Evaluation of Speaker Localization methods for Vehicle Interior Applications

Mattes Ohlenbusch[1], Andreas Volgenandt[1], Stephanus Volke[1], Christian Rollwage[1],
Joerg Bitzer[1,2]

[1] *Fraunhofer IDMT / Hör-, Sprach- und Audiotechnologie; 26129 Oldenburg, Deutschland*

[2] *Jade Hochschule Oldenburg, Insitut für Hörtechnik und Audiologie, 26129 Oldenburg, Deutschland*

Contact: mattes.ohlenbusch@idmt.fraunhofer.de

Introduction

Following the commercial advent of advanced driver-assistance systems (ADAS), the scope of automotive human-machine interfaces has grown considerably during recent years. As a consequence, operations carried out by users have become more complex structures, thus amplifying the appeal of intuitive voice control-based interaction. Spatial signal enhancement methods, e.g. beamforming, require acoustic source localization. In order to ensure the viability of speech-based human machine-interfaces, robust algorithms for speaker localization are required.

Previous research on this topic has been published in [1], where the Generalized Cross-Correlation method (GCC) was utilized in combination with the Phase Transform (PHAT) [2]. The approach based on time delay estimation was shown to be viable in a parking vehicle. A successive publication observed the overall superiority of the Steered Response Power algorithm (SRP) in respect to the GCC [3]. The authors also found that algorithmic performance differs based on head orientation. In [4], additional evaluation was carried out within a driving vehicle. The SRP-PHAT method was also applied in different, automotive-related environmental noises. Microphones mounted on the cabin ceiling proved to be superior to a line-array in front of the cabin for the speaker localization task. Hu et al. proposed a different method in which Gaussian Mixture Models were used to differentiate between speaker locations based on phase differences [5]. In experimental evaluation, their method was able to outperform the Multiple Signal Classification algorithm (MUSIC), which is known to be asymptotically efficient under certain circumstances [6]. This paper presents a comparison of recent methods which are to be considered for in-car speaker localization. Specifically, different algorithms are evaluated based on their performance in both synthetic and real disturbances.

Methods

Since the localization of speakers inside moving vehicles presents a challenging task, different approaches have to be considered. For an algorithm to be applicable, it is required to be robust against various disturbances. These can be manifold, such as motor and street noises or reflections from surfaces, e.g. car window panes. In this section, a signal model is established first, followed by algorithmic descriptions.

Signal Transfer Model

Considering a microphone array of M receivers, the discrete input signal at the mth receiver is denoted as $x_m(k)$, where k is the sample index and $m \in 1, \ldots, M$ the microphone index. Its Short-Time Fourier Transform (STFT) is referred to as $X_m(n)$, where n denotes the frequency bin index. Using the STFT spectra, power spectral density (PSD) estimates can be made using recursive Welch periodograms. They will be referred to as $\Phi_{ml}(n)$, with PSD estimates being cross spectral densities (CSD) for $m \neq l$. The $M \times M$ PSD matrix $\mathbf{\Phi}(n)$ contains all possible CSDs for the considered array.

A signal $s(n)$ arriving at the array from an azimuth angle of θ is included in all signal channels, but with varying time delay $d_m(\theta)$ in fractional samples due to geometrical distances. Assuming uncorrelated noise $n_m(k)$ and ignoring inter-microphone level differences, the input signal at microphone m is written as

$$x_m(k) = s(k + d_m(\theta)) + n_m(k). \tag{1}$$

Therefore, it is reasonable for localization methods to employ cross-correlation measures. All of these methods require a broadband estimate of spatial pseudo-power, over which a maximum search is then carried out. The highest peaks in the respective spatial powers correspond to source direction estimates.

Global Coherence Field

The Global Coherence Field (GCF) algorithm employs GCC measures to perform a spatial grid search over positions $\mathbf{x}$, where $\mathbf{x}$ is a coordinate vector [7]. It is computed as

$$P_{\mathrm{GCF}}(\mathbf{x}) = \sum_{(m,l)} r_{m,l}(\lfloor \tau_{m,l}(\mathbf{x}) \rceil), \tag{2}$$

where $r_{m,l}$ denotes the generalized cross-correlation for microphones m and l, $\tau_{m,l}(\mathbf{x})$ the delay corresponding to position $\mathbf{x}$ and $\lfloor \cdot \rceil$ rounding to the nearest integer. Utilizing the Phase Transform (PHAT), it can be written as

$$r_{m,l}(\kappa) = \mathcal{F}^{-1} \left\{ \frac{\Phi_{m,l}(n)}{|\Phi_{m,l}(n)|} \right\} \tag{3}$$

for a time shift κ, where $\mathcal{F}^{-1}\{\cdot\}$ denotes the inverse STFT.

Steered Response Power

The SRP can be formulated as a beamforming algorithm, where the steering vector is defined as

$$\mathbf{a}(n,\theta) = \left[1, e^{\frac{j2\pi n}{N},\cdot\tau_1(\theta)}, \ldots, e^{\frac{j2\pi n}{N}\cdot\tau_M(\theta)}\right]^T, \tag{4}$$

using an STFT size of N and delays $\tau_m(\theta)$ corresponding to the respective microphone and direction. An easily comprehensible way to express the SRP is

$$P_{\mathrm{SRP}}(n,\theta) = \mathbf{a}^H(n,\theta)\mathbf{\Phi}(n)\mathbf{a}(n,\theta), \tag{5}$$

although formulations exist which enable more efficient, vectorized computation [8].

MUSIC

The Multiple Signal Classification (MUSIC) algorithm can be used for both frequency and direction of arrival estimation [6]. It is based on the separation of signal and noise within the PSD matrix, which is carried out through an eigenvalue decomposition. Assuming spatially uncorrelated noise, the signal component is associated with the largest eigenvalue. For additional sources, the corresponding number of largest values is selected. This requires the number of sources to be known in advance, although a criterion where all eigenvalues larger than their overall mean are selected is also possible. The eigenvectors paired with the selected eigenvalues span the signal subspace $\mathbf{U}_{\mathrm{s}}$, others the noise subspace $\mathbf{U}_{\mathrm{n}}$. To compute the pseudo-power for a given location, the following expression has to be computed [9]:

$$P_{\mathrm{MUSIC}}(n,\theta) = \frac{1}{\mathbf{a}(n,\theta)^H\mathbf{U}_{\mathrm{n}}\mathbf{U}_{\mathrm{n}}^H\mathbf{a}(n,\theta)} \tag{6}$$

Under certain conditions, this method provides optimal estimation, although this comes with high computational complexity.

Diagonal Unloading Beamforming

Analogous to the SRP, the Diagonal Unloading (DU) Beamformer [9] is defined as

$$P_{\mathrm{DU}}(n,\theta) = \frac{1}{\mathbf{a}^H(n,\theta)(\mathrm{tr}(\mathbf{\Phi}(n)\mathbf{I} - \mathbf{\Phi}(n))\mathbf{a}(n,\theta)}. \tag{7}$$

Here, $\mathrm{tr}(\cdot)$ denotes the trace operator and $\mathbf{I}$ the identity matrix. The trace of a matrix contains the sum of its eigenvalues, invariant to a change of its basis. This unloading operation along the main diagonal of the PSD matrix achieves the attenuation of the signal subspaces compared the noise subspace, achieving a similar effect as the MUSIC algorithm [9].

Frequency Fusion

In order to utilize the majority of the aforementioned algorithms, a frequency fusion algorithm combining all subband components has to be chosen. Within this paper, a Normalized Arithmetic Mean (NAM) frequency fusion is employed [10]. It is defined as

$$P_{\mathrm{NAM}}(\theta) = \sum_n \frac{P(n,\theta)}{\max_\theta[P(n,\theta)]}. \tag{8}$$

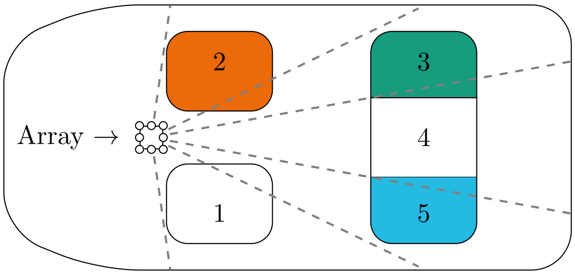

Figure 1: Schematic structure of the proposed setup. The microphone array is attached to the interior ceiling in front of the front row.

Other than achieving spatial normalization for the observed frequency range, this post-filtering technique has the important property of attenuating noise when a input signal component only occurs in few or even a single beamformers. This presents an advantage to the localization of acoustic sources when used with beamforming algorithms.

Support Vector Localization

A data-driven approach computes localization estimates based on GCC features, which are assigned to a feature class corresponding to a directional range [11]. Instead of taking into account the transfer path model for a specific acoustic environment, training data incorporating the acoustic conditions is used. On this basis, class borders in a multidimensional GCC feature space are established using support vector classification. For the specific scenario of in-car speaker localization, the limited amount of possible speaker positions may provide an advantage as relatively few feature classes need to be distinguished within the feature space.

Evaluation Methods

In order to evaluate the methods considered, an experimental setup shown in Figure 1 is proposed. Instead of computing position estimates, algorithms are applied to produce directional estimates. For evaluation purposes, spatial regions are attributed to their specific seats. Data for experimental evaluation was recorded from within a Mitsubishi Colt Z30, where a total of three speakers with their individual heights ranging from 1.75 m to 2 m. Additionally, noise disturbances within the driving vehicle were recorded at approximate velocities of 30, 50, 70 and 100 km/h in the absence of rainfall during afternoon commuter traffic. In order to generate synthetic training data for the SVM approach, car impulse responses for each speaker position were measured. The data was then generated from white noise convoluted with the respective impulse responses, where uncorrelated noise was added to result in Signal to Noise-Ratios (SNRs) of 0, 5 10 and 15 dB. In addition to real recordings, synthetic diffuse noise was generated using an approach published in [12].

Results

The evaluation consists of four steps: first, a conversation within a standing vehicle is considered. A simulated situation in diffuse noise follows, after which evaluation in acoustic disturbances in different driving speeds is carried out. The localization performance with a varying number of microphones utilized is also analyzed.

Conversation Within Standing Vehicle

During conversation of multiple passengers, the location of each speaker's mouth may change within the seat region boundary. An example conversation is visualized in Figure 2, where an audio signal and corresponding localization estimates are shown. Individual estimates are mostly close to or directly at the approximate speaker direction, but small variations occur.

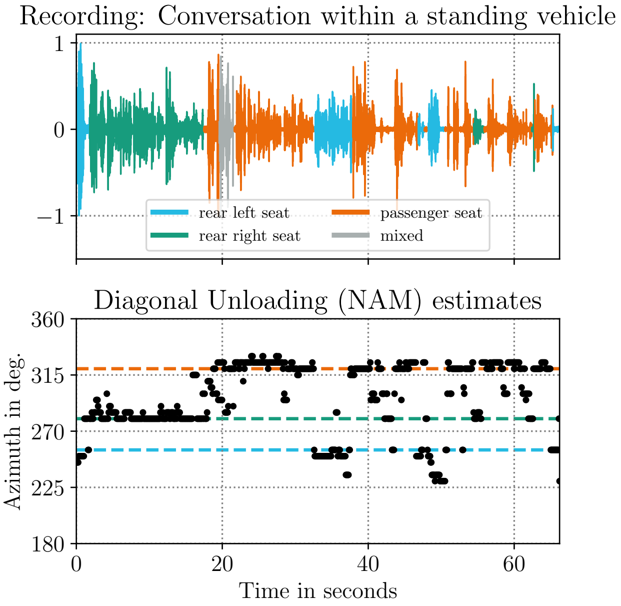

Figure 2: Audio signal of a conversation and the resulting Diagonal Unloading-estimates over time. In the top subplot, different speakers are color-coded, their respective approximate directions are shown as dashed lines in the bottom subplot.

Evaluation in Diffuse Noise

In order to evaluate the theoretical performance of individual algorithms, pink diffuse noise was used. The results for varying SNRs are shown in Figure 3. Accuracy is defined as the relative amount of estimates within the correct directional boundaries. In this situation, the MUSIC and the SVM method show a higher accuracy in low SNRs while SRP, DU and MUSIC are most accurate for higher SNRs.

Influence of Driving Noise

Recordings of acoustic disturbances and noise for different vehicle velocities were used to recreate realistic acoustic situations for in-car speaker localization. Different from the diffuse noise scenario, this situation features impulse-like noises as well as motor noise and emissions from other passing vehicles nearby. Quantitative results for this experiment are illustrated in Figure 4. With increasing velocity, the performance of all algorithm worsens, but SRP, MUSIC and DU appear to be the most robust methods.

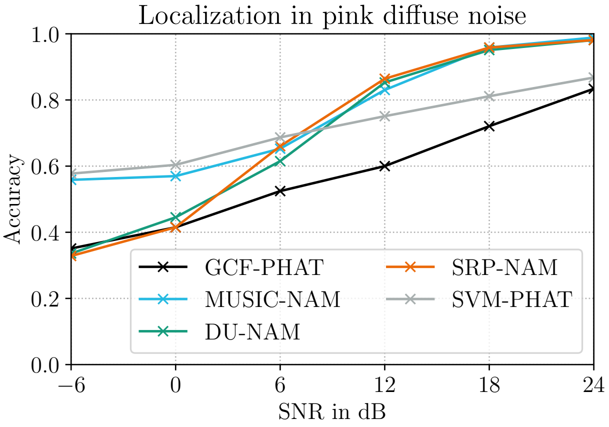

Figure 3: Localization accuracy for real speech recordings with added synthetic noise.

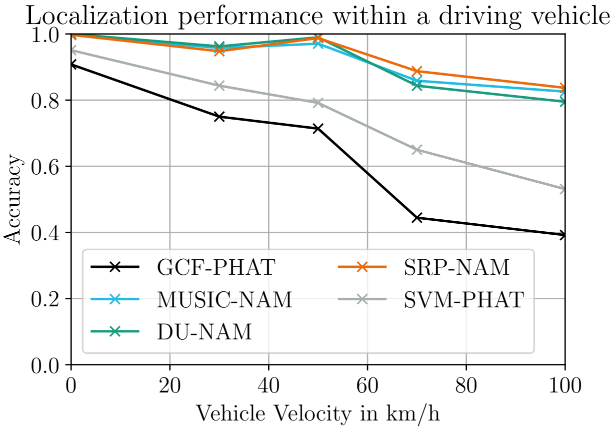

Figure 4: Accuracy of localization experiments using real recordings of driving noise. Signal to Noise-Ratio for speech varies, but remains mostly within -5 to 5 dB.

Reducing the Microphone Count

Since real applications for vehicle interior speaker localization feature numerous constrains when it comes to material and computational costs, an additional measure for the suitability of algorithms is their performance with a reduced count of receivers. As an experiment, The driving noise originating from a vehicle driving with 30 km/h were used to analyze this aspect of in-car speaker localization.

Results for this experiment are shown in Figure 5. In this context, the DU Beamformer shows greater accuracy compared to other algorithms with a low microphone count. It should be noted that when the number

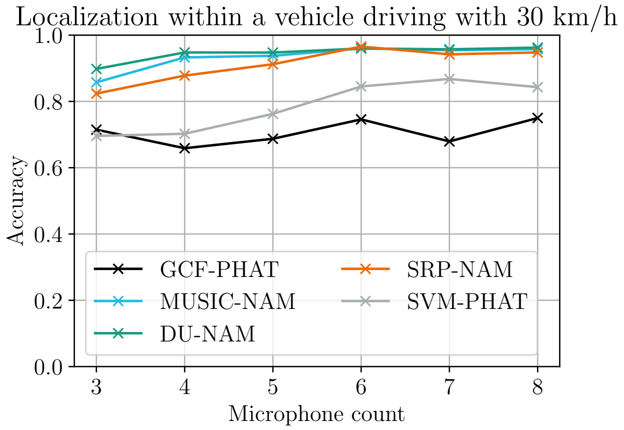

Figure 5: Localization accuracy within a driving vehicle at varying microphone counts.

of active sources is equal or greater than the microphone count, MUSIC and DU-Beamforming will fail to utilize the subspace orthogonality.

Conclusions and Outlook

While the computational cost for a concrete implementation varies and depends on the algorithm choice, it can be said that real-time in-car speaker localization is feasible, as also noted in [4]. Since interior structure between different vehicle types varies, optimal microphone count placement needs to be optimized accordingly for specific models. Generally speaking, a trade-off between accuracy, robustness and computational and material costs is to be expected. Following from the results presented in this paper, Diagonal Unloading Beamforming and the Steered Response Power Algorithm are assumed to provide the best performance in terms of noise and with low microphone count. Both these factors are vital to practical solutions for in-car speaker localization systems. What has yet to be considered are impulse-like noises appearing during regular traffic. Possible solutions include the use of preprocesssing algorithms that apply knowledge of the acoustic surroundings and spectral properties of the target signal. Practical implementations could also benefit from multi-speaker tracking algorithms to further improve voice control interaction. A viable solution for device selection based on head orientation as proposed in [13] might assist in enhancing the distinction of passenger-to-passenger from passenger-to-car system-communication.

References

[1] Alexej Swerdlow, Kristian Kroschel, and Timo Machmer. “Speaker localization in vehicles via acoustic analysis”. In: *34th Deutsche Jahrestagung für Akustik (DAGA’07), Stuttgart, Germany* (2007).

[2] Charles Knapp and G. Carter. “The generalized correlation method for estimation of time delay”. en. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.4 (Aug. 1976), pp. 320–327. ISSN: 0096-3518. DOI: 10.1109/TASSP.1976.1162830. URL: http://ieeexplore.ieee.org/document/1162830/ (visited on 08/01/2018).

[3] Alexej Swerdlow, Timo Machmer, and Kristian Kroschel. “Speaker position estimation in vehicles by means of acoustic analysis”. In: *35th Deutsche Jahrestagung für Akustik (DAGA’08), Dresden, Germany* (2008).

[4] Timo Machmer et al. “Position Estimation of Car Occupants by Means of Voice Analysis”. In: *International Conference on Acoustics NAG/DAGA, Rotterdam, The Netherlands* (2009).

[5] Jwu-Sheng Hu, Chieh-Cheng Cheng, and Wei-Han Liu. “Robust speaker’s location detection in a vehicle environment using GMM models”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36.2 (2006), pp. 403–412.

[6] Ralph Schmidt. “Multiple emitter location and signal parameter estimation”. In: *IEEE transactions on antennas and propagation* 34.3 (1986), pp. 276–280.

[7] Maurizio Omologo and Piergiorgio Svaizer. “Use of the crosspower-spectrum phase in acoustic event location”. In: *IEEE Transactions on Speech and Audio Processing* 5.3 (May 1997), pp. 288–292. ISSN: 1063-6676. DOI: 10.1109/89.568735.

[8] Bowon Lee and Ton Kalker. “A vectorized method for computationally efficient SRP-PHAT sound source localization”. In: *12th International workshop on acoustic echo and noise control (IWAENC 2010)*. 2010.

[9] Daniele Salvati, Carlo Drioli, and Gian Luca Foresti. “A low-complexity robust beamforming using diagonal unloading for acoustic source localization”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.3 (2018), pp. 609–622.

[10] Daniele Salvati, Carlo Drioli, and Gian Luca Foresti. “Incoherent frequency fusion for broadband steered response power algorithms in noisy environments”. In: *IEEE Signal Processing Letters* 21.5 (2014), pp. 581–585.

[11] Hendrik Kayser and Jörn Anemüller. “A discriminative learning approach to probabilistic acoustic source localization”. In: *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE. 2014, pp. 99–103.

[12] Jens-Alrik Adrian et al. “Synthesis of Perceptually Plausible Multichannel Noise Signals Controlled by Real World Statistical Noise Properties”. In: *Journal of the Audio Engineering Society* 65.11 (2017), pp. 914–928.

[13] Menno Müller, Steven van de Par, and Joerg Bitzer. “Head-Orientation-Based Device Selection: Are You Talking to Me?” In: *Speech Communication; 12. ITG Symposium*. VDE. 2016, pp. 1–5.