Influence of a Spherical Microphone Array on a Sound Source Number Estimator

Paul Kranzusch, Stephan Gerlach, Danilo Hollosi, Stefan Goetze

Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology 26129 Oldenburg, Germany

Email: {paul.kranzusch, stephan.gerlach, danilo.hollosi, s.goetze}@idmt.fraunhofer.de

Abstract

For acoustic sound source localization, e.g. in hands-free communication systems, speech recognition or acoustic event detection systems, knowledge of the number of concurrent sound sources is necessary. This paper analyses to what extent a spherical microphone array influences the estimation of the number of sound sources. State of the art approaches already exploit knowledge about head related transfer functions (HRTF) for performance improvement, inspired by findings in human auditory perception. Unfortunately, the influence of HRTFs on the estimation of the number of sound sources is rarely investigated. Therefore, an estimation algorithm based on information-theoretic criteria using multi-channel data is evaluated in this paper. We consider a uniform circular microphone array, either as open construction or for microphones attached to the surface of a rigid sphere. Simulated data as well as realistic acoustic transfer functions measured in anechoic room conditions are used to derive recommendations for improvement of the sound source number estimation.

Motivation

In the recent years, uniform circular microphone arrays (UCA) attached to the surface of rigid sphere, are often used for multi-channel-signal-processing approaches. These methods require information about the number of concurrent source signals. In this, paper we evaluate the influence of a rigid sphere on the performance of the source number estimation method based upon the minimum description length criterion (MDL).

Signal Model

We consider a scenario where Q speech sources are recorded using M microphones which, in vector notation, can be written as

$$\mathbf{y}_m(t) = \mathbf{H}(t) * \mathbf{s}_q(t) + \mathbf{v}_m(t) \tag{1}$$

where $\mathbf{y}_m(t)$ describes the $m = 1, 2, ..., M$ microphone signals at time t generated from $q = 1, 2, ..., Q$ source signals $\mathbf{s}_q(t)$ convolved with the $M \times Q$ mixing matrix $\mathbf{H}(t)$. The operator $*$ denotes the convolution operation, and $\mathbf{v}_m(t)$ is additional uncorrelated noise. In Figure 1, the schematic setup of the signal model is depicted.

This work was partly funded by the EU projects "Sounds for Energy-Efficient Buildings" (S4EcoB, project no. 284628) and "Experimenting Acoustics in Real environments using Innovative Testbeds" (EAR-IT, project no. 318381)

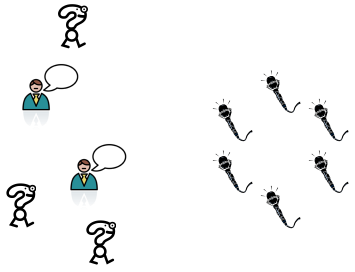

Figure 1: Signal model using a uniform cirular array with known number of microphones M and unknown number of sources Q.

Minimum Description Length

The minimum description length (MDL) is based on information theoretic criteria [1, 2]. The MDL can be used in time- or frequency-domain, while the frequency-domain has turned out to be more reliable in an acoustic setting. The flow chart of the used method with all processing steps is shown in Figure 2. As a first step,

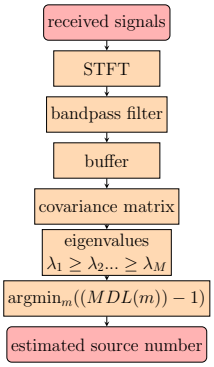

Figure 2: Proceeding steps for the estimation of the number of sound sources using MDL

the microphone signals are transformed into the time-frequency domain by the short-time-Fourier-transform (STFT). Only considering frequency bins of interest, the covariance matrix for each frequency bin is estimated using the $N \times Q$ matrices $\mathbf{Y}(\omega, \ell)$, with ω denoting the frequency bins at time block ℓ. The eigenvalues λ_m of the covariance matrix provide information about the power and distribution of the signals on the microphone channels. After sorting the eigenvalues in decreasing order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$, the eigenvalue $\lambda_{(M-1)}$ which is supposed to be dedicated to the strongest noise component, can be found using $\text{argmin}_M \{MDL(m)\}$. All larger

eigenvalues $\lambda_{\geq M-Q}$ can be associated with a source signal. Thus, the source number is estimated by

$$\hat{Q}_\omega = \text{argmin}_m \{MDL(\omega, m)\} - 1 \tag{2}$$

Based on this condition, the maximum estimated source number is limited by $M - 1$. The MDL criterion itself can be described as

$$MDL(\omega, m) = -\ln \left(\frac{\prod_{i=m+1}^{M} \lambda_i(\omega)^{1/(M-m)}}{\frac{1}{M-m} \sum_{i=m+1}^{M} \lambda_i(\omega)} \right)^{(M-m)N} + 1/2m(2m - M) \ln N \tag{3}$$

for every frequency bin ω and the eigenvalues λ_m of the covariance matrix [1, 2, 3].

Experiment

The MDL based source number estimator has been evaluated using simulated impulse responses, generated with the SMIRgen toolbox [4] and with impulse responses measured in an anechoic chamber. As a precondition, the number of microphones exceeds the number of sources $Q < M$ in all measurements. Three measuring conditions are tested: (i) a simulated scenario with an open UCA and (ii) an UCA on rigid sphere as well as (iii) impulse responses measured on a rigid sphere in an anechoic room. Four or eight equally circular distributed microphones with approx. 15 cm sphere radius. Up to six sources where positioned in a semicircle at a distance of 2.5 m to the center of the sphere and angles in steps of 24°. Uncorrelated, speech-shaped noise was added to each microphone channel at a SNR defined by the signal in the first microphone channel. The results are shown in Figures 3 and 4. The depicted detection rates are composed of the number of correct estimations in relation to the number of all estimations in a measurement.

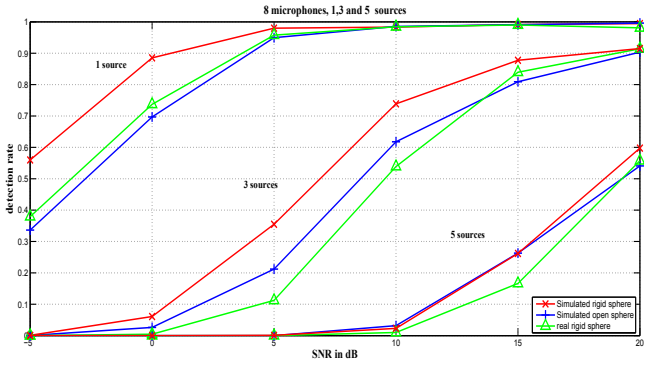

Figure 3: Detection rates for different SNRs for 1, 3 and 5 sources using a UCA with 8 microphones. The detection rate decreases with higher numbers of sources. The simulated rigid sphere provides the best results.

A detection rate exceeding 0.5 is an indicator for a good estimation, as one can get the proper result by calculating the median from the single detections. Figure 3 shows the detection rate in all setups with 8 microphones. With a higher number of sources, the algorithm needs a higher SNR for the correct estimation. It can be seen from Figures 3 and 4 that the simulated rigid sphere provides better results than the open sphere, while the simulated open sphere and the real rigid sphere are comparable in most cases.

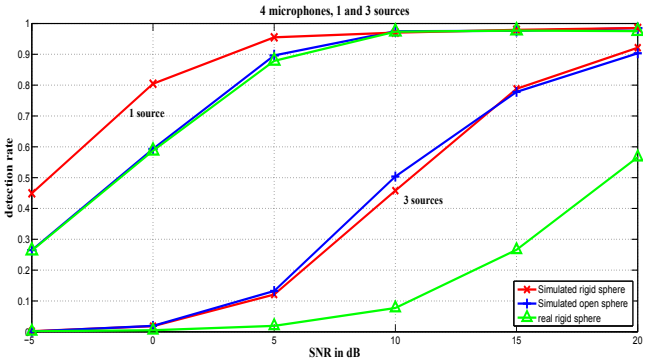

Figure 4: Detection rates for 1 and 3 sources using a UCA with 4 microphones. Detection rates are lower compared to the UCA 4 microphones (c.f. Figure 3).

In Figure 4, results with 4 microphones are shown. Compared to 8 microphones, the detection rate is decreased for less microphones, especially for a higher source number. The detection rate of the real rigid sphere is shifted significantly. This may occur from deviations in the RIR measurement. With more microphones the detection rate seems to be more reliable, especially for the real rigid sphere condition with multiple sources. In most cases the real rigid sphere condition leads to approx. the same results as the simulated open sphere.

Conclusion

In all experiments the simulated rigid sphere outperforms the simulated open sphere. Thus, the surface of the sphere seems to have a positive influence on the source number estimation algorithm. Furthermore, the estimations based on the measured real impulse response data provide decent performance compared to the simulated data.

References

[1] Kritchman, S., Nadler, B.: Non-Parametric Detection of the Number of Signals: Hypothesis Testing and Random Matrix Theory (2009)

[2] Wax, M., Kailath, T.: How to write a manuscript. Detection of signals by information theoretic criteria (1985), (pp. 387-392),Acoustics, Speech and Signal Processing, IEEE Transactions on Vol. 33(2)

[3] Nannen, V.: A Short Introduction to Model Selection, Kolgomorov Complexity and Minimum Description Length (MDL) (2003)

[4] Jarrett, D., Habets, E., Thomas, M., Naylor, P.: Rigid sphere room impulse response simulation: algorithm and applications, *http://home.tiscali.nl/ehabets/smirgen.html* (2012)