Thakallapalli *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:13
https://doi.org/10.1186/s13636-021-00201-y

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

NMF-weighted SRP for multi-speaker direction of arrival estimation: robustness to spatial aliasing while exploiting sparsity in the atom-time domain

Sushmita Thakallapalli[1*], Suryakanth V. Gangashetty[1,2] and Nilesh Madhu[3]

Abstract

Localization of multiple speakers using microphone arrays remains a challenging problem, especially in the presence of noise and reverberation. State-of-the-art localization algorithms generally exploit the sparsity of speech in some representation for this purpose. Whereas the broadband approaches exploit time-domain sparsity for multi-speaker localization, narrowband approaches can additionally exploit sparsity and disjointness in the time-frequency representation. Broadband approaches are robust to spatial aliasing but do not optimally exploit the frequency domain sparsity, leading to poor localization performance for arrays with short inter-microphone distances. Narrowband approaches, on the other hand, are vulnerable to spatial aliasing, making them unsuitable for arrays with large inter-microphone spacing. Proposed here is an approach that decomposes a signal spectrum into a weighted sum of *broadband* spectral components (atoms) and then exploits signal sparsity in the *time-atom* representation for simultaneous multiple source localization. The decomposition into atoms is performed in situ using non-negative matrix factorization (NMF) of the short-term amplitude spectra and the localization estimate is obtained via a broadband steered-response power (SRP) approach for each active atom of a time frame. This SRP-NMF approach thereby combines the advantages of the narrowband and broadband approaches and performs well on the multi-speaker localization task for a broad range of inter-microphone spacings. On tests conducted on real-world data from public challenges such as SiSEC and LOCATA, and on data generated from recorded room impulse responses, the SRP-NMF approach outperforms the commonly used variants of narrowband and broadband localization approaches in terms of source detection capability and localization accuracy.

Keywords: Sound source localization, Direction-of-arrival, Non-negative matrix factorization, Spatial aliasing, Speech sparsity

1 Introduction

Speech remains the natural mode of interaction for humans. Present day smart-home devices are, therefore, increasingly equipped with voice controlled personal assistants to exploit this for human-machine interfacing. The performance of such devices depends, to a large extent, on the performance of the localization techniques used in these systems. The term localization in this context implies the detection and spatial localization of a number of overlapping speakers, and it is usually the first stage in many speech communication applications. Accurate acoustic localization of multiple active speakers, however, remains a challenging problem—especially in the presence of background noise and room reverberation.

Localization is typically achieved by means of the spatial diversity afforded by microphone arrays. Large

*Correspondence: sushmita.t@research.iiit.ac.in
[1]Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India
Full list of author information is available at the end of the article

microphone arrays (inter-microphone spacing in the order of a meter) sample the sound fields at large spatial intervals, thereby reducing the effect of diffuse background noise in the localization. However, these arrays are increasingly prone to spatial aliasing at higher frequencies. Compact microphone arrays, with inter-microphone spacing of the order of a few centimeters, offer greater robustness to spatial aliasing, but are biased by diffuse background noise. The size of the chosen array is usually a trade-off between these two factors and, further, is often driven by practical considerations.

State-of-the-art algorithms for multi-speaker localization usually exploit the sparsity and *disjointness* [1] of speech signals. While some approaches exploit, mainly, temporal sparsity (i.e., speakers are not concurrently active at all times), others exploit the time-frequency (TF) sparsity (i.e., speakers are not concurrently active at all time and frequency points of the short-time frequency domain representation) of speech. Here, the short-time Fourier transform (STFT) representation is typically chosen because of its computational efficiency. The former approaches are categorized as broadband and the latter as narrowband. For both these approaches, the localization estimates over time and/or frequency are subsequently aggregated to obtain an estimate of the number of active sources and their respective locations.

Frequently used broadband methods are based on the generalized cross-correlation (GCC) [2] and its variants, e.g., the average magnitude difference function (AMDF) estimators [3], the adaptive eigenvalue decomposition approach [4], information theoretic criteria-based approaches [5], and the broadband steered-response power approaches [6]. Such approaches typically localize the *dominant* source in each time segment, thereby exploiting the temporal sparsity induced by natural pauses in speech. The GCC with phase transform (PHAT) weighting has proven to be the most robust among all the GCC weightings in low noise and reverberant environments [7]. However, in GCC-PHAT, the localization errors increase when the signal to noise ratio (SNR) is poor. To address this issue, researchers have proposed SNR-based weights on GCC-PHAT to highlight the speech dominant TF bins and to de-emphasize TF bins with noise or reverberant speech (see, e.g., [8–11]). A performance assessment of various GCC algorithms may be found in [12].

Narrowband frequency domain approaches, on the other hand, use the approximate disjointness of speech spectra in their short-time frequency domain representation to localize the dominant source *at each time-frequency point*. Multi-speaker localization is subsequently done by pooling the individual location estimates. In [13], for example, a (reliability-weighted) histogram is computed on the pooled DoA estimates, and the locations of peaks of the histogram yield the speaker location estimates. In [14], instead of a histogram, a mixture of Gaussians (MoG) model is applied to cluster the time-difference of arrival (TDoA) estimates. The approach of [15] is a generalization of [14] in which speaker *coordinates* are estimated and tracked, rather than speaker TDoAs. Similarly, in [16] the authors propose a MoG clustering of the direction of arrival (DoA) estimates obtained by a narrowband steered response power (SRP) approach. This is extended in [17], where a Laplacian mixture model is proposed for the clustering. In [18], source separation and localization are iteratively tackled: source masks are first estimated by clustering the TDoA estimates at each TF bin and subsequently SRP-PHAT is used to estimate the DoAs of the separated sources. The estimated DoAs are fed back to the cluster tracking approach for updating the cluster centers. Other recent works build upon this basic idea of exploiting the TF sparsity by introducing reliability weights on the time-frequency units before localization such as [19], which uses SNR-based weights, [20], which uses TF weights predicted by neural-networks, and [21], which considers a weighted histogram of the narrowband estimates, where the weights correspond to a heuristic measure of the reliability of the estimate in each TF bin. A comprehensive overview of the relations between the commonly used localization approaches is presented in [22].

When performing source localization independently at each time-frequency point, typical optimization functions for narrowband localization do not yield a unique DoA estimate above a certain frequency. This is due to the appearance of grating lobes, and the phenomenon is termed spatial aliasing. As the distance between the microphones in the array increases, the frequency at which spatial aliasing occurs reduces, leading to ambiguous DoA estimates across a larger band of frequencies. Broadband approaches circumvent this problem by summing the optimization function across the whole frequency band and computing a location estimate per time frame. Such averaging is indicated for arrays with large inter-element spacing. However, this constitutes a promiscuous averaging across frequencies, each of which may be dominated by a different speaker, leading to (weakened) evidence for only the strongest speaker in that time frame—i.e., only the location of the highest peak in the angular spectrum of the frame is considered as a potential location estimate and other peaks are usually ignored, since they may not reliably indicate other active speaker locations [23]. Multiple speaker localization is still possible in such cases by aggregating the results across different *time frames* but, by disregarding the frequency sparsity of speech signals, softer speakers (who may not be dominant for a sufficient number of time frames) may not be localized.

Instead of averaging across the whole frequency range, a compromise can be effected by only averaging across smaller, contiguous *sub-bands* of frequencies and computing a location estimate per time and sub-band region. By pooling the estimates across the various sub-bands, multi-speaker localization may still be achieved. Such bands may be either psycho-acoustically motivated (e.g., the Bark scale used in [24]) or heuristically defined. However, these are fixed frequency groupings and the previously described shortcomings with regard to such groupings still hold. Other approaches [25, 26] try to resolve the spatial aliasing problem by trying to *unwrap* the phase differences of spatially aliased microphone pairs. Initial (rough) estimates of the source locations are required to resolve the spatial aliasing, and it is assumed that at least a few non-aliased microphone pairs are available for this. Consequently, this requires arrays with several microphones at staggered distances such that multiple microphone pairs, aliasing at different frequencies, are available.

The key idea of our approach is to average the narrowband optimization function for localization only across frequency bins that show simultaneous excitation in speech (e.g., fundamental frequency and its harmonics for a voiced speech frame, etc.). Thereby the frequency grouping is not fixed, but data- and time frame dependent. Further, since the averaging is carried out across frequency bins that are simultaneously excited during the speech, the interference from other speakers should be minimal in these bins due to the sparsity and disjointness property. Thus, we can simultaneously exploit the time and frequency sparsity of speech while being robust to spatial aliasing—thereby overcoming the shortcomings of the previously mentioned approaches.

Non-negative matrix factorization (NMF) allows for the possibility to learn such typical groupings of the frequencies based on the magnitude spectrum of the microphone signal. These frequency groupings are termed atoms in our work. Thus we speak of localization based on time-atom sparsity, i.e., in any one time frame only a few atoms are active and each active atom only belongs to one speaker, and localizing across the different atoms in a time frame allows for multi-speaker localization. Since we use the SRP approach for localization, our algorithm is termed the SRP-NMF approach.

The rest of the paper is organized as follows: we first summarize prior approaches utilizing NMF for source localization and place our proposed approach in the context of these works. Next, in Section 3, we describe the signal model, followed by a review of the basic ideas underlying state-of-the-art narrowband and broadband SRP approaches. SRP-NMF is introduced and detailed in Section 5. In Section 6, the approach is thoroughly tested. The details of the databases, the comparison approaches and evaluation metrics, the method used to estimate SRP-NMF parameters, an analysis of the results and limitations of the approach are presented. Finally, we summarize the work and briefly mention the future scope.

2 Prior work using NMF for localization

NMF has previously been used for source localization and separation in several conceptually different ways. For example, in [27], NMF is applied to decompose the SRP-PHAT function (collated across all time-frequency points) into a combination of angular activity and source presence activity. This decomposition assumes unique maxima of the SRP-PHAT function (i.e., no spatial aliasing), allowing for a sparse decomposition using NMF.

In [28], on the other hand, NMF is used to decompose the GCC-PHAT *correlogram matrix* to a low-dimensional representation consisting of bases which are the GCC-PHAT correlation functions for each source location and weights (or activation functions) which determine which time frame is dominated by which speaker. Thus, this approach may be interpreted as a broadband GCC-PHAT approach assuming temporal sparsity. As it is a broadband approach, spatial aliasing is not a problem. However, simultaneous localization of multiple sources within a single time frame is not straightforward.

The approach of [29] is, again, fundamentally different from [27] and [28]. Here, *complex* NMF is used to decompose the multi-channel instantaneous spatial covariance matrix into a combination of weight functions that indicate which locations in a set of (pre-defined) spatial kernels are active (thus corresponding to localization). This approach is supervised—NMF basis functions of the individual source spectra (learnt in a training stage), as well as a pre-defined spatial dictionary are incorporated into the approach.

In a recent separation approach called GCC-NMF [30], GCC-PHAT is used for localization, and the NMF decomposition of the mixture spectrum is used for dictionary learning. Subsequently, the NMF atoms at each time instant are clustered, using the location estimates from GCC-PHAT, to separate the underlying sources. The results of this approach, along with the successful use of NMF in supervised single-channel source separation, indicate that an NMF-based spectral decomposition results in basis functions (atoms) that are sufficiently distinct for each source, and which do not overlap significantly in time—i.e., we have some form of disjointness in the time-*atom* domain. Thus, we hypothesise that using such atoms as weighting for the frequency averaging would allow for exploiting this time-atom sparsity and disjointness to simultaneously localize multiple sources within a single time frame while being robust to spatial aliasing due to the frequency averaging.

Specifically, we investigate the use of an *unsupervised* NMF decomposition as a weighting function for the SRP-based localization and apply it to the task of multi-speaker localization. Further, we also investigate modifications to the NMF atoms which lead to a better weighting for the purpose of localization, followed by a rigorous evaluation of NMF-weighted SRP for DoA estimation in various room acoustic environments, and with different array configurations. The proposed approach is comprehensively compared to (a) the state-of-the-art localization approaches for closely spaced microphones and (b) the state-of-the-art methods for widely spaced microphones.

3 Signal model

3.1 Spatial propagation model

Consider an array of M microphones that captures the signals radiated by Q broadband sound sources in the far field. The microphone locations may be expressed in 3D cartesian co-ordinates by the vectors as $\mathbf{r}_1, \ldots, \mathbf{r}_M$. Under the far field assumption, the DoA vector for source q in this co-ordinate system can be denoted as:

$$\mathbf{n}_q(\theta, \phi) = \left(\cos(\theta_q)\sin(\phi_q)\,, \sin(\theta_q)\sin(\phi_q)\,,\ \cos(\phi_q) \right)^T\,, \tag{1}$$

where $0 \leq \theta \leq 2\pi$ is the azimuth angle between the projection of $\mathbf{n}_q(\theta, \phi)$ on to the xy plane and the positive x-axis and $0 \leq \phi \leq \pi$ is the elevation angle with respect to the positive z-axis.

In the STFT domain, the image of source q at the array, in the kth frequency bin and bth time frame, can be compactly denoted as: $\mathbf{X}_q(k,b) = [X_{q,1}(k,b), \ldots, X_{q,M}(k,b)]^T$. If $\mathbf{V}(k,b)$ is the STFT-domain representation of the background noise at the array, the net signal captured by the array can be written as:

$$\mathbf{X}(k,b) = \sum_{q=1}^{Q} \mathbf{X}_q(k,b) + \mathbf{V}(k,b), \tag{2}$$

where $\mathbf{X}(k,b) = [X_1(k,b), \ldots, X_M(k,b)]^T$.

Under the common assumption of direct path dominance, and taking the signal at the first microphone as the reference, the image of source q at the array can be re-cast, *relative* to its image at the reference microphone, as:

$$\mathbf{X}_q(k,b) = \left(1\,, e^{J\,\Omega_k \mathbf{r}_{21}^T \mathbf{n}_q/c}, \ldots, e^{J\,\Omega_k \mathbf{r}_{M1}^T \mathbf{n}_q/c} \right)^T X_{q,1}(k,b)\,, \tag{3}$$

where $\Omega_k = \frac{2\pi k f_s}{K}$ is the kth discrete frequency, f_s is the sampling rate, K is the number of DFT points, $\mathbf{r}_{i\ell} = \mathbf{r}_i - \mathbf{r}_\ell$ is the position difference between microphones i and ℓ, and c is the speed of sound.

The term $\left(1\,, e^{J\,\Omega_k \mathbf{r}_{21}^T \mathbf{n}_q/c}, \ldots, e^{J\,\Omega_k \mathbf{r}_{M1}^T \mathbf{n}_q/c} \right)^T$ is often termed the *relative* steering vector $\mathbf{A}_q(k)$ in the literature. Further, it is also often assumed that each TF-bin is dominated by only one source based on W-disjoint orthogonality property [1]. Consequently, assuming source q is dominant in TF-bin (k,b), (2) can be simplified as:

$$\mathbf{X}(k,b) \approx \mathbf{X}_q(k,b) + \mathbf{V}(k,b)\,. \tag{4}$$

3.2 NMF model

Given the STFT representation $S_q(k,b)$ of a source signal q, computed over K discrete frequencies and B time frames, we denote the discrete magnitude spectrogram of this signal by the $(K \times B)$ non-negative matrix $|\mathbf{S}_q|$. We shall subsequently use the compact notation: $|\mathbf{S}_q| \in \mathbb{R}_+^{(K\times B)}$ to denote a non-negative matrix and its dimensions. The element (k,b) of the matrix $|\mathbf{S}_q|$ is denoted as $|S_q(k,b)|$.

A low rank approximation of $|\mathbf{S}_q|$ of rank D can be obtained using NMF as:

$$|\mathbf{S}_q| \approx \mathbf{W}_q \mathbf{H}_q\,, \tag{5}$$

where $\mathbf{W}_q \in \mathbb{R}_+^{(K\times D)}$ and $\mathbf{H}_q \in \mathbb{R}_+^{(D\times B)}$. Eq (5) implies that:

$$|S_q(k,b)| \approx \sum_{d=1}^{D} W_q(k,d) H_q(d,b). \tag{6}$$

The columns $\mathbf{w}_{d,q}$, $d = 1, 2, \ldots, D$, of $\mathbf{W}_q$ encode spectral patterns typical to the source q and are referred to as *atoms* in the ensuing. The rows of $\mathbf{H}_q$ encode the activity of the respective atoms in time. A high value of $H_q(d,b)$ for an atom d at frame b indicates that the corresponding atom is active in that time frame.

However, based on the assumption of signal sparsity in the time-atom representation, only the atoms whose activation values exceed a certain threshold value need be considered as contributing to the signal at a particular time frame. Let $\mathcal{D}_{b,q}$ be the set of atom indices whose activation values exceed the threshold at time frame b. Then, we can further simplify (6) as:

$$|S_q(k,b)| \approx \sum_{d\in\mathcal{D}_{b,q}} w_{d,q}(k) H_q(d,b)\,, \tag{7}$$

where $w_{d,q}(k) = W_q(k,d)$.

4 Steered response power beamformers

4.1 Narrowband SRP (NB-SRP)

To localize a source at any frequency bin k and time frame b, the NB-SRP approach basically steers a constructive beamformer towards each candidate DoA (θ, ϕ), in a predefined search space of candidate DoAs, and picks the candidate with the maximum energy as the location of the active source at the TF point (k,b). This assumes, implicitly, that the time-frequency bin in question

contains a directional source. Formally, this approach may be written as:

$$\left(\widehat{\theta}(k,b),\widehat{\phi}(k,b)\right) = \underset{\theta,\phi}{\operatorname{argmax}}\, \mathcal{J}_{\text{NB-SRP}}(k,b,\theta,\phi)\,, \tag{8}$$

where $(\widehat{\theta}(k,b),\widehat{\phi}(k,b))$ is the DoA estimate at each TF bin and $\mathcal{J}_{\text{NB-SRP}}(k,b,\theta,\phi)$ is the optimization function given by:

$$\mathcal{J}_{\text{NB-SRP}}(k,b,\theta,\phi) = |\mathbf{A}^H(k,b,\theta,\phi)\mathbf{X}(k,b)|^2\,. \tag{9}$$

In the above, $\mathbf{A}(k,b,\theta,\phi)$ can be any generic beamformer that leads to a constructive reinforcement of a signal along (θ,ϕ). In practice, the normalized delay-and-sum beamformer of (10) is widely used. Since this is similar to the PHAT weighting, this approach is called the NB-SRP-PHAT.

$$\mathbf{A}(k,b,\theta,\phi) = \left[\frac{1}{|X_1(k,b)|}, \frac{e^{J\,\Omega_k \mathbf{r}_{21}^T \mathbf{n}(\theta,\phi)/c}}{|X_2(k,b)|}, \ldots, \frac{e^{J\,\Omega_k \mathbf{r}_{M1}^T \mathbf{n}(\theta,\phi)/c}}{|X_M(k,b)|}\right]^T. \tag{10}$$

The source location estimates for the different TF bins, obtained as in (8), are subsequently clustered and the multi-speaker location estimates are obtained as the centroids of these clusters.

4.2 Broadband SRP (BB-SRP)

NB-SRP fails to provide a unique maximum for (8) for frequencies above the spatial aliasing frequency. As the inter-microphone distance increases, a larger range of frequencies are affected by spatial aliasing, and the efficacy of NB-SRP-based methods decreases. To overcome this problem, (9) is summed across the frequency range, leading to the broadband SRP (BB-SRP) optimization function [31]:

$$\mathcal{J}_{\text{BB-SRP}}(b,\theta,\phi) = \sum_k |\mathbf{A}^H(k,b,\theta,\phi)\mathbf{X}(k,b)|^2. \tag{11}$$

BB-SRP may be seen as a multi-channel analog of GCC-PHAT approach. Note that (11) yields a single localization result per time frame. The results from multiple time frames can then be clustered as in the NB case for multi-speaker localization. The broadband approach ameliorates spatial aliasing at the cost of un-utilized TF sparsity. Since only the dominant source is located in *each time frame*, softer speakers who are not dominant in a sufficient number of time frames may not be localized.

5 The SRP-NMF approach

As we shall now demonstrate, by incorporating the D_T basis functions $\mathbf{W} = \left[\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D_T}\right]$ obtained from an NMF decomposition of the microphone signal spectrum, we can exploit sparsity in what we term the 'time-atom' domain. For compactness of expression, and without loss of generality, we shall consider localization only in the azimuth plane (i.e., $\phi = \pi/2$) in the following.

In each time frame we compute a *weighted* version of (11) as:

$$\mathcal{J}_{\text{SRP-NMF}}(d,b,\theta) = \sum_k w_d(k)\left|\mathbf{A}^H(k,b,\theta)\mathbf{X}(k,b)\right|^2\,, \tag{12}$$

where $w_d(k)$ is the kth element of the dth atom $\mathbf{w}_d$. Based on (12), we obtain a DoA estimate *per active atom d* as:

$$\widehat{\theta}(d,b) = \underset{\theta}{\operatorname{argmax}}\, \mathcal{J}_{\text{SRP-NMF}}(d,b,\theta)\,. \tag{13}$$

As previously explained, we expect the atoms $\mathbf{w}_d$ to embody the spectral patterns typical to the underlying sources. Further, the time-frequency sparsity and disjointness of speech results in each atom being unique to a single source. Thus, the weighted sum in (12) only aggregates information across frequencies that are simultaneously excited by a source, yielding a spatial-aliasing robust location estimate for that source in (13). This is the rationale behind the weighting in (12). Multi-speaker localization is subsequently obtained by clustering the DoA estimates computed for all active atoms.

We present an intuitive idea of how this works using a toy example in Section 5.1.

5.1 Demonstration of the working principle of SRP-NMF

Consider two spatially separated, simultaneously active sources captured by microphones placed 12 cm apart. Each source is a harmonic complex of different fundamental frequencies. Figure 1 describes the two underlying source atoms $\mathbf{w}_d$. In this simple example, $w_1(k) = 1$ only at frequencies where the source 1 is active, and zero otherwise (the red lines in Fig. 1) and $w_2(k) = 1$ only at frequencies where the source 2 is active (the blue dashed lines in Fig. 1). Figure 2 depicts the BB-SRP optimization function $\mathcal{J}_{\text{BB-SRP}}(\theta)$ and the SRP-NMF optimization functions $\mathcal{J}_{\text{SRP-NMF}}(d,\theta)$, $d = 1,2$ for the two atoms, over the azimuthal search space. The dashed lines indicate the ground truth DoAs. The locations of the peaks of the optimization functions correspond to the respective DoA estimates. It is evident from this figure that the BB-SRP can localize only one source when considering the dominant peak (and even then with a large error). When considering the locations of the two largest peaks of $\mathcal{J}_{\text{BB-SRP}}(\theta)$ for estimating the two underlying source DoAs, both estimates are in error by more than 5°. This is quite large for such a synthetic example. In contrast, the SRP-NMF estimates (one each from the respective $\mathcal{J}_{\text{SRP-NMF}}(d,\theta)$) are much more accurate and localize both sources. This is because the each atom emphasizes frequency components specific to a single source in the weighted summation, while suppressing the other components.

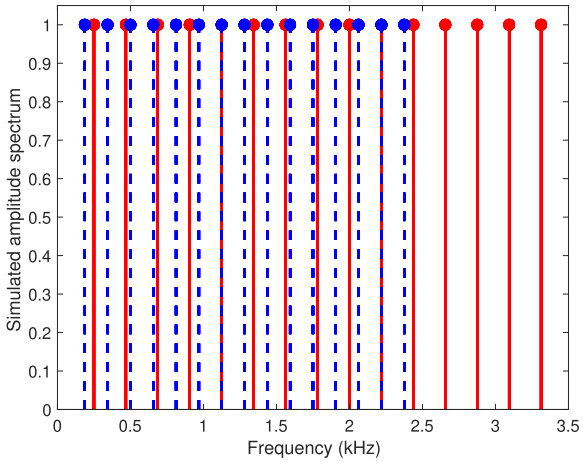

Fig. 1 Simulated amplitude spectrum of two spatially separated sources

5.2 SRP-NMF implementation

With the intuitive understanding from the previous section, we now focus on the implementation details. In a supervised localization approach, source-specific atoms can be easily obtained by NMF of the individual source spectra. However, we focus here on the *unsupervised* case, where no prior information of the sources to be localized is available. The atoms, therefore, are extracted from the mixture signal at the microphones. It has previously been demonstrated [32] that NMF of mixture spectra still results in atoms that correspond to the underlying source spectra. However, it is not possible to attribute the atoms to their corresponding sources without additional information. In our case, NMF is performed on the average of the magnitude spectrograms of the signals of the different microphones. Another possibility is a weighted average spectrogram where the weights could be estimated based, e.g., on some SNR measure [33, 34].

The steps in SRP-NMF localization are:

- Compute the average of the magnitude spectrograms of the signals at all microphones m:

$$\overline{|X(k,b)|} = \frac{1}{M}\sum_{m=1}^{M}|X_m(k,b)|. \tag{14}$$

  This yields the average magnitude spectrum matrix $\overline{|\mathbf{X}|} \in \mathbb{R}_+^{(K\times B)}$, where K and B indicate, again, the number of discrete frequencies and time frames of the STFT representation.
- Decompose $\overline{|\mathbf{X}|}$ using NMF into the matrix $\mathbf{W} \in \mathbb{R}_+^{(K\times D_T)}$, containing the D_T dictionary atoms, and the matrix $\mathbf{H} \in \mathbb{R}_+^{(D_T\times B)}$ containing the activations of these atoms for the different time frames:

$$\overline{|\mathbf{X}|} \approx \mathbf{WH}\,. \tag{15}$$

  The cost function used for NMF is the generalized KL divergence [35]:

$$D_{\mathrm{KL}}(\overline{|\mathbf{X}|}, \mathbf{WH}) = \sum_k\sum_b \left(\overline{|X(k,b)|}\log\left(\frac{\overline{|X(k,b)|}}{[\mathbf{WH}](k,b)}\right) - \overline{|X(k,b)|} + [\mathbf{WH}](k,b)\right), \tag{16}$$

  where $[\mathbf{WH}](k,b)$ indicates element (k,b) of the *product* $\mathbf{WH}$. The well-known multiplicative update rules are applied to estimate $\mathbf{W}$ and $\mathbf{H}$. Once the atoms are obtained, they can be used for the weighting in (12)
- We note that only the active atoms of each time frame are used in the localization. To obtain the active atoms for any frame b, they are sorted in decreasing order of their activations $H(d,b)$ in *that* frame. The first atoms that contribute to a certain percentage (here empirically set at 99 percent) of the sum of the activation values in that frame are considered as *active*.
  The SRP-NMF optimization function is, consequently,

$$\mathcal{J}_{\text{SRP-NMF}}(d_b,b,\theta) = \sum_k w_{d_b}(k)\left|\mathbf{A}^H(k,d_b,b,\theta)\mathbf{X}(k,b)\right|^2, \tag{17}$$

  where $\mathbf{w}_{d_b}$ is an active atom at frame b.
- By maximizing (17) with respect to θ, a DoA estimate is obtained for each active atom in frame b as:

$$\widehat{\theta}(d_b,b) = \underset{\theta}{\operatorname{argmax}}\,\mathcal{J}_{\text{SRP-NMF}}(d_b,b,\theta). \tag{18}$$

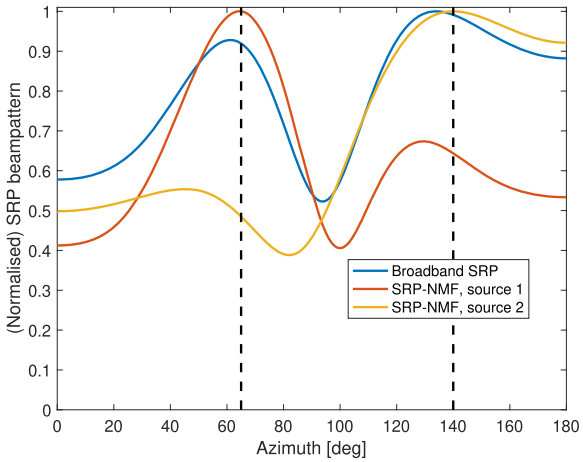

Fig. 2 Normalized SRP beampatterns of broadband SRP and SRP-NMF on a mixture of two sources. The dashed line indicates the ground truth of the source locations

- Lastly, we compute the histogram of the DoA estimates across all the time-atom combinations. The locations of peaks in the histogram correspond to DoA estimates of the active sources in the given mixtures.

5.2.1 NMF modifications

The NMF decomposition of speech spectra as in (15) results in dictionary atoms with higher energy at low frequencies than at high frequencies. This is because speech signals typically have a larger energy at the lower frequencies. Further, due to the large dynamic range of speech, the energy in high frequency components can be several decibels lower than that in low-frequency components [32]. This characteristic is, subsequently, also reflected in the NMF atoms. When these atoms are used as weighting functions, the resulting histogram of location estimates is biased towards the broadside of the array. We illustrate this on a 3 source stereo mixture (`dev1_male3_liverec_250ms_5cm_mix.wav`) from the SiSEC database. The details of the database are in Section 6.3. The ground truth DoAs of the 3 sources are $50°, 80°$ and $105°$. The histogram obtained by the SRP-NMF is shown in Fig. 3. The bias at the broadside of the array (around $90°$) is evident from the figure. While the second and third peaks near $90°$ are prominent, the first peak at $50°$, which is away from the broadside, is not clear.

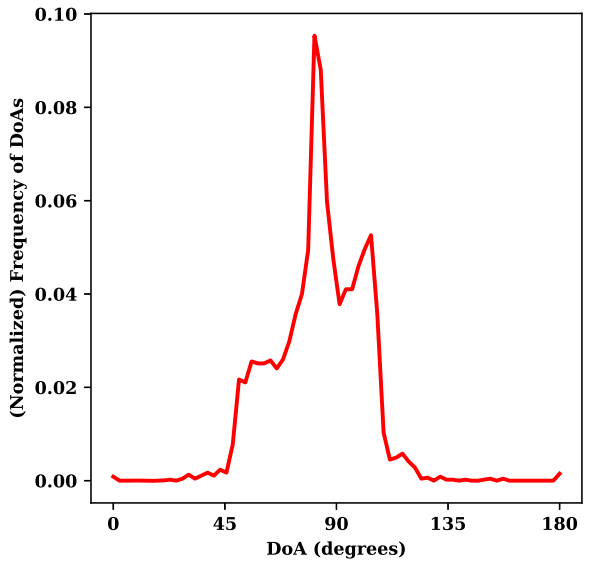

Fig. 3 Histogram of SRP-NMF on a mixture of 3 sources taken from the SiSEC database `dev1_male3_liverec_250ms_5cm_mix.wav`, with $\beta = 0$ and $D_T = 35$. The ground truth DoAs are $50°, 80°$ and $105°$. The estimate does not clearly present evidence for the 1st peak at $50°$

This broadside bias can be explained as follows: localization essentially exploits the inter-microphone phase difference (IPD), which is a linear function of frequency (with some added non-linearities in real scenarios due to reverberation [28]). This linear dependence implies that low frequencies have smaller IPDs (concentrated around 0), compared to high frequencies. This leads to localization around the broadside for the low frequencies. When using the weighted averaging, the dominant low frequency components in the atoms thereby emphasize the broadside direction.

To remove this bias, a penalty term [28, 36] is added to flatten the atoms, thereby reducing the dominance of low frequency components in the atoms. This penalty term is given by:

$$F(\mathbf{W}) = \sum_{d} \left[\mathbf{W}^T\mathbf{W}\right](d,d)\,, \tag{19}$$

where $\left[\mathbf{W}^T\mathbf{W}\right](d,d)$ indicates the elements along the main diagonal of $\mathbf{W}^T\mathbf{W}$. This leads to the constrained NMF (CNMF) cost function:

$$\mathcal{C}\left(|\overline{\mathbf{X}}|, \mathbf{WH}\right) = D_{\mathrm{KL}}(|\overline{\mathbf{X}}|, \mathbf{WH}) + \beta F(\mathbf{W})\,, \tag{20}$$

where β is the weighting factor of the penalty term. The multiplicative update equations subsequently become:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \frac{|\overline{\mathbf{X}}|}{\mathbf{WH}}}{\mathbf{W}^T\mathbf{1}} \qquad \text{and} \qquad \mathbf{W} \leftarrow \mathbf{W} \odot \frac{\frac{|\overline{\mathbf{X}}|}{\mathbf{WH}}\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T + 2\beta\mathbf{W}}, \tag{21}$$

where $\mathbf{1}$ represents a matrix of ones of the appropriate dimensions, $\odot$ represents the Hadamard product and the division is element-wise. This constrained decomposition favors atoms with a flat spectrum. Figure 4 shows the histogram of SRP-NMF when using the CNMF decomposition, where it may be observed that the broadside bias is overcome and azimuths of all the sources are correctly estimated.

6 Experimental evaluation

In this section, the performance of SRP-NMF is compared to the state-of-the-art localization approaches for closely spaced and widely spaced microphones. Since our approach is closely related to the SRP/GCC family of approaches (being, as it were, an intermediate between the broadband and narrowband versions of these), and because these are the typical, well-understood methods for source localization, these form the basis for our benchmark.

Specifically, we compare our approach to:

- The NB-SRP-PHAT according to Section 4.1;
- A sub-band variant of the above (termed Bark-SRP-PHAT), where the optimization function is

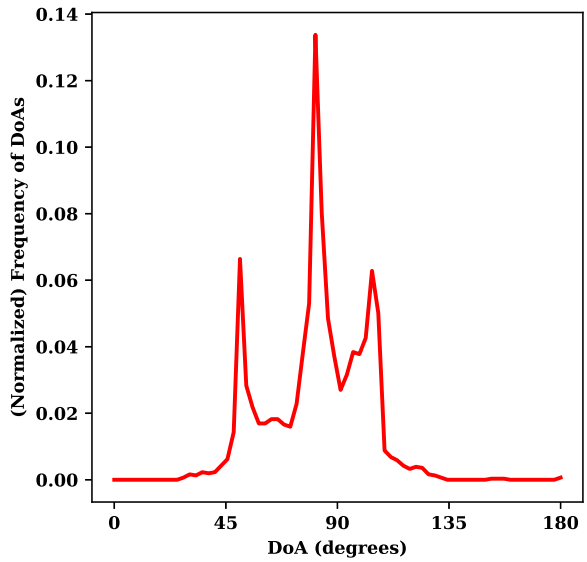

Fig. 4 Histogram of SRP-NMF (CNMF) on a mixture of 3 sources taken from the SiSEC database `dev1_male3_liverec_250ms_5cm_mix.wav`, with $\beta = 60$ and $D_T = 35$. The ground truth DoAs are 50°, 80° and 105°. The 3 peaks are clearly visible now

averaged over Sub-bands defined according to the Bark scale as in [24]; and

- Four other *best performing algorithms* among a broad variety of localization algorithms benchmarked in [19] and implemented within the open source Multichannel BSS-locate toolbox [37].

For completeness, a brief summary of Bark-SRP-PHAT and the approaches from the Multichannel BSS-locate toolbox is given in Section 6.1.

Tests are conducted on four different databases (three of which are openly available) in order to evaluate the approaches across different microphone arrays (different spacing and configurations) as well as in different acoustic environments, from relatively dry ($T_{60} \approx$ 130ms) to highly reverberant ($T_{60} \approx$ 660ms). The evaluation setup is described in Section 6.2, followed by the details of the databases used. The evaluation metrics are described in Section 6.4 and the method adopted for choosing NMF parameters is presented in Section 6.5.

Further, Section 6.6 presents a comparison of the proposed SRP-NMF to a supervised approach wherein the underlying sources at each microphone are first separated using NMF, and localization is subsequently performed on the separated sources.

Section 6.7 presents the results of the benchmarking.

6.1 Brief summary of benchmarked approaches

6.1.1 Bark-SRP-PHAT

NB-SRP and BB-SRP are, respectively, fully narrowband or fully broadband approaches. However, SRP-NMF only averages the optimization function over a (source-dependent) *subset* of frequencies. Thus, we include a comparison with a modified SRP approach where the optimization function is averaged along sub-bands, where the sub-bands are the critical bands defined according to the Bark scale. A single localization estimate is computed for each critical band within a time frame. These estimates are then pooled across all time frames in a manner similar to the narrowband SRP-PHAT approach, to obtain the final localization result. This approach thus exploits available sparsity and disjointness in time and sub-bands. This scale was chosen because of its psychoacoustical relevance, as seen in previous localization research (e.g., [24]).

6.1.2 MVDRW approaches

The MVDRW approaches [19] use minimum variance distortionless response (MVDR) beamforming to estimate, for each frequency bin k and each time frame b, the signal to noise ratio (SNR) in all azimuth directions. Since the spatial characteristics of the sound field are taken into account, the SNR indicates, effectively, time-frequency bins where the direct path of a single-source is dominant. The *MVDRWsum* variant averages the SNR across all time-frequency points and, subsequently, the DoA estimates are computed as the location of the peaks of this averaged SNR. When all sources are simultaneously active within the observation interval, this averaging is beneficial. However, when a source is active only for a few time frames, the averaging smooths out the estimate, thereby possibly not localizing the source. Hence [19] also proposes an alternative called *MVDRWmax*, where a *max* pooling of the SNR is performed over time.

6.1.3 GCC-variants

The two GCC-variants considered in [19] are the *GCC-NONLINsum* and *GCC-NONLINmax*. The key difference with the traditional GCC is the non-linear weighting applied to compensate for the wide lobes of the GCC for closely spaced microphones [38]. In GCC-NONLINsum and GCC-NONLINmax, respectively, the *sum* and *max* pooling of the GCC-PHAT, computed over the azimuthal space, is done across time.

As previously stated, these approaches were chosen for the benchmark because they have previously been demonstrated to be the best performing approaches among a broad variety of localization approaches. Further, since the implementation of these approaches is open source, it allows for a reproducible, fair benchmark against which new methods may be compared.

6.2 Evaluation setup

For all the experiments, the complex-valued short-term Fourier spectra were generated from 16 kHz mixtures using a DFT size of 1024 samples (i.e., $K = 512$) and a hop

size of 512 samples. A periodic square-root Hann window of size 1024 samples is used prior to computing the DFT.

The NMF parameters D_T and β are set to 55 and 60 respectively. These parameters are set based on preliminary experiments that are described in Section 6.5. The maximum number of NMF iterations is 200.

For all the approaches, the azimuth search space ($0° - 180°$) was divided into a uniformly spaced grid with a 2.5° spacing between adjacent grid points. Further, in all cases, it is assumed that the number of speakers in the mixture is known.

6.3 Data

The following four databases, covering a wide range of recording environments, are used for evaluations.

6.3.1 *Signal Separation and Evaluation Campaign (SiSEC) [39]*

The *dev1* and *dev2* development data of SiSEC, consisting of under-determined stereo channel speech mixtures, is used. The mixtures are generated by adding live recordings of static sources played through loudspeakers in a meeting room (4.45m x 3.55m x 2.5m) and recorded one at a time by a pair of omnidirectional microphones. Two reverberation times of 130 ms and 250 ms are considered.

Two stereo arrays are used: one with an inter-microphone spacing of 5cm (SiSEC1) and the other with spacing of 1m (SiSEC2). The speakers are at a distance of 0.80m or 1.20m from the array, and at azimuths between 30° and 150° with respect to the array axis. The data thus collected consists of twenty 10 s long mixtures of 3 or 4 simultaneous speakers (either all male or all female). The ground truth values of DoAs are provided. They were further verified by applying the GCC-NONLIN approach on the *individual* source images that are available in the data set.

Since the mixtures are generated by mixing live recordings from a real environment, they also contain measurement and background noise. Further, both closely spaced and widely spaced arrays can be evaluated in the same setting. This makes the SiSEC dataset ideal for the comparison of the various approaches.

6.3.2 *Challenge on acoustic source LOCalization And TrAcking (LOCATA) [40]*

LOCATA comprises multi-channel recordings in a real-world closed environment setup. Among several tasks that this challenge offers, we consider *Task1*: localization of a single, static speaker using a static microphone array and *Task2*: localization of multiple static speakers using a static microphone array.

The data consists of simultaneous recordings of static sources. Sentences selected from the CSTR VCTK database [41] are played back through loudspeakers in a computer laboratory (dimensions: 7.1m x 9.8m x 3 m, $T_{60} = 550$ms). These signals are recorded by a non-uniformly spaced linear array of 13 microphones [40]. In total, there are 6 mixtures of one to four speakers, and the mixtures are between 3 s to 7 s long. The ground truth values of the source locations are provided.

To evaluate different linear array configurations we consider 4 uniform sub-arrays: 3 mics with 4 cm inter-microphone spacing (LOCATA1), 3 mics with an 8 cm inter-microphone spacing (LOCATA2), 3 mics with 16 cm inter-microphone spacing (LOCATA3), and 5 mics with a 4 cm inter-microphone spacing (LOCATA4). This dataset is generated from live recordings in a highly reverberant room, which makes it interesting for benchmarking localization approaches.

6.3.3 *Aachen Multi-Channel Impulse Response Database (AACHEN) [42]*

This is a database of impulse responses measured in a room with configurable reverberation levels. Three configurations are available, with respective T_{60}s of 160 ms, 360 ms and 610 ms. The measurements were carried out for several source positions for azimuths ranging from 0° to 180° in steps of 15° and at distances of 1 m and 2 m from the microphone array. Three different microphone array configurations are available.

For this paper, we choose the room configuration with $T_{60} = 610$ ms. The impulse responses corresponding to sources placed at a distance of 2m from the 8 microphone uniform linear array with an inter-microphone spacing of 8 cm are selected. Multi-channel speech signals are generated by convolving the selected impulse responses with dry speech signals. Fifty mixtures, each 5 s long, and from 3 speakers (randomly chosen from the TSP database [43]), placed randomly at 3 different azimuths with respect to the array axis are generated.

6.3.4 *UGent Multi-Channel Impulse Response Database (UGENT)*

The impulse responses from the UGENT database were measured using exponential sine sweeps for azimuth angles varying from 15° to 175° with the source at a distance of 2m from the array. The recordings were conducted in a meeting room with a $T_{60} \approx 660$ms. The microphone array is a triangular array with the following microphone coordinates: (0m,0m,0m), (0.043m,0m,0m) and (0.022m, − 0.037m,0m). Fifty mixture files, each of 5 s duration, are generated with 3 speakers (randomly chosen from the TSP database) placed at random, different azimuths.

Except for the UGent database, all other databases are openly accessible.

6.4 Evaluation metrics

The evaluation measures chosen are a detection metric (F-measure) and a location accuracy metric (mean azimuth error - MAE). In a given dataset, let N be the total number of sources in all mixture files and N_e be the number of sources that are localized by an approach. The estimated source azimuths for each mixture are matched to the ground truth azimuths by greedy matching to ensure minimum azimuth error. If, after matching, the estimated source azimuth is within $\pm 7.5^\circ$ of the ground truth estimate then the source is said to be correctly estimated. Let N_c be the number of sources correctly localized for all mixtures. Then the F-measure is given by

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}, \tag{22}$$

where $\text{Recall} = N_c/N$ and $\text{Precision} = N_c/N_e$ The more the number of sources correctly localized, the higher the F-measure.

To quantify the localization accuracy, we present two error metrics: MAE and MAEfine. While MAE is the mean azimuth error between the estimated DoAs and true DoAs after greedy matching (irrespective of whether an approach managed to correctly localize all sources within the 7.5° tolerance), MAEfine is the mean error between the *correctly estimated DoAs* and true DoAs. Thus, while MAE gives location accuracy over all the sources in the mixture, MAEfine gives location accuracy of only the correctly detected sources. The former may, therefore, be seen as a global performance metric whereas the latter indicates a local performance criterion with respect to correctly detected sources.

6.5 Selecting suitable NMF parameters

To obtain suitable values of the flattening penalty term β and the dictionary size D_T, the localization performance of SRP-NMF is evaluated on a small dataset over a range of β and D_T.

Table 1 shows the F-measure obtained by SRP-NMF on SiSEC1 data for β varying from 0 to 80 and D_T from 15 to 55. It may be seen that with β fixed, as the dictionary size increases, the localization performance initially improves and later saturates. A similar trend is observed when D_T is fixed and β is increased. The pairs of β and D_T that yield an F-measure $\geq$0.95 (in bold) have similar performance and can be chosen as the NMF parameters. While a lower D_T leads to less computational complexity, a lower β leads to a lower residual error in the NMF approximation (i.e., a better approximation of the magnitude spectrum). Therefore, among various combinations of β and D_T that yield a comparable F-score, a lower β (such as 30) and lower D_T (such as 25) are preferred. However, we choose slightly higher parameter values to ensure robust performance and to allow generalization to other datasets with possibly more reverberation and/or noise. Hence, in the subsequent experiments, the values of β and D_T are set to 60 and 55 respectively.

The trends in Table 1 are illustrated in Figs. 5 and 6 for a mixture of 4 concurrent speakers. Figure 5 depicts the histogram plots of SRP-NMF with β ranging between 0 and 80 and $D_T = 35$. It is evident from the figure that when β=0, the peaks further away from the broadside direction are not prominent. The reason for this was explained in Section 5.2.1. As β increases, the peaks become increasingly prominent and can be easily detected.

Figure 6 presents the effect of varying D_T on the SRP-NMF outcome. Here, β is fixed at 60 and D_T increases from 5 to 55. It may be seen that as the dictionary size increases, the histogram peaks become increasingly distinct.

6.6 Experiment with supervised separation and localization

The basic idea for the proposed approach has its roots in the successful use of NMF for supervised source separation. Hence, we compare, here, the performance of SRP-NMF against a *supervised* variant where the microphone signals are first decomposed into their underlying sources using NMF [44] and the localization is then performed on the separated sources using the broadband SRP approach. This approach is termed *SNMF-SRP*, and is implemented as follows:

1 First, for any test case, the magnitude spectrum $|\mathbf{S}_q|$ of each individual source q in the mixture is

Table 1 The detection metric, F-measure, obtained by SRP-NMF on SiSEC1 data for β and D_T varying from 0 to 80 and 15 to 55 respectively

$\frac{\beta}{D_T}$	0	10	20	30	40	50	60	70	80
15	0.72	0.83	0.80	0.85	0.87	0.89	0.88	0.94	**0.95**
25	0.83	0.89	0.92	**0.95**	**0.98**	**0.98**	**0.98**	**0.99**	**0.98**
35	0.78	0.93	**0.99**	**0.96**	**0.97**	**0.99**	**0.98**	**0.98**	**0.99**
45	0.84	**0.95**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
55	0.85	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**0.98**

Values of F-measure $\geq$ 0.95 are in bold, indicating good performance

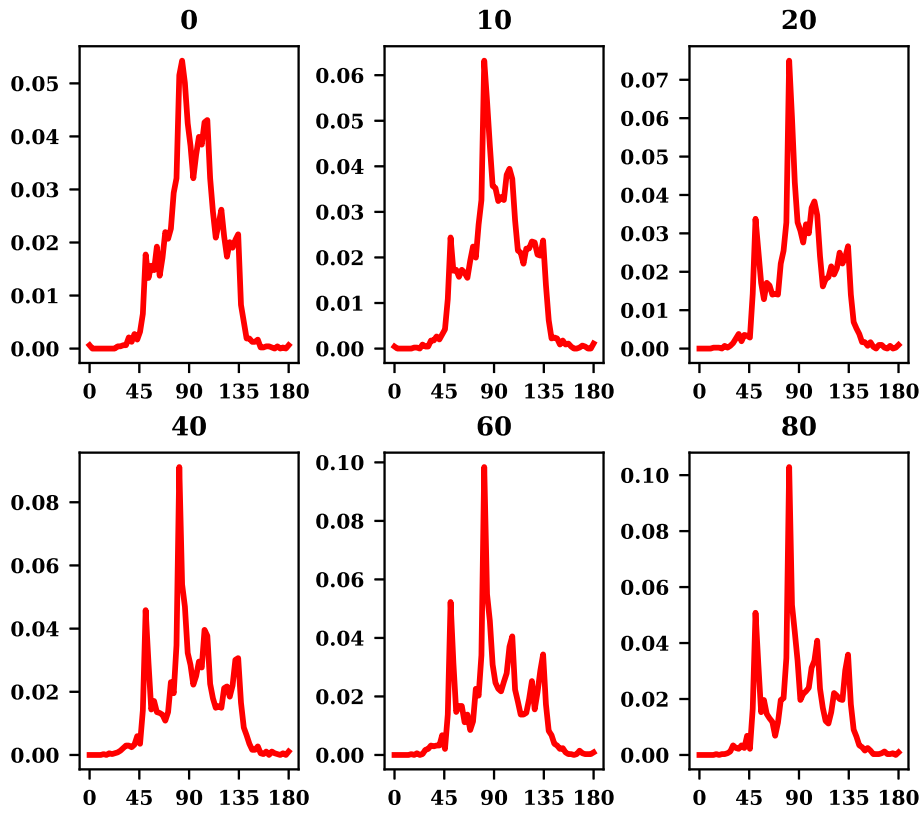

Fig. 5 Effect of β on SRP-NMF performance. Subplots show the histograms of SRP-NMF on a mixture of 4 sources taken from SiSEC database (`dev1_male4_liverec_250ms_5cm_mix.wav`) with $D_T = 35$ and β varying from 0 to 80. β is displayed on each of the subplots. The *x*-axis shows DoA (degrees). The *y*-axis is the (normalized) frequency of DoAs

decomposed using constrained NMF. This results in the $\mathbf{W}_q \in \mathbb{R}_+^{(K \times D_q)}$ basis function matrix for that source, where D_q is the number of atoms for source q. We assume that the number of atoms is the same for all sources, i.e., $D_q = D \, \forall q$. The basis functions for all sources are then concatenated into a matrix $\mathbf{W}$ as:

$$\mathbf{W} = \left[\mathbf{W}_1 , \mathbf{W}_2 , \ldots , \mathbf{W}_Q \right] \in \mathbb{R}_+^{(K \times QD)} \tag{23}$$

2 NMF is next used to decompose the magnitude spectrogram of the mixture at any one *reference* microphone m as $|\mathbf{X}_m| \approx \mathbf{WH}$. In this step, $\mathbf{W}$ is kept fixed and only the *activation matrix* $\mathbf{H}$ is adapted. This matrix can then be partitioned into the activations of the individual sources as:

$$\mathbf{H} = \left[\mathbf{H}_1^T , \mathbf{H}_2^T , \ldots , \mathbf{H}_Q^T \right] \in \mathbb{R}_+^{(QD \times B)} , \tag{24}$$

where B is the total number of frames in the mixture spectrogram.

3 The spectral magnitude estimates for each source can then be obtained as: $|\widehat{\mathbf{S}}_q| = \mathbf{W}_q \mathbf{H}_q$. These estimates are used to define binary masks for each source, whereby each TF point is allocated to the source with the maximum contribution (i.e., the dominant source) at that TF-point.

4 The binary masks belonging to each source are, finally, applied to the complex mixture spectrograms at all microphones, and the broadband SRP-PHAT approach is used to obtain the source location estimate.

Since SNMF-SRP first separates the sources before localizing them, the interference from the other sources is minimized in the localization. Further, a binary mask attributes a time-frequency point (k, b) to only the dominant source at that point. Due to this "winner-takes-all" strategy, only the dominant source components are preserved at each time-frequency point. Consequently, the effect of the interference on the SRP-PHAT function is further reduced, resulting in more accurate DoA estimates as compared to when continuous masks are used. This experiment with oracle knowledge of the underlying sources should, therefore, give a good indication of the possible upper bound of our proposed approach.

We note that an alternative to the SNMF-SRP would be *unsupervised* NMF-based separation approaches. Such

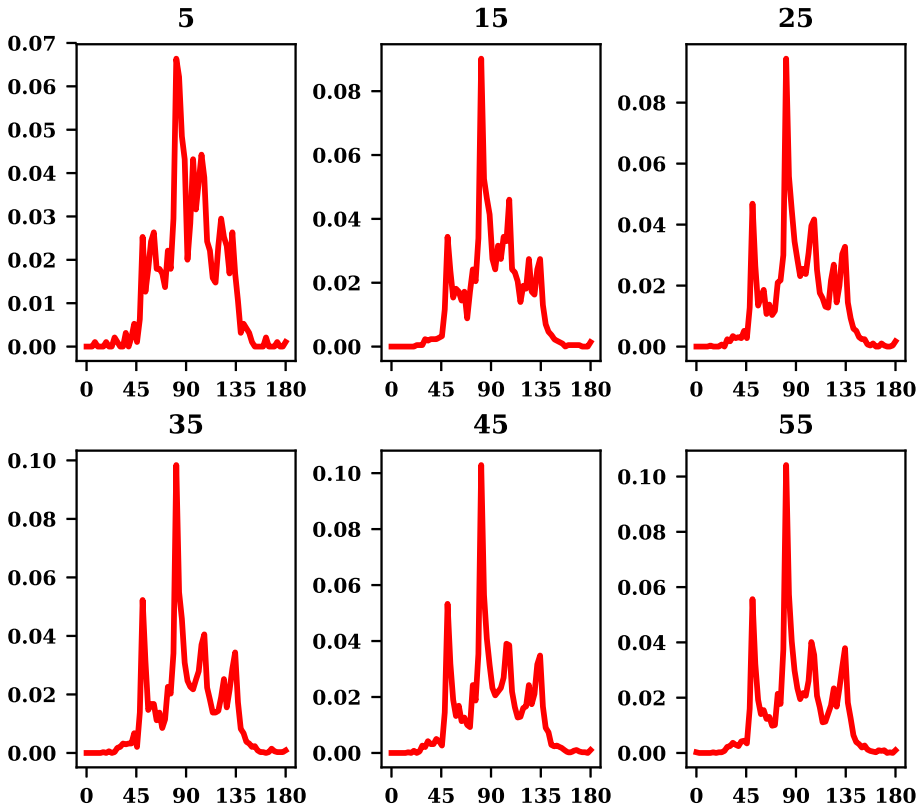

Fig. 6 Effect of dictionary size D_T on SRP-NMF performance. Subplots show the histograms of SRP-NMF on a mixture of 4 sources taken from SiSEC database (`dev1_male4_liverec_250ms_5cm_mix.wav`) with $\beta = 60$ and D_T varying from 5 to 55. D_T is displayed on each of the subplots. The *x*-axis shows DoA (degrees). The *y*-axis is the (normalized) frequency of DoAs

approaches may be seen as comprising the following two steps: (a) decomposing the mixture spectrum into basis functions and their corresponding activations, and, (b) grouping (clustering) the basis functions according to the sources they belong to, to generate the separated source signals. Usually, some additional signal knowledge or signal model needs to be incorporated into the approach to perform the clustering and the quality of the source separation is, consequently, dependent on the kind of clustering approach. Typically, these steps are not performed independently, and the clustering model is often incorporated (explicitly or implicitly) as a set of additional constraints in the decomposition step. If one neglects the additional step (and associated effort) of grouping the basis components and simply uses the obtained basis functions as a weighting within the SRP-PHAT approach, then there is no conceptual difference between our proposed approach and the use of unsupervised NMF-based separation followed by localization.

6.6.1 Experimental set-up

We compare, first, the SNMF-SRP and SRP-NMF. For this purpose, fifty mixtures, each 5 s long and comprising 3 sound sources at randomly chosen azimuths, ranging from 15° to 175°, are generated using room impulse responses from the AACHEN database. The responses corresponding to the room configuration with $T_{60} = 610$ ms are used. Two arrays are considered: the 8 microphone uniform linear array with 8 cm inter-microphone spacing, and a 4-microphone uniform linear sub-array with 4 cm inter-microphone spacing (this is part of a larger 8-mic array with spacing 4-4-4-8-4-4-4). The position of the speakers was also randomly chosen for each test file. The optimal dictionary size and weighting factor for the SNMF-SRP approach are first determined in a manner similar to that for SRP-NMF, and using data from the 3-mic sub-array with inter-microphone spacing of 8 cm.

Dictionary sizes $D_{\text{SNMF-SRP}}$ of 50, 90, and 130 and weighting factors $\beta_{\text{SNMF-SRP}}$ of 0, 20, and 40 are evaluated. The F-measure and MAE obtained for each case are reported in Table 2, from where it is observed that a dictionary size $D_{\text{SNMF-SRP}}$ of 130 and $\beta_{\text{SNMF-SRP}}$ of 20 give the best results in terms of the chosen metrics. These are consequently fixed for the subsequent evaluation of the SNMF-SRP approach.

Table 2 F-measure and mean azimuth error (MAE) for the supervised NMF-SRP (SNMF-SRP) for varying $D_{\text{SNMF-SRP}}$ and $\beta_{\text{SNMF-SRP}}$

$\frac{\beta_{\text{SNMF-SRP}}}{D_{\text{SNMF-SRP}}}$	0	20	40
50	0.9/7.69	0.95/6.22	0.93/6.57
90	0.91/7.66	0.97/4.95	0.94/7.15
130	0.92/6.82	**0.98/4.5**	0.97/4.85

Figures 7 and 8 depict the performance of SNMF-SRP compared to SRP-NMF.

Since we can expect the best localization performance in the absence of reverberation and interfering sources, we simulate this case as well and include it in the comparison (this is termed direct-path (DP) single-source-SRP-PHAT). To obtain this result, each source in the mixture is *individually* simulated at the arrays. Further, for generating the source image, the room impulse response is limited to only the filter taps corresponding to the direct path and 20 ms of early reflections. Then, a DoA estimate is obtained by the broadband SRP-PHAT. This corresponds to the localization of a *single* source in the near absence of reverberation and noise and, thus forms a further performance upper bound for all the approaches.

The figures show that, especially for a smaller number of microphones and lower inter-microphone spacing, the supervised NMF-SRP approach is significantly better than the proposed unsupervised SRP-NMF. The SRP-NMF has the lowest F-measure and the largest MAE. This indicates that incorporating the knowledge of the underlying sources may be beneficial when the spatial diversity is limited and cannot be fully exploited. As the spatial diversity increases, the performance of the unsupervised method begins to converge to that of the supervised approach. As expected, the performance of both these approaches are upper bounded by the DP-single-source SRP-PHAT approach.

6.7 Results and discussion

The benchmarking results, in terms of F-measure and mean azimuth errors, for the various datasets are plotted in Fig. 9. We start with the MAEfine metric, which focusses on the average localization error for sources that have been *correctly* localized. The chosen margin for a correct localization implies that the MAEfine is necessarily $\leq 7.5°$. Figure 9 further indicates that the MAEfine metric is comparable among all the approaches, with a difference of only about 1 deg or less (except for the GCC-NonLinsum and MVDRWsum of LOCATA1 and MVDRWmax of LOCATA2, where it is slightly higher). Thus, we may not claim, categorically, that any particular approach is better than the other in terms of this metric. More indicative metrics for the performance of any approach would be the MAE and F-measure, which are discussed next.

NB-SRP-PHAT localizes well with closely spaced stereo microphones and its performance deteriorates with larger inter-microphone spacing due to spatial aliasing. This is clearly seen from the SiSEC results, where its performance is better in SiSEC1 (5 cm spacing) than in SiSEC2 (1 m spacing). Furthermore, in the case of multiple microphones, it performs poorly in LOCATA1 and UGENT. The reason for the poor performance may be explained as follows: both LOCATA1 and UGENT have only 3 microphones that are very closely spaced ($\approx$4 cms apart) and high reverberation ($T_{60} \approx 600$ ms). We hypothesize that the TF bins in which noise or reverberant components are dominant are allocated to spurious locations and, since NB-SRP-PHAT pools the decisions per TF bin, these spurious locations mask the source locations in the histogram. This behavior is worse in closely spaced arrays, as the beam pattern of the SRP optimization function has wide main lobes. Increasing the microphone separation or the number of microphones, narrows the main lobes thus improving the performance - as is evident in LOCATA2/3 and LOCATA4 respectively.

Among the GCC-NONLIN approaches, *max* pooling performs better than *sum* pooling, which verifies the conclusions in [19]. Further, due to the non-linearity introduced to improve the performance in microphone arrays with short inter-microphone spacing (cf. Section 6.1), the GCC-NONLINmax performs reasonably well in almost all datasets and microphone configurations.

Between the MVDRW methods, *max* and *sum* pooling give similar results for the smaller array of SiSEC1. In SiSEC2, *sum* pooling is superior, which is consistent with [19]. However, for a larger number of microphones *max* pooling performs better in all microphone configurations. In LOCATA1 and UGENT, though the beampattern of MVDR has wide lobes due to closely spaced microphones, the performance of the MVDRW-based approaches is better than that of NB-SRP-PHAT. We reason that this is because the MVDRW approaches factor in the sound field characteristics and introduce a frequency weighting that emphasizes the time-frequency bins that are dominated by the direct sound of a single source (cf. Section 6.1).

Figure 9 also indicates that SRP-NMF performs consistently well across the various databases. In terms of MAE and F-measure, the scores of SRP-NMF is among the top two for each tested case. The atom weighting highlights time-frequency bins consisting of information relating to a single source, similar to SNR weighting, thus exploiting time-atom sparsity and leading to superior performance in short arrays. In large arrays, averaging the optimization function across the frequency axis ensures robustness to spatial aliasing, thus leading to good performance.

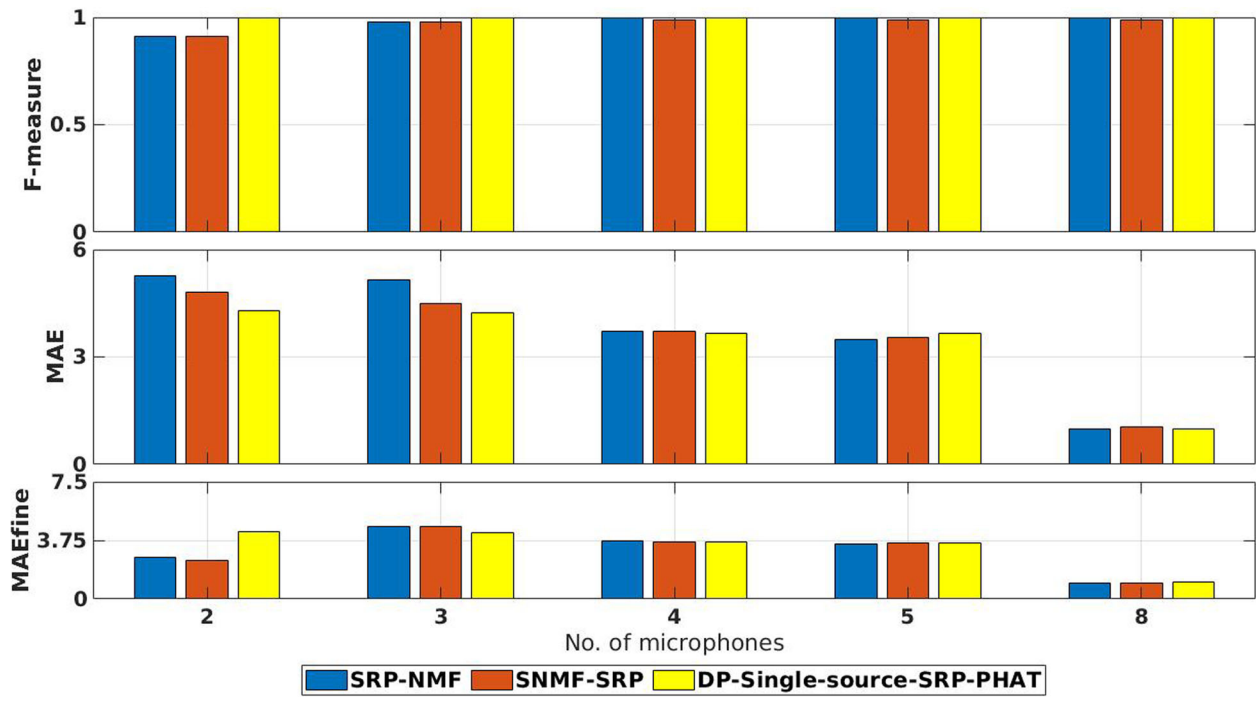

Fig. 7 Performance comparison of SNMF-SRP, DP-single-source-SRP-PHAT, and the proposed SRP-NMF in terms of the chosen instrumental metrics on the AACHEN dataset, on various subsets of the 8-microphone uniform linear array with an inter-microphone spacing of 8 cm

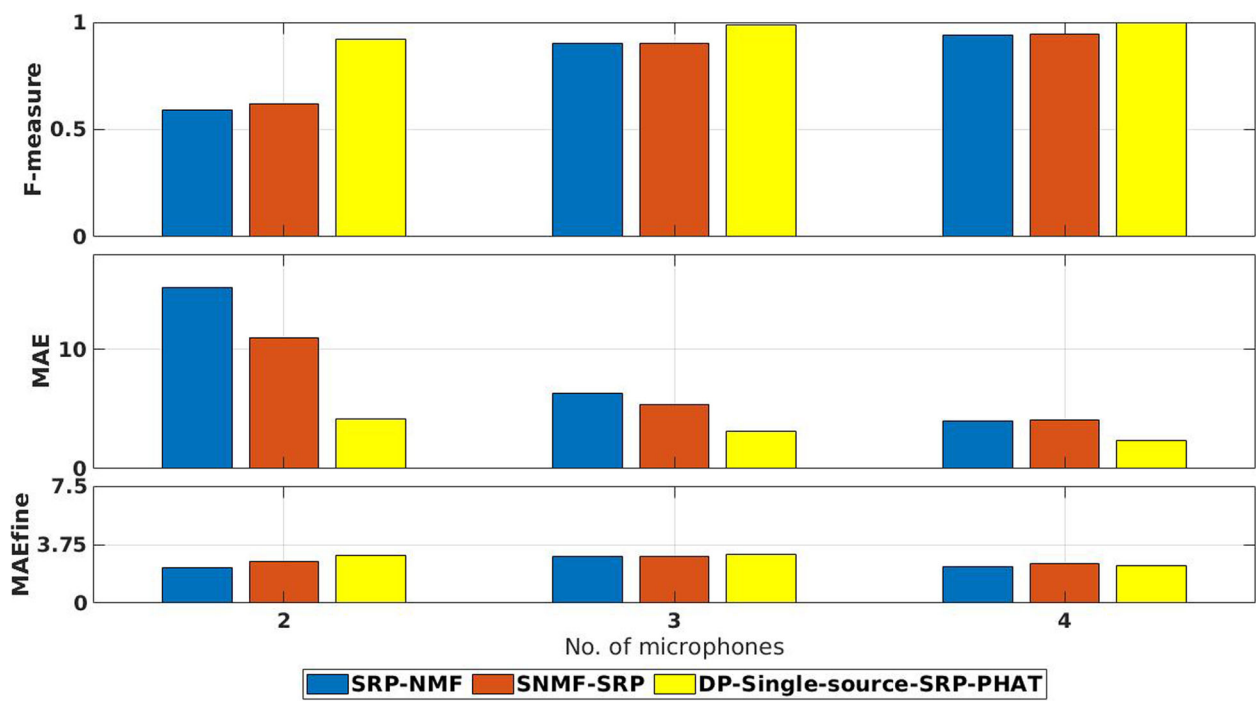

Fig. 8 Performance comparison of SNMF-SRP, DP-single-source-SRP-PHAT, and the proposed SRP-NMF in terms of the chosen instrumental metrics on the AACHEN dataset. The results are depicted for subsets of a uniform linear array with an inter-microphone spacing of 4 cm

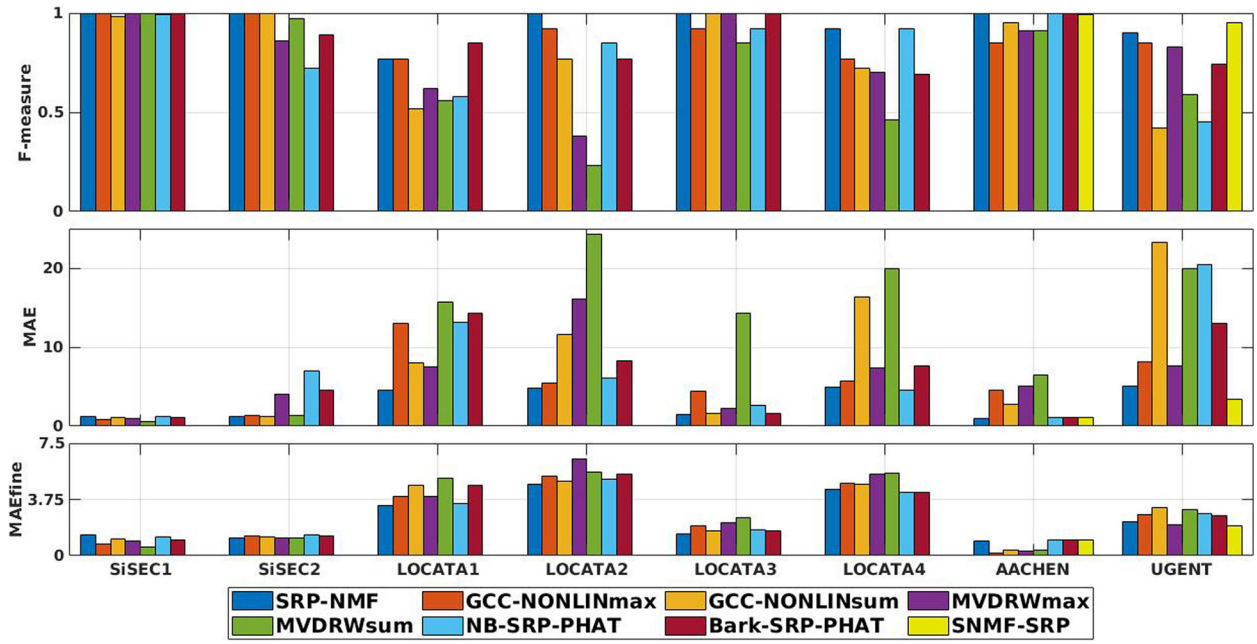

Fig. 9 F-measure, MAE and MAEfine of SRP-NMF and comparison approaches on various datasets. Mean azimuth error between ground truth azimuths and the detected azimuths (MAE) is the gross location accuracy metric. F-measure is the detection metric. Mean azimuth error between ground truth azimuths and the correctly detected azimuths (MAEfine) is the fine location accuracy metric

Further, the performance of SRP-NMF is consistently better than (or comparable to) that of the Bark-SRP-PHAT, indicating the benefit of the data-dependent weighted frequency averaging, as compared to a fixed frequency averaging.

Lastly, we also include a comparison with the SNMF-SRP (cf. Section 6.6) for the AACHEN and UGENT data. It may be seen, then, that this *supervised* approach outperforms all the other *unsupervised* approaches—which is expected, based on the results in Section 6.6 and Figs. 7 and 8. We note that since SNMF-SRP is based on the availability of the underlying source signals, it could not be applied to the LOCATA data, where this information is not consistently available. Further, we chose not to report performance metrics of this approach on the SiSEC data, since all approaches perform well in this case, and the performance of SNMF-SRP would add no value in a comparative analysis of the performances.

While the evaluation conclusively demonstrates the benefit of the proposed SRP-NMF approach, this comes at the cost of increased computational complexity. Its complexity is more than that of NB-SRP-PHAT and depends on the number of active atoms per frame. Further, we empirically observe that SRP-NMF gives good DoA estimates if the data segments are long ($> 3s$). We hypothesize, consequently, that the NMF dictionary atoms extracted from short segments may not be accurate. Therefore, in the current form, SRP-NMF is not suitable for real-time applications. However, with pre-trained dictionaries, the requirement of long data segments can be relaxed and SRP-NMF can be explored for real-time localization.

In order to better appreciate the benefits of the SRP-NMF approach, a graphical comparison of SRP-PHAT and SRP-NMF is presented in Figs. 10 and 11. These depict the histogram plots obtained by SRP-PHAT and

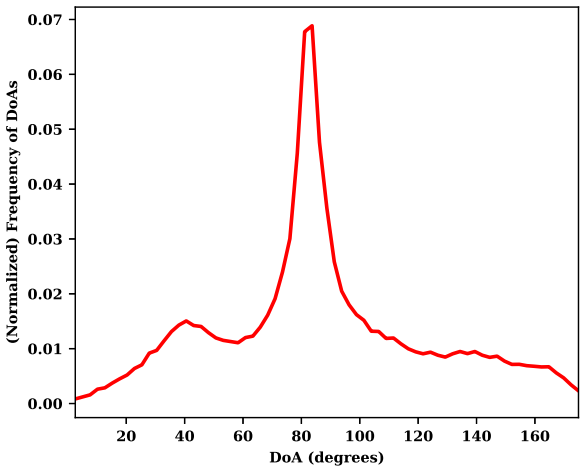

Fig. 10 Histogram of DoA estimates obtained by SRP-PHAT on *recording 2* (*Task2*) from LOCATA1 containing four concurrent speakers

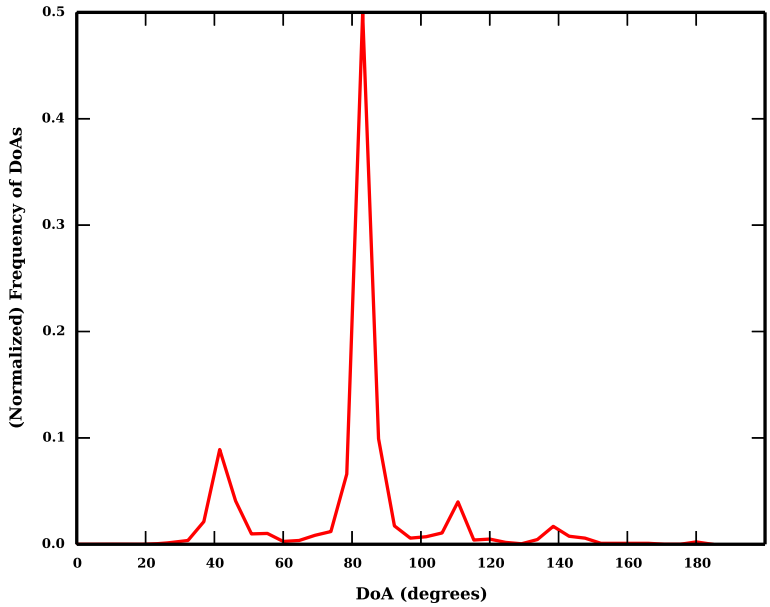

Fig. 11 Histogram of DoA estimates obtained by SRP-NMF on *recording 2* (*Task 2*) from LOCATA1 containing four concurrent speakers

SRP-NMF on a real-room mixture consisting of 4 concurrent speakers. Note that SRP-NMF clearly indicates the presence of the 4 sources, whereas the histogram of the SRP-PHAT approach (Fig. 10) does not present clear evidence of all 4 sources. The histogram plot in Fig. 11 can be further improved if subsampling is performed. Subsampling is an approach borrowed from *Word Embedding* in the field of NLP. Based on the observation that words with high frequency of occurrence do not contribute as much information as the words that occur more rarely, the frequent words are subsampled [45] to counter the imbalance between the frequent and rare words. In a similar manner, in the histogram of estimated DoAs, to counter the imbalances between frequent and occasional DoA estimates (e.g., due to a speaker being only active for a short while), the frequently occurring DoAs are subsampled after crossing a certain threshold. The subsampled version of Fig. 11 is shown in Fig. 12, where the benefit of subsampling is clearly visible.

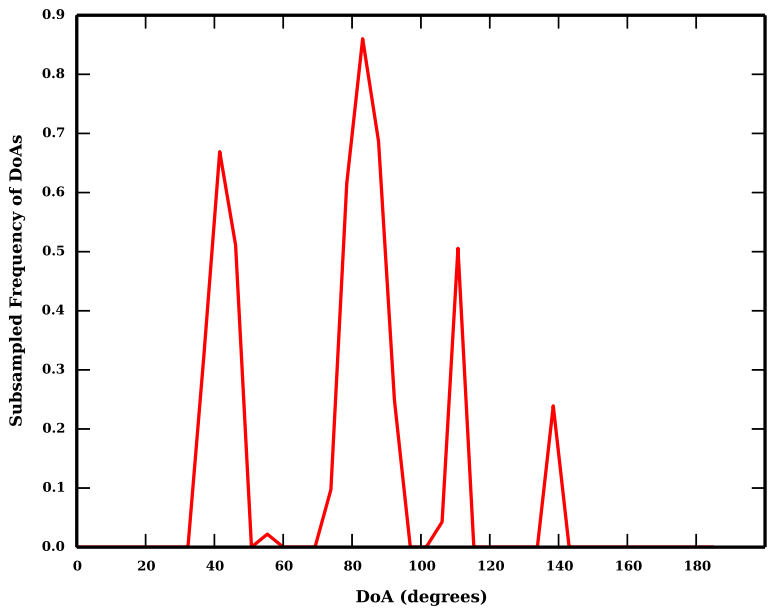

Fig. 12 Subsampled histogram of DoA estimates obtained by SRP-NMF on *recording 2* (*Task 2*) from LOCATA1 containing four concurrent speakers

7 Conclusions

SRP-NMF is a localization approach that uses the NMF atoms of the underlying sources to obtain a broadband localization estimate for each atom. By exploiting the sparsity of the sources in the time-atom domain, this still allows for the simultaneous localization of multiple sources in a time frame. Thereby the proposed approach combines the benefits of standard broadband and narrowband localization approaches. It can, therefore, be used with compact and large array configurations. Compared to the state-of-the-art narrowband and broadband approaches on data collected in natural room acoustic environments, and with various microphone configurations, the proposed approach can reliably localize the active sources in all cases, and with *a comparable* or *lower* localization error. The use of such an NMF-based decomposition and subsequent frequency grouping can be seamlessly extended in a variety of ways. For example, it can be combined with extant methods that improve the robustness of localization approaches to noise (e.g., in combination with the SNR weighting of the MVDR-based approaches), or it can be combined with a priori knowledge in the form of speaker-specific NMF atoms to localize only a specific speaker in the mix. It may also be modified for real-time applications with pre-learned universal NMF dictionary and online estimation of activation coefficients. We intend to address these extensions in future work.

Abbreviations

NMF: Non-negative matrix factorization; SRP: Steered-response power; TF: Time-frequency; STFT: Short-time Fourier transform; GCC: Generalized cross-correlation; AMDF: Average magnitude difference function; PHAT: Phase transform; DP: Direct-path; SNR: signal to noise ratio; TDoA: Time-difference of arrival; DoA: Direction of arrival; IPD: Inter-microphone phase difference; NB-SRP: Narrowband SRP; BB-SRP: Broadband SRP; CNMF: Constrained NMF; SNMF: Supervised NMF; MoG: Mixture of Gaussians; MBSS: Multi-channel BSS locate; SiSEC: Signal separation and evaluation campaign; LOCATA: Localization and tracking; MAE: Mean azimuth error; MVDR: Minimum variance distortionless response; MVDRW: Weighted MVDR; NLP: Natural language processing; AACHEN: RWTH AACHEN university; UGENT: Ghent University

Acknowledgements

Not applicable.

Authors' contributions

NM + ST: conceptualized SRP-NMF; ST implemented the approaches, introduced improvements, and conducted the experiments under the supervision of NM and SVG. NM and ST were involved in the writing. All the authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

SiSEC1 and SiSEC2 are publicly available at: *https://sisec.inria.fr/sisec-2016/2016-underdetermined-speech-and-music-mixtures*
LOCATA data is available at https://www.locata.lms.tf.fau.de/
AACHEN impulse responses are at http://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/multi-channel-impulse-response-database/
UGENT Multi-Channel Impulse Response Database is not publicly available but is available from the last author on reasonable request.
Details of the speech database used in the evaluations may be found in [43].

Competing interests

The authors state that they have no competing interests.

Author details

[1]Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India. [2]Present address: K L University, Guntur, Andhra Pradesh, India. [3]IDLab, Dept. Electronics & Information Systems, Ghent University - imec, Ghent, Belgium.

Received: 26 August 2020 Accepted: 2 February 2021
Published online: 03 March 2021

References

1. S. Rickard, O. Yilmaz, in *2002 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. C. Knapp, G. Carter, The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Proc. (TASSP). **24**(4), 320–327 (1976)
3. G. Jacovitti, G. Scarano, Discrete time techniques for time delay estimation. IEEE Trans. Signal Proc. (TSP). **41**(2), 525–533 (1993)
4. J. Benesty, Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. J. Acoust. Soc. Am. **107**(1), 384–391 (2000)
5. F. Talantzis, A. G. Constantinides, L. C. Polymenakos, Estimation of direction of arrival using information theory. IEEE Signal Proc. Lett. **12**, 561–564 (2005)
6. J. DiBiase, H. F. Silverman, M. S. Brandstein, in *Microphone arrays: signal processing techniques and applications*, ed. by M. Brandstein, D. Ward. Robust localization in reverberant rooms (Springer, New York, 2001), pp. 157–180
7. C. Zhang, D. Florencio, Z. Zhang, in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Why does PHAT work well in lownoise, reverberative environments? (2008), pp. 2565–2568
8. J. Valin, F. Michaud, J. Rouat, D. Letourneau, in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)*. Robust sound source localization using a microphone array on a mobile robot, vol. 2, (2003), pp. 1228–1233
9. Y. Rui, D. Florencio, in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Time delay estimation in the presence of correlated noise and reverberation, vol. 2, (2004), p. 133
10. H. Kang, M. Graczyk, J. Skoglund, in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. On pre-filtering strategies for the GCC-PHAT algorithm, (2016), pp. 1–5
11. Z. Wang, X. Zhang, D. Wang, Robust speaker localization guided by deep learning-based time-frequency masking. IEEE Trans. Audio Speech Lang. Process. (TASLP). **27**(1), 178–188 (2019)
12. J. M. Perez-Lorenzo, R. Viciana-Abad, P. Reche-Lopez, F. Rivas, J. Escolano, Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments. Appl. Acoust. **73**(8), 698–712 (2012)
13. B. Loesch, B. Yang, in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. Source number estimation and clustering for underdetermined blind source separation, (2008), pp. 1–4
14. M. I. Mandel, D. P. W. Ellis, T. Jebara, in *Proceedings of the Annual Conference on Neural Information Processing Systems*. An em algorithm for localizing multiple sound: sources in reverberant environments, (2006), pp. 953–960
15. O. Schwartz, S. Gannot, Speaker tracking using recursive EM algorithms. IEEE Trans. Audio Speech Lang. Process. (TASLP). **22**(2), 392–402 (2014)
16. N. Madhu, R. Martin, in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. A scalable framework for multiple speaker localization and tracking, (2008), pp. 1–4
17. 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)
18. M. Swartling, B. Sällberg, N. Grbić, Source localization for multiple speech sources using low complexity non-parametric source separation and clustering. Signal Process. **91**(8), 1781–1788 (2011)

19. C. Blandin, A. Ozerov, E. Vincent, Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Signal Process. **92**(8), 1950–1960 (2012)
20. P. Pertilä, Online blind speech separation using multiple acoustic speaker tracking and time-frequency masking. Comput. Speech Lang. **27**(3), 683–702 (2013)
21. E. Hadad, S. Gannot, in *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*. Multi-speaker direction of arrival estimation using SRP-PHAT algorithm with a weighted histogram, (2018), pp. 1–5
22. 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 (John Wiley & Sons, Ltd., New York, USA, 2008), pp. 135–170
23. D. Bechler, K. Kroschel, in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. Considering the second peak in the GCC function for multi-source TDOA estimation with a microphone array, (2003), pp. 315–318
24. C. Faller, J. Merimaa, Source localization in complex listening situations: selection of binaural cues based on interaural coherence. J. Acoust. Soc. Am. **116**(5), 3075–3089 (2004)
25. M. Togami, T. Sumiyoshi, A. Amano, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*. Stepwise phase difference restoration method for sound source localization using multiple microphone pairs, vol. 1, (2007), pp. 117–120
26. M. Togami, A. Amano, T. Sumiyoshi, Y. Obuchi, in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOA estimation method based on sparseness of speech sources for human symbiotic robots, (2009), pp. 3693–3696
27. J. Traa, P. Smaragdis, N. D. Stein, D. Wingate, in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Directional NMF for joint source localization and separation, (2015), pp. 1–5
28. H. Kayser, J. Anemüller, K. Adiloğlu, in *2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. Estimation of inter-channel phase differences using non-negative matrix factorization, (2014), pp. 77–80
29. A. Muñoz-Montoro, V. Montiel-Zafra, J. Carabias-Orti, J. Torre-Cruz, F. Canadas-Quesada, P. Vera-Candeas, in *Proceedings of the International Congress on Acoustics (ICA)*. Source localization using a spatial kernel based covariance model and supervised complex nonnegative matrix factorization, (2019), pp. 3321–3328
30. S. U. N. Wood, J. Rouat, S. Dupont, G. Pironkov, Blind speech separation and enhancement with GCC-NMF. IEEE Trans. Audio Speech Lang. Process. (TASLP). **25**(4), 745–755 (2017)
31. J. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments. Ph.D. dissertation*. (Brown University, Providence RI, USA, 2000)
32. T. Virtanen, J. F. Gemmeke, B. Raj, P. Smaragdis, Compositional models for audio processing: uncovering the structure of sound mixtures. IEEE Signal Process. Mag. **32**, 125–144 (2015)
33. J. Tchorz, B. Kollmeier, SNR estimation based on amplitude modulation analysis with applications to noise suppression. IEEE Trans. Speech Audio Process. (TSAP). **11**(3), 184–192 (2003)
34. S. Elshamy, N. Madhu, W. Tirry, T. Fingscheidt, Instantaneous a priori SNR estimation by cepstral excitation manipulation. IEEE Trans. Audio Speech Lang. Process. (TASLP). **25**(8), 1592–1605 (2017)
35. D. D. Lee, H. S. Seung, in *Advances in Neural Information Processing Systems 13*, ed. by T. K. Leen, T. G. Dietterich, and V. Tresp. Algorithms for non-negative matrix factorization, (2001), pp. 556–562
36. V. P. Pauca, J. Piper, R. J. Plemmons, Nonnegative matrix factorization for spectral data analysis. Linear Algebra Appl. **416**(1), 29–47 (2006). Special Issue devoted to the Haifa 2005 conference on matrix theory
37. R. Lebarbenchon, E. Camberlein, Multi-Channel BSS Locate (2018). https://bass-db.gforge.inria.fr/bss-locate/bss-locate. Accessed 4 2020
38. B. Loesch, B. Yang, in *9th International Conference on Latent variable analysis and signal separation (LVA/ICA)*. Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions (Springer, Berlin, Heidelberg, 2010), pp. 41–48
39. N. Ono, Z. Koldovský, S. Miyabe, N. Ito, in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. The 2013 signal separation evaluation campaign, (2013), pp. 1–6
40. H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, W. Kellermann, in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. The LOCATA challenge data corpus for acoustic source localization and tracking, (2018), pp. 410–414
41. C. Veaux, J. Yamagishi, K. MacDonald, *English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*. (University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019). https://doi.org/10.7488/ds/2645
42. Multi-channel impulse response database. https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/multi-channel-impulse-response-database/. Accessed 12 2020
43. P. Kabal, *TSP speech database. Technical report*. (Telecommunications and Signal Processing Laboratory, McGill University, Canada, 2002)
44. C. Févotte, E. Vincent, A. Ozerov, in *Audio Source Separation*, ed. by S. Makino. Single-channel audio source separation with NMF: divergences, constraints and algorithms (Springer, Cham, 2018), pp. 1–24
45. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, in *Advances in neural information processing systems 26*, ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Distributed representations of words and phrases and their compositionality, (2013), pp. 3111–3119

Publisher's Note