are identified by local maximum searching or the hill-climbing method

E. Sound source tracking

Subsubsection text.

F. Sound source separation

Subsubsection text.

[1] explicar que es wavefront

G. Sound source recognition

Speech recognition in HARK consists of two main processes. First, feature extraction from an audio signal and second, speech recognition using JuliusMFT or KaldiDecoder.

- **Feature extraction** Hark supports Mel-Scale Log Spectrum (MSLS) and Mel-frequeny cepstal coefficients (MFCCS) feature extraction. The MSLS is defined by the log power spectrum of a short time sound signal on the Mel-frequency domain.
- **Speech recognition** Both decoder, JuliusMFT as well as KaldiDecoder are compatible to MSLS and MFCCS as input. JuliusMFT is a modified version of Julius which is a open source high-performance continuous speech recognition decoder software. It has a large vocabulary database and can perform almost real-time decoding in 60.000 word dictation tasks. But despite it's performance in speech recognition it's not designed for classification of environmental sounds. Built with the connection to a large database to recognize human speech in mind, this approach might not be suitable for our purpose.
  KaldiDecoder is an acoustic model decoder developed for HARK. It was designed using libraries from Kaldi which is a deep learning speech recognition toolkit. It is written is C++, and the core library supports modeling of arbitrary phonetic-context sizes, acoustic modeling with subspace Gaussian mixture models as well as standard Gaussian mixture models, together with all commonly used linear and affine transforms ($https://infoscience.epfl.ch/record/192584$). Such models proved to be usefull tool for environmental sound detection as Li et all. ($https://arxiv.org/pdf/1703.06902.pdf$) showed in recent studies. They compared the performance of different kind of neural networks challenging them with fifteen different classifiable common indoor and outdoor acoustic scenes, such as bus, cafe, car, city center, forest path, train etc. They presented a gaussian mixture model which achived 77.2% test accuracy with MFCC features. The best performance achived a hierachical deep neural network with Smile6k features with a test accuracy of 88.2%. The problem with such a deep neural network is the small amount of labeled data available which is necessarry for sufficiant training. The used models were trained with in total, 13 hours of stereo audio recordings, making it one of the largest datasets available. Considering this the gaussian mixture model seems to be a sufficiant apprach and the Kaldi decoder overall favourable over the JuliusMFT decoder.

III. Experiment and result

A. Results

IV. Zusammenfassung

Anhang I
Optionaler Titel

Anhang eins.

Anhang II

Anhang zwei.

Danksagung

Wenn ihr jemanden danken wollt, der Euch bei der Arbeit besonders unterstützt hat (Korrekturlesen, fachliche Hinweise,...), dann ist hier der dafür vorgesehene Platz.

Literaturverzeichnis

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.
[2] Deutsche Forschungsgemeinschaft, *Vorschläge zur Sicherung guter wissenschaftlicher Praxis*, Denkschrift, Weinheim: Wiley-VCH, 1998.
[3] M. Durkovic: *Localization, Tracking, and Separation of Sound Sources for Cognitive Robots*, PhD thesis, Technische Universi/ät M/ünchen (2012).
[4] J.-M. Valin , F. Michaud , J. Rouat , D. Letourneau, *Robust sound source localization using a microphone array on a mobile robot*, DOI: 10.1109/IROS.2003.1248813, Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003).
[5] J.-M. Valin, F. Michaud, J. Rouat, *Robust Localization and Tracking of Simultaneous Moving Sound Sources Using Beamforming and Particle Filtering*, DOI: 10.1016/j.robot.2006.08.004, Robotics and Autonomous Systems Journal 55(3), 216-228 (2007).
[6] J. Liu, H. Erwin, S. Wermter, M. Elsaid, *A Biologically Inspired Spiking Neural Network for Sound Localisation by the Inferior Colliculus*, Artificial Neural Networks - ICANN, Springer Berlin Heidelberg (2008).
[7] J. C. Murray, H. Erwin, S. Wermter, *A Hybrid Architecture Using Cross-Correlation and Recurrent Neural Networks for Acoustic Tracking in Robots*, Biomimetic Neural Learning for Intelligent Robots, Lecture Notes in Computer Science, vol 3575, Springer Berlin Heidelberg (2005).
[8] M. Murase, S. Yamamoto, J.-M. Valin, K. Nakadai, K. Yamada, K. Komatani?, T. Ogata, H. G. Okuno, *Multiple Moving Speaker Tracking by Microphone Array on Mobile Robot*, Interspeech (2015).
[9] B. Fazenda, H. Atmoko, F. Gu, L. Guan, A. Ball: Acoustic based safety emergency vehicle detection for intelligent transport systems, ICCAS-SICE (2009).
[10] G. Shen, Q. Nguyen, J. Choi, *An Environmental Sound Source Classification System Based on Mel-Frequency Cepstral Coefficients and Gaussian Mixture Models*, IFAC proceedings, volume 45, Issue 6, pages 1802-1807 (2012).
[11] K. J. Piczak, *Environmental sound classification with convolutional neural networks*, DOI: 10.1109/MLSP.2015.7324337, IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP) (2015).
[12] M. Baelde, C. Biernacki, R. Greff, *A mixture model-based real-time audio sources classification method*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017).