This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing



# Direction of Arrival with One Microphone, a few LEGOs, and Non-Negative Matrix Factorization

Dalia El Badawy and Ivan Dokmanić, *Member, IEEE*

***Abstract*—Conventional approaches to sound source localization require at least two microphones. It is known, however, that people with unilateral hearing loss can also localize sounds. Monaural localization is possible thanks to the scattering by the head, though it hinges on learning the spectra of the various sources. We take inspiration from this human ability to propose algorithms for accurate sound source localization using a single microphone embedded in an arbitrary scattering structure. The structure modifies the frequency response of the microphone in a direction-dependent way giving each direction a signature. While knowing those signatures is sufficient to localize sources of white noise, localizing speech is much more challenging: it is an ill-posed inverse problem which we regularize by prior knowledge in the form of learned non-negative dictionaries. We demonstrate a monaural speech localization algorithm based on non-negative matrix factorization that does not depend on sophisticated, designed scatterers. In fact, we show experimental results with ad hoc scatterers made of LEGO bricks. Even with these rudimentary structures we can accurately localize arbitrary speakers; that is, we do not need to learn the dictionary for the particular speaker to be localized. Finally, we discuss multi-source localization and the related limitations of our approach.**

***Index Terms*—direction-of-arrival estimation, group sparsity, monaural localization, non-negative matrix factorization, sound scattering, universal speech model**

## I. Introduction

In this paper, we present a computational study of the role of scattering in sound source localization. We study a setting in which localization is a priori not possible: that of a single microphone, referred to as monaural localization. It is well established that people with normal hearing localize sounds primarily from binaural cues—those that require both ears. Different directions of arrival (DoA) result in different interaural time differences which are the dominant cues for localization at lower frequencies, as well as in interaural level differences (ILD) which are dominant at higher frequencies [1]. The latter are linked to the head-related transfer function (HRTF) which encodes how human and animal heads, ears, and torsos scatter incoming sound waves. This scattering results in direction-dependent filtering whereby frequencies are selectively attenuated or boosted; the exact filtering depends on the shape of the head and ears and therefore varies for different people and animals. Thus the same mechanism responsible for frequency-dependent ILDs in the HRTF also provides monaural cues. The question is then, can these monaural cues embedded in the HRTF be used for localization?

Indeed, monaural cues are known to help localize in elevation [1] and resolve the front/back confusion [2]: two cases where binaural cues are not sufficient. Additionally, studies on the HRTFs of cats [3] and bats [4] also reveal their use for localization in both azimuth and elevation, albeit in a binaural setting. This implies that the directional selectivity of the HRTF i.e., the monaural cues, is sufficient to enable people with unilateral hearing loss to localize sounds, though with a reduced accuracy compared to the binaural case [5].

### A. Related Work

Combining HRTF-like directional selectivity with source models has already been explored in the literature [6], [7], [8], [9]. For example, in one study [8], a small microphone enclosure was used to localize one source with the help of a Hidden Markov Model (HMM) trained on a variety of sounds including speech. In another study [7], a metamaterial-coated device with a diameter of 40 cm and a dictionary of noise prototypes were used to localize known noise sources. In our previous work [9], we used an omnidirectional sensor surrounded by cubes of different sizes and a dictionary of spectral prototypes to localize speech sources.

A single omnidirectional sensor can also be used to localize sound sources inside a known room [10]. Indeed, in place of the head, the scattering structure is then the room itself and the localization cues are provided by the echoes from the walls [11]. The drawback is that the room should be known with considerable accuracy—it is much more realistic to assume knowing the geometry of a small scatterer.

As for source models, those used in previous work on monaural localization rely on full complex-valued spectra [7]. Other approaches to multi-sensor localization with sparsity constraints also operate in the complex frequency domain [12], [13], [14]. In this paper, we choose to work with non-negative data which in this case corresponds to the power or magnitude spectra of the audio. We highlight two reasons for this choice. First, unlike the multi-sensor case, the monaural setting generates fewer useful relative phase cues. Second, if *prototypes*—that is, the exact source waveform—are assumed to be known as in [7], there are no modeling errors or challenges associated with the phase information. We, however, assume much less, namely only that the source is speech. It is then natural to leverage the large body of work that addresses dictionary learning with real or non-negative values as opposed

In line with the philosophy of reproducible research, code and data to reproduce the results of this paper are available at http://github.com/swing-research/scatsense.

D. El Badawy is a student at EPFL, Switzerland, e-mail: dalia.elbadawy@epfl.ch.

I. Dokmanić is with ECE Illinois, e-mail: dokmanic@illinois.edu.

Manuscript received January xx, 2018; revised Month xx, 2018.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing



to complex values. In particular, we consider models based on non-negative matrix factorization (NMF). NMF results in a parts-based representation of an input signal [15] and can for instance identify individual musical notes [16]. Thus with training data, NMF can be used to learn a representation for each source [17], [18]. For more flexibility, it can also be used to learn an overcomplete dictionary where each source admits a sparse representation [17], [18]. For the latter, either multiple representations are concatenated [17] or the learning is modified by including sparsity penalties [18], [19].

To solve the localization problem, we first fit the postulated non-negative model to the observed measurements. The cost functions previously used often involve the Euclidean distance [7], [9], [13], [12], [14]. Non-negative modeling lets us use other measures more suitable for speech and audio such as the Itakura–Saito divergence [16]. While NMF is routinely used in single-channel source separation [17], [20], [21], [22], speech enhancement [23], polyphonic music transcription [24], and has been used in a multichannel joint separation and localization scenario [25], the present work is to the best of our knowledge the first time NMF is used in single-channel source localization. Finally, when the localization problem is ill-posed, as is the case for the monaural setting, various regularizations are utilized. Typical regularizers promote sparsity [7], group sparsity [13], [14] or a combination thereof [9].

### B. Contributions & Outline

The current paper extends our previous work [9] in several important ways. We summarize the contributions as follows:

- We derive an NMF formulation for monaural localization via scattering;
- We formulate two different regularized cost functions with different distance measures in the data fidelity term to solve the localization based on either universal or speaker-dependent dictionaries;
- We present extensive numerical evidence using simple "devices" made from LEGO®bricks;
- For the sake of reproducibility, we make freely available the code and data used to generate the results.

Unlike [8], the source model we present easily accommodates more than one source. And unlike [6] or [7], we present localization of challenging sources such as speech without the need for metamaterials or accurate source models—we only use ad hoc scatterers and NMF. In this paper we limit ourselves to anechoic conditions and localization in the horizontal plane as our goal is to assess the potential of this simple setup.

In the following, we first lay down an intuitive argument for how monaural cues help as well as a simple algorithm for localizing white sources. We then formulate the localization problem using NMF and give an algorithm for general colored sources in Section III. In Section IV, we describe our devices and results for localizing white noise and speech.

## II. Background

The sensor we consider in this work is a microphone, possibly omnidirectional, embedded in a compact scattering structure; we henceforth refer to it as "the device". We discretize the azimuth into $D$ candidate source locations $\Omega = \{\theta_1, \theta_2, \ldots, \theta_D\}$ and consider the standard mixing model in the time domain for $J$ sources incoming from directions $\Theta = \{\theta_j\}_{j\in\mathcal{J}}$,

$$y(t) = \sum_{j\in\mathcal{J}} s_j(t) * h_j(t) + e(t), \tag{1}$$

where $\mathcal{J} \subseteq \{1, 2, \ldots, D\} \stackrel{\text{def}}{=} \mathcal{D}$, $|\mathcal{J}| = J$, $*$ denotes convolution, $y$ is the observed signal, $s_j$ is the $j^{\text{th}}$ source signal, $h_j(t) \stackrel{\text{def}}{=} h(t;\theta_j)$ is the impulse response of the directionally-dependent filter, and $e$ is additive noise. The goal of localization is then to estimate the set of directions $\Theta$ from the observed signal $y$. Note that in general we could also include the elevation by considering a set of $D$ directions in 3D, though this would likely yield many additional ambiguities.

The mixing (1) can be approximated in the short-time Fourier transform (STFT) domain as

$$Y(n,f) = \sum_{j\in\mathcal{J}} S_j(n,f)H_j(f) + E(n,f), \tag{2}$$

where $n$ and $f$ denote the time and frequency indices. This so-called narrowband approximation holds when the filter $h_j$ is short enough with respect to the STFT analysis window [26], [27]. For reference, the impulse response corresponding to an HRTF is around 4.5 ms long [28], while the duration of the STFT window for audio is commonly anywhere between 5 ms and 128 ms during which the signal is assumed stationary. Finally, the mixture's spectrogram with $N$ time frames and $F$ frequency bins can be written as

$$\mathsf{Y} = \sum_{j\in\mathcal{J}} \operatorname{diag}(\mathsf{H}_j)\mathsf{S}_j + \mathsf{E}, \tag{3}$$

where $\mathsf{Y} \in \mathbb{C}^{F\times N}$, $\mathsf{S}_j \in \mathbb{C}^{F\times N}$ the spectrogram of the source impinging from $\theta_j$, $\mathsf{H}_j \in \mathbb{C}^F$ is the frequency response of the directionally-dependent filter, $\mathsf{E} \in \mathbb{C}^{F\times N}$ is the spectrogram of the additive noise, and $\operatorname{diag}(\mathbf{v})$ is a matrix with $\mathbf{v}$ on the diagonal.

At least conceptually, monaural localization is a simple matter if the source is always the same: for each direction the HRTF imprints a distinct spectral signature onto the sound which can be detected through correlation. In reality, the sources are diverse but this fixed-source case lets us develop a good intuition.

### A. Intuition

To see how scattering helps, suppose the sources are white and a set of $D$ directional transfer functions $\{\mathsf{H}_d\}_{d=1}^D$ of our device is known. The power spectral density (PSD) of a white source is flat and scaled by the source's power: $\mathbb{E}[|\mathsf{S}_j|^2] = \sigma_j^2$. Assuming the noise has zero mean, the PSD of the observation is

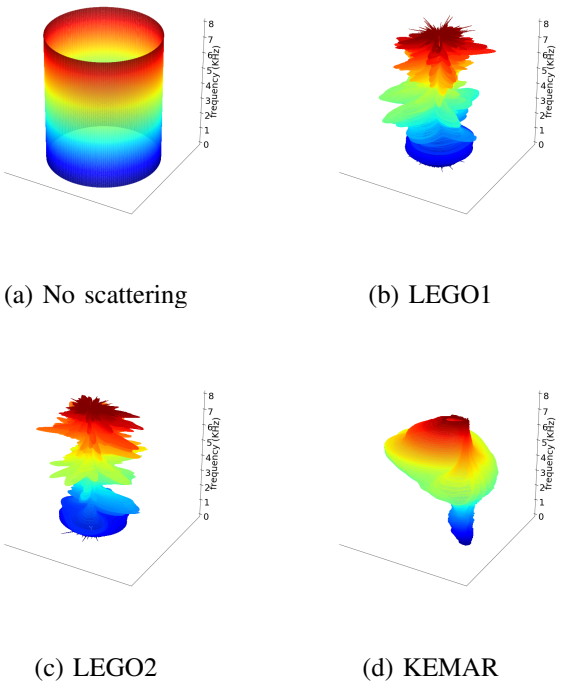
$$\mathbb{E}[|\mathsf{Y}|^2] = \sum_{j\in\mathcal{J}} \sigma_j^2 |\mathsf{H}_j|^2, \tag{4}$$

which is a positive linear combination of the squared magnitudes of the transfer functions. In other words, $\mathbb{E}[|\mathsf{Y}|^2]$ belongs to a cone defined as

$$C_{\mathcal{J}} = \{\mathbf{x} : \mathbf{x} = \sum_{j\in\mathcal{J}} c_j |\mathsf{H}_j|^2,\ c_j > 0\}, \tag{5}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing





(a) No scattering (b) LEGO1

(c) LEGO2 (d) KEMAR

Fig. 1. Directional frequency magnitude response for different devices. Each horizontal slice is the polar pattern at the corresponding frequency between 0-8000 Hz from bottom to top. The colors only aid visualization.

Each configuration of sources $\mathcal{J}$ results in a different cone $C_{\mathcal{J}}$. For $D$ directions and $J$ white sources, there are $\binom{D}{J}$ possible cones which are known a priori since we assume knowing the scatterer. These cones reside in an $F$-dimensional space of direction-dependent spectral magnitude responses, $\mathbb{R}_+^F$, rather than the physical scatterer space $\mathbb{R}^3$. While the arrangement of cones in $\mathbb{R}_+^F$ is indeed determined by the geometry of the device in $\mathbb{R}^3$, the relation is complicated and nonlinear, namely it requires solving a boundary value problem for the Helmholtz equation at each frequency.

Thus, we have $\mathbb{E}[|\mathsf{Y}|^2] \in \bigcup_{\mathcal{J}} C_{\mathcal{J}}$, and in theory, the localization problem becomes one of identifying the correct cone

$$\widehat{\mathcal{J}} = \arg\min_{\mathcal{J}} \operatorname{dist}\left(\widehat{\mathbb{E}}[|\mathsf{Y}|^2], C_{\mathcal{J}}\right), \tag{6}$$

where $\widehat{\mathbb{E}}\big[|\mathsf{Y}|^2\big]$ denotes the empirical estimate of the corresponding expectation from observed measurements. We discuss this further in the next section where we give the complete algorithm.

Testing for cone membership results in correct localization when $C_{\mathcal{J}_1} = C_{\mathcal{J}_2}$ implies $\mathcal{J}_1 = \mathcal{J}_2$ (distinct direction sets span distinct cones)—a condition that is loosely speaking more likely to hold the more *diverse* $\mathsf{H}_j$ are. Examples of $|\mathsf{H}_j|$ are illustrated in Figure 1. In particular, Figure 1(a) corresponds to an omnidirectional microphone with a flat frequency response and no scattering structure. In this case $C_{\mathcal{J}} = \{\sigma^2\mathbf{1} : \sigma \geq \mathbf{0}\}$ and monaural localization is impossible. Figure 1(d) corresponds to an HRTF which features relatively smooth variations. Finally, Figures 1(b) and 1(c) correspond to our devices constructed using LEGO bricks whose responses have more fluctuating variations. In a nutshell, scattering induces a union-of-cones structure that enables us to localize white sources using a single sensor; stronger and more diverse scattering implies easier localization.

**Algorithm 1** White Noise Localization

**Input:** Number of sources $J$, magnitudes of directional transfer functions $\{|\mathsf{H}_j|^2\}_{j\in\mathcal{D}}$, $N$ audio frames $\mathsf{Y} \in \mathbb{C}^{F\times N}$.
**Output:** Directions of arrival $\widehat{\Theta} = \{\widehat{\theta}_1, \ldots, \widehat{\theta}_J\}$.
Compute the empirical PSD $\mathbf{y} = \frac{1}{N}\sum_{n=1}^{N} |\mathsf{Y}_n|^2$
**for** every $\mathcal{J} \subseteq \mathcal{D}$, $|\mathcal{J}| = J$ **do**
  $\mathbf{B}_{\mathcal{J}} \leftarrow \left[|\mathsf{H}_j|^2\right]_{j\in\mathcal{J}}$
  $\mathbf{P}_{\mathcal{J}} \leftarrow \mathbf{B}_{\mathcal{J}}\mathbf{B}_{\mathcal{J}}^{\dagger}$
**end for**
$\widehat{\mathcal{J}} \leftarrow \arg\min_{\mathcal{J}} \|(\mathbf{I} - \mathbf{P}_{\mathcal{J}})\mathbf{y}\|$
$\widehat{\Theta} \leftarrow \left\{\theta_j \mid j \in \widehat{\mathcal{J}}\right\}$

### *B. White Noise Localization*

In this section we describe a simple algorithm for localizing noise sources based on the intuition provided in the previous section[1]. Our experiments with white noise localization will provide us with an ideal case baseline.

First, we need to replace the expected value $\mathbb{E}[|\mathsf{Y}|^2]$ by its empirical mean computed from $N$ time frames. For many types of sources this approximation will be accurate already with a small number of frames by the various concentration of measure results [29]; we corroborate this claim empirically.

Second, for simplicity, we replace each cone $C_{\mathcal{J}}$ by its smallest enclosing subspace $\mathcal{S}_{\mathcal{J}} = \operatorname{span}\left\{|\mathsf{H}_j|^2\right\}_{j\in\mathcal{J}}$ represented by a matrix

$$\mathbf{B}_{\mathcal{J}} \stackrel{\text{def}}{=} \left[\ |\mathsf{H}_{j_1}|^2,\ \ldots,\ |\mathsf{H}_{j_J}|^2\ \right],\ j_k \in \mathcal{J}.$$

This way the closest cone can be approximately determined by selecting $\mathcal{J} \subseteq \mathcal{D}$ such that the subspace projection error is the smallest possible. The details of the resulting algorithm are given in Algorithm 1; note the implicit assumption that $J < F$ as otherwise all cones lie in the same subspace.

The robustness of Algorithm 1 to noise largely depends on the angles between pairs of subspaces $\mathcal{S}_{\mathcal{J}}$ for different configurations $\mathcal{J}$, with smaller angles implying a higher likelihood of error. Intuitively, a transfer function that varies smoothly across directions is unfavorable as it yields smaller subspace angles (more similar subspaces).

We now turn our attention to the realistic case where sound sources are diverse: how can we determine whether an observed spectral variation is due to the directivity of the sensor or a property of the sound source itself? In fact, localization of unfamiliar sounds degrades not only for monaural but also binaural listening [30]. It has also been found that older children with unilateral hearing loss perform better in localization tasks than younger children [31]. We

[1]This algorithm appears in our previous conference publication [9].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing



can thus conclude that both knowledge and experience allow us to dissociate source spectra from directional cues. Once the HRTF and the source spectra have been learned, it becomes possible to differentiate directions based on their modifications by the scatterer.

## III. Method

We can think of an ideal white source as belonging to the subspace span$\{\mathbf{1}\}$ since $|\mathsf{S}|^2 = \mathbf{1}\sigma^2$. In the following, we generalize the source model to more interesting signals such as speech. For those signals, testing for cone membership the same way we did for white sources is not straightforward. We can, however, take advantage of the non-negativity of the data to design efficient localization algorithms based on NMF. Instead of continuing to work with power spectra $|\mathsf{S}|^2$, we switch to magnitude spectra $|\mathsf{S}|$: prior work [20], [23] and our own experiments found that magnitude spectra perform better in this context.

### A. Problem Statement

We adopt the usual assumption that magnitude spectra are additive [20], [21]. Then the magnitude spectrogram of the observation (3) can be expressed as

$$\mathbf{Y} = \sum_{j \in \mathcal{J}} \operatorname{diag}(\mathbf{H}_j)\mathbf{S}_j + \mathbf{E}, \tag{7}$$

for $\mathbf{Y} = |\mathsf{Y}|$, $\mathbf{H} = |\mathsf{H}|$, $\mathbf{S}_j = |\mathsf{S}_j|$, and $\mathbf{E} = |\mathsf{E}|$. We further model the source $\mathbf{S}_j$ as a non-negative linear combination of $K$ atoms $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ such that $\mathbf{S}_j = \mathbf{W}\mathbf{X}_j$. The atoms in $\mathbf{W}$ can correspond to either spectral prototypes of the sources to be localized or they can be learned from training data. Using this source model, we rewrite (7) as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}, \tag{8}$$

where $\mathbf{Y} \in \mathbb{R}_+^{F \times N}$ is the observation,

$$\mathbf{A} = \left[\, \operatorname{diag}(\mathbf{H}_1)\mathbf{W}, \ldots, \operatorname{diag}(\mathbf{H}_D)\mathbf{W} \,\right] \in \mathbb{R}_+^{F \times KD}$$

is the mixing matrix, and

$$\mathbf{X} = \left[\, \mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_D^{\mathrm{T}} \,\right]^{\mathrm{T}} \in \mathbb{R}_+^{KD \times N}$$

are the dictionary coefficients. Each group $\mathbf{X}_d \in \mathbb{R}_+^{K \times N}$ corresponds to the set of coefficients for one source at one direction $d$.

For localization, we wish to recover $\mathbf{X}$; however, we are not interested in the coefficient values themselves but rather whether given coefficients are active or not—the activity of a coefficient indicates the presence of a source. In other words, we are only concerned with identifying the support of $\mathbf{X}$. Localization is achieved by selecting the $J$ directions whose corresponding groups $\mathbf{X}_d$ have the highest norms.

### B. Regularization

Still, recovering $\mathbf{X}$ from (8) is an ill-posed problem. To get a reasonable solution, we must regularize by prior knowledge about $\mathbf{X}$. We thus make the following two assumptions. First, the sources are few ($J \ll D$), which means that most groups $\mathbf{X}_d$ are zero. Second, each source has a sparse representation in the dictionary $\mathbf{W}$. These assumptions are enforced by considering the solution to the following penalized optimization problem

$$\underset{\mathbf{X} \geq 0}{\arg\min}\; D(\mathbf{Y} \,\|\, \mathbf{A}\mathbf{X}) + \lambda \Psi_g(\mathbf{X}) + \gamma \Psi_s(\mathbf{X}), \tag{9}$$

where $D(\cdot \,\|\, \cdot)$ is the data fitting term, $\Psi_g$ is a group-sparsity penalty to enforce the first assumption, and $\Psi_s$ is a sparsity penalty to enforce the second assumption. The parameters $\lambda > 0$ and $\gamma > 0$ are the weights given to the respective penalties.

A common choice of $D(\cdot \,\|\, \cdot)$ for speech is the Itakura–Saito divergence [16], which for strictly positive scalars $v$ and $\hat{v}$, is defined as

$$d_{IS}(v \,\|\, \hat{v}) = \frac{v}{\hat{v}} - \log \frac{v}{\hat{v}} - 1, \tag{10}$$

so that $D(\mathbf{V} \,\|\, \hat{\mathbf{V}}) = \sum_{fn} d_{IS}(v_{fn} || \hat{v}_{fn})$. Another option is the Euclidean distance

$$D(\mathbf{V} \,\|\, \hat{\mathbf{V}}) = \frac{1}{2} \sum_{fn} (v_{fn} - \hat{v}_{fn})^2. \tag{11}$$

Both the Itakura–Saito divergence and the Euclidean distance belong to the family of $\beta$-divergences with $\beta = 0$ and $\beta = 2$ respectively [32]. The former is scale-invariant and is thus preferred for audio which has a large dynamic range [16].

To promote group sparsity, we choose $\Psi_g$ to be the $\log/\ell_1$ penalty [33] defined as

$$\Psi_g(\mathbf{X}) = \sum_{d=1}^{D} \log(\epsilon + \|\operatorname{vec}(\mathbf{X}_d)\|_1), \tag{12}$$

where $\operatorname{vec}(\cdot)$ is a vectorization operator. To promote sparsity of the dictionary expansion coefficients, we choose $\Psi_s$ to be $\ell_1$-norm [34] as

$$\Psi_s(\mathbf{X}) = \|\operatorname{vec}(\mathbf{X})\|_1. \tag{13}$$

The combination of sparsity and group-sparsity penalties results in a small number of active groups that are themselves sparse. Thus the joint penalty is known as sparse-group sparsity [35].

We note that our main optimization (9) is performed only over the latent variables $\mathbf{X}$; the non-negative dictionary $\mathbf{A}$, which is constructed by merging a source dictionary learned by off-the-shelf implementations of standard algorithms with the direction-dependent transfer functions as described in Section III-A, is taken as input. We thus avoid the joint optimization over $\mathbf{A}$ and $\mathbf{X}$ which is a major source of non-convexity. However, our choices for non-convex functionals like the Itakura-Saito divergence and the $\log/\ell_1$ penalty (although the latter is quasi-convex) render the whole optimization (9) non-convex.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing



### C. Derivation

The minimization (9) can be solved iteratively by multiplicative updates (MU) which preserve non-negativity when the variables are initialized with non-negative values. The update rules for $\mathbf{X}$ are derived using maximization-minimization for the group-sparsity penalty in [33] and for the $\ell_1$-penalty in [32]. They amount to dividing the negative part of the gradient by the positive part and raising to an exponent. In the following we derive the MU rules for our objective (9).

Note that the objective is separable over the columns of $\mathbf{X}$

$$C(\mathbf{x}) \;=\; D(\mathbf{y} \,\|\, \mathbf{A}\mathbf{x}) + \lambda \sum_{d=1}^{D} \log(\epsilon + \|\mathbf{x}_d\|_1) + \gamma \|\mathbf{x}\|_1, \quad (14)$$

where $\mathbf{y} \in \mathbb{R}_+^F$, $\mathbf{x} \in \mathbb{R}_+^{FK}$ are columns of $\mathbf{Y}$ and $\mathbf{X}$ respectively. With $\mathbf{x}^{(i)}$ as the current iterate, the gradient of (14) with respect to one element $x_k$ of $\mathbf{x}$ when $D(\cdot \,\|\, \cdot)$ is the Itakura–Saito divergence is given by

$$\begin{aligned} \nabla_{x_k} C(\mathbf{x}^{(i)}) = &- \sum_f y_f (\mathbf{A}\mathbf{x}^{(i)})_f^{-2} a_{fk} \\ &+ \sum_f (\mathbf{A}\mathbf{x}^{(i)})_f^{-1} a_{fk} + \lambda \frac{1}{\epsilon + \|\mathbf{x}_d^{(i)}\|_1} + \gamma, \end{aligned} \quad (15)$$

where $a_{fk} = [\mathbf{A}]_{fk}$ are entries of $\mathbf{A}$. The update rule is then given as

$$\begin{aligned} x_k^{(i+1)} \;&=\; x_k^{(i)} \left( \frac{\nabla_{x_k}^- C(\mathbf{x}^{(i)})}{\nabla_{x_k}^+ C(\mathbf{x}^{(i)})} \right)^{\frac{1}{2}} \\ &=\; x_k^{(i)} \left( \frac{\sum_f y_f (\mathbf{A}\mathbf{x}^{(i)})_f^{-2} a_{fk}}{\sum_f (\mathbf{A}\mathbf{x}^{(i)})_f^{-1} a_{fk} + \lambda \frac{1}{\epsilon + \|\mathbf{x}_d^{(i)}\|_1} + \gamma} \right)^{\frac{1}{2}}, \end{aligned} \quad (16)$$

where $\frac{1}{2}$ is a corrective exponent [32]. The updates in matrix form are shown in Algorithm 2 where the multiplication $\odot$, division, and power operations are elementwise and $\mathbf{P}$ is a matrix of the same size as $\mathbf{X}$. Also shown are the updates for using the Euclidean distance following [32], [36] where $[v]_\epsilon = \max\{v, \epsilon\}$ is a thresholding operator to maintain non-negativity with $\epsilon = 10^{-20}$.

### D. Algorithm

The discretization of the azimuth into $D$ evenly-spaced directions has a direct correspondence with the localization errors. On the one hand, a course discretization limits the localization accuracy to approximately the size of the discretization bin $\frac{360}{D}^\circ$. On the other hand a fine discretization may warrant a smaller error floor, but it implies a model matrix with a higher coherence only worsening the ill-posedness of the optimization problem (9). It additionally results in a larger matrix which hampers the matrix factorization algorithms that are of complexity $\mathcal{O}(FKDN)$ per iteration [16], [33]. A common compromise is the multiresolution approach [12], [8] in which position estimates are first computed on a coarse grid, and then subsequently refined on a finer grid concentrated around the initial guesses. We test the following strategy:

**Algorithm 2** MU for NMF with Sparse-group Sparsity

**Input:** $\mathbf{Y}$, $\mathbf{A}$, $\lambda$, $\gamma$
**Output:** $\mathbf{X}$
  Initialize $\mathbf{X} = \mathbf{A}^\mathsf{T}\mathbf{Y}$
  $\widehat{\mathbf{Y}} \leftarrow \mathbf{A}\mathbf{X}$
  **repeat**
    **for** $d = 1, \ldots, D$ **do**
      $\mathbf{P}_d \leftarrow \dfrac{1}{\epsilon + \|\mathrm{vec}(\mathbf{X}_d)\|_1}$
    **end for**
    **if** Itakura–Saito **then**
      $\mathbf{X} \leftarrow \mathbf{X} \odot \left( \dfrac{\mathbf{A}^T(\mathbf{Y} \odot \widehat{\mathbf{Y}}^{-2})}{\mathbf{A}^T\widehat{\mathbf{Y}}^{-1} + \lambda\mathbf{P} + \gamma} \right)^{\frac{1}{2}}$
    **else if** Euclidean **then**
      $\mathbf{X} \leftarrow \mathbf{X} \odot \left[ \dfrac{\mathbf{A}^T\mathbf{Y} - \lambda\mathbf{P} - \gamma}{\mathbf{A}^T\widehat{\mathbf{Y}}} \right]_\epsilon$
    **end if**
    $\widehat{\mathbf{Y}} \leftarrow \mathbf{A}\mathbf{X}$
  **until** convergence

1) Attempt localization on a coarse grid,
2) Identify the top $T$ direction candidates,
3) Construct the model matrix using the $T$ candidates and their neighbors at a finer resolution,
4) Rerun the NMF localization.

The final algorithm for source localization by NMF with and without multiresolution is shown in Algorithm 3. Since (9) is non-convex, different initializations of $\mathbf{X}$ might lead to different results. We thus later run an experiment to test the influence on the actual localization performance in Section IV.

**Algorithm 3** Direction of Arrival Estimation by NMF

**Input:** Observation $y(t)$, Number of sources $J$, Parameter for group sparsity $\lambda$, Parameter for $\ell_1$ sparsity $\gamma$, magnitudes of directional transfer functions $\{\mathbf{H}_j\}_{j\in\mathcal{D}}$, source model $\mathbf{W}$
**Output:** Directions of arrival $\widehat{\Theta} = \{\widehat{\theta}_1, \ldots, \widehat{\theta}_J\}$
  Construct $\mathbf{A} \leftarrow \big[\mathrm{diag}(\mathbf{H}_1)\mathbf{W}, \ldots, \mathrm{diag}(\mathbf{H}_D)\mathbf{W}\big]$
  Construct $\mathbf{Y} \leftarrow |\mathrm{STFT}\{y\}|$
  Factorize $\mathbf{Y} \approx \mathbf{A}\mathbf{X}$ using Algorithm 2
  Calculate $\mathcal{D} = \{\|\mathrm{vec}(\mathbf{X}_d)\|_1 \text{ for } d = 1, 2, \ldots, D\}$
  **if** Multiresolution **then**
    Identify $T$ candidates and their $RT$ neighbors $\{\mathbf{H}_{t,r}\}_{t=1,r=0}^{t=T,r=R}$
    Construct $\widetilde{\mathbf{A}} \leftarrow \big[\mathrm{diag}(\mathbf{H}_{1,0})\mathbf{W}, \ldots, \mathrm{diag}(\mathbf{H}_{T,R})\mathbf{W}\big]$
    Factorize $\mathbf{Y} \approx \widetilde{\mathbf{A}}\widetilde{\mathbf{X}}$ using Algorithm 2
    Calculate $\mathcal{D} = \{\|\mathrm{vec}(\widetilde{\mathbf{X}}_d)\|_1 \text{ for } d = 1, 2, \ldots, (R+1)T\}$
  **end if**
  $\widehat{\mathcal{J}} \leftarrow \{\text{Indices of the } J \text{ largest elements in } \mathcal{D}\}$
  $\widehat{\Theta} \leftarrow \left\{\theta_j \mid j \in \widehat{\mathcal{J}}\right\}$

## IV. Experimental Results

### A. Devices

We ran experiments using three different devices:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing



*a) LEGO1 and LEGO2:* The first two devices are structures composed of LEGO bricks as shown in Figure 2. Since we aimed for diverse random-like scattering, we stacked haphazard brick constructions on a base plate of size 25 cm $\times$ 25 cm along with one omnidirectional microphone. The heights of the different constructions vary between 4 and 12.5 cm. We did not attempt to optimize the layout. The only assumption we make regarding the dimensions of the device is that some energy of the target source resides at frequencies where the device observably interacts with the acoustic wave. We note that the problem of designing and optimizing the structure to get a desired response is that of inverse obstacle scattering which is a hard inverse problem in its own right [37], [38]. For the present work, we simply observe that our random structures result in the desired random-like scattering.

The directional impulse response measurements were then done in an anechoic chamber where the device was placed on a turntable as shown in Figure 2(c) and a loudspeaker at a distance of 3.5 m emitted a linear sweep. We note that the turntable is symmetric, so its effect on localization in the horizontal plane, if any, is negligible. The duration of the measured impulse responses averages around 20 ms. Figures 1(b) and 1(c) show the corresponding magnitude response for the two devices. Due to their relatively small size, they mostly scatter high frequency waves and so the response at lower frequencies is comparably flat. We thus expect that only sources with enough energy in the higher range of frequencies can be accurately localized.

*b) KEMAR:* The third device is KEMAR [39] which is modeled after a human head and torso so that its response accurately approximates a human HRTF. The mannequin's torso measures $44 \times 24 \times 73$ cm and the head's diameter is 18 cm. The duration of the impulse response is 10 ms. Figure 1(d) shows the corresponding magnitude response. As can be seen, the variation across the directions is very smooth which we expect to result in worse monaural localization performance.

### B. Data and parameters

The mixtures are created by first convolving the source signals with the impulse responses and then corrupting the result by additive white Gaussian noise at various levels of signal-to-noise ratio defined as

$$\text{SNR} = 20 \log \frac{\| \sum_j s_j(t) * h_j(t) \|_2}{\|e(t)\|_2} \text{ dB}.$$

We use frame-based processing using the STFT with a Hann window of length 64 ms, with a 50% overlap. The number of iterations in NMF (Algorithm 2) was set to 100.

The test data contains 10 speech sources (5 female, 5 male) from TIMIT [40] sampled at 16000 Hz. The duration of the speech varies between 3.1 and 4.5 s and the maximum amplitude is normalized to 1 so that all sources have the same volume. No preprocessing of the sources such as silence removal was done; when mixing two sources, the longest one was truncated.

A separate validation set was used to select the best sparsity parameters for each device. The parameters that gave the best performance averaged for one and two sources were chosen. We additionally tested whether the lower frequencies can be ignored in localization since, as mentioned before, for the relatively small scatterers the lower frequency range lacks variation and is thus uninformative. Moreover, truncating the lower frequencies would help reduce coherence between the directional transfer functions. The final parameters and used frequency range are summarized in Table I.

TABLE I
PARAMETERS PER DEVICE.

| | LEGO1 | LEGO2 | KEMAR |
|---|---|---|---|
| Frequency | 3000-8000 Hz | 3000-8000 Hz | 0-8000 Hz |
| Prototypes | $\lambda = 10$, $\gamma = 10$ | $\lambda = 10$, $\gamma = 1$ | $\lambda = 10$, $\gamma = 0.1$ |
| USM ($\beta = 0$) | $\lambda = 0.1$, $\gamma = 10$ | $\lambda = 10$, $\gamma = 1$ | $\lambda = 100$, $\gamma = 10$ |
| USM ($\beta = 2$) | $\lambda = 1$, $\gamma = 1$ | $\lambda = 1$, $\gamma = 1$ | $\lambda = 1$, $\gamma = 1$ |
| Multiresolution | $\lambda = 0.1$, $\gamma = 1$ | $\lambda = 100$, $\gamma = 0.1$ | - |

*Source Dictionary:* For speech localization, we test two source dictionaries. For the first experiment, we use a dictionary of prototypes of magnitude spectra from 4 speakers (2 female, 2 male) in the test set.

For the second experiment, we use a more general universal speech model (USM) [17] learned from a training set of 25 female and 25 male speakers, also from TIMIT. We use a random initialization for the NMF when learning the USM. Each speaker in the training set is modeled using $K = 10$ atoms, thus the final USM is $\mathbf{W} \in \mathbb{R}_+^{F \times 500}$. In total, we use four versions of the USM in the experiments. Two versions correspond to learning the model by minimizing either the Itakura–Saito divergence or the Euclidean distance. The other two versions correspond to learning the model using only the subset of frequencies to be utilized in the localization.

### C. Evaluation

We estimate the azimuth of the sources in the range $[0^\circ, 360^\circ)$. The model (8) assumes a discrete set of 36 evenly spaced directions while the sources are randomly placed on a finer grid of 360 directions. Given the estimated directions $\hat{\Theta} = \{\hat{\theta}_1, \ldots, \hat{\theta}_J\}$ and the true directions $\Theta = \{\theta_1, \ldots, \theta_J\}$, the localization error is computed as the average absolute difference modulo $360^\circ$ as

$$\min_{\pi} \frac{1}{J} \sum_{j \in \mathcal{J}} \left| (\hat{\theta}_{\pi(j)} - \theta_j + 180) \bmod 360 - 180 \right|, \qquad (17)$$

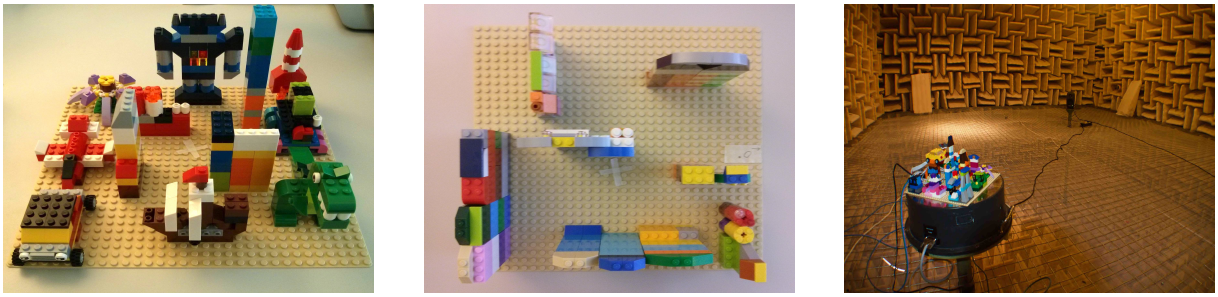
where $\pi : \mathcal{J} \to \mathcal{J}$ is a permutation that best matches the ordering in $\hat{\Theta}$ and $\Theta$.

For each experiment, we test 5000 random sets of directions. We emphasize that we have been careful to avoid an inverse crime, and we produced the measurements by convolution in the time domain, not by multiplication in the STFT domain. Thus in this set up, the reported errors also reflect the modeling mismatch.

Following [41], we report the *accuracy* defined as the percentage of sources localized to their closest $10^\circ$-wide bin as well as the mean error for those accurately localized sources. For 36 bins, there is an inherent average error of $2.5^\circ$. Thus, ideally the accuracy would be 100% and the error $2.5^\circ$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing





(a) (b) (c)

Fig. 2. Sensing devices made of LEGO bricks. The location of the microphone is marked by an "x". (a) LEGO1. (b) LEGO2. (c) Calibration setup in an anechoic chamber.

Additionally, we report the accuracy per source, that is, the rate at which a source is correctly localized regardless of the other sources.

### D. NMF Initialization

Since in a non-convex problem different initializations might lead to different results, we run an experiment to test the effect of the initialization of $\mathbf{X}$ on the localization performance. The experiment consists of 300 tests for localizing one female speaker using LEGO2 and a USM. We compare the initialization mentioned in Algorithm 2 ($\mathbf{X} = \mathbf{A}^{\mathrm{T}}\mathbf{Y}$)[2] to different random initializations. The estimated DoAs were in agreement for both initializations 98.67% of the time with Itakura-Saito and 97% with Euclidean distance. We show in Table II the localization accuracy rates for that experiment which are comparable. This means that there are either "hard" situations where localization fails regardless of the initialization or "easy" situations where it succeeds regardless of the initialization. Certainly, tailor-made initializations in the spirit of [42], [43] may work slightly better, but such constructions are outside the scope of this paper. Additionally, we note that in these works initializations are constructed for the basis matrix. In our case, this matrix is $\mathbf{A}$ which is given as input to the algorithm.

TABLE II
LOCALIZATION ACCURACY FOR DIFFERENT NMF INITIALIZATIONS.

| | $\mathbf{A}^{\mathrm{T}}\mathbf{Y}$ | Random |
|---|---|---|
| Itakura-Saito | 93.00% | 93.33% |
| Euclidean | 89.67% | 90.00% |

### E. White Noise Localization

We first test the localization of one and two white sources at various levels of SNR using Algorithm 1. Each source is 0.5 s of white Gaussian noise. We compare the performance using the three devices LEGO1, LEGO2, and KEMAR described above. For white sources, using the full range of frequencies, not a subset, was found to perform better.

[2]We use a deterministic initialization to facilitate reproducibility and multithreaded implementations.

The accuracy rate and the mean localization error for the different devices are shown in Table III. In the one source case, all devices perform well. The mean error achieved by the devices for one white source is close to the ideal grid-matched $2.5°$ which is better than the reported $4.3°$ and $8.8°$ in [8] using an HMM. For two sources, the accuracy of the LEGO devices is still high, though lower than for one source. At the same time the accuracy of KEMAR deteriorates considerably. This is consistent with the intuition that interesting scattering patterns such as those of the LEGO devices result in better localization.

We also test the effect of the discretization on the localization performance. In Table IV, we report the localization errors using LEGO1 at three different resolutions: $2°$, $5°$, and $10°$. We find that improving the resolution results in more accurate localization for both one and two sources but the average error is still larger than the ideal $0.5°$ and $1.25°$ for the $2°$ and $5°$ resolutions respectively, especially for two sources. Since white sources are flat, this observation highlights a limitation of the device itself in terms of coherent or ambiguous directions.

### F. Speech Localization with Prototypes

We now turn to speech localization which is considerably more challenging than white noise, especially in the monaural setting. Using the three devices, we test the localization of one and two speakers at 30 dB SNR. In this first experiment, we use a subset of 4 speakers from the test data (two female, two male) and consider an easier scenario where we assume knowing the exact magnitude spectral prototypes of the sources. Still, localization with colored prototypes is harder compared to noise prototypes (as in [7]). This scenario serves as a gauge for the quality of the sensing devices for localizing speech sources. We organize the results by the number of sources as well as by whether the speaker is male or female. We expect the localization of female speakers to be more accurate since they have relatively more energy in the higher frequency range where the device responses are more informative.

The results for the three devices are shown in Table V. As expected the overall localization performance by the less smooth LEGO scatterers is significantly better than by KEMAR. Also as expected, the localization of male speech is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing



TABLE III
ERROR FOR WHITE NOISE LOCALIZATION AT A DISCRETIZATION OF $10^\circ$

| | | LEGO1 | | LEGO2 | | KEMAR | |
|---|---|---|---|---|---|---|---|
| | SNR | Accuracy | Mean | Accuracy | Mean | Accuracy | Mean |
| One source | 30 dB | 99.56% | $2.63^\circ$ | 96.64% | $2.54^\circ$ | 92.06% | $2.72^\circ$ |
| | 20 dB | 99.58% | $2.63^\circ$ | 96.54% | $2.53^\circ$ | 92.12% | $2.71^\circ$ |
| | 10 dB | 99.60% | $2.60^\circ$ | 96.42% | $2.53^\circ$ | 91.78% | $2.73^\circ$ |
| Two sources | 30 dB | 94.72% | $2.75^\circ$ | 83.64% | $2.62^\circ$ | 25.22% | $3.44^\circ$ |
| | 20 dB | 94.54% | $2.75^\circ$ | 83.34% | $2.62^\circ$ | 25.48% | $3.45^\circ$ |
| | 10 dB | 92.32% | $2.73^\circ$ | 81.52% | $2.62^\circ$ | 21.20% | $3.59^\circ$ |

TABLE IV
DISCRETIZATION COMPARISON FOR WHITE NOISE LOCALIZATION USING LEGO1.

| | | $2^\circ$ | | $5^\circ$ | | $10^\circ$ | |
|---|---|---|---|---|---|---|---|
| | SNR | Accuracy | Mean | Accuracy | Mean | Accuracy | Mean |
| One source | 30 dB | 100.0% | $0.52^\circ$ | 100.0% | $1.27^\circ$ | 99.56% | $2.63^\circ$ |
| | 20 dB | 100.0% | $0.52^\circ$ | 100.0% | $1.27^\circ$ | 99.58% | $2.63^\circ$ |
| | 10 dB | 100.0% | $0.54^\circ$ | 100.0% | $1.26^\circ$ | 99.60% | $2.60^\circ$ |
| Two sources | 30 dB | 98.56% | $0.70^\circ$ | 98.78% | $1.43^\circ$ | 94.72% | $2.75^\circ$ |
| | 20 dB | 98.50% | $0.71^\circ$ | 98.70% | $1.43^\circ$ | 94.54% | $2.75^\circ$ |
| | 10 dB | 97.30% | $0.82^\circ$ | 97.32% | $1.47^\circ$ | 92.32% | $2.73^\circ$ |

worse than female speech except for LEGO1. Similar to the white noise case, the accuracy for localizing two sources is lower in comparison to one source. Moreover, we find that the presence of one female speaker improves the accuracy for LEGO2 and KEMAR, most likely due to the spectral content.

### G. Speech Localization with USM

In this experiment, we switch to a more realistic and challenging setup where we use a learned universal speech model. We compare the performance of the Itakura–Saito divergence to that of the Euclidean distance in the cost function (9). The accuracy and mean error for the three devices are shown in Table VI. We observe that using the Itakura–Saito divergence results in better performance in a majority of cases which is in line with the recommendations for using Itakura–Saito for audio.

Similar observations as in the previous experiment hold with the LEGO scatterers offering better localization than KEMAR. We find that localizing one female speaker is successful with 93% accuracy. Compared to the use of prototypes, the source model is here speaker-independent and the test set is larger containing 10 speakers; however, the accuracy is still only lower by 3-5%. We also note that the mean localization error is $2.5^\circ$ which is smaller than the reported $7.7^\circ$ in [8] with an HMM though at a lower SNR of 18 dB.

As expected, the localization accuracy for male speakers is lower than for female speakers. Since the mean errors are however not much larger than the ideal $2.5^\circ$, the lower accuracy points to the presence of outliers. We thus plot confusion matrices in Figures 4 and 3 for female and male speakers respectively. On the horizontal axis, we have the estimated direction which is one of 36 only. First, we look at the single source case in Figures 3(a) and 4(a) where we can clearly see the few outliers away from the diagonal. The number of outliers is larger for male speakers which is a direct result of the absence of spectral variation for male speech in the used higher frequency range.

For two sources, the number of outliers increases for both types as seen in Figure 3(b). We also plot in Figure 3(a) the confusion matrix for the case of using prototypes which has less outliers in comparison due to the stronger model. Note that outliers exist even with white sources as shown in Figure 3(c), which points to a deficiency of the device itself as mentioned before. However, we note that while the reported accuracy corresponds to correctly localizing the two sources simultaneously, the average accuracy per source which reflects the number of times at least one of the sources is correctly localized is often higher. For instance for female speakers, the accuracy is 53.52% while the average accuracy per source is higher at 73.93%. The overall best performance is achieved by LEGO2 with Itakura–Saito divergence.

*1) Finer resolution:* As mentioned, one straightforward improvement to our system is to increase the resolution. We show in Table VII the result of doubling the resolution from $10^\circ$ to $5^\circ$. For a single female speaker, the error is slightly higher than the ideal average of $1.25^\circ$ and the accuracy is improved relative to the initial bin size of $10^\circ$. While some improvement is apparent for the localization of one male speaker as well, the mismatch between the useful scattering range and source spectrum still prevents good performance. However, in line with the discussion in Section III-D, localization of two sources is worse than at a coarser grid due to the increased matrix coherence, with the accuracy dropping from 55% to 45% for two female speakers.

*2) Multiresolution:* Next we tested the multiresolution strategy where we refine the top estimates on the coarse grid using a search on a finer grid. We arbitrarily use the best 7 candidates at the $10^\circ$ grid spacing, and redo the localization at a finer $2^\circ$ grid centered around the 7 initial guesses. The hyperparameters for localization on the finer grid were tuned on a separate validation set and are given in Table I.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing



TABLE V
ERROR FOR SPEECH LOCALIZATION USING PROTOTYPES AT A DISCRETIZATION OF 10°

| | LEGO1 | | | LEGO2 | | | KEMAR | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Mean | Per Source | Accuracy | Mean | Per Source | Accuracy | Mean | Per Source |
| female speech | 98.48% | 2.53° | 98.48% | 96.94% | 2.51° | 96.94% | 79.74% | 3.42° | 79.74% |
| male speech | 98.76% | 2.56° | 98.76% | 96.00% | 2.53° | 96.00% | 72.06% | 3.35° | 72.06% |
| female/female | 75.24% | 2.46° | 87.07% | 78.28% | 2.40° | 88.31% | 11.66% | 3.50° | 46.70% |
| female/male | 76.60% | 2.44° | 87.79% | 74.36% | 2.41° | 86.17% | 10.90% | 3.59° | 44.47% |
| male/male | 80.24% | 2.43° | 89.82% | 74.22% | 2.39° | 86.04% | 9.24% | 3.91° | 43.09% |

TABLE VI
ERROR FOR SPEECH LOCALIZATION USING A USM AT A DISCRETIZATION OF 10°

| | LEGO1 | | | LEGO2 | | | KEMAR | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Mean | Per Source | Accuracy | Mean | Per Source | Accuracy | Mean | Per Source |
| *Itakura–Saito* | | | | | | | | | |
| female speech | 93.20% | 2.67° | 93.20% | 93.72% | 2.54° | 93.72% | 46.56% | 3.33° | 46.56% |
| male speech | 89.80% | 2.74° | 89.80% | 87.70% | 2.66° | 87.70% | 35.56% | 3.46° | 35.56% |
| female/female | 26.38% | 2.64° | 54.65% | 53.52% | 2.42° | 73.93% | 7.60% | 3.90° | 35.29% |
| female/male | 24.76% | 2.77° | 54.42% | 49.22% | 2.49° | 70.93% | 7.40% | 4.01° | 35.56% |
| male/male | 19.78% | 3.02° | 50.61% | 39.54% | 2.63° | 65.45% | 7.44% | 4.36° | 33.76% |
| *Euclidean* | | | | | | | | | |
| female speech | 85.60% | 2.79° | 85.60% | 91.26% | 2.57° | 91.26% | 29.26% | 3.75° | 29.26% |
| male speech | 76.00% | 2.78° | 76.00% | 86.74% | 2.65° | 86.74% | 23.24% | 3.78° | 23.24% |
| female/female | 29.34% | 2.88° | 56.66% | 46.86% | 2.48° | 69.89% | 4.62% | 4.40° | 23.75% |
| female/male | 30.62% | 2.88° | 57.55% | 42.28% | 2.58° | 66.40% | 3.36% | 4.34° | 21.19% |
| male/male | 23.72% | 2.96° | 52.67% | 35.50% | 2.74° | 62.71% | 2.80% | 3.97° | 18.60% |

TABLE VII
ERROR FOR SPEECH LOCALIZATION AT A RESOLUTION OF 5°.

| | LEGO1 | | | LEGO2 | | |
|---|---|---|---|---|---|---|
| | Accuracy | Mean | Per Source | Accuracy | Mean | Per Source |
| female speech | 97.08% | 1.59° | 97.08% | 99.72% | 1.41° | 99.72% |
| male speech | 93.26% | 1.76° | 93.26% | 92.68% | 1.57° | 92.68% |
| female/female | 22.24% | 1.95° | 55.25% | 43.26% | 1.47° | 71.23% |
| female/male | 21.60% | 2.14° | 55.33% | 39.66% | 1.61° | 68.82% |
| male/male | 15.42% | 2.47° | 50.38% | 29.72% | 1.87° | 63.31% |

As before, multiresolution localization results in some improvement for one source but not for two sources (Table VIII). We show the relevant confusion matrices in Figure 5: the lack of increase in performance can be explained by the fact that in the second round of localization the included directions are still strongly correlated and the only way to resolve the resulting ambiguities is through more constrained source models. Additionally, the set of correlated directions are not necessarily concentrated around the true direction which might explain the drop in accuracy for LEGO1. Overall, it seems the extra computation for the multiresolution approach does not bring about significant improvements compared to using a finer discretization.

Finally, in Figure 6, we show a summary of the performance of the different methods for localizing one or two female speakers using LEGO2 along with the average accuracy and error. Note that the results for prototypes use a smaller test set and that the error is lower bounded by the grid size. We also show the size of the model matrix $\mathbf{A}$ from (8) which contributes to the overall complexity of NMF as well as the actual runtime which depends on the machine. The figure suggests that overall using a USM and a 10° resolution works well. For two-source localization, however, a good source model like prototypes is required.

## V. CONCLUSION

Any scattering that causes spectral variations across directions enables monaural localization of one white source. On the other hand, more complex and interesting scattering patterns are needed to localize multiple sources. As shown by our "random" LEGO constructions, interesting scattering is not hard to come by. In order to localize general, non-white sources, one further requires a good source model.

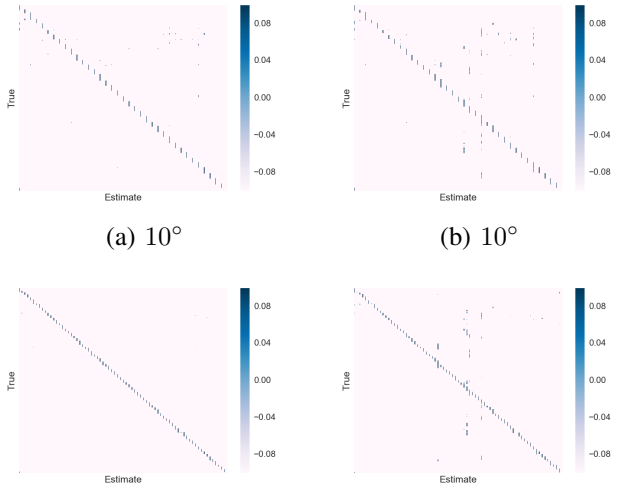
We demonstrated successful localization of one speaker using regularized NMF and a universal speech model. Both our LEGO scatterers were found to be superior in localization to a mannequin's HRTF. Finally, we stress that speech localization is challenging and note that the fundamental frequency of the human voice is below 300 Hz while the range of usable frequencies for our devices is above 3000 Hz. This discrepancy is responsible for outliers when localizing multiple speakers, a problem that can potentially be alleviated by increasing the size of the device or using sophisticated metamaterial-based

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing



TABLE VIII
ERROR FOR SPEECH LOCALIZATION WITH A MULTIRESOLUTION APPROACH.

| | LEGO1 | | | LEGO2 | | |
|---|---|---|---|---|---|---|
| | Accuracy | Mean | Per Source | Accuracy | Mean | Per Source |
| female speech | 96.94% | 1.15° | 96.94% | 99.08% | 0.70° | 99.08% |
| male speech | 86.00% | 1.26° | 86.00% | 90.62% | 0.95° | 90.62% |
| female/female | 17.88% | 1.80° | 56.66% | 32.26% | 1.08° | 65.39% |
| female/male | 17.64% | 1.87° | 56.17% | 29.06% | 1.33° | 63.47% |
| male/male | 13.84% | 2.19° | 52.72% | 20.22% | 1.64° | 57.68% |




(a) 10° (b) 10°

(c) 5° (d) 5°

Fig. 3. Confusion matrices for localizing one speaker using LEGO2. Female speech has less outliers and improving the resolution decreases the number of outliers. Left: Female speech. Right: Male speech.

designs. Perhaps a source model other than the universal dictionary could approach the performance of using prototypes.

Finally, we presented our results for anechoic conditions. Preliminary numerical experiments show that the current approach underperforms in a reverberant setting. This shortcoming is partly due to violations of our modeling assumptions. For example, in Eq. (1), the noise is assumed independent of the sources which is no longer true in the presence of reverberation. For practical scenarios it is thus necessary to extend the approach to handle reverberant conditions as well as to test the localization performance in 3D i.e., estimate both the azimuth and the elevation. For accurate localization in elevation, we expect that a taller device with more variation along the vertical axis would perform better. Since we only use one microphone, the number of ambiguous directions would likely grow considerably in 3D making the problem comparably harder. Other interesting open questions include blind learning of the directional transfer functions and understanding the benefits of scattering in the case of multiple sensors.

## VI. ACKNOWLEDGMENT

We thank Robin Scheibler and Mihailo Kolundžija for help with experiments and valuable comments. We also thank Martin Vetterli for numerous insights and discussions, and for suggesting Figure 1. This work was supported by the Swiss National Science Foundation grant number 20FP-1 151073, Inverse Problems regularized by Sparsity.

## VII. DISCLAIMER

LEGO®is a trademark of the LEGO Group which does not sponsor, authorize or endorse this work.

## REFERENCES

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, The MIT Press, 1997.

[2] A. D. Musicant and R. A. Butler, "The Influence of Pinnae-based Spectral Cues on Sound Localization," *J. Acoust. Soc. Am.*, vol. 75, no. 4, pp. 1195–1200, 1984.

[3] J. J. Rice, B. J. May, G. A. Spirou, and E. D. Young, "Pinna-based Spectral Cues for Sound Localization in Cat," *Hearing Research*, vol. 58, no. 2, pp. 132–152, 1992.

[4] M. Aytekin, E. Grassi, M. Sahota, and C. F. Moss, "The Bat Head-related Transfer Function Reveals Binaural Cues for Sound Localization in Azimuth and Elevation," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3594–3605, 2004.

[5] S. R. Oldfield and S. P. A. Parker, "Acuity of Sound Localisation: A Topography of Auditory Space. III. Monaural Hearing Conditions," *Perception*, vol. 15, no. 1, pp. 67–81, 1986, PMID: 3774479.

[6] J. G. Harris, C.-J. Pu, and J. C. Principe, "A Monaural Cue Sound Localizer," *Analog Integrated Circuits and Signal Processing*, vol. 23, no. 2, pp. 163–172, May 2000.

[7] Y. Xie, T. Tsai, A. Konneker, B. Popa, D. J. Brady, and S. A. Cummer, "Single-sensor Multispeaker Listening with Acoustic Metamaterials," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 112, no. 34, pp. 10595–10598, Aug. 2015.

[8] A. Saxena and A.Y. Ng, "Learning Sound Location from a Single Microphone," in *Proc. IEEE Int. Conf. on Robotics and Automation*, 2009, pp. 1737–1742.

[9] D. El Badawy, I. Dokmanić, and M. Vetterli, "Acoustic DoA Estimation by One Unsophisticated Sensor," in *13th Int. Conf.on Latent Variable Analysis and Signal Separation - LVA/ICA*, P. Tichavský, M. B. Zadeh, O. Michel, and N. Thirion-Moreau, Eds. 2017, vol. 9237 of *Lecture Notes in Computer Science*, pp. 489–496, Springer.

[10] I. Dokmanić, *Listening to Distances and Hearing Shapes: Inverse Problems in Room Acoustics and Beyond*, Ph.D. thesis, École polytechnique fédérale de Lausanne, 2015.

[11] I. Dokmanić and M. Vetterli, "Room Helps: Acoustic Localization with Finite Elements," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, Mar. 2012, pp. 2617–2620.

[12] D. Malioutov, M. Cetin, and A. S. Willsky, "A Sparse Signal Reconstruction Perspective for Source Localization with Sensor Arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.

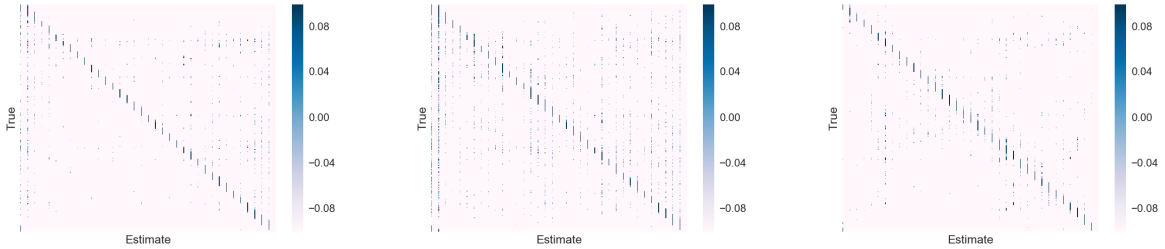
[13] P. T. Boufounos, P. Smaragdis, and B. Raj, "Joint Sparsity Models for Wideband Array Processing," in *SPIE*, 2011, vol. 8138, pp. 81380K–81380K–10.

[14] E. Cagli, D. Carrera, G. Aletti, G. Naldi, and B. Rossi, "Robust DOA Estimation of Speech Signals via Sparsity Models Using Microphone Arrays," in *Proc. IEEE Workshop on Applications of Signal Process. Audio Acoust.*, Oct. 2013, pp. 1–4.

[15] D. D. Lee and H. S. Seung, "Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

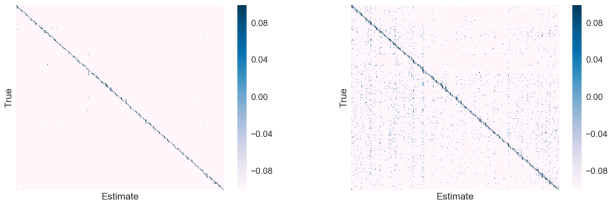
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2867081, IEEE/ACM Transactions on Audio, Speech, and Language Processing





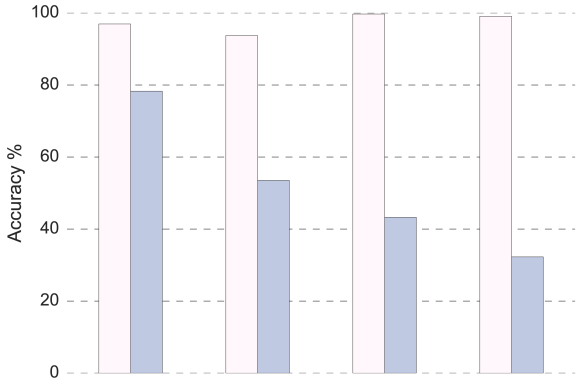

(a) (b) (c)

Fig. 4. Confusion matrices for localizing two sources using LEGO2 at a resolution of $10^\circ$. (a) With prototypes. (b) With a USM. (c) White sources.




(a) One speaker. (b) Two speakers.

Fig. 5. Confusion matrices for localizing female speech with LEGO2 using a multiresolution approach. Improving the resolution decreases the number of outliers in the one-speaker case but not the two-speaker case.




| | Prototypes | USM | USM | USM |
|---|---|---|---|---|
| Resolution | $10^\circ$ | $10^\circ$ | $5^\circ$ | Multiresolution |
| Size of **A** | $321 \times 17172$ | $321 \times 18000$ | $321 \times 36000$ | $321 \times 18000(17500)$ |
| Runtime | 3-6 s | 4-8 s | 6-17 s | 8-19 s |
| Avg. Accuracy | 87.61% | 73.62% | 71.49% | 65.67% |
| Avg. Error | $2.46^\circ$ | $2.48^\circ$ | $1.44^\circ$ | $0.89^\circ$ |

Fig. 6. Summary of localizing one (left) or two (right) female speakers using LEGO2.


[16] C. Févotte, N. Bertin, and J. Durrieu, "Non-negative Matrix Factorization with the Itakura-Saito Divergence. With Application to Music Analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[17] D. L. Sun and G. J. Mysore, "Universal Speech Models for Speaker Independent Single Channel Source Separation," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, 2013, pp. 141–145.

[18] M. N. Schmidt and R. K Olsson, "Single-channel Speech Separation using Sparse Non-negative Matrix Factorization," in *Interspeech*, 2006, pp. 2614–2617.

[19] J. Le Roux, F. J. Weninger, and J. R. Hershey, "Sparse NMF – Half-baked or Well Done?," Tech. Rep. TR2015-023, Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA, Mar. 2015.

[20] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[21] P. Smaragdis, "Convolutive Speech Bases and Their Application to Supervised Speech Separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[22] O. Dikmen and A. T. Cemgil, "Unsupervised Single-channel Source Separation using Bayesian NMF," in *Proc. IEEE Workshop on Applications of Signal Process. Audio Acoust.*, Oct. 2009, pp. 93–96.

[23] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[24] P. Smaragdis and J. C. Brown, "Non-negative Matrix Factorization for Polyphonic Music Transcription," in *Proc. IEEE Workshop on Applications of Signal Process. Audio Acoust.*, Oct. 2003, pp. 177–180.

[25] J. Traa, P. Smaragdis, N. D. Stein, and D. Wingate, "Directional NMF for Joint Source Localization and Separation," in *Proc. IEEE Workshop on Applications of Signal Process. Audio Acoust.*, 2015, pp. 1–5.

[26] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the Narrowband Approximation: Wideband Convex Methods for Under-Determined Reverberant Audio Source Separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1818–1829, Sep. 2010.

[27] L. Parra and C. Spence, "Convolutive Blind Separation of Non-stationary Sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[28] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF Database," in *Proc. IEEE Workshop on Applications of Signal Process. Audio Acoust.*, 2001, pp. 99–102.

[29] M. Ledoux, *The Concentration of Measure Phenomenon*, Math. Surveys Monogr. American Mathematical Society, Providence (R.I.), 2001.

[30] J. Hebrank and D. Wright, "Are Two Ears Necessary for Localization of Sound Sources on the Median Plane?," *J. Acoust. Soc. Am.*, vol. 56, no. 3, pp. 935–938, 1974.

[31] R. M. Reeder, J. Cadieux, and J. B. Firszt, "Quantification of Speech-in-Noise and Sound Localisation Abilities in Children with Unilateral Hearing Loss and Comparison to Normal Hearing Peers," *Audiology and Neurotology*, vol. 20(suppl 1), no. Suppl. 1, pp. 31–37, 2015.

[32] C. Févotte and J. Idier, "Algorithms for Non-negative Matrix Factorization with the Beta-divergence," *Neural Comput.*, vol. 23, no. 9, pp. 2421–2456, Sep. 2011.

[33] A. Lefèvre, F. Bach, and C. Févotte, "Itakura–Saito Non-negative Matrix Factorization with Group Sparsity," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, May 2011, pp. 21–24.

[34] D. L. Donoho, "For Most Large Underdetermined Systems of Linear Equations the Minimal l1-norm Solution is also the Sparsest Solution," *Comm. Pure Appl. Math*, vol. 59, pp. 797–829, 2004.

[35] J. Friedman, T. Hastie, and R. Tibshirani, "A Note on the Group Lasso and a Sparse Group Lasso," *arXiv*, 2010.

[36] A. Cichocki, R. Zdunek, and S. Amari, "New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, May 2006, vol. 5, pp. V621–V624.

[37] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Applied Mathematical Sciences. Springer, New York, NY, 3 edition, 2013.

[38] D. Colton, J. Coyle, and P. Monk, "Recent Developments in Inverse Acoustic Scattering Theory," *SIAM Review*, vol. 42, no. 3, pp. 369–414, 2000.

[39] H. Wierstorf, A. Geier, M.and Raake, and S. Spors, "A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances," June 2016.

[40] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA TIMIT: Acoustic-phonetic Continuous Speech Corpus," Tech. Rep., NIST, 1993, distributed with the TIMIT CD-ROM.

[41] J. Woodruff and D. Wang, "Binaural Localization of Multiple Sources in Reverberant and Noisy Environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1503–1512, July 2012.

[42] D. Kitamura and N. Ono, "Efficient Initialization for Nonnegative Matrix Factorization based on Nonnegative Independent Component Analysis," in *Proc. IEEE Int. Workshop on Acoustic Signal Enhancement*, Sep. 2016, pp. 1–5.

[43] A. N. Langville, C. D. Meyer, R. Albright, J. Cox, and D. Duling, "Algorithms, Initializations, and Convergence for the Nonnegative Matrix Factorization," *arXiv*, 2014.