DIRECTIONAL NMF FOR JOINT SOURCE LOCALIZATION AND SEPARATION

Johannes Traa[1]*, Paris Smaragdis*[1,2]

[1] University of Illinois at Urbana-Champaign
[2] Adobe Systems, Inc.
traa2@illinois.edu , paris@illinois.edu

Noah Stein, David Wingate

Lyric Labs
Analog Devices, Inc.
noah.stein@analog.com , david.wingate@analog.com

ABSTRACT

We propose a method for simultaneously localizing and separating speech signals by factorizing a non-negative matrix of Steered Response Power (SRP) measurements. We use a probabilistic interpretation of the SRP function to compute a wideband SRP matrix. Non-negative Matrix Factorization (NMF) is used to decompose it into three terms that describe (1) the source distributions over spatial directions, (2) the overall source activations, and (3) the source activations over the time-frequency (TF) plane. The first term indicates the sources' Directions-of-Arrival (DOA) and the latter two terms provide TF weights for separating the sources. Experiments show that this joint approach out-performs a sequential SRP localization + beamforming method.

***Index Terms*—** steered response power, source localization, beamforming, nonnegative matrix factorization

1. INTRODUCTION

This paper focuses on the localization and separation of speakers recorded with a compact microphone array. Beamforming [1] is often applied to enhance directional signals in the presence of interferers and noise. This involves minimizing the output power of a linear filter subject to distortionless and/or null constraints that protect the target signal and block undesired signals. Well-known examples are the Minimum-Variance Distortionless Response (MVDR) and Linearly-Constrained Minimum-Variance (LCMV) beamformers [2]. These both assume that estimates of the sources' Directions-Of-Arrival (DOA) are available.

Steered Response Power (SRP) localization attempts to identify the DOAs by computing the output power of a beamformer over all DOAs and locating peaks in the resulting SRP function. A related method searches for peaks in the Generalized Cross-Correlation (GCC) function [3]. The authors in [4] used this approach to localize speech sources with a microphone array.

The authors in [4] described the relationship between the GCC and SRP functions and how they relate to a probabilistic model for the observed signal. The authors in [5] and [6] describe the relationship between SRP-based localization using an MVDR beamformer and a maximum-likelihood formulation. In [7], the authors model the source DFT coefficients as zero-mean Gaussian random variables. They use an EM algorithm to learn the sources' covariance parameters and apply multichannel Wiener filtering to separate the sources. A supervised Bayesian approach was proposed in [8] to jointly solve the localization and separation problems.

Nonnegative Matrix Factorization [9] is a popular algorithm for decomposing a spectrogram into a set of spectral templates and their activations in time. This approach was applied in [10] to transcribe polyphonic piano music. In [11], the authors used NMF to decompose a matrix of GCC functions in order to estimate non-linear phase difference patterns in a speech mixture.

This paper combines techniques from beamforming, SRP localization, and NMF to simultaneously localize and separate multiple speakers in an unsupervised setting. We call this approach Directional NMF (D-NMF). We compute a non-negative, wideband SRP matrix from the observed data and factorize it into three terms that describe the spatial locations of the sources and their activations across the Time-Frequency (TF) plane. The latter information is used to form soft TF masks and separate the sources. TF masking [12] is motivated by the disjointness of speech in the TF domain [13]. We show that the proposed method outperforms a traditional sequential approach.

2. ARRAY PROCESSING

2.1. Data model

Consider a noisy, convolutive model of an audio signal recorded at M microphones:

$$x_m[t] = a_m[t] * s[t] + n_m[t] \quad , \tag{1}$$

where $x_m[t]$ is the recorded sample at the m^{th} channel, $a_m[t]$ is the room impulse response (RIR) from the source to the m^{th} channel, $n_m[t]$ is a Gaussian noise process, and $*$ denotes convolution. We apply the Short-Time Fourier Transform (STFT) to de-couple the signal components across frequency. At frequency $f \in [1, F]$ and time frame $t \in [1, T]$, we have:

$$\mathbf{x}_{ft} = \mathbf{a}_f \, s_{ft} + \mathbf{n}_{ft} \quad , \quad \mathbf{n}_{ft} \sim \mathcal{N}\left(\mathbf{0}, \sigma_f^2 \, \mathbf{I}\right) \quad , \tag{2}$$

where $\mathbf{x}_{ft} \in \mathbb{C}^M$ is an observed data vector, $\mathbf{a}_f \in \mathbb{C}^M$ is a mixing vector, $s_{ft} \in \mathbb{C}$ is the source coefficient, and $\mathbf{n}_{ft} \in \mathbb{C}^M$ contains the noise coefficients. The source and noise coefficients are assumed to be statistically independent.

If the signal propagates from a point source in the far field of the array in an anechoic environment, we can express the mixing vector $\mathbf{a}_f$ in terms of the source's Direction-Of-Arrival (DOA) vector $\boldsymbol{\phi} \in \mathbb{R}^3, \|\boldsymbol{\phi}\|_2 = 1$. We define the unit steering vector:

$$\mathbf{a}_f(\boldsymbol{\phi}) = \frac{1}{\sqrt{M}} \exp\left(j \frac{2\pi l_f}{u} \mathbf{m}^\top \boldsymbol{\phi}\right) \quad , \tag{3}$$

where $\mathbf{m} \in \mathbb{R}^{3 \times M}$ denotes the matrix of M microphone positions, u is the speed of sound, and l_f is the center frequency of

the f^{th} band. The far field assumption implies that the largest inter-microphone spacing is small relative to the distances between the array and the sources.

When multiple signals are active simultaneously, we can still apply the one-source model given that the signals are approximately disjoint in the TF plane [13]. Formally, the disjointness condition:

$$\forall\, f,t,k \neq k' \quad |s_{ft}^{(k)}| \cdot |s_{ft}^{(k')}| \approx 0 \ . \tag{4}$$

says that at most one source has appreciable energy in any TF bin. It has served as the foundation for many effective separation algorithms such as DUET [14]. In the rest of this paper, we will assume that (4) holds for speech signals.

2.2. Beamforming

Linear filtering algorithms in the frequency domain are often used for enhancing/separating directional signals. Consider a linear filter described by a weight vector $\mathbf{w}_f \in \mathbb{C}^M$ that aims to reconstruct a source coefficient s_{ft} via $\widehat{s}_{ft} = \mathbf{w}_f^H \mathbf{x}_{ft}$.

We can define the optimal $\mathbf{w}_f$ as that which minimizes the expected output power:

$$P = \mathrm{E}\left[\, |\widehat{s}_{ft}|^2 \,\right] = \mathbf{w}_f^H \, \mathbf{R}_f \, \mathbf{w}_f \ , \tag{5}$$

of the filter without affecting the signal at DOA $\boldsymbol{\phi}$:

$$\widehat{\mathbf{w}}_f = \underset{\mathbf{w}_f}{\operatorname{argmin}} \ \mathbf{w}_f^H \mathbf{R}_f \, \mathbf{w}_f \ , \tag{6}$$

$$\text{s.t.} \quad \mathbf{a}_f^H \left(\boldsymbol{\phi}\right) \mathbf{w}_f = 1 \ , \tag{7}$$

for a channel correlation matrix $\mathbf{R}_f = \mathrm{E}\left[\mathbf{x}_{ft}\, \mathbf{x}_{ft}^H\right]$.

The solution is the MVDR beamformer:

$$\widehat{\mathbf{w}}_f^{MVDR} = \frac{\mathbf{R}_f^{-1} \mathbf{a}_f \left(\boldsymbol{\phi}\right)}{\mathbf{a}_f^H \left(\boldsymbol{\phi}\right) \mathbf{R}_f^{-1} \mathbf{a}_f \left(\boldsymbol{\phi}\right)} \ . \tag{8}$$

If $\mathbf{R}_f = \mathbf{I}$, this reduces to the data-independent Delay-and-Sum (DS) beamformer:

$$\widehat{\mathbf{w}}_f^{DS} = \mathbf{a}_f \left(\boldsymbol{\phi}\right) \ . \tag{9}$$

The MVDR beamformer provides better noise suppression when $\mathbf{R}_f \propto \mathbf{I}$. It can be generalized for multiple targets and/or interferers via the multiply-constrained optimization problem:

$$\widehat{\mathbf{w}}_f = \underset{\mathbf{w}_f}{\operatorname{argmin}} \ \mathbf{w}_f^H \mathbf{R}_f \, \mathbf{w}_f \ , \tag{10}$$

$$\text{s.t.} \quad \mathbf{A}_f^H \left(\boldsymbol{\Phi}\right) \mathbf{w}_f = \mathbf{u} \ , \tag{11}$$

where $\mathbf{u} \in \mathbb{R}^K$ contains desired gains at each DOA. The solution is the Linearly-Constrained Minimum-Variance (LCMV) beamformer:

$$\widehat{\mathbf{w}}_f^{LCMV} = \mathbf{R}_f^{-1} \mathbf{A}_f \left(\boldsymbol{\Phi}\right) \left[\mathbf{A}_f^H \left(\boldsymbol{\Phi}\right) \mathbf{R}_f^{-1} \mathbf{A}_f \left(\boldsymbol{\Phi}\right)\right]^{-1} \mathbf{u} \ . \tag{12}$$

2.3. Steered Response Power (SRP) localization

The beamformers described in the previous section can estimate the source coefficients if the true source DOA(s) $\boldsymbol{\Phi}$ are known. A simple way to estimate a source's DOA is to look for peaks in the output power of the DS or MVDR beamformers. This approach is referred to as Steered Response Power (SRP) localization.

Consider the DS output power for a single data vector and look direction $\boldsymbol{\theta}$ in the presence of one source with DOA $\boldsymbol{\phi}$. We use (2) to write:

$$P\left(\boldsymbol{\theta}\right) = \frac{1}{T} \sum_{t=1}^{T} |\, \mathbf{a}_f^H \left(\boldsymbol{\theta}\right) \mathbf{x}_{ft}|^2 \leq \frac{1}{T} \sum_{t=1}^{T} |s_{ft}|^2 + C \ , \tag{13}$$

where $C \xrightarrow{T\to\infty} \sigma_f^2$ is due to the additive noise term and equality is achieved only if $\boldsymbol{\theta} = \boldsymbol{\phi}$. This suggests that we can identify $\boldsymbol{\phi}$ by scanning over all feasible $\boldsymbol{\theta}$'s and choosing the one with the largest value of $P\left(\boldsymbol{\theta}\right)$. The Phase Transform (PHAT) [3] is often used to enhance the SRP function by setting all the magnitudes of the components of $\mathbf{x}_{ft}$ to 1. This isolates the data's phase information that is crucial to differentiating DOAs.

2.4. Probabilistic SRP Model

The single-source propagation model described in (2) corresponds to a Gaussian likelihood for a data vector:

$$\mathcal{L}_{ft}\left(\boldsymbol{\theta}\right) = \mathcal{N}\left(\mathbf{x}_{ft}\,;\, \boldsymbol{\mu}_{ft}\,,\, \sigma_f^2 \mathbf{I}\right) \ . \tag{14}$$

The mean $\boldsymbol{\mu}_{ft}$ encodes the expected value of $\mathbf{x}_{ft}$:

$$\boldsymbol{\mu}_{ft} = \mathrm{E}\left[\mathbf{x}_{ft}\right] = \mathbf{a}_f \left(\boldsymbol{\theta}\right) \mathrm{E}\left[s_{ft}\right] \ , \tag{15}$$

for a hypothesized DOA $\boldsymbol{\theta}$. Since the source coefficients are unavailable (we are trying to recover them), we replace the expectation in (15) with a least-squares estimate of s_{ft} to write:

$$\widehat{\boldsymbol{\mu}}_{ft} = \mathbf{a}_f \left(\boldsymbol{\theta}\right) \widehat{s}_{ft} = \mathbf{a}_f \left(\boldsymbol{\theta}\right) \mathbf{a}_f^H \left(\boldsymbol{\theta}\right) \mathbf{x}_{ft} \ . \tag{16}$$

Substituting (16) into (14) and expanding, we can write:

$$\log \mathcal{L}_{ft}\left(\boldsymbol{\theta}\right) \propto -\frac{1}{2\sigma_f^2} \left(\|\mathbf{x}_{ft}\|_2^2 - |\mathbf{a}_f^H \left(\boldsymbol{\theta}\right) \mathbf{x}_{ft}|^2\right) \ . \tag{17}$$

This is equivalent to the SRP function defined in (13) in terms of identifying the true DOA. As in [4], we have seen that the localization task can be cast as a maximum-likelihood (ML) problem.

3. DIRECTIONAL NMF

In this section, we describe how the SRP matrix is calculated and how NMF can be used to factorize it.

3.1. SRP Matrix

We evaluate the likelihood $\mathcal{L}_{ft}\left(\boldsymbol{\theta}\right)$ for each data vector $\mathbf{x}_{ft}$ (after PHAT weighting [3]) over a set of D DOAs of interest. In this paper, we sample DOAs over the unit hemisphere. This results in FT SRP vectors $\mathbf{L}_{ft} \in \mathbb{R}^D$. Concatenating these vectors, we form the $D \times FT$ SRP matrix $\mathbf{L}$.

3.2. Matrix Factorization

We use NMF [9] to decompose the SRP matrix $\mathbf{L} \in \mathbb{R}^{D \times FT}$ into the product of three terms: $\mathbf{W}^{D \times K}$ with source DOA activations in the columns, $\mathbf{A} \in \mathbb{R}^{K \times K}$ with source weights on the diagonal, and $\mathbf{H}^{K \times FT}$ with source TF activations in the rows. Formally, the NMF problem is stated as:

$$\{\widehat{\mathbf{W}, \mathbf{A}, \mathbf{H}}\} = \underset{\mathbf{W},\mathbf{A},\mathbf{H}}{\operatorname{argmin}} \ KL\left(\mathbf{L} \| \mathbf{W}\, \mathbf{A}\, \mathbf{H}\right) \tag{18}$$

$$\text{s.t.} \quad \mathbf{W} \geq \mathbf{0} \ , \quad \mathbf{A} \geq \mathbf{0} \ , \quad \mathbf{H} \geq \mathbf{0} \ . \tag{19}$$

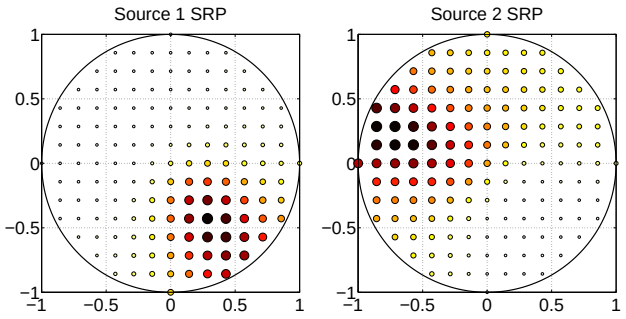

Figure 1: SRP distributions (from $\mathbf{W}_{:j}, j = 1, 2$) for two sources located on the DOA hemisphere. The hemisphere is flattened such that (azimuth,zenith) points map to (argument,modulus) points. Larger/darker circles denote areas of higher probability mass.

We apply multiplicative updates like those proposed in [9] to iteratively solve for the factors:

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left(\left(\mathbf{L} \oslash \widehat{\mathbf{L}} \right) \mathbf{H}^\top \mathbf{A}^\top \right) \oslash \left(\mathbf{J}\mathbf{H}^\top \mathbf{A}^\top \right) \quad , \tag{20}$$

$$\mathbf{A} \leftarrow \mathbf{A} \odot \left(\mathbf{W}^\top \left(\mathbf{L} \oslash \widehat{\mathbf{L}} \right) \mathbf{H}^\top \right) \oslash \left(\mathbf{W}^\top \mathbf{J}\mathbf{H}^\top \right) \quad , \tag{21}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\mathbf{A}^\top \mathbf{W}^\top \left(\mathbf{L} \oslash \widehat{\mathbf{L}} \right) \right) \oslash \left(\mathbf{A}^\top \mathbf{W}^\top \mathbf{J} \right) \quad , \tag{22}$$

where $\odot$ and $\oslash$ denote element-wise multiplication and division, $\mathbf{J}$ is a $D \times FT$ matrix of ones, and $\widehat{\mathbf{L}} = \mathbf{WAH}$ is a reconstruction of the SRP matrix. To avoid scale ambiguities, we normalize the columns of $\mathbf{W}$ and the rows of $\mathbf{H}$:

$$\mathbf{A} \leftarrow \text{diag}\left(\mathbf{W}^\top \mathbf{1} \right) \mathbf{A}\, \text{diag}\left(\mathbf{H}\, \mathbf{1} \right) \quad , \tag{23}$$

$$\mathbf{W} \leftarrow \mathbf{W}\, \text{diag}\left(\mathbf{W}^\top \mathbf{1} \right)^{-1} \quad , \tag{24}$$

$$\mathbf{H} \leftarrow \text{diag}\left(\mathbf{H}\, \mathbf{1} \right)^{-1} \mathbf{H} \quad , \tag{25}$$

where $\mathbf{1}$ is a $K \times 1$ ones vector.

We can interpret the columns of $\mathbf{W}$ as distributions over DOA space $p\left(\boldsymbol{\theta}_k\right), k = 1, \ldots, K$, and the rows of $\mathbf{H}$ as time-frequency distributions $p\left(f, t\right)$. We can easily induce sparsity on $\mathbf{A}$ to automatically estimate the number of sources (this is left for future work). Figure 1 shows two SRP distributions found by NMF for a mixture of two speakers.

3.3. Dictionary Constraints

We can also interpret $\mathbf{W}$ as representing activation weights over a set of SRP function templates. We introduce a dictionary matrix $\mathbf{D} \in \mathbb{R}^{D \times D}$ whose columns represent the ideal SRP functions in (17) over DOAs for each steering angle. The NMF problem becomes $\mathbf{L} \approx \mathbf{DWAH}$ with multiplicative updates very similar to those in (20)-(22). This formulation constrains the (effective) SRP basis vectors in $\mathbf{DW}$ to be consistent with our data model.

3.4. SRP Matching Across Frequency

In [15], a wideband beamformer is described that has a constant beam width at all frequencies. Similarly, we would like all the SRP functions to be matched. As the frequency f increases, the DS beam pattern gets increasingly sharp, resulting in a non-uniformity of the SRP vectors corresponding to a single source. This degrades the factorization considerably. We can choose the variances σ_f^2 (and normalization constants) wisely to counter-act this. For simplicity, we will optimize the precision $\lambda_f = 1/\sigma_f^2$.

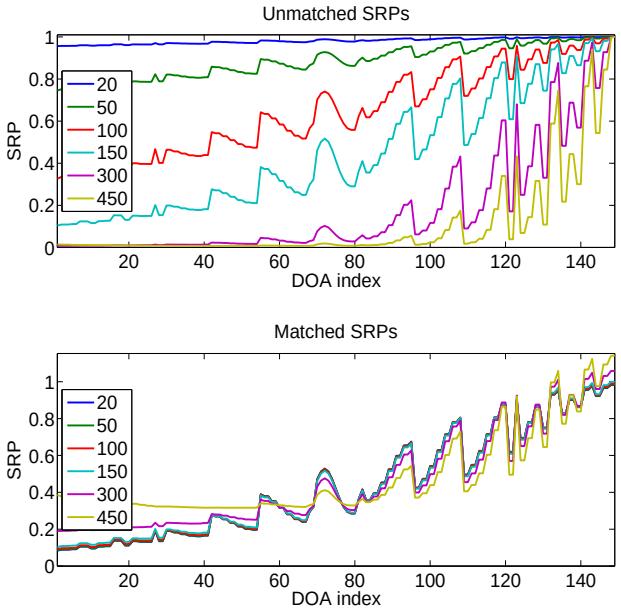

Figure 2: SRP functions at various frequencies evaluated over the DOA grid in Figure 1 for $\boldsymbol{\phi} = [0\,1\,0]^\top$ and a 4×4 centimeter square, 8-channel array. (Top) No matching ($\lambda_{f_{ref}} = 1$). (Bottom) Matched SRPs with $f_{ref} = 150$ and $\lambda_{f_{ref}} = 10$.

We minimize the squared error between the SRP function at a frequency f and the SRP function at a reference frequency f_{ref}, averaged over all source DOAs $\boldsymbol{\phi}$ and directions $\boldsymbol{\theta}$:

$$\widehat{\lambda}_f = \underset{\lambda_f}{\text{argmin}}\ e\left(\lambda_f\right) \quad , \tag{26}$$

$$e\left(\lambda_f\right) = \iint \left[\mathcal{L}_f\left(\boldsymbol{\theta}, \boldsymbol{\phi}\right) - \mathcal{L}_{f_{ref}}\left(\boldsymbol{\theta}, \boldsymbol{\phi}\right) \right]^2 \, d\boldsymbol{\theta}\, d\boldsymbol{\phi} \quad . \tag{27}$$

We use SRP values corresponding to a noiseless data model with $|s_{ft}|^2 = 1$ so that we can write[1]:

$$\log \mathcal{L}_f\left(\boldsymbol{\theta}, \boldsymbol{\phi}\right) = \log c_f - \frac{\lambda_f}{2} \left(1 - |\mathbf{a}_f^H\left(\boldsymbol{\theta}\right) \mathbf{a}_f\left(\boldsymbol{\phi}\right)|^2 \right) . \tag{28}$$

The optimization can be solved off-line with coordinate descent using a Newton step for λ_f and a least-squares solution for c_f. We initialize at $(\lambda_f, c_f) = (0, 1)$ and use a discretized set of directions. Given these initial conditions, the problem is locally convex. The reference precision $\lambda_{f_{ref}}$ is used to adjust the sharpness of the matched SRPs. Without loss of generality, we set $c_{f_{ref}} = 1$. SRP functions with and without matching are shown in Figure 2. This shows that the matching is especially important at lower frequencies where important speech information is present.

We set the reference frequency f_{ref} as high as possible while avoiding aliasing effects:

[1] This corresponds to PHAT weighting on the observed data vectors: $\mathbf{x} \leftarrow \frac{1}{\sqrt{M}} \mathbf{x}./|\mathbf{x}|$. The same weighting is used to compute the SPR matrix $\mathbf{L}$.

$$f_{ref} = \min\left(\left\lfloor \frac{F \cdot u}{d_{max} \cdot f_s} \right\rfloor, F\right) \quad , \tag{29}$$

where u is the speed of sound, d_{max} is the maximum inter-sensor distance, and f_s is the sampling rate. We use the reference precision $\lambda_{f_{ref}}$ to adjust the sharpness of the peaks. Sharper SRP functions lead to sparsity in the spatial activations (i.e. $\mathbf{W}$), which typically induces sparsity in the TF activations (i.e. $\mathbf{H}$). As NMF is a local optimization procedure, excessively sparse SRP functions may cause issues with local optima. However, SRPs that are not sparse enough result in less precise localization and, therefore, spatial isolation of the sources. In practice, we tune $\lambda_{f_{ref}}$ to achieve a desired balance between precision and algorithmic robustness.

4. SOURCE LOCALIZATION AND SEPARATION

Once the NMF procedure has converged, we can use the learned factors to localize and separate the sources. Maximum-likelihood DOA vectors are estimated from $\mathbf{W}$ as:

$$\widehat{\boldsymbol{\theta}}^{(k)} = \sum_{d=1}^{D} W_{dk}\boldsymbol{\theta}_d \Big/ \left\| \sum_{d=1}^{D} W_{dk}\boldsymbol{\theta}_d \right\|_2 \quad , \tag{30}$$

and soft TF masks are derived from $\mathbf{A}$ and $\mathbf{H}$ as:

$$w_{ft}^{(k)} \propto A_{kk} H_{kI_{ft}} \quad , \tag{31}$$

where I_{ft} is an indexing function that returns the column index in $\mathbf{H}$ corresponding to the $(f,t)^{\text{th}}$ data point. We approximate each source's STFT matrix as:

$$\widehat{\mathbf{S}}^{(k)} = \mathbf{w}^{(k)} \odot \mathbf{X} \quad , \tag{32}$$

and use the overlap-add algorithm to reconstruct the corresponding time-domain source signals.

5. EXPERIMENTS

We ran experiments with a 4×4 centimeter square, 8-channel array located in the middle of a $5 \times 5 \times 5$ meter room simulator. We chose $K = 4$ source DOAs $\boldsymbol{\phi}_k$ uniformly at random on a hemisphere of radius 1 meter above the array subject to a minimum angular separation, i.e. $\forall\, i \neq j \;\; \boldsymbol{\phi}_i^\top \boldsymbol{\phi}_j \geq \cos\left(2\pi/(K+2)\right)$. STFTs were computed with window and hop sizes of 1024 and 256. We used 3-second-long speech sentences from the TSP corpus [16] (down-sampled to 16 kHz) for the source signals. We corrupted the recorded audio with white Gaussian noise for an input SNR of 5.4 dB and simulated reverberation with a T_{60} time of 270 milliseconds using the image method [17].

We compared four methods: (1) D-NMF with a dictionary, (2) D-NMF without a dictionary, (3) D-NMF with a dictionary initialized with the output of the dictionary-free method, and (4) standard SRP localization. All SRP s were calculated along the grid shown in Figure 1. To form a TF mask for the last method, we applied K LCMV beamformers to isolate each source and normalized the resulting output energies in each TF bin. We ran all NMF algorithms for 50 iterations and evaluated the localization error (i.e. average angular error) as:

$$e = \min_{\mathcal{P}} \frac{1}{K} \sum_{k=1}^{K} \cos^{-1}\left(\boldsymbol{\phi}_k^\top \widehat{\boldsymbol{\theta}}^{\mathcal{P}(k)}\right) \quad , \tag{33}$$

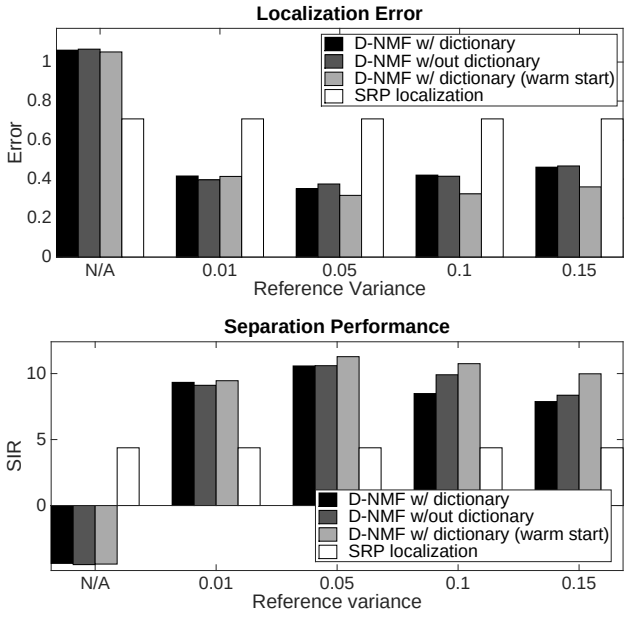

Figure 3: Localization error and source separation results for various algorithms and reference variances ($f_{ref} = 192$). The N/A group shows results without SRP matching.

where $\mathcal{P} : \{1, K\} \rightarrow \{1, K\}$ is a permutation mapping. We evaluated the Signal-to-Interference Ratio (SIR) of the separated signals with the BSSEval toolbox [18].[2], averaged over 100 trials.

Figure 3 shows localization and separation results. Without matching the SRP functions across frequencies, D-NMF does quite poorly. After matching, all variants of D-NMF out-perform standard SRP localization and beamforming. Furthermore, we find that a "warm start" is necessary for D-NMF to take advantage of the SRP dictionary. This is due to the presence of more local optima and a slower convergence rate when D-NMF is constrained by the dictionary. We also note that there appears to be an optimal value for the reference variance $\sigma^2_{f_{ref}} = 1/\lambda_{f_{ref}}$. For our experimental set-up, this is $\widehat{\sigma}^2_{f_{ref}} \approx 0.1$ with a reference frequency index of $f_{ref} = 192$. Overall, the best results are given by dictionary-constrained D-NMF initialized with unconstrained D-NMF.

6. CONCLUSIONS

We have presented a method for simultaneously localizing and separating multiple sources that involves factorizing a non-negative matrix of Steered Response Power (SRP) measurements. We showed that a naive approach is sub-optimal in that the SRP functions for a fixed source direction vary across frequencies. A convex optimization procedure that matched the SRP functions across frequencies was used to remedy this, leading to significantly better separation and localization performance. We showed that the proposed approach out-performs a sequential SRP localization method. We note that our approach could be extended in various ways to the on-line setting via on-line dictionary learning techniques [19, 20].

[2]Other BSSEval metrics qualitatively mirrored the SIR results.

7. REFERENCES

[1] H. K. van Trees, *Detection, Estimation, and Modulation Theory: Optimum Array Processing (Part IV)*, Wiley, 2002.

[2] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[3] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[4] S. T. Birchfield and D. K. Gillmor, "Fast Bayesian acoustic localization," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 1793–1796, 2002.

[5] K. Harmanci, J. Tabrikian, and J.L. Krolik, "Relationships between adaptive minimum variance beamforming and optimal source localization," *IEEE Transactions on Signal Processing*, vol. 48, no. 1, 2000.

[6] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 528–548, 2008.

[7] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[8] K. H. Knuth, "Bayesian source separation and localization," *Proceedings of SPIE: Bayesian Inference for Inverse Problems*, vol. 3459, pp. 147–158, 1998.

[9] D. L. Daniel and H. S. Seung, "Algorithms for non-negative matrix factorization," *Conference on Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

[10] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[11] H. Kayser, H. Anemuller, and K. Adiloglu, "Estimation of inter-channel phase differences using non-negative matrix factorization," *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 77–80, 2014.

[12] Y. Li and D Wang, "On the optimality of ideal binary time-frequency masks," *Elsevier - Speech Communication*, vol. 51, pp. 230–239, 2009.

[13] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," *IEEE Conference on Acoustics, Speech, and Signal Processing (CASSP)*, vol. 1, pp. 529–532, 2002.

[14] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[15] J. Benesty, J. Chen, and Y. Huang, *Topics in Signal Processing: Microphone Array Signal Processing*, vol. 1, Springer, 2008.

[16] P. Kabal, "TSP speech database," 2002, Telecommunications and Signal Processing Lab, McGill University.

[17] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[18] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462 –1469, 2006.

[19] Zhiyao Duan, Gautham J. Mysore, and Paris Smaragdis, "Online plca for real-time semi-supervised source separation," in *Proceedings of the 10th International Conference on Latent Variable Analysis and Signal Separation*. 2012, pp. 34–41, Springer-Verlag.

[20] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 19–60, 2010.