Bohlender *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2022) 2022:16
https://doi.org/10.1186/s13636-022-00246-7

EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY Open Access

DOA-guided source separation with direction-based initialization and time annotations using complex angular central Gaussian mixture models

Alexander Bohlender*, Lucas Van Severen, Jonathan Sterckx and Nilesh Madhu

Abstract

By means of spatial clustering and time-frequency masking, a mixture of multiple speakers and noise can be separated into the underlying signal components. The parameters of a model, such as a complex angular central Gaussian mixture model (cACGMM), can be determined based on the given signal mixture itself. Then, no misfit between training and testing conditions arises, as opposed to approaches that require labeled datasets to be trained. Whereas the separation can be performed in a completely unsupervised way, it may be beneficial to take advantage of a priori knowledge. The parameter estimation is sensitive to the initialization, and it is necessary to address the frequency permutation problem. In this paper, we therefore consider three techniques to overcome these limitations using direction of arrival (DOA) estimates. First, we propose an initialization with simple DOA-based masks. Secondly, we derive speaker specific time annotations from the same masks in order to constrain the cACGMM. Thirdly, we employ an approach where the mixture components are specific to each DOA instead of each speaker. We conduct experiments with sudden DOA changes, as well as a gradually moving speaker. The results demonstrate that particularly the DOA-based initialization is effective to overcome both of the described limitations. In this case, even methods based on normally unavailable oracle information are not observed to be more beneficial to the permutation resolution or the initialization. Lastly, we also show that the proposed DOA-guided source separation works quite robustly in the presence of adverse conditions and realistic DOA estimation errors.

Keywords: Guided source separation, Spatial clustering, Direction of arrival, Time-frequency masks

1 Introduction

The extraction of clean speech from a mixture with unwanted components, such as background noise, is an important task in the context of applications like speech enhancement for human-to-human communication and automatic speech recognition. If the mixture contains multiple concurrently active speakers, however, algorithms that rely solely on spectro-temporal information may fail due to the similarity of the underlying source signal characteristics. In this case, spatial information, which is available when a microphone array is used, may be exploited to distinguish between the signal components.

Because speech is characterized by a high degree of sparsity in the short-time Fourier transform (STFT) domain, an effective separation can be realized with the help of masks that identify the dominant signal component in each time-frequency (TF) bin [1]. Supervised learning approaches, particularly based on deep neural networks (DNNs), are commonly employed to obtain such TF masks. For example, permutation invariant training (PIT) [2] can be incorporated to enable the separation of multiple talkers in this case. Other approaches distinguish

*Correspondence: alexander.bohlender@ugent.be
IDLab, Department of Electronics and Information Systems, Ghent University - imec, Ghent, Belgium

the sources based on their directions of arrival (DOAs), either by estimating these along with the corresponding masks [3, 4], or by assuming DOA estimates to be available in advance [5, 6]. Rather than first computing TF masks, [7] proposes a beamforming-based speaker separation with an implicitly performed broadband DOA estimation. Deep clustering, which was originally proposed for single-channel mixtures [8] but has since been extended to microphone arrays as well [9], represents another class of approaches. A DNN is trained to return high-dimensional embeddings before the application of a clustering algorithm such as k-means. Deep attractor networks [10] are an extension of deep clustering, where the embeddings are optimized by minimizing the reconstruction error, thereby making the training end-to-end.

The main drawback of these supervised methods is the need for a (large) set of labeled training data, i.e., noisy mixtures and the corresponding clean source signals. If the clean signals are not available, or if there is a mismatch between training and testing conditions, the resulting performance may be suboptimal. In contrast, (spatial) clustering approaches that directly model the given signals (or features extracted therefrom) by a mixture of components that each follow a different distribution, do not require such representative training sets. The parameters of this mixture model are determined e. g., using the expectation-maximization (EM) algorithm, from which posteriors that serve as TF masks can then be extracted. In this work, specifically, we describe the normalized vector of microphone signals in the STFT domain with a complex angular central Gaussian mixture model (cACGMM), as originally proposed in [11]. The normalization effectively discards the single-channel (magnitude) information, so that only the inter-channel differences, which represent the spatial information, are retained.

Two main problems are characteristic of the spatial clustering approach, regardless of the specific choice of the mixture model. Typically, the separation is performed independently for each frequency. This leads to the well-known frequency permutation problem, where the same component index may correspond to a different speaker for every frequency bin. Secondly, the iterative model parameter estimation is sensitive to the initialization.

To address the frequency permutation problem, cross-frequency information may be incorporated. This can be done by resolving the permutation ambiguity in the end, or within the parameter estimation itself. Particularly the method proposed in [12], which is based on the correlation of the posteriors between frequency bins, is commonly employed to perform a manual permutation alignment. One way to introduce a dependence between the optimization problems otherwise solved for each frequency independently is the use of time-variant but frequency independent mixture weights, in order to enforce a consistent permutation [13]. However, this link between different frequencies might not be sufficient to prevent the occurrence of permutation errors. For this reason, a more advanced approach is adopted in [14], where the DOAs are integrated as hidden variables into the model for the spatial covariance matrices of the employed complex Gaussian mixture model. As all parameters are estimated jointly, no prior knowledge of the source locations is required. The exploitation of prior knowledge can, however, be an effective alternative when the requirement of a completely *blind* source separation is relaxed. The source separation (GSS) approach proposed in [15], for example, incorporates time annotations into the mixture model that indicate when each source is active.

On the other hand, it is reported in [16, 17] that the availability of suitable initial masks alone can be sufficient to mitigate the need for additional measures to address the permutation problem. These can be used to initialize the EM algorithm accordingly (*weak integration*, e. g., [16]), or by incorporating them into the model in the form of fixed mixture weights (*tight integration*, e. g., [17, 18]). A similar notion is adopted in [19], where embeddings acquired by means of deep clustering are integrated into the model instead of initial masks.

The fact that initial masks can be used to address both shortcomings, frequency permutation problem and sensitivity to initialization, makes them a valuable tool. A wide variety of techniques have been proposed in this context. For example, a scheme to initialize the mixing matrix of a blind source separation problem was proposed in [20]. More recently, particularly the use of spatial clustering in conjunction with DNN-based methods for initial mask estimation has received a lot of attention. TF masks for the integration into the mixture model are obtained with a bidirectional long short-term memory (LSTM) network in [18]. Both [16] and [21] take advantage of spatial clustering methods to train neural networks in an unsupervised way, as well as to compute the final masks in the end. In [17], a convolutional neural network (CNN) with utterance-level PIT is employed prior to the mixture model-based mask estimation. For all of these approaches alike, it is reported that the ultimate spatial clustering step improves the performance compared to using the output of the respective DNNs directly.

Thus, although spatial clustering can be used in a completely unsupervised fashion, we note that the incorporation of a priori knowledge can improve the speaker separation significantly. In this work, we focus on the GSS approach [15], which takes advantage of time annotations to address the permutation problem. Whereas ground truth annotations are already available for the CHiME-5 dataset [22], to which the GSS was originally applied, this is not the case in general.

In this paper, we therefore propose to use broadband DOA estimates to guide the cACGMM-based source separation. We generically refer to such approaches as DOA-GSS. The usefulness of DOA information in the context of otherwise blind source separation algorithms has previously been demonstrated, e. g., for independent component analysis in [23], where an initial unmixing matrix is obtained by means of null beamforming. For the GSS approach, in particular, the advantage of using DOA estimates, instead of estimating time annotations directly, is that they are helpful in the acquisition of initial masks as well.

Specifically, the aim of this work is to determine how DOA knowledge can be exploited most effectively. For this purpose, we consider three different methods: (i) the initialization of the EM algorithm with DOA-based masks, (ii) the inclusion of time annotations derived from the same initial masks, and (iii) the use of DOA-based (rather than speaker-based) mixture components to reflect that the cACGMM models *spatial* signal characteristics. In the evaluation, we compare different combinations of these techniques. By considering oracle initialization, oracle time annotations, and oracle permutation alignment as baselines, we show that the proposed initial masks, despite being relatively simple, are sufficient to avoid the frequency permutation problem, and to cope with the inherent sensitivity of the approach to the initialization. This suggests that it may be unnecessary to resort to one of the previously proposed more elaborate schemes, such as the estimation of initial masks using a DNN. Only for the case where the parameter estimation is performed on very short signal segments, the performance is observed to degrade significantly due to the lack of sufficient data to improve upon the initialization.

In Section 2, we first introduce the source separation problem, and outline how it can be addressed with the help of TF masks. The GSS, which the proposed approach is an extension of, is reviewed in Section 3. Subsequently, Section 4 describes the DOA-GSS in detail, including the derivation of DOA-based initial masks, and the extraction of speaker or direction specific time annotations. Based on the experiments in Section 5, we then evaluate which setup is the best to make use of the DOA estimates. Section 6 concludes the paper.

2 Problem statement

The vector $\mathbf{Y}(f,t) = \left[Y_1(f,t), \ldots, Y_N(f,t)\right]^T$ contains the STFT domain signals captured by an array of N microphones. The length of the discrete Fourier transform (DFT) and the number of frames are denoted by F and T, respectively, so that the frequency index is $f \in \{0, \ldots, F-1\}$ and the frame index is $t \in \{0, \ldots, T-1\}$. We assume that the microphone signals are an additive mixture

$$\mathbf{Y}(f,t) = \sum_j \mathbf{S}_j(f,t) \; + \; \mathbf{V}(f,t), \tag{1}$$

which is composed of the contributions $\mathbf{S}_j(f,t)$ of sound sources $j \in \{1, \ldots, J\}$ and noise $\mathbf{V}(f,t)$. The focus of this work is on speech, which implies that each of the J sources is one talker. Further, the microphone signal contribution of the j-th source is composed of a direct-path component $\mathbf{S}'_j(f,t)$ and a reverberation component $\mathbf{S}''_j(f,t)$ i.e.,

$$\begin{aligned} \mathbf{S}_j(f,t) &= \mathbf{S}'_j(f,t) &&+ \mathbf{S}''_j(f,t) \\ &= \mathbf{A}_j(f,t)\, S'_j(f,t) &&+ \mathbf{S}''_j(f,t), \end{aligned} \tag{2}$$

where $\mathbf{A}_j(f,t)$ is the direct-path propagation vector. Our aim is to extract the anechoic (dry) source signals at the reference microphone

$$S'_j(f,t) = \mathbf{u}_{n_\mathrm{r}}^T \mathbf{S}'_j(f,t) \tag{3}$$

for all j from the microphone signal mixture. In Eq. 3, $\mathbf{u}_{n_\mathrm{r}}$ is a unit vector where only the element corresponding to the reference microphone is 1, and all other entries are 0. In the following, the reference is arbitrarily set to $n_\mathrm{r} = 1$. For a source with DOA vector

$$\mathbf{n} = \left[\, \cos(\varphi)\cos(\vartheta) \;\; \sin(\varphi)\cos(\vartheta) \;\; \sin(\vartheta) \,\right]^T, \tag{4}$$

which is located in the far field, the propagation vector is given by

$$\mathbf{A}(f) = \left[\, 1 \;\; e^{\jmath\kappa(f)\mathbf{r}_{21}^T\mathbf{n}} \;\; \cdots \;\; e^{\jmath\kappa(f)\mathbf{r}_{N1}^T\mathbf{n}} \,\right]^T. \tag{5}$$

In the above, $\kappa(f) = \frac{2\pi}{c} f_s \frac{f}{F}$ is the wavenumber, c is the speed of sound, f_s is the sampling rate, and $\mathbf{r}_{mn} = \mathbf{r}_m - \mathbf{r}_n$ is the difference between the positions of two microphones, where $\mathbf{r}_n = (x_n, y_n, z_n)^T$ are the coordinates of the nth microphone. Further, φ and ϑ denote the azimuth and elevation angles of arrival, respectively. This definition of the DOA vector is illustrated in Fig. 1.

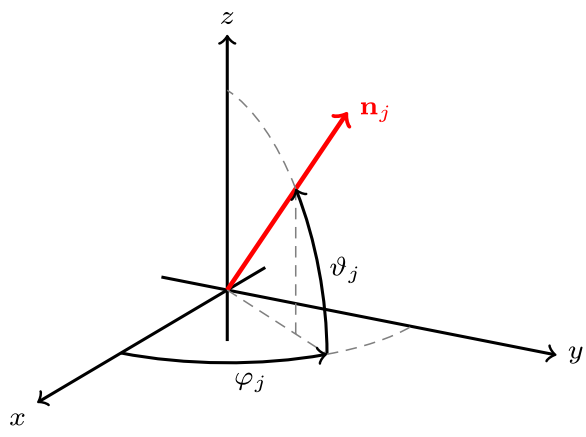

Fig. 1 Definition of the azimuth and elevation angles of arrival, φ_j and ϑ_j, as well as the corresponding DOA vector $\mathbf{n}_j$

Since speech may be considered sparse in the STFT domain, the sources can be separated by attenuating TF bins that are dominated by unwanted components. This can be realized by multiplying the microphone signals with a TF mask $\mathcal{M}_j(f,t) \in [0,1]$. This yields

$$\widehat{\mathbf{S}}'_j(f,t) = \mathcal{M}_j(f,t)\,\mathbf{Y}(f,t), \tag{6}$$

which serves as an estimate of $\mathbf{S}'_j(f,t)$. Consequently,

$$\widehat{S}'_j(f,t) = \mathbf{u}_{n_\mathrm{r}}^T \widehat{\mathbf{S}}'_j(f,t), \tag{7}$$

represents the corresponding target signal estimate. Alternatively, as proposed in [24, 25], the masks can be used in the estimation of the power spectral density (PSD) matrices $\widehat{\mathbf{\Phi}}_{S'_j}(f,t) = E\left\{\mathbf{S}'_j(f,t)(\mathbf{S}'_j(f,t))^H\right\}$ required for beamforming. For this purpose, Eq. 6 is inserted for the unknown $\mathbf{S}'_j(f,t)$, and the expectation $E\{\cdot\}$ can be replaced by, e. g., recursive averaging [26]. This yields

$$\begin{aligned}\widehat{\mathbf{\Phi}}_{S'_j}(f,t) = \alpha\,\widehat{\mathbf{\Phi}}_{S'_j}(f,t-1) + \\ (1-\alpha)\,\mathcal{M}_j^2(f,t)\,\mathbf{Y}(f,t)\mathbf{Y}^H(f,t),\end{aligned} \tag{8}$$

where α is an averaging parameter. Additionally, we define

$$\mathbf{V}_j(f,t) = \mathbf{Y}(f,t) - \mathbf{S}'_j(f,t) \tag{9}$$

as the mixture of all unwanted components with respect to the j-th source, and use $\mathbf{\Phi}_{V_j}(f,t)$ to denote the corresponding PSD matrix. An estimate thereof is given by

$$\begin{aligned}\widehat{\mathbf{\Phi}}_{V_j}(f,t) = \alpha\,\widehat{\mathbf{\Phi}}_{V_j}(f,t-1) + \\ (1-\alpha)\,(1-\mathcal{M}_j(f,t))^2\,\mathbf{Y}(f,t)\mathbf{Y}^H(f,t).\end{aligned} \tag{10}$$

These PSD matrices can be used to cancel noise and interference by an appropriate beamforming operation. Here, we select the minimum variance distortionless response (MVDR) beamformer [27]

$$\mathbf{W}_j(f,t) = \frac{\widehat{\mathbf{\Phi}}_V^{-1}(f,t)\widehat{\mathbf{\Phi}}_{S'_j}(f,t)}{\mathrm{trace}\left\{\widehat{\mathbf{\Phi}}_V^{-1}(f,t)\widehat{\mathbf{\Phi}}_{S'_j}(f,t)\right\}}\,\mathbf{u}_{n_\mathrm{r}}. \tag{11}$$

A target signal estimate is then obtained as

$$\widehat{S}'_j(f,t) = \mathbf{W}_j^H(f,t)\mathbf{Y}(f,t). \tag{12}$$

For both approaches, the direct application of Eq. 7 and the mask-based beamforming of Eq. 12, the source separation problem reduces to the estimation of TF masks.

3 Guided source separation

This section presents a summary of the guided source separation (GSS) proposed in [15]. The approach makes use of cACGMM-based TF masking [11], but additionally incorporates time annotations to constrain the mixture components.

The normalized vector of microphone signals defines the *directional statistics*

$$\mathbf{Z}(f,t) = \frac{\mathbf{Y}(f,t)}{\|\mathbf{Y}(f,t)\|_{\ell_2}}. \tag{13}$$

As shown in [11], these can be modeled by a mixture of K complex angular central Gaussian (cACG) components. For the source separation problem formulated in Section 2, we have $K = J$ in the simplest case, i.e., each component is used to describe one speaker. The probability density function of the cACG distribution with parameter matrix $\mathbf{B}$ is given by

$$\mathcal{P}(\mathbf{Z};\mathbf{B}) = \frac{(N-1)!}{2\pi^N \det\mathbf{B}}\,\frac{1}{(\mathbf{Z}^H\mathbf{B}^{-1}\mathbf{Z})^N}. \tag{14}$$

Consequently, we obtain the cACGMM

$$p\left(\mathbf{Z}(f,t);\Theta(f)\right) = \sum_k \psi_k(f)\,\mathcal{P}\left(\mathbf{Z}(f,t);\mathbf{B}_k(f)\right) \tag{15}$$

with mixture weights $\psi_k(f)$. The set $\Theta(f)$ contains the parameters $\mathbf{B}_k(f)$ and $\psi_k(f)$ for all components $k \in \{1,\ldots,K\}$, which can be estimated using the EM algorithm. As each frequency is considered independently, however, the same index k may correspond to a different source j at different frequencies.

To cope with the resulting frequency permutation problem, the GSS [15] takes advantage of time annotations $\beta_k(t) \in \{0,1\}$ that indicate whether the source that corresponds to the k-th component is active in frame t. To integrate these into the cACGMM, the mixture weights $\psi_k(f)$ in Eq. 15 are replaced by $\psi_k(f)\beta_k(t)$. With the proper normalization, this leads to the mixture model

$$\begin{aligned}&p\left(\mathbf{Z}(f,t);\Theta(f)\right) = \\ &\frac{\sum_k \psi_k(f)\beta_k(t)\mathcal{P}\left(\mathbf{Z}(f,t);\mathbf{B}_k(f)\right)}{\sum_k \psi_k(f)\beta_k(t)}.\end{aligned} \tag{16}$$

The EM algorithm can be reformulated accordingly, so that the permutation problem is inherently addressed [15]. The E-step is

$$\mathcal{N}_k(f,t) \leftarrow \frac{\psi_k(f)\beta_k(t)\,\mathcal{P}\left(\mathbf{Z}(f,t);\mathbf{B}_k(f)\right)}{\sum_{k'}\psi_{k'}(f)\beta_{k'}(t)\,\mathcal{P}\left(\mathbf{Z}(f,t);\mathbf{B}_{k'}(f)\right)}, \tag{17}$$

where the posterior $\mathcal{N}_k(f,t)$ may be interpreted as a TF mask for the k-th component. The M-step is given by

$$\psi_k(f) \leftarrow \frac{1}{T}\sum_t \mathcal{N}_k(f,t) \tag{18a}$$

$$\mathbf{B}_k(f) \leftarrow N\frac{\sum_t \mathcal{N}_k(f,t)\frac{\mathbf{Z}(f,t)\mathbf{Z}^H(f,t)}{\mathbf{Z}^H(f,t)\mathbf{B}_k^{-1}(f)\mathbf{Z}(f,t)}}{\sum_t \mathcal{N}_k(f,t)}. \tag{18b}$$

To obtain the masks $\mathcal{M}_j(f,t)$ from the posteriors $\mathcal{N}_k(f,t)$ after the algorithm has converged, it is only necessary to determine the frequency dependent mapping between the K cACGMM components and the J sources. Using the time annotations $\beta_k(t)$, a *fixed* (frequency independent) mapping can be enforced. Then, additional measures to resolve the permutation problem

are not required. To achieve this, the following must be ensured: (i) the annotations must correlate well with the true source activity, and (ii) they must be unique in the sense that the annotations for any pair of two components k_1 and k_2 must not be too similar (in particular, the annotations are not useful when $\beta_{k_1} \equiv \beta_{k_2}$).

As [15] proposes, an additional component, which is assumed to be active at all times ($\beta_K(t) = 1$ for all t), can be used to account for noise. Then, the total number of components is $K = J + 1$.

4 DOA-guided source separation

Two fundamental limitations of the GSS approach are that (i) the cACGMM parameter estimation is sensitive to the initialization, and (ii) time annotations first have to be estimated when they are not available in advance. To address these, despite spatial clustering being an unsupervised approach, it can be advantageous to incorporate a priori knowledge.

In this work, we propose a direction of arrival-guided source separation (DOA-GSS). It is assumed that the broadband source DOAs, or equivalently the DOA vectors $\mathbf{n}_j(t)$, have been estimated in advance. Numerous estimators that can be used for this purpose are available. An overview of statistical model-based methods can be found, e. g., in [28]. In order to estimate the DOAs of multiple concurrent sources, typically, a narrowband approach is applied first, followed by a clustering of the estimates across all frequencies. Among the most widely used methods are narrowband realizations of steered response power (SRP) [29, 30], where the direction is determined by maximizing the output power of a beamformer, as well as subspace decomposition-based methods like MUSIC [31]. Alternatively, a deep learning approach can also be used [32].

A block diagram of the resulting DOA-GSS system, which consists of DOA estimation, the derivation of DOA-based prior information, the cACGMM method, and the mask-based source separation, is shown in Fig. 2.

We consider three different techniques to take advantage of DOA estimates. Section 4.1 discusses DOA-based masks, which can be used to initialize the EM algorithm. Secondly, to replace the oracle time annotations, we propose to extract source time annotations (STAs) from the initial masks in Section 4.2. Thirdly, instead of using one cACG component for each speaker, an approach with DOA specific components could be adopted as described in Section 4.3, whereby DOA time annotations (DTAs) are obtained.

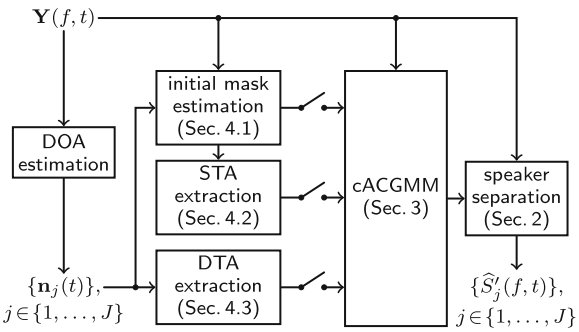

Fig. 2 DOA-GSS block diagram. We consider three different ways to make use of DOA knowledge: a DOA-based initialization of the cACGMM parameter estimation (see Section 4.1), extracting time annotations from the initial masks (Section 4.2), or using one mixture component for each discrete DOA (Section 4.3). These can be used independently or jointly

4.1 DOA-based initial masks

In the following, we introduce DOA-based masks for the initialization of the model parameter estimation. We would like to stress that it is unnecessary for these initial masks to already separate the components perfectly. Rather, they should be simple to compute, but should already distinguish sufficiently well between the signal components to improve the separation realized by the resulting cACGMM-based masks. In this work, we first perform a separation which focuses on the target (direct-path) components, and disregards all other signal contributions. Then, residual unwanted components are suppressed under the assumption of their spatial diffuseness. Specifically, for each of the J sources, we consider a cascade of two single-channel Wiener filters [26]. For both, we require an estimate of the auto-PSD $\Phi_{S'_j}(f,t)$ of the target signal $S'_j(f,t)$ defined in Eq. 3. To realize the source separation and noise suppression, the PSD estimates for the first and the second step are, however, obtained under different assumptions. This will be discussed in Sections 4.1.1 and 4.1.2, respectively.

The initial source separation and the residual noise suppression can both be expressed in terms of TF masks that will be denoted by $\mathcal{M}_j^{\text{sep}}(f,t)$ and $\mathcal{M}_j^{\text{noi}}(f,t)$. Because the two steps are applied sequentially, the initial mask for the j-th source is given by the multiplicative combination

$$\mathcal{M}_j^{\text{init}}(f,t) = \mathcal{M}_j^{\text{sep}}(f,t)\,\mathcal{M}_j^{\text{noi}}(f,t). \tag{19}$$

Rather than separating the sources directly, we use $\mathcal{M}_j^{\text{init}}(f,t)$ *only* to initialize the cACGMM parameter estimation. Thus, we set

$$\mathcal{N}_j^{\text{init}}(f,t) = \mathcal{M}_j^{\text{init}}(f,t) \quad \forall j \in \{1,\ldots,K-1\}, \tag{20}$$

for the $J = K - 1$ components that correspond to the target sources. For the noise, which is represented by the K-th component, the initialization is given by

$$\mathcal{N}_K^{\text{init}}(f,t) = 1 - \sum_j \mathcal{M}_j^{\text{init}}(f,t). \tag{21}$$

After the initialization, the model parameter estimation can be performed, starting with the M-step given by Eqs. 18a and 18b. The unprocessed microphone signals $\mathbf{Y}(f,t)$ are still used to define the directional statistics (Eq. 13), and to perform the final separation with the masks obtained from the cACGMM.

The initial mask estimation is performed independently for each frequency and frame. In the remainder of Section 4.1, the corresponding indices will therefore be omitted to simplify the notation.

4.1.1 Source separation

First, we focus on the separation of the direct-path signals. For this purpose, the strongly simplified signal model

$$\mathbf{Y} = \sum_j \mathbf{S}'_j = \sum_j \mathbf{A}_j S'_j \tag{22}$$

is considered, where the j-th propagation vector $\mathbf{A}_j$ is given with the corresponding DOA vector $\mathbf{n}_j$ according to Eq. 5. With the propagation matrix $\boldsymbol{\mathcal{A}} = [\mathbf{A}_1, \ldots, \mathbf{A}_J]$, and the vector of direct-path components at the reference microphone $\boldsymbol{\mathcal{S}} = [S'_1, \ldots, S'_J]^T$, we can reformulate Eq. 22 as a matrix-vector product

$$\mathbf{Y} = \boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{S}}. \tag{23}$$

For the special case where the number of sources is equal to the number of microphones ($J = N$), Eq. 23 can straightforwardly be solved for $\boldsymbol{\mathcal{S}}$ by left multiplying with $\boldsymbol{\mathcal{A}}^{-1}$. When $J \neq N$, we can obtain an approximation by using the Moore-Penrose pseudoinverse $\boldsymbol{\mathcal{A}}^{\dagger}$ instead [33]. The resulting estimate is

$$\widehat{\boldsymbol{\mathcal{S}}} = \boldsymbol{\mathcal{A}}^{\dagger}\mathbf{Y}. \tag{24}$$

Whereas Eq. 24 could be used to separate the sources directly, the usefulness of such an approach is limited due to the strongly simplified signal model employed. Instead, we use Eq. 24 to obtain a Wiener filter [26]. Under the assumption of the source signals being mutually uncorrelated, it is given by

$$\mathcal{M}_j^{\text{sep}} = \frac{\widehat{\Phi}_{S'_j}}{\sum_{j'} \widehat{\Phi}_{S'_{j'}}} = \frac{\left\{\boldsymbol{\mathcal{A}}^{\dagger}\boldsymbol{\Phi}_Y(\boldsymbol{\mathcal{A}}^{\dagger})^H\right\}_{jj}}{\text{trace}\left\{\boldsymbol{\mathcal{A}}^{\dagger}\boldsymbol{\Phi}_Y(\boldsymbol{\mathcal{A}}^{\dagger})^H\right\}}, \tag{25}$$

where the PSD matrix $\boldsymbol{\Phi}_Y = E\{\mathbf{Y}\mathbf{Y}^H\}$ can be estimated from the microphone signals, and $\{\cdot\}_{jj}$ is the j-th diagonal entry of this matrix.

4.1.2 Noise suppression

Thus far, only the direct-path contributions of the J sources are accounted for. Based on the definition of the source separation masks (Eq. 25), we note that $\sum_j \mathcal{M}_j^{\text{sep}} = 1$. If these were used for the initialization i.e., $\mathcal{M}_j^{\text{init}} = \mathcal{M}_j^{\text{sep}}$, Eq. 21 would produce an all-zero initialization for the noise component ($\mathcal{N}_K^{\text{init}} = 0$). Because this would result in a 0 in the denominator of Eq. 18b, this is not a valid choice. To obtain a suitable initialization for *all* components, we require a second step that addresses late reverberation, and additive noise that has no pronounced directivity. The simplified signal model for this step is therefore

$$\widetilde{\mathbf{Y}}_j = S'_j + \widetilde{\mathbf{V}}_j, \tag{26}$$

where $\widetilde{\mathbf{Y}}_j$ denotes the output of the initial source separation step for the j-th source, and $\widetilde{\mathbf{V}}_j$ is the corresponding residual of the unwanted components (Eq. 9), which will simply be referred to as *noise* for conciseness. To comply with this signal model, a time alignment with respect to the target DOA is additionally required, such that the same desired signal is present in each channel. Consequently, we define

$$\widetilde{\mathbf{Y}}_j = \left(\mathcal{M}_j^{\text{sep}}\mathbf{Y}\right) \odot \mathbf{A}_j^*, \tag{27}$$

where $\odot$ is the Hadamard (elementwise) product, and $(\cdot)^*$ is the complex conjugate.

Now, for the noise suppression, we make use of the Wiener postfilter proposed in [34], which permits a specific noise field coherence to be incorporated. Here, we consider a spherically isotropic (diffuse) noise field, such that the coherence function for microphone pair (m,n) is given by

$$\Gamma_{mn} = \text{sinc}\left(\kappa \|\mathbf{r}_{mn}\|_{\ell_2}\right). \tag{28}$$

It is assumed that target signal and residual noise are mutually uncorrelated, and that the noise auto-PSD is the same for all channels. As proposed in [34, 35], a target signal PSD estimate

$$\widehat{\Phi}_{S'_j}^{(mn)} = \frac{\Re\{\widehat{\Phi}_{\widetilde{Y}_j,mn}\} - \Re\{\Gamma_{mn}\}\frac{\widehat{\Phi}_{\widetilde{Y}_j,mm} + \widehat{\Phi}_{\widetilde{Y}_j,nn}}{2}}{1 - \Re\{\Gamma_{mn}\}} \tag{29}$$

can then be extracted based on each microphone pair, where $\Re\{\cdot\}$ denotes the real part, and $\widehat{\Phi}_{\widetilde{Y}_j,mn}$ is the (m,n)th entry of the estimated PSD matrix $\widehat{\boldsymbol{\Phi}}_{\widetilde{Y}_j}$. Subsequently, an improved estimate is obtained by averaging Eq. 29 over all unique microphone pairs. Similarly, instead of considering only the reference channel, the same averaging technique can be adopted to acquire an improved estimate of the signal-plus-noise auto-PSD. The resulting Wiener filter

$$\mathcal{M}_j^{\text{noi}} = \frac{\frac{2}{N(N-1)}\sum_{n=1}^{N-1}\sum_{m=n+1}^{N}\widehat{\Phi}_{S'_j}^{(mn)}}{\frac{1}{N}\sum_{n=1}^{N}\widehat{\Phi}_{\widetilde{Y}_j,nn}} \tag{30}$$

then serves as the noise suppression mask. Note that, due to the time alignment (Eq. 27), this mask is indirectly dependent on the respective DOA as well. The estimation of $\mathcal{M}_j^{\text{noi}}$ for all $j \in \{1, \ldots, J\}$ based on the output of the initial source separation is illustrated in Fig. 3.

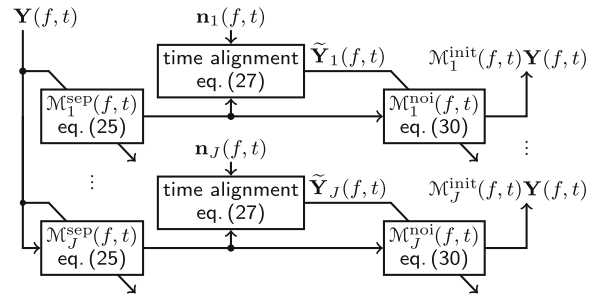

Fig. 3 The noise suppression masks $\mathcal{M}_j^{\text{noi}}(f,t)$ are determined based on each of the J outputs of the initial source separation after time aligning the channels for the respective target DOA. The combination of these steps yields the initial masks

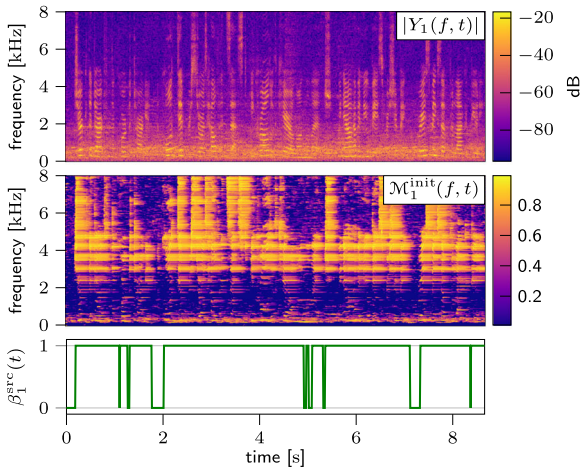

Fig. 4 Extraction of source time annotations (STAs) $\beta_j^{\text{src}}(t)$ for $J=1$ static source from the corresponding initial mask ($P=10$). The mask only captures the rough outline of the target speech, but this is sufficient to detect silent segments

4.2 Source time annotations (STAs)

As observed in [16], additional measures to address the permutation problem may not be required if the employed initial masks are already sufficiently reliable. With regard to the proposed initialization, however, this is not always the case. At low frequencies, in particular, it is difficult to distinguish between different sources based on spatial information, and the quality of the DOA-based masks deteriorates. Time annotations may, therefore, still be helpful.

To determine when each speaker is active, we propose to derive STAs $\beta_j^{\text{src}}(t)$ from the (DOA-based) initial masks according to

$$\beta_j^{\text{src}}(t) = \begin{cases} 1, & \sum_f \mathcal{M}_j^{\text{init}}(f,t) \geq \delta_j \\ 0, & \text{else.} \end{cases} \tag{31}$$

Here, the activity thresholds δ_j are chosen as the P-th percentile of $\sum_f \mathcal{M}_j^{\text{init}}(f,t)$ i.e., each source is assumed to be inactive in a total of $PT/100$ frames. Note that the STAs are used *only* in the cACGMM parameter estimation. After convergence, they are omitted in Eq. 17, so that the final masks can be non-zero for all frames t.

As opposed to a voice activity detection-based approach, the STAs given by Eq. 31 could also be used e. g., to explicitly take into account (localized) background noise sources, although this is not considered in this work. Further, by defining a fixed percentile P, it is ensured that the STAs remain distinctive, even when a speaker is active during the entire sequence. It may then still be appropriate to consider the corresponding mixture component to be *inactive* during brief speech pauses, such as between two words.

The extraction of STAs from an initial mask is illustrated in Fig. 4, where a scenario with $J=1$ speaker in the presence of noise (SNR $=5$ dB) is considered. The DOA-based masks are sufficient to identify frames with low speech activity.

4.3 DOA time annotations (DTAs)

Alternatively, the DOAs can be used to obtain time annotations directly. This is achieved by using different components to represent the same moving speaker at different times, depending on their current location. Given that the cACGMM (Eq. 15) models *spatial* information, using multiple components for the same source could be beneficial since the spatial signal properties are also time dependent.

Let $D \geq J$ be the total number of unique DOAs for which there is an active source at least once across the considered T frames. To acquire DTAs, the number of cACG components can then be set to $K = D+1$ i.e., one component is used for each direction rather than each speaker. To limit the total number of components, and ensure that a sufficient amount of data is available for each, the DOAs are discretized with a finite resolution.

The DTAs $\beta_k^{\text{dir}}(t)$ are defined based on the DOA estimates alone: a component is only active ($\beta_k^{\text{dir}}(t)=1$) while there is a source in this direction, otherwise it is considered to be inactive ($\beta_k^{\text{dir}}(t)=0$). This is illustrated in Fig. 5, where a gradual movement of $J=1$ source is assumed, starting from an azimuth angle of arrival $\varphi = 40°$ up to $\varphi = 60°$. For a discretization of φ in 10° steps, this results in a total of $K=4$ components, which correspond to the angles $\varphi \in \{40°, 50°, 60°\}$ and noise, respectively. As the figure shows, the DTAs are unique in the described scenario for all $k \in \{1,2,3,4\}$, so that no additional information is required to distinguish between the components.

For a static source, however, the DTAs are not helpful. This problem is illustrated in Fig. 6, where there is $J=1$ speaker with a constant DOA of $\varphi = 50°$. Consequently, the resulting DTAs coincide with the annotations for the

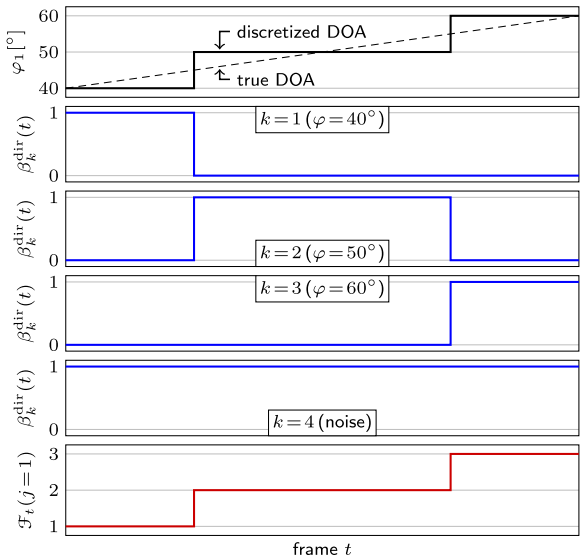

Fig. 5 DOA time annotations (DTAs) $\beta_k^{\text{dir}}(t)$ for $J=1$ speaker moving from $\varphi=40°$ to 60°. This results in $D=3$ discrete directions, and thus $K=4$ components. The function $\mathcal{F}_t(j)$ specifies, for each frame, which component corresponds to the j-th speaker. Note that DTAs and $\mathcal{F}_t(j)$ are determined solely by the DOA estimates. Nevertheless, the DTAs are unique, and thereby enable to enforce a fixed permutation

noise component. In this case, only the STAs can resolve the frequency permutation problem.

Therefore, we can also consider combined annotations. Whereas the DTAs are specific to each of the $K=D+1$ (DOA-based) components, the STAs are specific to each of the J sources. With the function $\mathcal{F}_t : \{1,\dots,J\} \to \{1,\dots,K-1\}$, that specifies which source index j corresponds to which component index $k=\mathcal{F}_t(j)$ in frame t, we define combined annotations that are 1 for component k when there is at least one *active* source ($\beta_j^{\text{src}}(t)=1$) in the associated direction ($\mathcal{F}_t(j)=k$), i.e.,

$$\beta_k(t) = \beta_k^{\text{dir}}(t) \max_{j:\mathcal{F}_t(j)=k} \beta_j^{\text{src}}(t) \quad \forall k < K. \tag{32}$$

The bottom plot of Fig. 5 shows $\mathcal{F}_t(j)$ for the considered example. Like the DTAs, this mapping between the source and component indices is only dependent on the current DOA estimates.

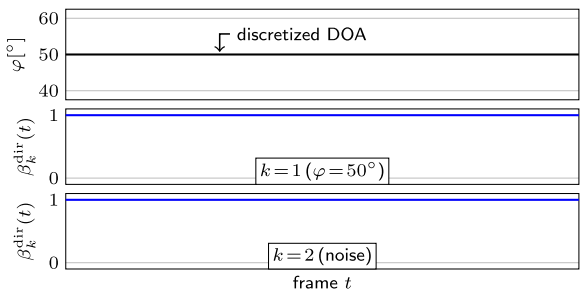

Fig. 6 DTAs $\beta_k^{\text{dir}}(t)$ for $J=1$ static speaker at $\varphi=50°$. The resulting DTAs are not unique in this case

Note that since the computed initial masks are source-based as well,

$$\mathcal{N}_{\mathcal{F}_t(j)}^{\text{init}}(f,t) = \mathcal{M}_j^{\text{init}}(f,t) \tag{33}$$

is used instead of Eq. 20 to initialize $\mathcal{N}_k^{\text{init}}(f,t)$ when the components are specific to each DOA rather than each speaker.

In summary, we have proposed three ways to take advantage of the availability of DOA estimates. To enable a good performance despite the approach being sensitive to the initialization, we can make use of DOA-based masks, like the ones presented in Section 4.1, to initialize the EM algorithm. Time annotations can be integrated into the model to avoid the frequency permutation problem. On the one hand, STAs can be derived directly from the initial masks. Alternatively, or in combination with the STAs, the mapping between the components and the sources can be defined based on the DOA estimates in order to generate DTAs. We will refer to this as the *approach with DOA-based components* to distinguish it from the speaker-based approach ($K=J+1$), where DTAs are *not* available. Equivalently, we will specify that the DTAs are used or omitted to indicate that the DOA-based or the speaker-based approach are used, respectively.

5 Results and discussion

To establish how DOA information can *best* be incorporated in the GSS, taking into account all of the introduced methods, we conduct a series of experiments. First, in Section 5.1, we focus on scenarios where the DOAs are static while the respective speaker is active. In this context, we aim to (i) determine how the parameter P can be chosen, (ii) verify that the proposed initialization and time annotations are effective, and (iii) assess the robustness of the DOA-GSS to DOA estimation errors. Subsequently, in Section 5.2, we evaluate the approach based on a gradually moving speaker. The goal of this experiment is to individually examine the usefulness of both types of annotations, STAs and DTAs. Additionally, we address the question whether the time annotations can be omitted entirely, and review the need for a manual permutation alignment. Finally, we use our findings to select one suitable DOA-GSS setup, based on which the performance for conditions of varying difficulty is evaluated in Section 5.3.

An overview of the different GSS setups that will be considered in the following is presented in Table 1. These will be explained in more detail in Sections 5.1.1 and 5.2.1. Figure 2 illustrates how the different components tie into the complete system.

Table 1 Overview of different options for various components of the system considered throughout the experiments. If nothing else is explicitly stated, the underlined default is used

DOA estimates	Oracle, *or* estimated using the DNN approach from [43]
Initial masks	(proposed) DOA-based (Eq. 19), *or* oracle (Eq. 34), *or* random
STAs (Eq. 31)	Extracted from (proposed) DOA-based initial masks (Eq. 19), *or* extracted from oracle initial masks (Eq. 34), *or* none
DTAs	One mixture component for each direction ($K = D + 1$: DTAs are available, see Sec. 4.3), *or* one mixture component for each speaker ($K = J + 1$: DTAs are *not* available)
Permutation alignment	No manual alignment, *or* oracle alignment (as explained in Section 5.2.1)
Speaker separation	Mask-based MVDR beamforming (Eq. 12), *or* direct application of the masks (Eq. 7)

5.1 Static speakers

For the experiments conducted in this section, the locations of the talkers are fixed during each utterance. Between two utterances, however, a new angle is selected with a probability of 50%. To cope with these sudden DOA changes, the approach introduced in Section 4.3 is employed, where each component corresponds to one discrete DOA ($K = D + 1$). The setup is explained in detail in Section 5.1.1, followed by the discussion of the results in Sections 5.1.2, 5.1.3, 5.1.4, and 5.1.5.

5.1.1 Experimental setup

Microphone signals are generated by additively mixing $J = 2$ speech signals and additive noise. We make use of the TSP speech database [36], which consists of anechoic recordings of the Harvard sentences [37] for 24 different speakers (a total of 1 444 utterances with an average duration of 2.4 s). The source signals are assembled by concatenating 5 utterances of the same speaker. For the first utterance, and for every instance where the DOA is changed at the end of an utterance, an azimuth angle of arrival is selected at random, under the constraint that different speakers are never at the same location at the same time. Consequently, to obtain the corresponding microphone signal component, the dry signal is convolved with one of the room impulse response that we recorded for azimuth angles $\varphi \in \{0°, 20°, \ldots, 180°\}$ with the miniDSP UMA-16 array [38] ($\vartheta \approx 0°$). The recordings were made in a meeting room with a reverberation time of about $T_{60} = 660$ ms (approximate dimensions: 7.50 m × 5.00 m × 2.65 m), for a source-array distance of 2 m. A relatively diffuse recording of the pub noise signal from the ETSI background noise database [39], that serves as the additive noise, was obtained with the same array in a room with $T_{60} \approx 1$ s.

Out of the available microphones, we consider a subarray of 9 microphones. As can be seen in Fig. 7, these form a uniform rectangular array (URA) with an element spacing of 4.2 cm. The sampling rate is $f_s = 16$ kHz. For the STFT, the frame length, as well as the transform size F, are set to 512 samples (32 ms). With a frame shift of 160 samples, we obtain 100 frames per second. A square-root Hann window is used in analysis and synthesis.

As in [15], we make use of the weighted prediction error (WPE)-based dereverberation [40] implemented in [41] prior to the (initial) mask estimation. To perform the cACGMM parameter estimation, we use the Python-based source code of [42]. In practice, to obtain locally optimal parameters and limit the required number of components, a new cACGMM may be computed periodically, or the model can be updated adaptively. In this work, for simplicity, we only compute a single model for each mixture, based on the entire signal. Subsequently, to perform the separation, we consider the direct application of the masks (Eq. 7), as well as mask-based MVDR beamforming (Eq. 12). To reduce artifacts such as musical tones, which are introduced particularly when the masks are applied directly, the final masks are lower bounded by 0.01. For the recursive averaging used in Eqs. 8 and 10 to estimate the PSD matrices needed for the MVDR beamformer, we set the averaging parameter to $\alpha = 0.90$, which corresponds to a time constant of 100 ms.

For the DOAs, we first assume that the true (oracle) source locations are known. Later, starting in Section 5.1.3, we also use realistic DOA estimates in order to test the robustness of the approach to DOA inaccuracies. For this purpose, we make use of the CNN/LSTM broadband DOA estimator from [43], which is an extension of the CNN proposed in [44]. The network is trained to return, for each discrete DOA (resolution 5°), a frame-wise probability that indicates when there is an active source in this direction. The *phases* of the microphone signals are the input to the network. Training data are generated using simulated RIRs, and datasets that do *not* overlap with our experimental setup. In practice, a simpler DOA estimation method may be preferred given that the aim is only to generate a priori information. Since source localization is not the focus of this work, however, the

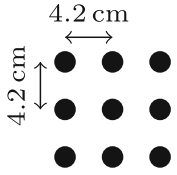

Fig. 7 A uniform rectangular array (URA) comprising 9 microphones is considered

selection of the algorithm is arbitrary. Note that DOA estimation errors have an impact on the initial masks and the annotations. Additionally, they are relevant in the selection of the number of components K when the approach with DOA-based components is used, for which the resolution is set to 20°. Then, no additional cACG components are introduced for errors $\Delta\varphi < 10°$.

In contrast to the mask-based adaptation of the MVDR beamformer, the signal components are not yet (well) separated when estimating the PSD matrices $\mathbf{\Phi}_Y(f,t)$ and $\mathbf{\Phi}_{\widetilde{Y}_j}(f,t)$ required for the initial masks. In this case, it is therefore beneficial to use a shorter averaging duration to take advantage of the signal components being (relatively) sparse in the time-frequency domain. Here, we use recursive averaging with an empirically chosen time constant of 40 ms ($\alpha = 0.78$) for the estimation of the PSD matrices that are needed for the initial mask computation only.

As an upper bound for the performance, we consider the initialization with an oracle mask

$$\mathcal{M}_j(f,t) = \min\left\{\frac{\left|\gamma_j S'_{j,1}(f,t)\right|^2}{\left|Y_1(f,t)\right|^2}; 1\right\}, \tag{34}$$

and STAs extracted therefrom with Eq. 31. For the direct-path component in Eq. 34, we use a delayed version of the dry source signal. In order to prevent a reverberation dependent attenuation of the signal, the scaling factor γ_j is set such that $\gamma_j \mathbf{S}'_j(f,t)$ has the same energy as the reverberant signal $\mathbf{S}_j(f,t)$. As lower bounds, we consider random initialization, and the omission of the STAs. The DTAs, however, are used for all configurations. Since the speaker locations only change between two utterances, the DTAs mainly distinguish different utterances here. Note that the GSS is also applied on an utterance-level in [15], although only a limited context around each considered utterance is taken into consideration in the cACGMM computation. Therefore, the configuration where only the DTAs are used (omission of the STAs) may be seen as representative of the GSS baseline for the particular experimental setup considered throughout Section 5.1, disregarding the effect of DOA errors.

As the instrumental metrics on which the approach is benchmarked, we use STOI [45], wideband PESQ [46] on a MOS-LQO scale, as well as the segmental SDR, SIR, and SNR [47]. For the latter metrics, $\widehat{S}'_j(f,t)$ is decomposed, in the time domain, into components that represent filtered target $s(i,t)$, residual interference $\varepsilon_{\mathrm{i}}(i,t)$, noise $\varepsilon_{\mathrm{n}}(i,t)$, and artifacts $\varepsilon_{\mathrm{a}}(i,t)$, respectively, where i indexes the samples within one frame. For all performance metrics, we report the improvement (Δ) compared to the noisy reference microphone signal. The clean target for the computation of all metrics is again the delayed source signal that is also used for the oracle masks (Eq. 34). We average the results for 25 independently generated sets of microphone signals for low-noise (mixing SNR of 30 dB), and for noisy (5 dB) conditions.

5.1.2 Selection of the percentile parameter

In Fig. 8, the results are displayed as a function of the percentile P that is used to set the thresholds δ_j for the STAs (Eq. 31). First, we note that even when the STAs are disabled ($P = 0$) and the masks are initialized randomly, the signal components are separated relatively well (e. g., for SNR = 5 dB and the MVDR beamformer: ΔSTOI = 0.15 and ΔPESQ = 0.13). This is because the speakers can be distinguished based on the DTAs alone when different utterances come from different directions.

Nevertheless, it remains beneficial to also incorporate STAs in this case: a maximum of ΔSTOI (0.17 for the same conditions as above) and ΔPESQ (0.17) is achieved around $P = 10$ for the random initialization, before these metrics start to deteriorate. This behavior for $P > 0$ may be explained with a larger portion of the signal being attributed to the additive noise, and thus being suppressed, when the speakers are assumed to be inactive part of the time. Although this assumption is plausible in light of the presence of speech pauses, the target signal might not be *entirely* absent during these defined periods of inactivity. Consequently, while the ΔSNR score increases monotonically with P, the speech distortion also becomes more considerable.

Regardless of the choice of P, the proposed initialization with the DOA-based masks boosts the achieved performance significantly. Moreover, these initial masks can guide the EM algorithm towards a solution with a permutation that is consistent across frequency, so that the STAs are no longer needed: ΔSTOI and ΔPESQ, in particular, are relatively stable for $P \leq 10$ (a maximum of 0.21 is obtained for both metrics using mask-based beamforming in noisy conditions), and start to degrade for $P > 10$. For a high input SNR, the degradation is more pronounced since there is then little benefit in increasing P.

For the considered setup, we conclude that the STAs are not needed when the permutation problem can be addressed with the initial masks alone. Here, this is the case for the DOA-based, but not for the random initialization. On the other hand, P can still be a useful trade-off parameter, in order to control how aggressively noise is suppressed. Note that the question whether time annotations may be dropped entirely (STAs *and* DTAs) was not addressed here. We empirically found that it is not reasonable to use the same mixture component for utterances impinging from completely different directions, and that the corresponding results are therefore not meaningful. Instead, a dedicated evaluation of the need for STAs and DTA will be performed with a different setup in Section 5.2.

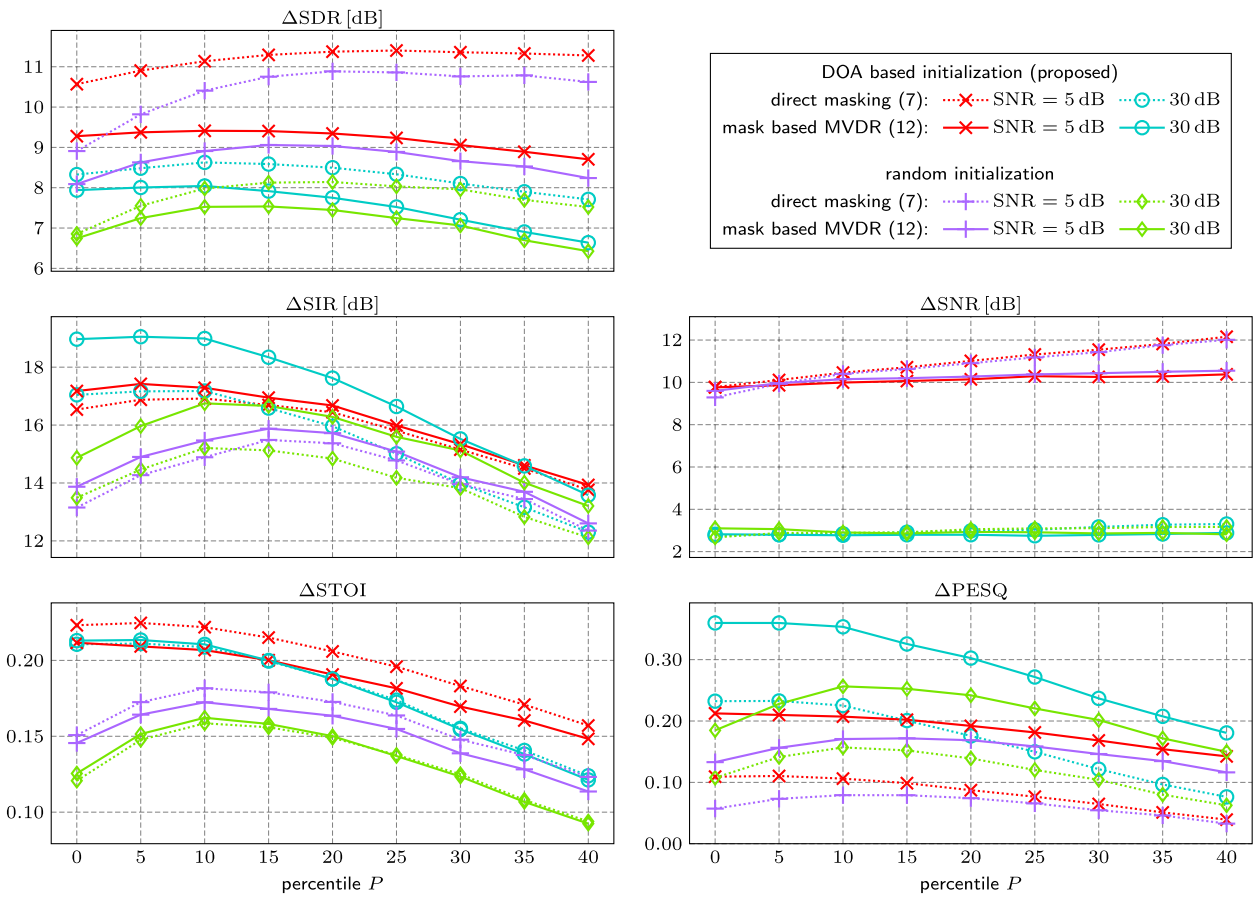

Fig. 8 Evaluation results as a function of the parameter P that is used in the computation of the STAs. When the masks are initialized randomly, the performance peaks around $P = 10$, whereas the proposed initialization permits the STAs to be omitted ($P = 0$) for the considered evaluation setup

Based on the above findings, we set $P = 10$ in the following, since this choice leads to a near-optimal performance for all considered configurations.

5.1.3 Impact of initialization and STAs

Next, we examine the influence of the initialization and the STAs more closely. The results obtained with the ground truth DOAs can be found in Table 2. The first two rows (labeled "0D" and "0B") correspond to the initialization according to Eq. 33 and STAs according to Eq. 31 using the proposed DOA-based masks. These are given by Eq. 19 with Eqs. 25 and 30. In the row labels, "D" indicates direct application of the masks (Eq. 7), and "B" mask-based MVDR beamforming (Eq. 12). The remaining rows (1 to 8) show the results for all other combinations of initialization and STAs (see Table 1 for an overview of all options). Groups of three different rows, where either the STAs *or* the initialization are fixed (e. g., rows 2, 5, and 8), can be considered to understand their effect on the performance.

Generally, the direct masking tends to yield higher ΔSDR and ΔSTOI scores, whereas the beamformer is superior regarding ΔSIR and ΔPESQ. This is because the direct masking permits an effective suppression of unwanted components regardless of their spatial properties. In the process, however, artifacts such as musical tones are introduced, which are detrimental to the speech quality. By inherently steering spatial nulls in the right directions, the beamformer, in contrast, can remove localized interferers effectively without distorting the target signal, but does not suppress diffuse components such as background noise and reverberation equally well. Since it is application dependent which method is preferred, we will compare the results based on the best scores obtained with either, direct masking *or* mask-based beamforming.

Even when the STAs are omitted and the initialization is random (row 8), the performance is still decent (for noisy conditions: ΔSTOI = 0.15 and ΔPESQ = 0.13). This is because the DTAs are still available given that the approach with DOA-based components is used. With the STAs derived from the proposed initial masks (row 2), the scores increase by an additional 0.03 in terms of ΔSTOI and 0.04 in terms of ΔPESQ. The difference regarding ΔPESQ is more significant (0.08) under low-noise conditions. Furthermore, we can compare row 2 with row 5

Table 2 Results with *oracle* DOAs for low-noise (left) and noisy (right) conditions. Each row corresponds to a fixed combination of STAs (Eq. 31), initialization (proposed DOA-based (Eq. 19), oracle (Eq. 34), or random), and mask application ("D" for direct (Eq. 7), "B" for beamforming (Eq. 12)). The rows are numbered from 0 to 8 for ease of reference. Boldface is used to indicate the highest scores for each metric. We note that the STAs are helpful on their own, but not needed if the proposed initialization is used. The difference between the results for proposed and oracle initialization is mainly due to the respective behavior at low frequencies

row	STAs	init	mixture SNR = 30 dB				mixture SNR = 5 dB				
			ΔSDR	ΔSIR	ΔSTOI	ΔPESQ	ΔSDR	ΔSIR	ΔSNR	ΔSTOI	ΔPESQ
0D)	prop.	prop.	8.6 dB	17.1 dB	**0.21**	0.22	11.1 dB	16.9 dB	10.5 dB	0.22	0.10
0B)			8.0 dB	18.9 dB	**0.21**	0.35	9.4 dB	17.2 dB	10.0 dB	0.21	0.20
1D)	prop.	oracle	8.7 dB	16.9 dB	0.19	0.22	11.5 dB	18.1 dB	10.7 dB	0.18	0.17
1B)			8.2 dB	18.6 dB	0.19	0.34	10.0 dB	**18.6 dB**	**11.0 dB**	0.17	**0.29**
2D)	prop.	rand.	8.0 dB	15.2 dB	0.16	0.16	10.4 dB	14.9 dB	10.4 dB	0.18	0.08
2B)			7.5 dB	16.7 dB	0.16	0.26	8.9 dB	15.4 dB	10.2 dB	0.17	0.17
3D)	oracle	prop.	8.7 dB	17.5 dB	**0.21**	0.24	11.3 dB	17.3 dB	10.3 dB	**0.23**	0.11
3B)			8.1 dB	**19.3 dB**	**0.21**	**0.36**	9.5 dB	17.7 dB	9.9 dB	0.21	0.21
4D)	oracle	oracle	**8.8 dB**	17.2 dB	0.20	0.23	**11.6 dB**	18.1 dB	10.7 dB	0.19	0.18
4B)			8.3 dB	19.0 dB	0.20	0.35	10.0 dB	**18.6 dB**	**11.0 dB**	0.17	**0.29**
5D)	oracle	rand.	8.2 dB	15.5 dB	0.16	0.16	10.7 dB	15.6 dB	10.3 dB	0.19	0.09
5B)			7.7 dB	17.1 dB	0.17	0.26	9.1 dB	16.0 dB	10.1 dB	0.18	0.18
6D)	none	prop.	8.3 dB	17.0 dB	**0.21**	0.23	10.6 dB	16.5 dB	9.8 dB	0.22	0.11
6B)			7.9 dB	18.8 dB	**0.21**	**0.36**	9.3 dB	17.2 dB	9.8 dB	0.21	0.21
7D)	none	oracle	8.3 dB	16.7 dB	0.19	0.21	11.0 dB	17.7 dB	10.3 dB	0.18	0.17
7B)			8.1 dB	18.6 dB	0.19	0.34	9.8 dB	**18.6 dB**	10.9 dB	0.17	**0.29**
8D)	none	rand.	6.9 dB	13.5 dB	0.12	0.11	8.9 dB	13.2 dB	9.3 dB	0.15	0.06
8B)			6.7 dB	14.9 dB	0.13	0.18	8.1 dB	13.9 dB	9.6 dB	0.15	0.13

(oracle mask-based STAs). The results are similar, which demonstrates that the proposed mask-based STAs are sufficient to address the permutation problem, at least when they are used in conjunction with the DTAs.

As already observed in Section 5.1.2, the need for STAs is mitigated by the proposed initialization for the considered evaluation setup: the differences between rows 6 (omission of the STAs), 3 (oracle STAs), and 0 (proposed DOA-based STAs) are minor. ΔSDR, ΔSIR, and ΔSNR indicate that the inclusion of STAs enables a slightly higher suppression of unwanted components (largest difference: 0.7 dB), but the ΔSTOI and ΔPESQ metrics barely reflect this. The same conclusions can be drawn based on the results obtained with the oracle initial masks (rows 7, 4, and 1).

The initialization has a greater impact on the results, but the trends resemble those for the STAs: Comparing rows 8, 7, and 6 (all for the case where the STAs are omitted), we observe that the proposed DOA-based initial masks (row 6) improve the performance considerably (for noisy conditions: an additional 0.07 and 0.08 in terms of ΔSTOI and ΔPESQ, respectively). The differences between the proposed initialization and oracle initialization (row 7) are inconsistent, however. Upon closer inspection, we find that this is due to the different behavior at low frequencies (particularly frequencies up to 400 Hz). This is a result of the poor quality of the DOA-based initial masks in this frequency range, as can also be seen in the example of Fig. 4. Whereas the oracle initialization enables a more effective suppression, the resulting masks still do not capture the target speech very well at low frequencies, which can be explained with the difficulty of separating components based on spatial signal characteristics when the phase and level differences between the microphones are small. Here, it seems that this dissimilarity in the generated masks favors the oracle initialization (due to more interference and noise suppression at the cost of an increased target speech distortion) in terms of ΔPESQ, and the DOA-based initialization in terms of ΔSTOI. For higher frequencies (above 400 Hz), however, the produced masks are very similar.

5.1.4 Robustness to DOA estimation errors

Table 3 shows the *difference* compared to the results in Table 2 when *estimated* DOAs are used i.e., negative numbers indicate a poorer performance due to erroneous

Table 3 Results with *estimated* DOAs in terms of the *difference* compared to the oracle DOA results displayed in Table 2. Here, the numbers printed in bold indicate when there is the most significant deterioration. Because $K = D + 1$ direction-based components are used, DOA errors here have an impact on the results regardless of the choice of STAs and initialization. On the other hand, based on the minor differences *between* the rows, we conclude that the incorporation of DOA-based a priori knowledge is quite robust to DOA errors

row	STAs	init	mixture SNR = 30 dB				mixture SNR = 5 dB				
			ΔSDR	ΔSIR	ΔSTOI	ΔPESQ	ΔSDR	ΔSIR	ΔSNR	ΔSTOI	ΔPESQ
0D)	prop.	prop.	−0.3 dB	−0.7 dB	**−0.02**	−0.04	**−0.5 dB**	−1.2 dB	0.1 dB	**−0.03**	−0.02
0B)			−0.3 dB	−0.7 dB	**−0.02**	−0.05	**−0.5 dB**	−1.2 dB	**−0.1 dB**	−0.02	−0.03
1D)	prop.	oracle	−0.3 dB	−0.6 dB	**−0.02**	−0.04	−0.3 dB	−1.1 dB	0.2 dB	−0.02	−0.03
1B)			−0.3 dB	−0.7 dB	**−0.02**	−0.05	−0.4 dB	−1.1 dB	**−0.1 dB**	−0.02	**−0.05**
2D)	prop.	rand.	−0.3 dB	−0.7 dB	−0.01	−0.03	−0.4 dB	−1.1 dB	0.1 dB	−0.02	−0.02
2B)			−0.3 dB	−0.8 dB	−0.01	−0.04	**−0.5 dB**	−1.1 dB	**−0.1 dB**	−0.02	−0.03
3D)	oracle	prop.	−0.3 dB	−0.8 dB	**−0.02**	−0.04	−0.4 dB	−1.3 dB	0.2 dB	**−0.03**	−0.02
3B)			**−0.4 dB**	−0.8 dB	**−0.02**	−0.05	**−0.5 dB**	−1.3 dB	0.0 dB	−0.02	−0.03
4D)	oracle	oracle	−0.3 dB	−0.6 dB	**−0.02**	−0.04	−0.2 dB	−1.0 dB	0.2 dB	−0.02	−0.03
4B)			**−0.4 dB**	−0.7 dB	**−0.02**	−0.05	−0.4 dB	−0.9 dB	**−0.1 dB**	−0.02	−0.04
5D)	oracle	rand.	−0.3 dB	−0.8 dB	−0.01	−0.03	−0.4 dB	**−1.5 dB**	0.3 dB	**−0.03**	−0.03
5B)			**−0.4 dB**	−0.8 dB	−0.01	−0.03	**−0.5 dB**	−1.4 dB	−0.0 dB	−0.02	−0.04
6D)	none	prop.	**−0.4 dB**	−0.7 dB	**−0.02**	−0.04	−0.4 dB	−1.2 dB	0.3 dB	**−0.03**	−0.03
6B)			**−0.4 dB**	−0.8 dB	**−0.02**	**−0.06**	**−0.5 dB**	−1.2 dB	**−0.1 dB**	−0.02	−0.04
7D)	none	oracle	−0.3 dB	−0.5 dB	**−0.02**	−0.04	−0.2 dB	−1.1 dB	0.3 dB	−0.02	−0.03
7B)			**−0.4 dB**	−0.5 dB	**−0.02**	−0.05	−0.4 dB	−1.1 dB	−0.0 dB	−0.02	−0.04
8D)	none	rand.	−0.3 dB	−1.0 dB	−0.01	−0.02	−0.1 dB	−0.8 dB	0.6 dB	−0.02	−0.01
8B)			**−0.4 dB**	**−1.1 dB**	−0.01	−0.02	−0.3 dB	−0.8 dB	0.2 dB	−0.01	−0.02

DOAs. For the considered conditions, the DOA error statistics are visualized in Fig. 9. The angular error is $\Delta\varphi \geq 10°$ in about 6% of the frames at SNR = 30 dB, and in about 9% of the frames at SNR = 5 dB.

The most considerable effect on the results comes from using the DOAs to assign (for each frame) which mixture component corresponds to which speaker. As a result, ΔSTOI deteriorates by −0.02 and ΔPESQ by −0.04 even when the oracle mask-based STAs and initialization are used (row 4, noisy conditions). The sensitivity of the DOA-based components to DOA estimation errors is controlled by the selected angular resolution (20° in this experiment). A finer resolution theoretically enables sources to be separated at a closer spacing, but increases the reliance on accurate DOA estimates. The proposed DOA-based STAs and initialization (row 0), in contrast, are quite robust to DOA estimation errors: the impact on the performance is only marginally higher than in row 4.

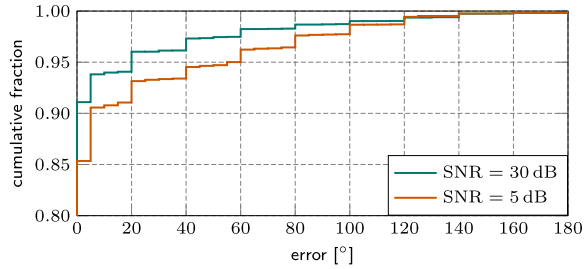

Fig. 9 Normalized cumulative histogram of the absolute DOA estimation error. Integer multiples of 20° are more common because of our setup, where only angles $\varphi \in \{0°, 20°, \ldots, 180°\}$ are available for the true source locations

Generally, we observe that particularly the ΔSIR score is affected by the imperfect source localization (with differences of up to 1.5 dB). This is to be expected, given that the DOAs essentially *define* target and interferers. Based on the ΔSNR metric, on the other hand, we conclude that the suppression of additive noise is not affected. The influence on the other metrics (ΔSDR, ΔSTOI, and ΔPESQ) is moderate because these account for all signal components.

5.1.5 Audio example

Audio files for one particular example (mixture SNR = 5 dB) are available at (Additional file 1)[1]. The corresponding azimuth angles of arrival (true and estimated) are

[1] https://users.ugent.be/~abohlend/DOA-GSS/

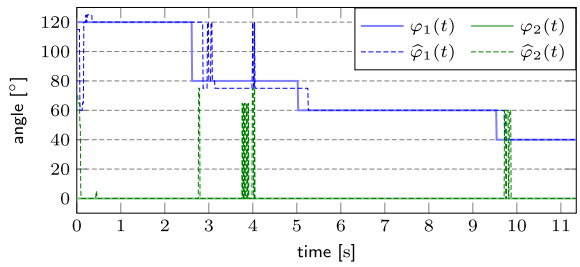

Fig. 10 Ground truth (solid lines) and estimated (dashed) azimuth angles of arrival for the considered audio example. Major errors occur especially after DOA changes and during speech pauses

shown in Fig. 10. At least for the first speaker, the output signal does not change fundamentally depending on the selected STAs and initialization. Rather, mask estimation errors that can manifest in the form of clearly audible artifacts occur in local time-frequency regions. Whereas the outputs again differ chiefly at low frequencies, where it is difficult to distinguish the signal components based on spatial information, some deviations can be observed across the entire spectrum. When (oracle or DOA-based) STAs or initial masks are used, the described mask errors become less common. However, as the comparison of oracle and DOA-based a priori information based on Table 2 has demonstrated, the benefit of an increasing quality of the incorporated prior knowledge saturates at some point.

Because the location of the second speaker is static in this case, the corresponding DTAs are not very useful, as in the example of Fig. 6. Consequently, due to frequency permutation errors, the differences between the output signals generated for various selections of STAs and initialization are more pronounced than for the first speaker. The proposed DOA-based initialization, in particular, remains sufficient to prevent the occurrence of permutation errors.

For a setup which, due to sudden changes of otherwise static DOAs, favors an approach with DOA specific components, we conclude that an initialization using DOA-based masks, combined with the DTAs, delivers the best results. The STAs, in contrast, are then not needed. Additionally, the DOA-GSS proves to be relatively robust to DOA inaccuracies, especially with regard to the use of the DOAs to obtain initial masks and STAs. A relevant deterioration is only observed because the DOA-based components are assigned to each speaker based on the respective DOA estimates.

5.2 Gradually moving speakers

5.2.1 Experimental setup

In the following, we consider a scenario where two speakers are simultaneously active (2 concatenated utterances per speaker, about 4.6 s in total), but one speaker moves around the array such that the corresponding azimuth angle of arrival changes linearly over time. For this setup, it is less straightforward to define time annotations that unambiguously identify each of the components. These conditions are, therefore, also suitable for comparing DTAs and STAs. Specifically, (a combination of) the following techniques can be used to address the frequency permutation problem: (i) incorporating the initial mask-based STAs, (ii) producing DTAs by using one cACG component for each discrete direction rather than each speaker, (iii) using appropriate initial masks, and (iv) performing a manual permutation alignment after the EM algorithm has converged.

We consider the moving and the static speaker to be the target and the interferer, respectively. An important parameter in the described scenario is the length of the trajectory of the target speaker during the signal i.e., the total movement in terms of the azimuth angle φ. On the one hand, if the speaker is (almost) static for the entire signal duration, no information can be gained from the DTAs (see Fig. 6). On the other hand, a large movement may be challenging for the speaker-based approach ($K = J + 1$ components, DTAs are unavailable), because the spatial signal characteristics change significantly over the course of the signal, as well as for the DOA-based approach ($K = D + 1$ components, DTAs are available), because less data are available to determine the optimal model parameters for each component.

Therefore, we consider the results as a function of the total movement. For this purpose, we use simulated microphone signals, where the contributions of the 2 speakers have been obtained with the signal generator [48], which makes use of the image source method [49]. In the simulation, the room dimensions are 6.0 m × 5.0 m × 2.7 m, with a reverberation time of $T_{60} = 0.5$ s. The microphone array, which is arranged in a plane that is parallel to the ground, is positioned near the center of the room, at a height of 1 m. Initially, the speakers are located in a distance of ±1.5 m from the array in x-direction. The height of the sound sources used to represent the speakers is 1.5 m at all times. Thus, the fixed azimuth angle of arrival of the (static) interferer is $\varphi = 180°$, whereas the target speaker moves on an arc towards the interfering speaker starting at $\varphi = 0°$. Empirically, we choose a resolution of 30° for the DOA-based components (only relevant when $K = D + 1$).

The setup is otherwise unchanged compared to Section 5.1.1. For conciseness, we only consider noisy conditions (SNR = 5 dB) with mask-based MVDR beamforming, and use the estimated DOAs. To obtain an upper bound reference, the permutation alignment, when enabled, is performed by selecting (for each frequency) the permutation that minimizes the mean squared error

(MSE) between the estimated masks and the ideal masks (Eq. 34). The scaling factor γ_j is omitted in this case. Its incorporation would attenuate the ideal mask for the noise component, which leads to unexpected permutations where the MSE is minimized by using the noise component for one of the speakers at some frequencies.

5.2.2 Evaluation

The achieved ΔSTOI and ΔPESQ scores with regard to the target speaker are displayed in Fig. 11. First, we consider the case where no DTAs are available (only $K = J + 1 = 3$ components that correspond to the two speakers and the noise, respectively), and no manual permutation alignment is performed (first row in the figure).

With the random initialization, the use of STAs again leads to improved results. However, the improvement is mostly below 0.10 in terms of both ΔSTOI and ΔPESQ even when the oracle STAs are used. The reason for the comparatively poor performance is that the STAs alone are insufficient to fully resolve the permutation problem for $P = 10$. Whereas it would be possible to further increase P, a different approach for addressing the permutation problem may be preferred to avoid adding to the target speech distortion.

Combined with the proposed DOA-based initialization, the STAs are again no longer useful. This suggests that the considered initial masks alone are sufficient to address the permutation problem, so that no time annotations are needed in addition. Further, when a manual (oracle) permutation alignment is performed (second row), we observe that it is even detrimental to include STAs. This is because they provide no added benefit when the permutation can be resolved correctly anyway, but the increased target speech distortion inherent to the incorporation of these annotations can lead to a poorer speech quality.

As the results in Section 5.1 have already shown, the DOA-based initialization improves the performance considerably compared to random initialization, especially when no manual permutation alignment is performed (first row). The difference remains evident even after the permutation alignment (second row) e. g., for a total movement of 80°, no time annotations: an additional 0.03 in terms of ΔSTOI, 0.06 in terms of ΔPESQ.

Moreover, the results indicate that the DOA-based initial masks deliver a seemingly better performance than the oracle masks in terms of ΔSTOI. As previously noted in the context of similar trends observed in Table 2, this is primarily related to the different behavior at the lower end of the spectrum (especially frequencies up to 400 Hz). The produced masks are otherwise similar except for occasional permutation errors.

Regardless of the need for time annotations, it may be reasonable to use direction-based (instead of speaker-based) components e. g., when two different sources are located in the same direction at different times. Since the cACGMM itself is time-invariant, the resulting similarity of the spatial signal characteristics could be problematic for an approach with speaker specific components. Although an evaluation of this scenario is beyond the scope of this work, the evaluation setup considered here still permits assessing the practicability of an approach with direction specific components. The corresponding results are shown in the third and fourth row of Fig. 11, without and with manual permutation alignment, respectively.

To ensure that the DTAs are meaningful, the speaker must move sufficiently far to be covered by at least two different components given the selected resolution of 30°. Even for random initialization and without permutation alignment, the improvement under these conditions is only marginal, however. Upon examining the results more closely, we find that this is because, for a signal duration of no more than 2 utterances, the performance is still strongly dependent on the amount of data available to determine the optimal parameters for each mixture component. However, the more components are needed to encompass the entire trajectory of the speaker with a fixed angular resolution, the further the signal is subdivided into short segments. As a result of the inherent increase of the degrees of freedom in the mixture model parameter estimation, the produced final masks increasingly resemble the employed initial masks. This limits the performance resulting from random initialization, in particular, but also causes a degradation of the results obtained with the proposed DOA-based initialization the further the speaker moves.

The problem also becomes apparent when looking at Fig. 12, which shows a different representation of the same results, where the DOA-based initialization is used for all configurations, but the STAs are omitted. Clearly, the DOA-GSS performs best under the considered conditions when the components are speaker-based (no DTAs). For a moving speaker trajectory covering 120°, without permutation alignment, the difference is 0.07 in terms of ΔSTOI and 0.06 in terms of ΔPESQ. Thus, we find that a DOA-based subdivision of the signal into multiple segments is only sensible when the cACGMM is used to describe a longer signal, where each of the resulting segments retains a length of several seconds. In practice, this can be achieved e. g., by adaptively selecting an appropriate resolution based on the considered signal and the corresponding (estimated) DOAs.

Finally, based on Fig. 12, we can determine whether it is still beneficial to apply an additional manual permutation alignment when the proposed initialization is used. Again, ΔPESQ and ΔSTOI paint a contradicting picture. Similar

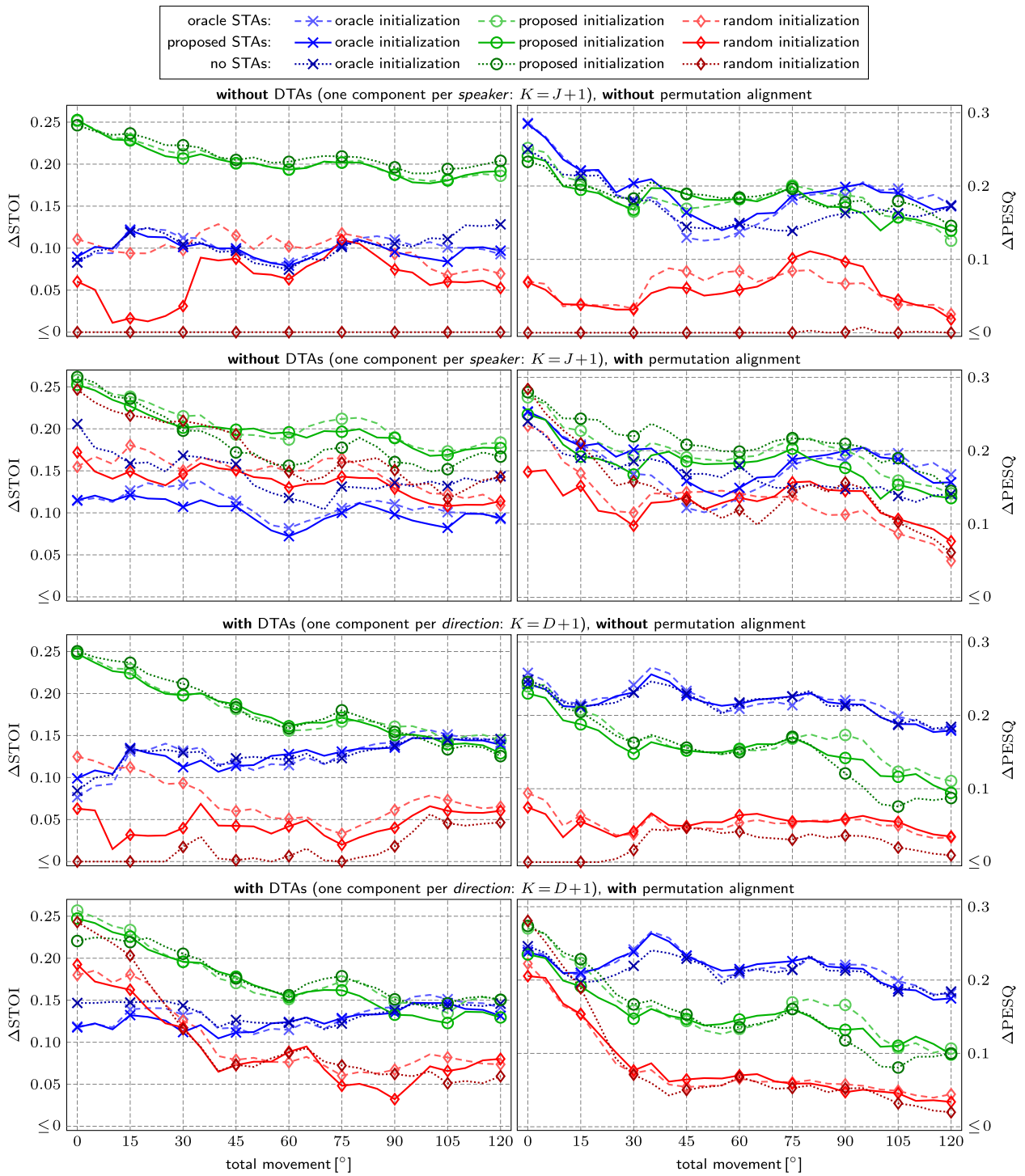

Fig. 11 ΔSTOI and ΔPESQ scores for a moving speaker as a function of the total movement over the entire signal duration in terms of the azimuth angle φ. To obtain a clearer representation, a moving average was applied (averaging over all results that deviate by up to ±10° compared to the values indicated on the *x*-axis), and negative results were replaced by 0. For this setup, too, the initialization remains the most effective way to benefit from DOA information, rendering time annotations (STAs or DTAs) as well as manual permutation alignment unnecessary

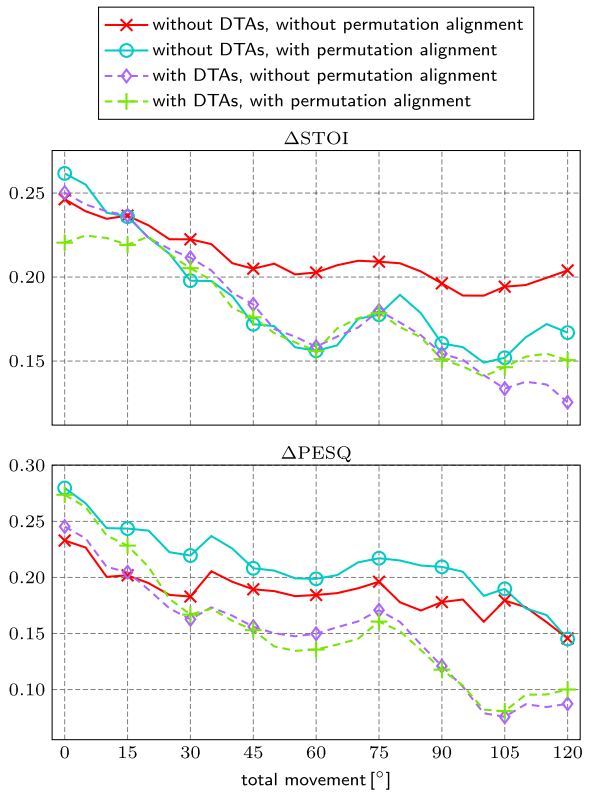

Fig. 12 Subset of the results from Fig. 11 (proposed initialization, omit STAs). The performance of the approach with DTAs deteriorates with an increasing length of the speaker trajectory due to the lack of data to determine optimal parameters for all direction specific mixture components

to the oracle initialization, the oracle permutation alignment leads to a stronger suppression at low frequencies, which appears to be favorable in terms of ΔPESQ, but deteriorates the ΔSTOI score. When comparing the spectra of the separated signals, we note that the differences at higher frequencies, in contrast, are marginal.

To conclude, the availability of estimated source DOAs can be exploited to derive initial masks which make time annotations and permutation alignment unnecessary. An approach with DOA specific components may be of interest e. g., for sources with overlapping trajectories, or when a greater number of sources intensifies the permutation problem. Additionally, it could be practical for the purpose of only extracting sources in a specified target direction. However, in the selection of the corresponding angular resolution, it must be taken into account that the performance clearly deteriorates when mixture components are optimized based on signal segments that are not at least a few seconds long. The STAs, on the other hand, provided no added benefit compared to the initial masks that they are derived from, but could be used to enforce a stronger noise suppression.

5.3 Performance in different conditions

To conclude the experiments, it is evaluated in this section how the performance of the DOA-GSS is dependent on the experimental conditions. We consider the setup of Section 5.1.1, where the speaker locations are static for the duration of an utterance. That being the case, each of the corresponding DOA-based mixture components is active for a reasonably long time, so that we can make use of the approach with DTAs. The angular resolution is again 20°. Given the findings from previous experiments, the proposed DOA-based initialization is used, but the STAs are omitted.

The Mix-mask estimator (Mix-MEst) approach proposed in [6] is considered as a baseline. Using the spatial information given by the microphone signal phases, the employed CNN produces TF masks for each of 72 discrete directions $\varphi \in \{0°, 5°, \ldots, 355°\}$, for the purpose of extracting a hypothetical source from any one direction. The DOA estimates are then used to select the right mask for each of the J sources. Thus, the approach is also DOA-based, but the DOA information is not taken into account in the mask estimation itself. This puts it at a disadvantage compared to e. g., the DOA-GSS, where DOA estimates are available as prior information. Nevertheless, it is interesting to consider Mix-MEst as a reference, since the DOAs are also used to define which part of the signal to extract.

The ΔSTOI and ΔPESQ scores are shown in Fig. 13 as a function of various parameters specifying the experimental conditions. In the first row, the mixture SNR is varied from −5 dB to 20 dB for an otherwise fixed setup. We used estimated DOAs in the generation of all results, which plays an important role particularly under the most adverse of the considered conditions. In the presence of strong noise, both approaches perform similarly. Mix-MEst is trained to cope with adverse conditions and, as the DOA estimates are only used to select the masks in the end, a higher robustness to DOA estimation errors may be expected. In contrast, when the mixture SNR is higher, we obtain better results with DOA-GSS than with Mix-MEst (e. g., at SNR = 10 dB: ΔSTOI = 0.20 with the DOA-GSS, compared to 0.16 with Mix-MEst when the sources are separated by direct masking).

It is interesting to note that the masks obtained from the DOA-GSS are particularly suitable for mask-based beamforming (solid lines). With Mix-MEst, better ΔSTOI and ΔPESQ scores are typically obtained by applying the masks directly (dotted lines), whereas the MVDR beamforming approach significantly increases the ΔPESQ scores when the DOA-GSS is used. This can also be seen in the second row of the figure, where the mixture SIR is varied from −15 dB to +15 dB, and in the third row, where different source-array distances are considered. For example, ΔPESQ increases from 0.13 to 0.24 when the

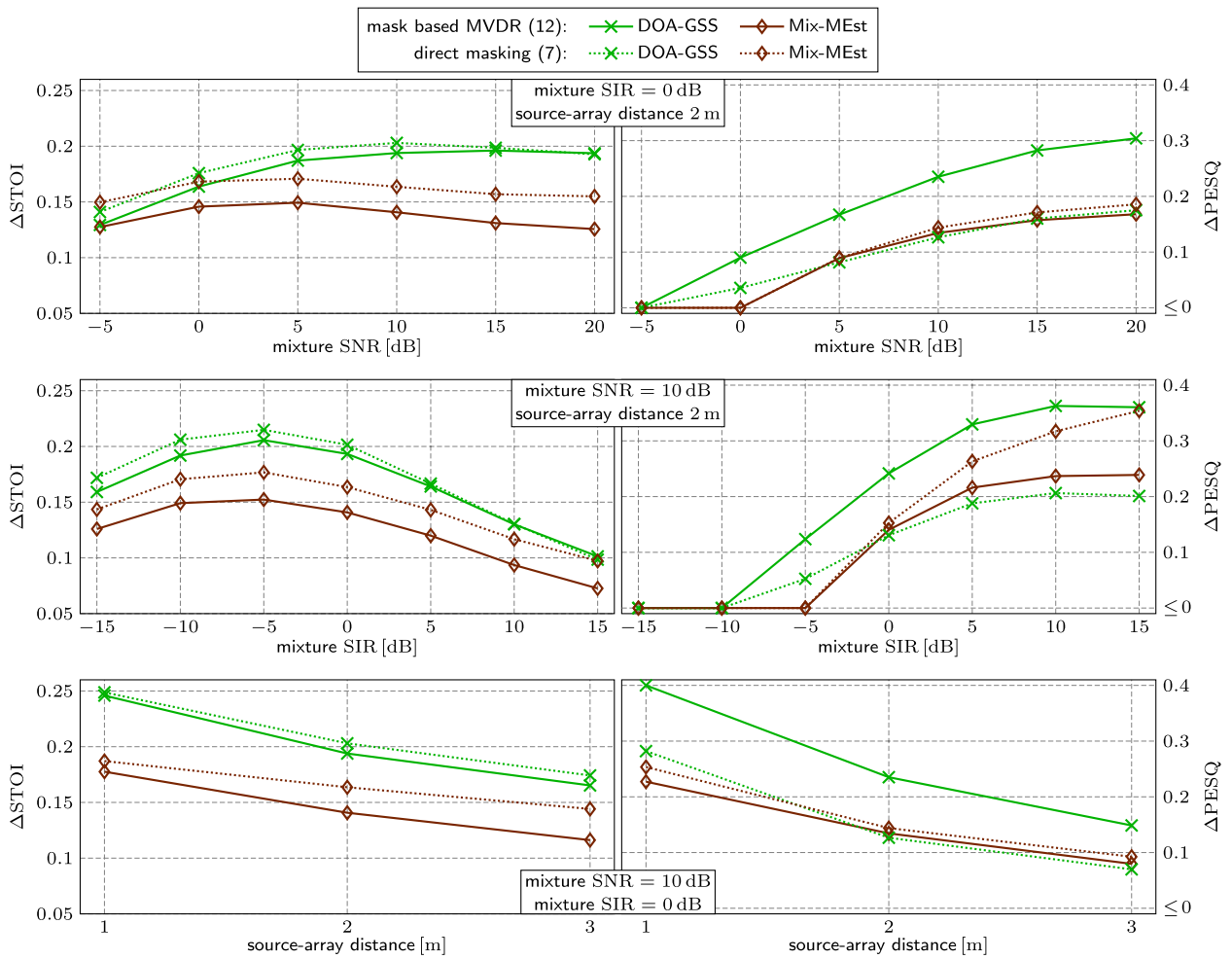

Fig. 13 Performance of the DOA-GSS and the neural network-based Mix-MEst approach of [6] as a function of the mixture SNR (top), SIR (middle), and the source-array distance (bottom). For the DOA-GSS, the DOA-based initialization and the DTAs are used, but the STAs are omitted

beamformer is used instead of direct masking for the DOA-GSS at SIR = 0 dB and SNR = 10 dB, but Mix-MEst achieves around 0.15 in both cases. Note that for a distance of 3 m, the size of the room restricted us to the recording of RIRs for angles $\varphi \in \{40°, 60°, \ldots, 140°\}$, so that the spacing between the sources is smaller on average.

Generally, we observe a favorable robustness of the DOA-GSS to adverse conditions, provided that the DOA estimation does not break down completely. Although the improvement is no longer reflected in the ΔPESQ score when the signal is dominated by unwanted components, the results remain decent in terms of ΔSTOI, which shows that the speakers can still be separated. For a mixture SNR of −5 dB, the improvement compared to the noisy mixture is still ΔSTOI = 0.14 (SIR = 0 dB, 2 m distance), 0.17 for SIR = −15 dB (SNR = 10 dB, 2 m distance), and 0.17 for a source-array distance of 3 m (SNR = 10 dB, SIR = 0 dB).

6 Conclusions

We compared various methods to take advantage of DOA estimates in probabilistic mixture model-based TF mask estimation for source separation. These clustering approaches suffer from the sensitivity to the initialization of the iterative model parameter estimation, and the need to address the frequency permutation problem. Therefore, incorporating additional information is helpful to fully exploit the potential of the approach.

Specifically, we considered the previously proposed GSS, which models the directional statistics of the microphone array signals by a cACGMM. The need for a permutation alignment is avoided by means of a tight integration of annotations that indicate when each speaker is active. To this end, we proposed to derive suitable STAs from simple DOA-based initial masks. Whereas experiments verify that these limit the occurrence of

permutation errors, an increased distortion of the target signals is observed as well.

In contrast, the weak integration by means of an initialization of the EM algorithm with the same DOA-based masks was found to be sufficient to address both of the described shortcomings. Compared to an ideal initialization and permutation alignment, significant deviations were only observed at low frequencies, where the lack of reliable spatial information, as given by the phase and level differences between the microphones, prevents high-quality results.

Finally, we considered the use of DOA-based components, where the correct component for each speaker is selected depending on their current location. Whereas this represents an alternative to acquire annotations, increasing the number of mixture components also implies that less data are available to determine the optimal model parameters for each individual component. To make better use of this approach, and to enable a realtime application thereof, an adaptive strategy where DOA specific components are updated continuously may therefore be considered in future work.

Abbreviations

cACG: Complex angular central Gaussian; cACGMM: Complex angular central Gaussian mixture model; cGMM: Complex Gaussian mixture model; CNN: Convolutional neural network; DFT: Discrete Fourier transform; DNN: Deep neural network; DOA: Direction of arrival; DTA: DOA time annotation; EM: Expectation-maximization; GSS: Guided source separation; LSTM: Long short-term memory; MEst: Mask estimator; MSE: Mean squared error; MUSIC: Multiple signal classification; MVDR: Minimum variance distortionless response; PESQ: Perceptual evaluation of speech quality; PIT: Permutation invariant training; PSD: Power spectral density; RIR: Room impulse response; SDR: Source-to-distortion ratio; SIR: Source-to-interferences ratio; SNR: Sources-to-noise ratio; SRP: Steered response power; STA: Source time annotation; STFT: Short-time Fourier transform; STOI: Short-time objective intelligibility measure; TF: Time-frequency; URA: Uniform rectangular array; WPE Weighted prediction error

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13636-022-00246-7.

Additional file 1: The audio examples have been enclosed as supplementary material with the submission. The files are identical to those accessible at https://users.ugent.be/~abohlend/DOA-GSS/. Along with the speaker ID, the file name indicates which type of STAs and initialization were used, and how the mask was applied.

Acknowledgements

Not applicable.

Authors' contributions

AB and NM contributed to motivating the study and defining the research questions addressed therein. AB developed and implemented the method, and carried out the experiments and analysis. LVS and JS supported the experiments and performed hyperparameter tuning. AB and NM contributed to the written manuscript. The authors read and approved the final manuscript.

Funding

This work is supported by the Research Foundation - Flanders (FWO) under grant numbers 11G0721N and G081420N.

Availability of data and materials

The clean speech files used in the generation of the analyzed microphone signals are taken from the TSP database [36], which is available at http://www-mmsp.ece.mcgill.ca/Documents/Data/. The diffuse noise is based on the pub noise signal from the ETSI background noise database [39], which is available at https://docbox.etsi.org/stq/Open/EG%20202%20396-1%20Background%20noise%20database. The corresponding diffuse noise recording, as well as the recorded impulse responses, are available from the corresponding author on reasonable request. Moving speakers were simulated using the signal generator [48], which is available at https://github.com/ehabets/Signal-Generator.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 3 January 2022 Accepted: 14 May 2022
Published online: 18 June 2022

References

1. S. Rickard, O. Yilmaz, in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. On the approximate W-disjoint orthogonality of speech, vol. 1, (2002), pp. 529–532. https://doi.org/10.1109/ICASSP.2002.5743771
2. D. Yu, M. Kolbæk, Z.-H. Tan, J. Jensen, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Permutation invariant training of deep models for speaker-independent multi-talker speech separation, (2017), pp. 241–245. https://doi.org/10.1109/ICASSP.2017.7952154
3. Y. Yu, W. Wang, P. Han, Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks. EURASIP J. Audio Speech Music Process. **1**, 1–18 (2016). https://doi.org/10.1186/s13636-016-0085-x
4. S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, S. Gannot, in *Proc. 27th European Signal Processing Conference (EUSIPCO)*. Multi-microphone speaker separation based on deep DOA estimation, (2019), pp. 1–5. https://doi.org/10.23919/EUSIPCO.2019.8903121
5. Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, Y. Gong, in *Proc. IEEE Spoken Language Technology Workshop (SLT)*. Multi-channel overlapped speech recognition with location guided speech extraction network, (2018), pp. 558–565. https://doi.org/10.1109/SLT.2018.8639593
6. A. Bohlender, A. Spriet, W. Tirry, N. Madhu, in *Proc. 29th European Signal Processing Conference (EUSIPCO)*. Neural networks using full-band and subband spatial features for mask based source separation, (2021), pp. 346–350. https://doi.org/10.23919/EUSIPCO54536.2021.9616138
7. A. Aroudi, S. Braun, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DBnet: Doa-driven beamforming network for end-to-end reverberant sound source separation, (2021), pp. 211–215. https://doi.org/10.1109/ICASSP39728.2021.9414187
8. J. R. Hershey, Z. Chen, J. Le Roux, S. Watanabe, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep clustering: Discriminative embeddings for segmentation and separation, (2016), pp. 31–35. https://doi.org/10.1109/ICASSP.2016.7471631
9. Z.-Q. Wang, J. Le Roux, J. R. Hershey, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation, (2018), pp. 1–5. https://doi.org/10.1109/ICASSP.2018.8461639
10. Z. Chen, Y. Luo, N. Mesgarani, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep attractor network for single-microphone speaker separation, (2017), pp. 246–250. https://doi.org/10.1109/ICASSP.2017.7952155

11. N. Ito, S. Araki, T. Nakatani, in *Proc. 24th European Signal Processing Conference (EUSIPCO)*. Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing, (2016), pp. 1153–1157. https://doi.org/10.1109/EUSIPCO.2016.7760429
12. H. Sawada, S. Araki, S. Makino, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. Audio Speech Lang. Process. **19**(3), 516–527 (2011)
13. N. Ito, S. Araki, T. Nakatani, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors, (2013), pp. 3238–3242. https://doi.org/10.1109/ICASSP.2013.6638256
14. J. Azcarreta, N. Ito, S. Araki, T. Nakatani, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Permutation-free cGMM: Complex gaussian mixture model with inverse wishart mixture model based spatial prior for permutation-free source separation and source counting, (2018), pp. 51–55. https://doi.org/10.1109/ICASSP.2018.8461934
15. C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, R. Haeb-Umbach, in *Proc. 5th International Workshop on Speech Processing in Everyday Environments (CHiME)*. Front-end processing for the chime-5 dinner party scenario, (2018). https://doi.org/10.21437/CHiME.2018-8
16. L. Drude, D. Hasenklever, R. Haeb-Umbach, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Unsupervised training of a deep clustering model for multichannel blind source separation, (2019), pp. 695–699. https://doi.org/10.1109/ICASSP.2019.8683520
17. T. Nakatani, R. Takahashi, T. Ochiai, K. Kinoshita, R. Ikeshita, M. Delcroix, S. Araki, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation, (2020), pp. 6399–6403. https://doi.org/10.1109/ICASSP40776.2020.9053343
18. T. Nakatani, N. Ito, T. Higuchi, S. Araki, K. Kinoshita, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming, (2017), pp. 286–290
19. L. Drude, R. Haeb-Umbach, Integration of neural networks and probabilistic spatial models for acoustic blind source separation. IEEE J. Sel. Top. Signal Process. **13**(4), 815–826 (2019). https://doi.org/10.1109/JSTSP.2019.2912565
20. D. H. T. Vu, R. Haeb-Umbach, in *Proc. 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. On initial seed selection for frequency domain blind speech separation, (2011), pp. 1757–1760. https://doi.org/10.21437/Interspeech.2011-494
21. Y. Bando, Y. Sasaki, K. Yoshii, in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. Deep bayesian unsupervised source separation based on a complex gaussian mixture model, (2019), pp. 1–6. https://doi.org/10.1109/MLSP.2019.8918699
22. J. Barker, S. Watanabe, E. Vincent, J. Trmal, in *Proc. 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines, (2018), pp. 1561–1565. https://doi.org/10.21437/Interspeech.2018-1768
23. H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, K. Shikano, Blind source separation based on a fast-convergence algorithm combining ICA and beamforming. IEEE Trans. Audio Speech Lang. Processing. **14**(2), 666–678 (2006). https://doi.org/10.1109/TSA.2005.855832
24. J. Heymann, L. Drude, R. Haeb-Umbach, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Neural network based spectral mask estimation for acoustic beamforming, (2016), pp. 196–200. https://doi.org/10.1109/ICASSP.2016.7471664
25. T. Higuchi, N. Ito, T. Yoshioka, T. Nakatani, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise, (2016), pp. 5210–5214. https://doi.org/10.1109/ICASSP.2016.7472671
26. P. Vary, R. Martin, *Digital Speech Transmission - Enhancement, Coding & Error Concealment*. (Wiley, Chichester, 2006)
27. M. Souden, J. Benesty, S. Affes, On optimal frequency-domain multichannel linear filtering for noise reduction. IEEE Trans. Audio Speech Lang. Process. **18**(2), 260–276 (2010). https://doi.org/10.1109/TASL.2009.2025790
28. N. Madhu, R. Martin, in *Advances in Digital Speech Transmission*, ed. by R. Martin, U. Heute, and C. Antweiler. Acoustic source localization with microphone arrays (Wiley, New York, 2008), pp. 135–170
29. J. H. DiBiase, A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays. PhD thesis, Brown University, Providence, RI, USA (2000)
30. M. Cobos, J. J. Lopez, D. Martinez, Two-microphone multi-speaker localization based on a Laplacian mixture model. Digit. Signal Process. **21**(1), 66–76 (2011). https://doi.org/10.1016/j.dsp.2010.04.003
31. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986). https://doi.org/10.1109/TAP.1986.1143830
32. P.-A. Grumiaux, S. Kitić, L. Girin, A. Guérin, A survey of sound source localization with deep learning methods (2021). arXiv:2109.03465. http://arxiv.org/abs/2109.03465. Accessed 28 May 2022
33. G. Strang, *Introduction to Linear Algebra, 5th ed.* (Wellesley-Cambridge Press, Wellesley, 2016)
34. I. A. McCowan, H. Bourlard, Microphone array post-filter based on noise field coherence. IEEE Trans. Speech Audio Process. **11**(6), 709–716 (2003)
35. R. Zelinski, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms, (1988), pp. 2578–2581. https://doi.org/10.1109/ICASSP.1988.197172
36. P. Kabal, TSP speech database. Technical report, McGill University, Montreal, Quebec, Canada (2002)
37. E. H. Rothauser, W. D. Chapman, *et al*, IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoustics. **17**(3), 225–246 (1969). https://doi.org/10.1109/TAU.1969.1162058
38. miniDSP, UMA-16 USB microphone array. https://www.minidsp.com/products/usb-audio-interface/uma-16-microphone-array. Accessed22 Feb 2022
39. European Telecommunications Standards Institute, *Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database*. (Standard, ETSI ES 202 396-1, 2005). Current version (V1.8.1, published in 2022). https://www.etsi.org/deliver/etsi_es/202300_202399/20239601/01.08.01_60/es_20239601v010801p.pdf. Accessed 28 May 2022
40. T. Yoshioka, T. Nakatani, Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. IEEE Trans. Audio Speech Lang. Process. **20**(10), 2707–2720 (2012). https://doi.org/10.1109/TASL.2012.2210879
41. L. Drude, J. Heymann, C. Boeddeker, R. Haeb-Umbach, in *Proc. 13th ITG Conference on Speech Communication*. NARA-WPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing, (2018), pp. 1–5. https://ieeexplore.ieee.org/document/8578026
42. R. Haeb-Umbach, et al., Blind Source Separation (BSS) algorithms. https://github.com/fgnt/pb_bss. Accessed 21 May 2021
43. A. Bohlender, A. Spriet, W. Tirry, N. Madhu, Exploiting temporal context in CNN based multisource DOA estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 1594–1608 (2021)
44. S. Chakrabarty, E. A. P. Habets, Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. IEEE J. Sel. Top. Signal Process. **13**(1), 8–21 (2019). https://doi.org/10.1109/JSTSP.2019.2901664
45. C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011). https://doi.org/10.1109/TASL.2011.2114881
46. International Telecommunication Union, *Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs*. (Standard, ITU-R P.862.2, Geneva, 2007). https://www.itu.int/rec/T-REC-P.862.2-200711-I/en. Accessed 28 May 2022
47. E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
48. E. A. P. Habets, Signal Generator. https://github.com/ehabets/Signal-Generator. Accessed 28 Oct 2021
49. J. B. Allen, D. A. Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979). https://doi.org/10.1121/1.382599