STFT Bin Selection for Localization Algorithms based on the Sparsity of Speech Signal Spectra

Andreas Brendel, Chengyu Huang, and Walter Kellermann
Multimedia Communications and Signal Processing,
Friedrich-Alexander-Universität Erlangen-Nürnberg,
Cauerstr. 7, D-91058 Erlangen, Germany
{Andreas.Brendel, Chengyu.Huang, Walter.Kellermann}@FAU.de

Summary
Many algorithms for localizing, tracking or Direction of Arrival (DOA) estimation of speech sources, rely on the so-called W-disjoint orthogonality, i.e., only one speaker is assumed to be active at a certain time-frequency bin. Based on this assumption, bin-wise DOA estimates can be computed from pairwise phase differences of each time-frequency bin and clustered afterwards. Averaging the estimates of each cluster, i.e., computing the cluster centroids, increases the robustness of the localization estimate. However, clustering can be computationally demanding due to the large amount of DOA estimates, and at the same time highly sensitive to errors as potentially many of them may not be reliable due to noise and reverberation. Therefore, an efficient selection algorithm for reliable Short-Time Fourier Transform (STFT) bins is desirable that aims at increasing the accuracy of the estimate while simultaneously reducing the computational complexity. In this contribution, we investigate different selection methods for STFT bins as suitable for localization algorithms for speech sources, which are based on the W-disjoint orthogonality, and exploit bin-wise speech signal power, Coherent-to-Diffuse Power Ratio (CDR), and Speech Presence Probability (SPP). The effectiveness of the selection processes is studied for different localization algorithms.

PACS no. 43.60.+d

1. Introduction

Localization and DOA estimation of one or multiple speakers in an acoustic environment is an important preprocessing step for many signal processing algorithms, e.g., steering a beam to the direction of a desired source [1] or pointing the camera in a video conferencing scenario to the active speaker [2].

Especially DOA estimation has attracted much interest in the last decades and many approaches have been developed to address this problem, e.g., Generalized Cross Correlation and Steered Response Power [3], Blind Source Separation (BSS)-based DOA estimation [4] or narrowband DOA estimation in combination with BSS [5, 6].

Many state-of-the-art approaches rely on W-disjoint orthogonality [7] of speech sources in the STFT domain, i.e., it is assumed that only one source is active at each STFT bin. Therefore, a narrowband DOA estimate computed from an STFT bin corresponds to a single source under the assumption of W-disjoint orthogonality. These narrowband estimates can be combined afterwards to obtain an overall localization result or DOA estimate. For DOA estimation, [8] used the k-means algorithm to cluster the observed narrowband estimates to obtain a global DOA estimate. The phase differences of the STFT bins between the observed microphone signals are used for DOA estimation in [9]. The same feature was used for localization in an Acoustic Sensor Network (ASN) in [10] and a distributed localization algorithm in [11]. Data fusion of the observed narrowband DOA estimates has been done by triangulation and subsequent clustering in the Cartesian space in [12] for localization and similarly for tracking in [13] for the estimation of the position of multiple speakers in an enclosure.

There are a few approaches to improve the performance of these narrowband localization algorithms, e.g., enhancing narrowband DOA estimation by replacing the microphone signals by the parameters of a complex Watson distribution [14] or using outlier rejection [12] to improve the localization performance. The signal power was used for weighting bin-wise estimates in [6], similarly the CDR was used in [15] and the SPP in [13]. However, a comparative study of the efficacy of such methods is still missing.

In this contribution, we investigate different observation-based parameter estimates, namely bin-wise signal power, CDR, and SPP, for STFT bin selection as a preprocessing step for DOA estimation. Thus, bins corresponding to low values for these parameters are discarded before clustering. Most of the narrowband DOA estimation or localization algorithms are based on clustering these bin-wise estimates. Hence, the selection aims at improving the localization results as well as reducing the computational complexity of the clustering. Finally, a simulation study is performed to show the beneficial effects of the selection process followed by a discussion w.r.t. the application in a simple DOA and a localization algorithm.

2. Signal Model

In the following, the frequency bins are indexed by $k = 0, \ldots, K-1$ and the time frames by $l = 0, \ldots, L-1$. We consider a microphone array of arbitrary shape comprising M microphones lying in a plane and observing a static acoustic scene containing S spatially separated speech sources. In the STFT domain, the m-th microphone signal in time-frequency bin (l, k) can be described by

$$x_m(l,k) = \sum_{s=1}^{S} h_{sm}(k) v_s(l,k) + w_m(l,k), \qquad (1)$$

where h_{sm} is the acoustic transfer function from source s to microphone m, v_s the source signal of index s and w_m additive noise present at microphone m. With the assumption of W-disjoint orthogonality, this signal model can be simplified to

$$x_m(l,k) = h_{sm}(k) v_s(l,k) + w_m(l,k), \qquad (2)$$

i.e., only a single source s contributes to the observed mixture in STFT bin (l, k).

3. DOA Estimation

In the following, an algorithm for DOA estimation in 2D with an array of arbitrary geometry is described [8], which will be a fundamental building block for the rest of the paper. To this end, a vector of unit length pointing to the direction of the source active at time-frequency bin (l, k) is defined by

$$\frac{\mathbf{q}(l,k)}{\|\mathbf{q}(l,k)\|_2} = \begin{bmatrix} \cos\theta_s(l,k) \\ \sin\theta_s(l,k) \end{bmatrix}, \qquad (3)$$

where $\theta_s(l, k)$ denotes the azimuth of the source s active in (l, k). With these preliminary steps, the acoustic transfer function in STFT bin (l, k) can be approximated under the free-field assumption as

$$h_m(l,k) = \exp\left[-j\frac{2\pi k}{K}\frac{f_s}{c}\mathbf{d}_m^{\mathrm{T}}\frac{\mathbf{q}(l,k)}{\|\mathbf{q}(l,k)\|_2}\right],$$

with the position of microphone m at $\mathbf{d}_m$, the sampling frequency f_s and the sound velocity c. The vectors pointing from the reference microphone R to the other microphones of the array are collected in the matrix

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1 - \mathbf{d}_R \\ \cdots \\ \mathbf{d}_{R-1} - \mathbf{d}_R \\ \mathbf{d}_{R+1} - \mathbf{d}_R \\ \cdots \\ \mathbf{d}_M - \mathbf{d}_R \end{bmatrix} \qquad (4)$$

where the pair $\mathbf{d}_R - \mathbf{d}_R$ is discarded. The unit vector pointing to the source, active in bin (l, k), can be expressed as [8]

$$\frac{\mathbf{q}(l,k)}{\|\mathbf{q}(l,k)\|_2} = \mathbf{D}^{+}\boldsymbol{\tau}_{mR}\, c, \qquad (5)$$

where $(\cdot)^+$ denotes the pseudoinverse. The Time Difference of Arrival (TDOA) vector of all microphones m w.r.t. the reference microphone $\boldsymbol{\tau}_{mR}$ can be computed by calculation of the phase differences between microphone signal m and reference microphone signal R in STFT bin (l, k) and normalization by the frequencies f_k corresponding to the respective time-frequency bin

$$\tau_{mR}(l,k) = \frac{\arg\{x_m(l,k)x_R^*(l,k)\}}{2\pi f_k}. \qquad (6)$$

The ratio of the second and the first element of the directional vector in (3) is the tangent of the respective DOA

$$\tan\theta_s(l,k) = \frac{\sin\theta_s(l,k)}{\cos\theta_s(l,k)} = \frac{[\mathbf{q}(l,k)]_2}{[\mathbf{q}(l,k)]_1}. \qquad (7)$$

Therefore, the DOA, i.e., the azimuth angle w.r.t. the microphone array reference system can be computed by $\hat{\theta}(l,k) = \mathrm{atan}\left(\frac{[\mathbf{q}(l,k)]_2}{[\mathbf{q}(l,k)]_1}\right)$.

4. STFT Bin-wise Parameter Estimates

The following section introduces parameters for selecting or discarding STFT bins for DOA estimation or localization.

4.1. Bin-wise Signal Power

The bin-wise signal power has been used for weighting an EM algorithm for acoustic source separation in [6]. Here, we determine the bin-wise power of microphone m as $P_m(l,k) = |x_m(l,k)|^2$.

$$\mathrm{CDR}_{mR} = \frac{\Gamma_w^{mR}\,\mathrm{Re}\left\{\hat{\Gamma}_x^{mR}\right\} - \left|\hat{\Gamma}_x^{mR}\right|^2 - \sqrt{(\Gamma_w^{mR})^2\,\mathrm{Re}\left\{\hat{\Gamma}_x^{mR}\right\}^2 - (\Gamma_w^{mR})^2\left|\hat{\Gamma}_x^{mR}\right|^2 + (\Gamma_w^{mR})^2 - 2\,\Gamma_w^{mR}\,\mathrm{Re}\left\{\hat{\Gamma}_x^{mR}\right\} + \left|\hat{\Gamma}_x^{mR}\right|^2}}{\left|\hat{\Gamma}_x^{mR}\right|^2 - 1} \tag{8}$$

4.2. Coherent-to-Diffuse Power Ratio

The CDR was used for dereverberation of speech signals by spectral subtraction [16]. In the context of DOA estimation, it has been applied to weighting of narrowband DOA estimates [15].

The auto Power Spectral Densities (PSDs) $\hat{\Phi}_{x_m x_m}(l,k)$, $\hat{\Phi}_{x_R x_R}(l,k)$ and the cross PSDs $\hat{\Phi}_{x_m x_R}(l,k)$ can be estimated by recursive averaging of the instantaneous signal power

$$\begin{aligned}\hat{\Phi}_{x_i x_j}(l,k) = \lambda\hat{\Phi}_{x_i x_j}(l-1,k) + \\ \cdots + (1-\lambda)x_i(l,k)x_j^*(l,k),\end{aligned} \tag{9}$$

where $i,j \in \{m,R\}$. Exploiting (9), the microphone coherence between microphone m and the reference microphone can be estimated by

$$\hat{\Gamma}_x^{mR}(l,k) = \frac{\hat{\Phi}_{x_m x_R}(l,k)}{\sqrt{\hat{\Phi}_{x_m x_m}(l,k)\hat{\Phi}_{x_R x_R}(l,k)}}. \tag{10}$$

One building block to classify CDR estimators is their coherence models for the direct and reverberant acoustical path. In this contribution, we model the reverberant sound field to be diffuse, i.e., the model for the coherence is given by [17]

$$\Gamma_w^{mR}(f_k) = \frac{\sin\left(2\pi f_k d_{\mathrm{mic}}^{mR}/c\right)}{2\pi f_k d_{\mathrm{mic}}^{mR}/c}, \tag{11}$$

where d_{mic}^{mR} denotes the distance between microphone m and reference microphone R and f_k the frequency corresponding to frequency band k. In this contribution, we use the DOA-independent CDR estimator (8). Therefore, no model for the coherence of the direct path and the early reflections is needed. With regard to the definition of a discarding threshold, the CDR is converted into the diffuseness DIFF sensed by microphone m and reference microphone R. The diffuseness is defined by

$$\mathrm{DIFF}_m(l,k) = \frac{1}{\mathrm{CDR}_{mR}(l,k)+1}, \tag{12}$$

which yields values between zero and one, where zero means perfectly coherent noise and one means spatially white noise. Therefore, a low diffuseness corresponds to STFT bins, which contain less reverberation.

4.3. Speech Presence Probability

The SPP has been widely used in noise power estimation [18]. For the following, we assume two hypotheses: speech absence $\mathcal{H}_0$ and speech presence $\mathcal{H}_1$. The posterior density of the hypothesis $\mathcal{H}_1$ given $x_m(l,k)$ is defined as

$$\begin{aligned}P(\mathcal{H}_1|x_m(l,k) = \Bigg(1 + (1+\xi_{\mathrm{opt}})\ldots \\ \ldots\exp\left(-\frac{|x_m(l,k)|^2}{\hat{\sigma}_N^2}\frac{\xi_{\mathrm{opt}}}{1+\xi_{\mathrm{opt}}}\right)\Bigg)^{-1},\end{aligned} \tag{13}$$

where ξ_{opt} denotes the a priori Signal-to-Noise Ratio (SNR). ξ_{opt} is set to a fixed value to guarantee a specified performance of the SPP estimator. Here, we choose $10\log_{10}(\xi_{\mathrm{opt}}) = 15\mathrm{dB}$ as in [18]. The noise power $\hat{\sigma}_N^2$ is assumed to be constant and is estimated from a time frame containing only noise [18]. The beginning and duration of this frame is assumed to be known a priori. The SPP estimate is recursively averaged to increase its robustness

$$\overline{\mathcal{P}}(l,k) = \beta\,\overline{\mathcal{P}}\,(l-1,k) + (1-\beta)\;P\left(\mathcal{H}_1|x_m(l,k)\right).$$

Here, we choose $\beta = 0.9$ as in [18] for our experiments.

5. STFT Bin Selection

To select beneficial STFT bins with respect to DOA estimation or localization, we evaluate the impact of the parameters, proposed in Section 4. Thus, the SPP and power are computed for each microphone, whereas CDR is evaluated for each microphone pair consisting of reference microphone R and another microphone. Afterwards, the criteria are evaluated by checking the following inequalities for different selection strategies. The criterion for power-based selection is given as

$$P_m(l,k) > \gamma_P, \tag{14}$$

the criterion for CDR-based selection as

$$\mathrm{DIFF}_m(l,k) < \gamma_{\mathrm{DIFF}} \tag{15}$$

and for SPP-based selection as

$$\mathrm{SPP}_m(l,k) > \gamma_{\mathrm{SPP}}. \tag{16}$$

Here, γ_P, γ_{DIFF} and γ_{SPP} denote suitably chosen thresholds. We define a set of selected STFT bins $\mathcal{A}_P$, $\mathcal{A}_{\mathrm{DIFF}}$ and $\mathcal{A}_{\mathrm{SPP}}$ for each criterion, which contains STFT bin (l,k), if (14), (15) or (16) are fulfilled, respectively. The intersection of these sets

$$\mathcal{A}_{\mathrm{COMB}} = \mathcal{A}_P \cap \mathcal{A}_{\mathrm{DIFF}} \cap \mathcal{A}_{\mathrm{SPP}}. \tag{17}$$

describes the set of the bins selected by a logical 'and' combination of all criteria. Note that the selection

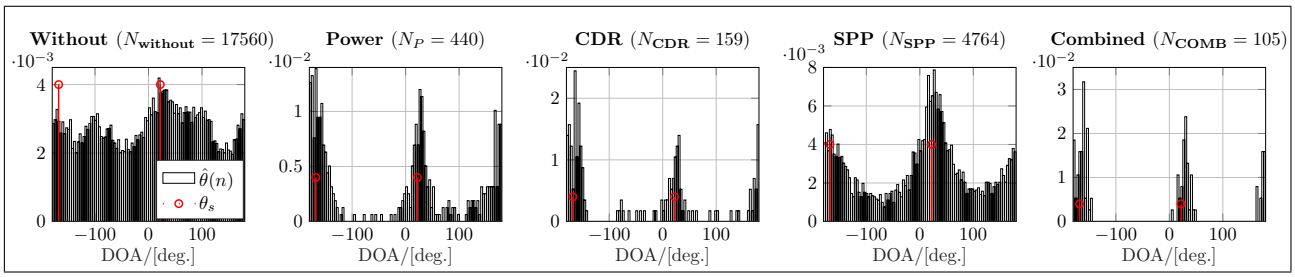

Figure 1. Examplary histograms of STFT bin-wise DOA estimates without selection, with power-based selection, with CDR-based selection, with SPP-based selection, and with the combination of all criteria. The histograms have been normalized. The number in the brackets in the titles of the figures denote the number of bins which are left after the selection process and the red lines denote the true source positions θ_s.

process and the evaluation of the criteria is computationally cheap, as can be seen by the description of the criteria.

In the following, we replace the dependency of the observations on the time-frequency bin (l, k) by a dependency on the data index n as the dependency on the time-frequency indices is not longer meaningful due to the discarding process. The number of selected STFT bin-wise estimates, i.e., the cardinality of the sets $\mathcal{A}_P$, $\mathcal{A}_{\text{DIFF}}$, $\mathcal{A}_{\text{SPP}}$ and $\mathcal{A}_{\text{COMB}}$ is denoted by N_P, N_{DIFF}, N_{SPP} and N_{COMB}, respectively. Exemplary histograms for illustration of the selection process based on the described criteria and their combination are depicted in Fig. 1.

6. Localization of Acoustic Sources Exploiting Sparsity of Speech Signal Mixtures

In the following subsections, we are going to describe algorithms for DOA estimation and localization which might be an application for the discussed selection strategies. The previously described selection strategies can be motivated as preprocessing step for, e.g., DOA estimation or tracking. In the following subsections, we are going to describe two exemplary algorithms, which have been used in similar forms in literature: Clustering with a wrapped Gaussian Mixture Model (GMM) has been used in [6] for blind source separation and source counting of speech sources. Clustering with a two-dimensional GMM has been used in [12] for localization and in [13] for tracking of multiple speakers in an enclosure.

6.1. Wrapped GMM Clustering

For the following basic DOA estimation approach, we model the observed bin-wise DOA estimates by a GMM, whose parameters have to be estimated. However, when the distribution of the bin-wise DOA estimates is modeled with a GMM, it can be observed that the distribution wraps if a mean of a Gaussian component is close to 0 or 2π and the respective Gaussian component becomes bimodal. To solve this problem, [19] proposed a wrapped phase GMM modeling

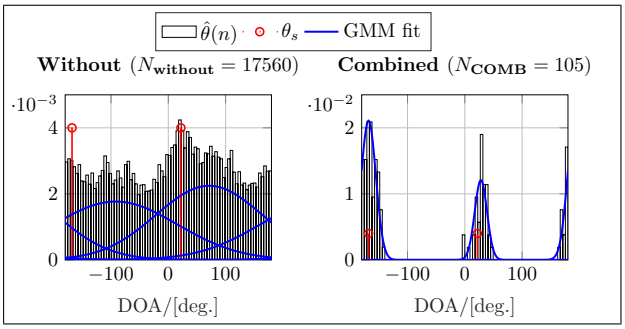

Figure 2. Fitting result of a wrapped GMM with and without STFT bin selection by the combination of power-, CDR- and SPP-based selection.

these effects

$$p(\boldsymbol{\theta}) = \prod_{n=1}^{N} \sum_{s=1}^{S} \alpha_s \sum_{\nu=-\infty}^{\infty} \mathcal{N}\left(\theta(n) + 2\pi\nu; \mu_s, \sigma_s^2\right).$$

Hereby, $\boldsymbol{\theta}$ denotes the vector of all STFT bin-wise DOA estimates, α_s the class probability, μ_s the mean, and σ_s^2 the variance of Gaussian component s. The parameters of the GMM can be estimated exploiting the EM algorithm [20]. The mean of Gaussian component s yields a Maximum Likelihood estimate of the DOA of source s after a maximum number of iterations or after convergence of the algorithm, i.e., $\theta_s = \mu_s$. An example for fitting of a wrapped GMM can be found in Fig. 2.

6.2. Narrowband Localization

In the following, we consider an ASN consisting of Q sensor nodes covering the area of interest for localization of multiple simultaneously active speakers. Hereby, the source locations can be estimated by combining the observations of the distributed microphones, e.g., triangulation. Due to the assumption of W-disjoint orthogonality, i.e., only one source is assumed to be active in a specific STFT bin, the problem of ghost sources, i.e., wrong combinations of DOA estimates within one bin are disregarded.

In the following, all positions and DOAs are expressed w.r.t. a common room coordinate system. The

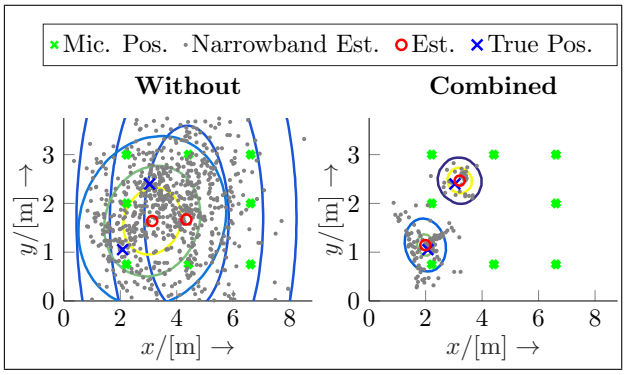

Figure 3. Clustering result of the narrowband localization with the 2D GMM without and with STFT bin selection. The lines denote positions with equal Mahalanobis distance to the mean of a Gaussian component.

vector corresponding to array q pointing into the direction of the source active in data point n can be expressed as $\mathbf{q}_q(n) = \mathbf{p}(n) - \mathbf{r}_q$ with the coordinates of the array's reference point $\mathbf{r}_q = \left[r_{x,q}, r_{y,q}\right]^{\mathrm{T}}$. The position of source $\mathbf{p}(n)$ dominant in data point n can be described using (7) by the following matrix-valued expression $\mathbf{A}(n)\mathbf{p}(n) = \mathbf{b}(n)$. Hereby, the matrix $\mathbf{A}(n)$ contains the STFT bin-wise DOA estimates $\hat{\theta}_q$ of data index n of the Q microphone arrays. The q-th row of $\mathbf{A}(n)$ is defined as

$$[\mathbf{A}(n)]_{q\in\{1,\ldots,Q\}} = \left[\sin\hat{\theta}_q(n) \quad -\cos\hat{\theta}_q(n)\right]$$

and the q-th element of the vector $\mathbf{b}(n)$ as

$$[\mathbf{b}(n)]_{q\in\{1,\ldots,Q\}} = r_{x,q}\sin\hat{\theta}_q(n) - r_{y,q}\cos\hat{\theta}_q(n).$$

The position of the source active in data point n can be estimated by a Least Squares (LS) approach $\hat{\mathbf{p}}(n) = \mathbf{A}^{+}(n)\mathbf{b}(n)$.

The obtained narrowband position estimates $\hat{\mathbf{p}}(n)$ have to be clustered similarly as for the DOA estimation in order to obtain reliable position estimates by using, e.g., a two-dimensional GMM

$$p\left(\mathbf{p}(1),\ldots,\mathbf{p}(N)\right) = \prod_{n=1}^{N}\sum_{s=1}^{S}\alpha_s\,\mathcal{N}\left(\mathbf{p}(n);\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s\right)$$

and the EM algorithm. The class probability of cluster s is denoted by α_s, the mean vector by $\boldsymbol{\mu}_s$ and the covariance matrix by $\boldsymbol{\Sigma}_s$. The localization result is represented by the estimated mean vectors of the Gaussian components, i.e., $\hat{\mathbf{p}}_s = \boldsymbol{\mu}_s$.

It is desirable to have narrowband position estimates scattered densely around the true source positions in order to get good clustering results and thus reliable estimates of the positions. Preprocessing the data by bin selection aims at enhancing the cluster structure as shown in the following sections.

An exemplary result for the case of narrowband localization and clustering is shown in Fig. 3.

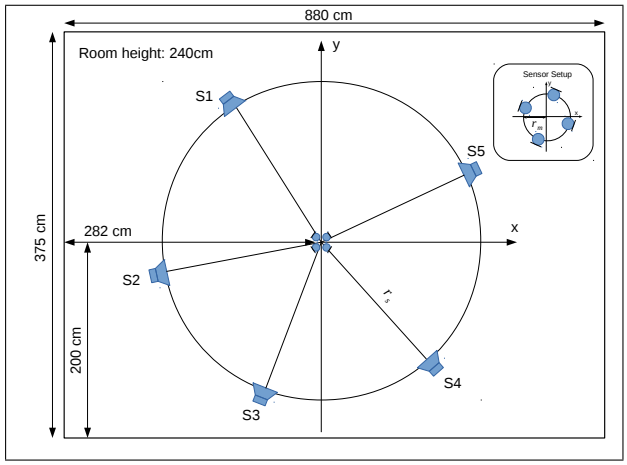

Figure 4. Experimental setup for DOA estimation. A circular array consisting of four microphones is located at $(2.82\,\mathrm{m}, 2\,\mathrm{m}, 1.2\,\mathrm{m})$ in an enclosure of dimensions $8.8\,\mathrm{m} \times 3.75\,\mathrm{m} \times 2.4\,\mathrm{m}$. The sources are located at radius $r_s = 0.8\,\mathrm{m}$ from the arrays reference point.

7. Experiments

The following subsections describe the experimental setup, results of bin selection strategies as preprocessing for the introduced DOA and localization algorithms.

7.1. Setup

Experiments are conducted in a simulated enclosure of dimensions $8.8\,\mathrm{m} \times 3.75\,\mathrm{m} \times 2.4\,\mathrm{m}$ by simulating Room Impulse Responses (RIRs) with the image-source method [21] and the RIR generator [22]. The sensor arrays and the sources are placed on a height of 1.2 m. Circular sensor arrays consisting of four microphones with radius 1.7 cm are employed. The sampling frequency is set to 16 kHz. The STFT of the observed microphone signals has been computed using a Hamming window, a Fast Fourier Transform (FFT) length of 1024 and a frame shift of 256. Furthermore, the frequency range for DOA estimation is limited to $[0.7\,\mathrm{kHz}, 1\,\mathrm{kHz}]$ for our experiments, which is common practice in DOA estimation and localization algorithms relying on the sparsity of speech signal spectra [10]. This is motivated by the fact that this frequency interval contains most of the acoustical energy emitted from a human speaker.

7.2. DOA Estimation

The sources are located at a distance of 0.8 m from the center of the microphone array for the DOA estimation task. The corresponding experimental setup is depicted in Fig. 4. The signal duration of the source signals has been chosen to be about 14 sec including an initial interval of 5 sec containing only noise.

In the following, we describe and discuss experiments which evaluate the bin selection performance of the considered strategies in general and investigate

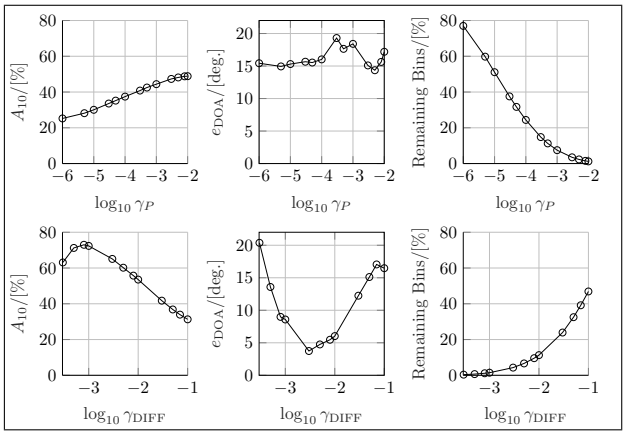

Figure 5. Performance of the different selection strategies in dependence of varying thresholds for 30 dB SNR and $T_{60} = 0.3\,\mathrm{sec}$.

the performance gain of DOA estimation compared to no bin discarding.

A histogram of STFT bin-wise DOA estimates which are clustered densely around the true source positions is desired. Therefore, a measure to quantify this cluster characteristic is introduced by defining

$$\mathcal{A}_{10} = \left\{ \hat{\theta}(n) \,\middle|\, \exists \theta_s : \mathrm{dist}_{\mathrm{wrap}}\left(\hat{\theta}(n), \theta_s\right) \leq 10^\circ \right\}.$$

which captures the set of STFT bin-wise DOA estimates, which are in an interval of $\pm 10^\circ$ around the true DOA of a source. Hereby, $\mathrm{dist}_{\mathrm{wrap}}(\cdot,\cdot)$ denotes the absolute difference between two DOAs which takes the wrapping at $\pm 180^\circ$ into account. To quantify the benefit of bin selection, the ratio of the number of elements of $\mathcal{A}_{10}$, i.e., its cardinality, and the total number of selected bins is considered. The result is averaged over J_{DOA} different realizations of the experiment with randomly chosen source DOAs

$$A_{10} = \frac{100\%}{J_{\mathrm{DOA}}} \sum_{j=1}^{J_{\mathrm{DOA}}} \frac{|\mathcal{A}_{10}^j|}{N_j}, \tag{18}$$

where N_j denotes the total number of estimates after selection in the j-th realization and $|\mathcal{A}_{10}^j|$ the number of bins in an interval $\pm 10^\circ$ around a true DOA. This yields the averaged relative amount of STFT bin-wise DOA estimates in proximity to the true DOAs w.r.t. all selected STFT bins. A_{10} takes on values in percentage, with high numbers characterizing a histogram with most STFT bin-wise estimates in proximity to a true DOA.

In addition to the A_{10} measure, the benefit of bin selection for DOA estimation is evaluated by application of the algorithm described in Sec. 6.1. To measure the performance of the algorithm, the absolute error of the DOA estimates w.r.t. the true DOAs is calculated. To this end, association between true DOAs $\theta_{s,j}$ and estimated DOAs $\hat{\theta}_{s,j}$ has to be done, which is accomplished by successively assigning the pairs of

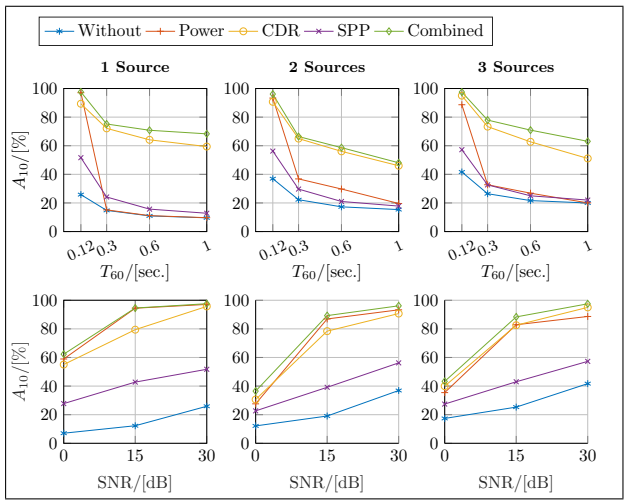

Figure 6. Relative amount of STFT bin-wise estimates in 10° intervals around the true source DOAs w.r.t. all estimates in dependence of the reverberation time T_{60} and the SNR.

estimated and true DOAs with smallest distance to each other. The overall performance of the DOA estimation is expressed by the absolute averaged error

$$e_{\mathrm{DOA}} = \frac{1}{J \cdot S} \sum_{s=1}^{S} \sum_{j=1}^{J_{\mathrm{DOA}}} \mathrm{dist}_{\mathrm{wrap}}\left(\hat{\theta}_{s,j}, \theta_{s,j}\right), \tag{19}$$

with the true DOA $\theta_{s,j}$ and estimated bin-wise DOA of realization j $\hat{\theta}_{s,j}$.

$J_{\mathrm{DOA}} = 100$ realizations of an experiment with 1, 2 or 3 sources with randomly chosen DOAs with minimum angular separation of 70° have been conducted. To investigate the influence of the choice of the thresholds, the A_{10} measure and the averaged DOA estimation error e_{DOA} are plotted together with the amount of remaining bins in Fig. 5 for the power- and CDR-based selection strategy for SNR $= 30\,\mathrm{dB}$ and $T_{60} = 0.3\,\mathrm{sec}$. The SPP-based method is not considered here as the threshold has to be chosen very tight ($\gamma_{\mathrm{SPP}} = 0.99$) in order to obtain satisfying results and does not allow for variation of it. By evaluating these figures it can be seen that for the CDR-based selection a clear minimum of the localization error e_{DOA} and a clear maximum of A_{10} is obtained, whereas the power-based selection has no distinct minimum of e_{DOA} nor a maximum of A_{10}. It can be concluded, that the CDR-based selection is beneficial for DOA estimation w.r.t. localization error as well as computational complexity and the power-based selection gives also a reduction in computational complexity but only marginal improvement of localization error in this exemplary scenario.

For power-based selection, thresholds $\gamma_P \in [10^{-7}, 0.04]$, for CDR-based selection thresholds $\gamma_{\mathrm{CDR}} \in [8 \cdot 10^{-4}, 0.1]$ and averaging factors $\lambda \in [0.1, 0.5]$, and for SPP-based selection the threshold $\gamma_{\mathrm{SPP}} = 0.99$ have been chosen according to

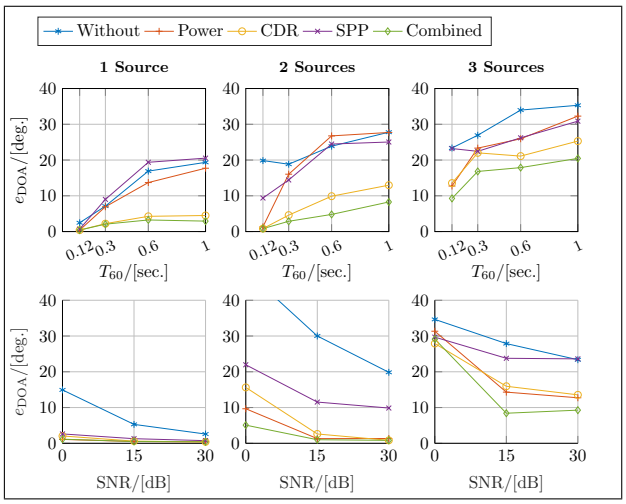

Figure 7. Absolute averaged localization error in degrees in dependence of reverberation time T_{60} and the SNR.

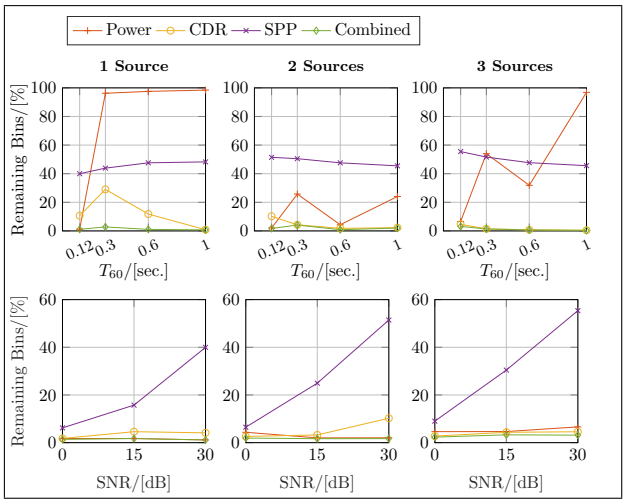

Figure 8. Averaged amount of bins left after selection in dependence of reverberation time T_{60} and SNR.

the best performance in each scenario w.r.t. (19). Two kind of experiments have been conducted. On the one hand, the reverberation time $T_{60} = 0.12\,\mathrm{sec}, 0.3\,\mathrm{sec}, 0.6\,\mathrm{sec}, 1\,\mathrm{sec}$ has been varied while keeping the SNR $= 30\,\mathrm{dB}$ fixed. On the other hand, the $T_{60} = 0.12\,\mathrm{sec}$ has been kept fixed while the SNR $= 0\,\mathrm{dB}, 15\,\mathrm{dB}, 30\,\mathrm{dB}$ has been varied. The results for both groups of experiments are shown in Fig. 6 for A_{10} and in Fig. 7 for e_{DOA}. Additionally, the reduction of data points by discarding is depicted in Fig. 8.

CDR-based selection is superior to all other methods w.r.t. A_{10} for high reverberation times, see Fig. 6. This can be explained by interpreting the CDR as a measure for the amount of reverberation in an STFT bin. The combination with other methods only slightly improves the results. For low reverberation time and varying SNRs, the CDR- and power-based selection show comparable results w.r.t. A_{10}. Similar results are obtained w.r.t. DOA estimation performance measured by e_{DOA}, see Fig. 7. The SPP-based

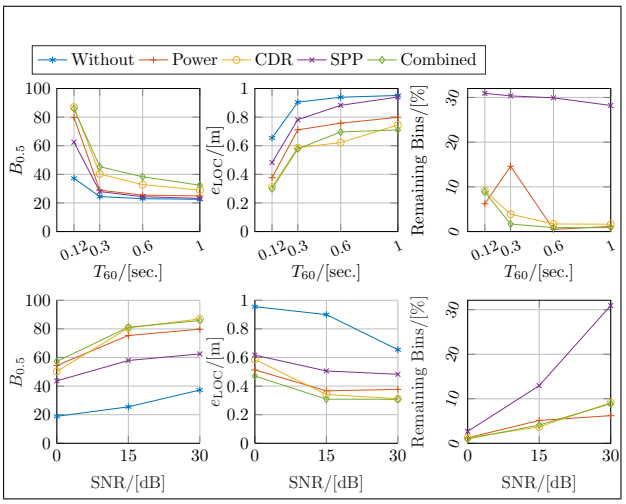

Figure 9. Localization performance for two sources with random positions in dependence of different measures for different reverberation times and SNRs.

method shows comparable results w.r.t. the power-based selection for high T_{60} and is clearly worse for low T_{60} and varying SNR values w.r.t. A_{10} as well as w.r.t. e_{DOA}. The number of bins remaining after selection is comparable for power-based and CDR-based selection w.r.t. the SNR, but only comparable for low T_{60}, see Fig. 8. The SPP-based method yields always the largest number of remaining STFT bins.

In summary, CDR-based selection demonstrates its benefits in scenarios with varying reverberation time as well as SNR and discards in all cases the largest amount of STFT bin-wise estimates (about 1% of the bin-wise estimates is left after selection). Therefore, this selection method yields the highest savings for computational power as well as the highest improvement in DOA estimation performance in varying acoustical scenarios.

7.3. Localization

For comparing the proposed strategies w.r.t. localization, eight arrays configured as shown in Fig. 4 are distributed in the room and narrowband localization and clustering as described in Sec. 6.2 have been used for localization of two speech sources. To calculate the localization error, the association between true and estimated positions has been done by first assigning the pair of estimate and true position with the smallest distance to each other. The absolute localization error

$$e_{\mathrm{LOC}} = \frac{1}{J \cdot S} \sum_{s=1}^{S} \sum_{j=1}^{J} \|\mathbf{p}_{s,j} - \hat{\mathbf{p}}_{s,j}\|_2 \qquad (20)$$

averaged over $J_{\mathrm{LOC}} = 20$ realizations of the experiment has been computed to quantify the results. Similar to the set $\mathcal{A}_{10}$, the set $\mathcal{B}_{0.5}$ of narrowband position estimates with a distance smaller than 0.5 m to a true source position is defined as

$$\mathcal{B}_{0.5} = \left\{ \hat{\mathbf{p}}(n) \,\middle|\, \exists \mathbf{p}_s : \left\| \hat{\mathbf{p}}(n) - \mathbf{p}_s \right\|_2 \leq 0.5\,\mathrm{m} \right\}.$$

The ratio of the number of position estimates in this set w.r.t. all position estimates averaged over the J_{LOC} random source positions is computed as

$$B_{0.5} = \frac{100\%}{J_{\mathrm{LOC}}} \sum_{j=1}^{J_{\mathrm{LOC}}} \frac{|\mathcal{B}_{0.5}^{j}|}{N_j}, \tag{21}$$

where N_j denotes the selected position estimates of the j-th realization. The results of these experiments are shown in Fig. 9 in dependence of the SNR and T_{60}. A similar trend as for the DOA estimation can be deduced: CDR-based selection works best for high reverberation times and is comparable with power-based selection for varying SNRs. SPP-based selection performs worse than the other selection strategies for all cases.

8. Conclusions

Different selection strategies for STFT bin-wise DOA estimates, namely power-, CDR-, SPP-based selection and the combination of these three strategies have been discussed in this contribution. It has been shown that STFT bin selection increases the performance of DOA estimation and localization while simultaneously reducing the computational complexity of the underlying algorithms at the same time. Under the investigated selection methods, CDR-based selection achieved best results for a broad range of acoustical scenarios.

Acknowledgement

This work was supported by DFG under contract no <Ke890/10-1> within the Research Unit FOR2457 "Acoustic Sensor Networks".

References

[1] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[2] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *IEEE ICASSP*, Munich, Germany, Oct. 1997, p. 4.

[3] J. H. DiBiase, "A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays," PHD Thesis, Brown University, Providence, Rhode Island, May 2000.

[4] A. Lombard et al., "TDOA Estimation for Multiple Sound Sources in Noisy and Reverberant Environments Using Broadband Independent Component Analysis," *IEEE TASLP*, vol. 19, no. 6, pp. 1490–1503, Aug. 2011.

[5] M. Mandel, R. Weiss, and D. Ellis, "Model-Based Expectation-Maximization Source Separation and Localization," *IEEE TASLP*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[6] S. Araki et al., "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *IEEE ICASSP*, Apr. 2009, pp. 33–36.

[7] S. Rickard, "Sparse Sources are Separated Sources," in *EUSPICO*, Florence, Italy, Sep. 2006.

[8] S. Araki et al., "DOA Estimation for Multiple Sparse Sources with Arbitrarily Arranged Multiple Sensors," *J. of Signal Process. Syst.*, vol. 63, no. 3, pp. 265–275, Jun. 2011.

[9] O. Schwartz et al., "DOA Estimation in Noisy Environment with Unknown Noise Power using the EM Algorithm," in *HSCMA*, San Francisco, CA, USA, Mar. 2017.

[10] O. Schwartz and S. Gannot, "Speaker Tracking Using Recursive EM Algorithms," *IEEE/ACM TASLP*, vol. 22, no. 2, pp. 392–402, Feb. 2014.

[11] Y. Dorfan and S. Gannot, "Tree-Based Recursive Expectation-Maximization Algorithm for Localization of Acoustic Sources," *IEEE/ACM TASLP*, vol. 23, no. 10, pp. 1692–1703, Oct. 2015.

[12] A. Alexandridis and A. Mouchtaris, "Multiple sound source location estimation and counting in a wireless acoustic sensor network," in *IEEE WASPAA*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[13] M. Taseska, G. Lamani, and E. A. P. Habets, "Online Clustering of Narrowband Position Estimates with Application to Multi-speaker Detection and Tracking," in *Advances in Machine Learning and Signal Process.* Cham: Springer International Publishing, 2016, vol. 387, pp. 59–69.

[14] A. Alexandridis and A. Mouchtaris, "Improving narrowband DOA estimation of sound sources using the complex Watson distribution," in *EUSIPCO*, Budapest, Hungary, Aug. 2016, pp. 1468–1472.

[15] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *IEEE WASPAA*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[16] A. Schwarz and W. Kellermann, "Coherent-to-Diffuse Power Ratio Estimation for Dereverberation," *IEEE/ACM TASLP*, vol. 23, no. 6, pp. 1006–1018, Jun. 2015.

[17] H. Kuttruff, *Room acoustics*, 5th ed. London & New York: Spon Press/Taylor & Francis, 2009.

[18] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *IEEE WASPAA*, New Paltz, NY, USA, Oct. 2011, pp. 145–148.

[19] P. Smaragdis and P. Boufounos, "Learning source trajectories using wrapped-phase hidden Markov models," in *IEEE WASPAA*, New Paltz, NY, Oct. 2005, pp. 114–117.

[20] C. M. Bishop, *Pattern recognition and machine learning*, ser. Information science and statistics. New York: Springer, 2006.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[22] E. A. P. Habets, "Room Impulse Response Generator," International Audio Laboratories, Tech. Rep., Sep. 2010.