Binaural Localization of Multiple Sound Sources by Non-Negative Tensor Factorization

Laurent Benaroya, Nicolas Obin, Marco Liuni, Axel Roebel, Wilson Raumel, Sylvain Argentieri

Binaural Localization of Multiple Sound Sources by Non-Negative Tensor Factorization

Laurent Benaroya†, Nicolas Obin, Marco Liuni, Axel Roebel,
Wilson Raumel, Sylvain Argentieri

***Abstract*—This paper presents non-negative factorization of audio signals for the binaural localization of multiple sound sources within realistic and unknown sound environments. Non-negative tensor factorization (NTF) provides a sparse representation of multi-channel audio signals in time, frequency, and space that can be exploited in computational audio scene analysis and robot audition for the separation and localization of sound sources. In the proposed formulation, each sound source is represented by mean of spectral dictionaries, temporal activation, and its distribution within each channel (here, left and right ears). This distribution, being dependent on the frequency, can be interpreted as an explicit estimation of the Head-Related Transfer Function (HRTF) of a binaural head which can then be converted into the estimated sound source position. Moreover, the semi-supervised formulation of the non-negative factorization allows to integrate prior knowledge about some sound sources of interest whose dictionaries can be learned in advance, whereas the remaining sources are considered as background sound which remains unknown and is estimated on-the-fly. The proposed NTF-based sound source localization is here applied to binaural sound source localization of multiple speakers within realistic sound environments.**

Index Terms: binaural localization, robot audition, computational audio scene analysis, non-negative tensor factorization.

I. Introduction

HUMANS have the ability to identify, separate, and localize sound sources while listening to complex sound environments. The objective of machine listening is to reproduce human's listening ability in order to analyze and understand automatically the content of a sound scene, presumably unknown in advance. Over the past decade, computational audio scene analysis (CASA) has grown in interest in the audio signal processing community, especially for the sound event detection of speech, music, and recently moving forward to all sort of environmental sounds [1]–[4]. Nowadays, CASA faces realistic, complex, and challenging sound environments: multiple sources are present simultaneously, in the presence of background noise and reverberant conditions, and at different and possibly moving positions. In this context, non-negative matrix factorization (NMF) [5], [6] has considerably gained in popularity in the recent times for audio scene analysis with competitive scores in international challenges [7], [8].

Humanoid robotics appears to be an ideal ground for the development of machine listening systems reproducing humans listening ability, and confronting realistic sound environments. Furthermore, modern robotics applications require them to be able to adapt to any environment unknown in advance. In contrast with traditional computational audio scene analysis, this can be achieved with a robot generally endowed with multiple microphones which allows him to exploit spatial information to localize, separate, and identify some sources of interest around him. In particular, a humanoid robot has, like humans, two ears to listen to its environment: consequently, binaural audition is commonly used for the localization of sound sources in robot audition. Binaural source localization is generally based on the estimation of binaural cues such as the Interaural Phase Difference (IPD) and the Interaural Level Difference (ILD), and their conversion to the corresponding sound source position (e.g., PHAT [9]–[11], DUET [12], [13], MESSL [14], [15], among others [16]–[18]). This conversion can be performed analytically through simple heuristics [13], statistically through machine learning [18], or by using the Head-Related Transfer-Function (HRTF) of the robot [16], [17]. HRTFs are commonly used in humanoid robotics, since it models the effect of the robot head and body on its sound perception, and thus can be simply used to construct a mapping between the sound source position and the corresponding binaural cues. Like for computational audio scene analysis, the challenge of robot audition has recently moved towards binaural source localization within realistic environments [19].

This paper presents a framework based on multi-channel non-negative factorization for the binaural localization of multiple and simultaneous sound sources in the context of humanoid robot audition and realistic sound environments. Initially utilized for single channel audio signals, NMF provides a sparse time-frequency representation of simultaneous sound sources by means of spectral dictionaries and temporal activation. More recently, its extension to multi-channel audio signals such as multi-channel NMF and non-negative tensor factorization (NTF) [20], [21] allows to integrate implicitly spatial information through considering a term of mixing of the sound sources over multiple channels. However, these mixing factors can not necessarily be translated into some explicit spatial information for sound source localization, such as the azimuth of the sound sources: for instance, the "spatial position" determined [22] is limited to the estimation of the mixing

†This research was conducted during the French EMERGENCE project ROUTE (RObot à l'écoUTE, 2015-2016) and funded by Sorbonne Universités.

L. Benaroya is with LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France. His contribution was mainly done while he was with IRCAM.

N. Obin, M. Liuni, and A. Roebel are with STMS, IRCAM (Institut de Recherche et Coordination Acoustique Musique), CNRS, Sorbonne Université, Paris, France.

W. Raumel and S. Argentieri are with Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France.

scalars, and evaluated on a single and simple study case. In more recent contributions, multi-channel NMF has been proposed for sound source localization based on beamforming from a microphone array [23], [24]. The main contribution of this paper is to present a NTF framework specifically designed for binaural sound source localization. It proposes the formulation of a binaural NTF, in which the mixing matrices are function of the frequency and interpreted as the head-related transfer function (HRTF) of the binaural head, which encodes the spatial information of the sound sources. Additionally, a generic binaural framework is presented to face realistic sound source localization, by integrating prior knowledge about the spectral content and the nature of the sound sources to be localized within some unknown background sound environment. Finally, this constitutes the first evaluation of NTF in a task of sound source localization, and especially for binaural sound source localization.

The remaining of the paper is organized as follows: the fundamentals of robot audition are presented in Section II, with a particular attention on binaural audition, binaural sound source localization, and head-related transfer function (HRTF). Then, the proposed NTF framework for binaural source localization is described in details in Section III. The proposed NTF-based localization method is evaluated in Section IV for the localization of multiple speakers within realistic environments and compared to State-of-the-Art (SoA) NMF separation and binaural localization methods.

II. Robot Audition and Binaural Sound Localization

Binaural audition has been receiving a growing attention recently, due to an increasing demand for humanoid robots endowed with bio-inspired perception and symbiotic interaction between humans and robots. This paper is rooted in the binaural paradigm [25], [26] where only two signals originating from two microphones placed on an anthropomorphic head must face complex auditory scenario, involving simultaneous spatialized sound sources under noisy and reverberant conditions [27]. This section introduces the fundamentals of binaural audition, including the notions of head-related transfer functions, binaural cues, and binaural sound source localization.

A. Binaural audition

Let us consider a single pointwise sound source emitting, at time t in samples, a signal $s(t)$ whose position is defined by its distance d (in m), azimuth θ (in rad) and elevation ψ (in rad) w.r.t. the head of a robot. Its frequency counterpart $S_{k,n}$ is obtained by a Short-Term Fourier transform (STFT), with the frequency index $k \in [1, \dots, F]$ in bins, and the time index $n \in [1, \dots, T]$ in frames. Let us define $y_l(t)$ and $y_r(t)$ the left and right ear signals perceived by the robot, with frequency counterparts denoted by the tensor $\mathcal{Y} \in \mathbb{C}^{2\times F\times T}$ and the left and right scalar $\mathcal{Y}_{l,k,n}$ and $\mathcal{Y}_{r,k,n}$ respectively.

The perceived binaural signals are obtained by the modification brought by the body of the robot to the incident source sound wave, including its torso, head and pinnas

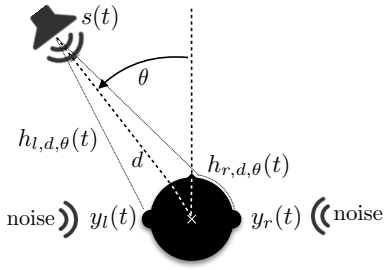


Fig. 1. Illustration of binaural sound source localization. A sound source emits a signal $s(t)$ from the position $(d, \theta, \psi = 0^\circ)$, with $\theta = 0^\circ$ corresponding to a source placed in front of the robot. The source signal propagates to the left and right ears of a binaural system, thus defining the left and right binaural signals $y_l(t)$ and $y_r(t)$, both of them being related to the source signal $s(t)$ thanks to the Head-Related Impulse Response $h_{l,d,\theta}(t)$ and $h_{r,d,\theta}(t)$ respectively. Additionally to the source signal, a diffuse noise field is simulated, producing some additive left and right noise signals.

effects. These effects are captured through the Head-Related Impulse Responses (HRIRs) $h_{l,.}(t)$ and $h_{r,.}(t)$, whose Fourier transforms define the left and right Head-Related Transfer Functions (HRTFs) $H_{l,.}$ and $H_{r,.}$, depending on the sound source position (see Figure 1). As most of the studies on sound source localization, this paper is strictly focused on the estimation of the azimuthal localization of θ, so that the elevation ψ and the distance d will be further ignored. Under the anechoic assumption, the relationship between the emitted source signal and the perceived binaural signals can be written as

$$\begin{cases} \mathcal{Y}_{l,k,n} = H_{l,\theta,k} S_{k,n} \\ \mathcal{Y}_{r,k,n} = H_{r,\theta,k} S_{k,n} \end{cases}. \tag{1}$$

In this paper, a KEMAR Head And Torso Simulator (HATS), together with its measured HRTFs provided by the MIT database [28], is used.

B. Binaural cues

Two primary auditory cues exploited by humans for binaural localization are introduced, inspired by the *duplex theory*. The Interaural Level Difference (ILD) is defined as the difference between the intensity of the left and right ears as

$$\mathrm{ILD}(k,n) = 20 \log \frac{|\mathcal{Y}_{l,k,n}|}{|\mathcal{Y}_{r,k,n}|}. \tag{2}$$

The Interaural Phase Difference (IPD)is defined by the path difference to be traveled by the source wave to reach the left and right ears as

$$\mathrm{IPD}(k,n) = \angle \frac{\mathcal{Y}_{l,k,n}}{\mathcal{Y}_{r,k,n}}, \tag{3}$$

where $\angle$ denotes the phase in radians of a complex number. By integrating the HRTFs as defined in Eq. (1) into Eq. (2) and (3), one can directly express the expected binaural cues as a function of the azimuth, as

$$\mathrm{ILD}_\theta^{hrtf}(k) = 20 \log \frac{|H_{l,\theta,k}|}{|H_{r,\theta,k}|}, \tag{4}$$

$$\mathrm{IPD}_\theta^{hrtf}(k) = \angle \frac{H_{l,\theta,k}}{H_{r,\theta,k}}. \tag{5}$$

This means that the position of the sound source can be determined from the binaural cues, by comparison to the expected values which can be derived from the HRTFs measurements. The expected binaural cues are represented in Figure 2, as a function of the source azimuth and the frequency in Hz. While IPD exhibits an almost linear dependency to the frequency, ILD shows more complex patterns, with frequency components being possibly amplified or attenuated by up to 40dB.

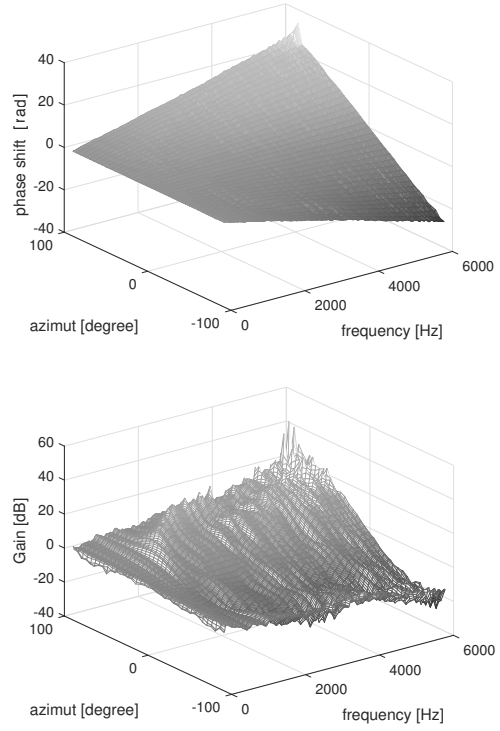


Fig. 2. Illustration of IPD and ILD cues, as a function of the source azimuth and the frequency as measured from the HRTFs in [28].

C. Binaural sound source localization

In a real environment, the single source is emitting in a possibly noisy and reverberated environment, so that Eq. (1) is no longer valid. Accordingly, the sound source is localized by comparison of the observed binaural cues ILD and IPD with the expected binaural cues as computed from Eq. (4) and (5) [29]. The estimated sound source position $\widehat{\theta}(n)$ can be then determined as

$$\widehat{\theta}_{\text{ILD}}(n) = \underset{\theta}{\text{argmin}} \ \|\text{ILD}(.,n) - \text{ILD}_{\theta}^{hrtf}(.)\|, \quad (6)$$

$$\widehat{\theta}_{\text{IPD}}(n) = \underset{\theta}{\text{argmin}} \ \|\text{IPD}(.,n) - \text{IPD}_{\theta}^{hrtf}(.)\|. \quad (7)$$

where $||.||$ is the Frobenius norm, calculated over all frequencies.

D. Issues and challenges

Since robots are more and more used in realistic environments, the source localization must face the presence of noise in the measurements. However, sound source localization exhibits a serious drop in performance in the presence of noisy environments [30], especially in the binaural context. Besides, binaural sound source localization using HRTF is generally limited to the localization of a single source, as Eq. (2) and (3) only apply to one source [16], [31]. Its extension to multiple sources is not straightforward, and would require a time-frequency decomposition of the sources. The proposed binaural framework assumes explicitly the presence of multiple sources and noise in the measurements, and will be able to face realistic sound environments.

III. NTF-based binaural localization

The main contribution of this paper is the use of non-negative factorization of audio signals for the binaural localization of multiple sound sources within unknown sound environment. The proposed system is based on a Non Negative Tensor Factorization (NTF) [21], which is specifically designed for binaural sound source localization. The proposed binaural NTF framework is presented for the representation of multiple sound sources, followed by a description on the main contribution of the paper: how the sound sources are localized from the NTF decomposition. In the following, vectors and matrices are denoted by bold upper case letters, and tensors by upper case letters with calligraphic letters.

A. NMF

Let $\boldsymbol{X} \in \mathbb{R}^{+F \times T}$ be the matrix of some observations, represented by the amplitude spectrogram of a signal $x(t)$, where F is the number of frequency bins and T is the number of frames. The standard approximation of $\boldsymbol{X}$ by Non-negative Matrix Factorization (NMF) can be written as:

$$\boldsymbol{X} \simeq \boldsymbol{V} = \boldsymbol{W}\boldsymbol{H} \quad (8)$$

where: $\boldsymbol{W} \in \mathbb{R}^{+F \times S}$ is a dictionary of spectral bases, and $\boldsymbol{H} \in \mathbb{R}^{+S \times T}$ is a matrix of their activations over time, S being the number of components. The NMF problem is then to determine the $(\boldsymbol{W}, \boldsymbol{H})$ parameters which minimize a given cost function $C(\boldsymbol{X}|\boldsymbol{V})$. Usual costs are Kullback-Leiber (KL) and Itakura-Saito (IS) divergences, which are both limit cases of the β-divergence d_β, as defined in [32] (in equation 5).

$$C(\boldsymbol{X}|\boldsymbol{V}) = D_\beta(\boldsymbol{X}|\boldsymbol{V}) = \sum_{k=1}^{F}\sum_{n=1}^{T} d_\beta(X_{k,n}|V_{k,n}) \quad (9)$$

The solution of the NMF problem, using β-divergence, can be efficiently obtained by applying an iterative algorithm, derived from a gradient step descent technique which leads to the heuristic Multiplicative Update rules (noted "MU") [32].

B. Non-Negative Tensor Factorization of Binaural Signals

The Non-negative Tensor Factorization (NTF) is a generalization of the NMF to multiple channels as provided by multiple microphones [20], [21]. This generalization assumes the sources are non-equally distributed over the channels and represented by mixing matrices, thus introducing implicitly spatial information within the NMF framework. This section presents the general NTF framework and the specific architecture designed for binaural listening and sound source localization.

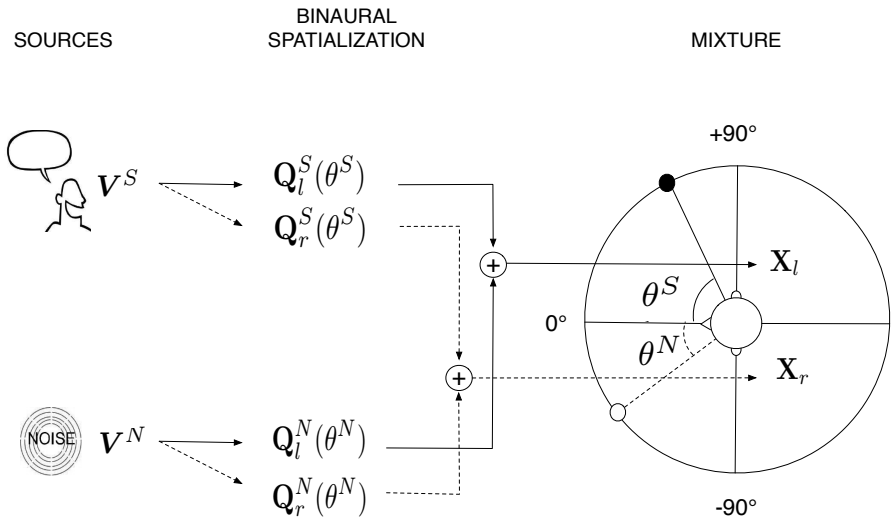


Fig. 3. Illustration of NTF for binaural source localization in the case of one speech source and one noise source. For the speech source: $\boldsymbol{V}^S$ is the amplitude spectrogram of the source, $\boldsymbol{Q}_l^S(\theta^S)$ and $\boldsymbol{Q}_r^S(\theta^S)$ are the HRTFs associated with the azimuth θ^S. Similarly, for the noise source, which is considered as a punctual source in this graphic, $\boldsymbol{V}^N$ is the amplitude spectrogram, $\boldsymbol{Q}_l^N(\theta^N)$ and $\boldsymbol{Q}_r^N(\theta^N)$ are the HRTFs associated with the azimuth θ^N.

The amplitude spectrogram $\mathcal{X}$ and its approximation $\mathcal{V}$ are tensors in $\mathbb{R}^{+C\times F\times T}$, where C is the total number of channels. Let $\boldsymbol{X}_c$ in $\mathbb{R}^{+F\times T}$ be the matrix corresponding to the channel c and $\boldsymbol{V}_c$ its approximation. Considering R sources, the contribution term $\mathcal{V}^{(r)}$ of the r^{th} source to the spectrogram approximation $\mathcal{V}$ is factorized by a mixing matrix $\boldsymbol{Q}^{(r)} \in \mathbb{R}^{+C\times F}$ and a spectro-temporal component $\boldsymbol{V}^{(r)} \in \mathbb{R}^{+F\times T}$. This can be written in matrix form as

$$\boldsymbol{X}_c \approx \boldsymbol{V}_c = \sum_{r=1}^{R} \widetilde{\boldsymbol{Q}}_c^{(r)}\, \boldsymbol{V}^{(r)} \tag{10}$$

where the matrix $\widetilde{\boldsymbol{Q}}_c^{(r)} = \operatorname{diag} \boldsymbol{Q}_c^{(r)}$ is the diagonal matrix in $\mathbb{R}^{+F\times F}$ with the coefficients set to $\boldsymbol{Q}_c^{(r)}$ on the diagonal, the vector $\boldsymbol{Q}_c^{(r)} \in \mathbb{R}^{+F}$ being the contribution of the mixing matrix of the r^{th} source to the channel c.

The NTF parameters are estimated so as to minimize the reconstruction cost, defined as:

$$\mathcal{D}_\beta(\mathcal{X}|\mathcal{V}) = \sum_{c=1}^{C} D_\beta(\boldsymbol{X}_c|\boldsymbol{V}_c) \tag{11}$$

where $D_\beta(\boldsymbol{X}_c|\boldsymbol{V}_c)$ is defined by equation (9). The contributions of each channel to the cost function are independent, i.e. there is no cross-channel term.

The proposed NTF architecture is specifically designed for the binaural localization of speech sources mixed with unknown noise sources, as illustrated in Figure 3. First, the proposed NTF model assumes that the amplitude spectrogram is a mixture of M speech sources and P noise sources. The channel amplitude spectrogram $\boldsymbol{X}_c$ is then approximated by $\boldsymbol{V}_c$:

$$\boldsymbol{V}_c = \sum_{m=1}^{M} \widetilde{\boldsymbol{Q}}_c^{S,(m)}\, \boldsymbol{V}^{S,(m)} + \sum_{p=1}^{P} \widetilde{\boldsymbol{Q}}_c^{N,(p)}\, \boldsymbol{V}^{N,(p)} \tag{12}$$

where $\boldsymbol{V}^{S,(m)} \in \mathbb{R}^{+F\times T}$ is the spectral component of the m^{th} *speech* source and $\boldsymbol{V}^{N,(p)} \in \mathbb{R}^{+F\times T}$ is the spectral component of the p^{th} *noise* source, and $\boldsymbol{Q}_c^{S,(m)}$ and $\boldsymbol{Q}_c^{N,(p)} \in \mathbb{R}^{+F}$ are the corresponding mixing vectors with the corresponding diagonal matrices $\widetilde{\boldsymbol{Q}}_c^{S,(m)}$ and $\widetilde{\boldsymbol{Q}}_c^{N,(p)} \in \mathbb{R}^{+F\times F}$. Second, the proposed NTF model also assumes spatial and spectral factorizations that are specific to the binaural factorization of speech and noise sources.

1) Spatial factorization: The binaural NTF is obtained by setting the channels to $c \in \{l, r\}$, with $c = l$ for the left ear and $c = r$ for the right ear, obtaining a binaural spectrogram $\mathcal{X} \in \mathbb{R}^{+2\times F\times T}$. Accordingly, the mixing matrices $\boldsymbol{Q}^{S,(m)} \in \mathbb{R}^{+2\times F}$ defined in Eq. (12) are considered as HRTF estimates depending on the position of the corresponding speech sources. The proposed binaural localization framework can be generalized to more than two sensors, which would require some additional measurements of the HRTF at specific locations (e.g., left/right ear, nose, back of the head, etc...), and the extension of binaural cues to multi-aural cues.

2) Spectral factorization: The spectral dictionaries of the speech sources $\boldsymbol{V}^{S,(m)}$ are learned or computed in advance and then fixed as priors, while the parameters of the noise sources $\boldsymbol{V}^{N,(p)}$ are presumably unknown and learned on-the-fly. Also, specific dictionaries are constructed depending on the nature of the source. The speech sources spectral components are factorized by using the source/filter factorization as proposed in [33], [34] and [35]. The NMF source/filter decomposition of the magnitude spectrogram of a speech signal can be expressed as:

$$\begin{aligned} \boldsymbol{V}^{S,(m)} &= \boldsymbol{V}_{\text{ex}}^{S,(m)} \odot \boldsymbol{V}_{\Phi}^{S,(m)} \\ &= \underbrace{\left(\boldsymbol{W}_{\text{ex}}^{S} \boldsymbol{H}_{\text{ex}}^{S,(m)}\right)}_{\text{excitation}} \odot \underbrace{\left(\boldsymbol{W}_{\Phi}^{S,(m)} \boldsymbol{H}_{\Phi}^{S,(m)}\right)}_{\text{filter}} \end{aligned} \tag{13}$$

where the symbol $\odot$ indicates the Hadamard product, i.e. point-wise multiplication of matrices; $\boldsymbol{V}_{\text{ex}}^{S,(m)}$ and $\boldsymbol{V}_{\Phi}^{S,(m)}$ are respectively the magnitude spectrogram of the excitation part and the filter part; $\boldsymbol{W}_{\text{ex}}^{S}$ and $\boldsymbol{H}_{\text{ex}}^{S,(m)}$ are the standard NMF decomposition for the speech excitations $\boldsymbol{V}_{\text{ex}}^{S,(m)}$, with $\boldsymbol{W}_{\text{ex}}^{S}$ being a fixed dictionary including periodic and noisy basis; $\boldsymbol{W}_{\Phi}^{S,(m)}$ and $\boldsymbol{H}_{\Phi}^{S,(m)}$ are the standard NMF decomposition

for the speech filters $\boldsymbol{V}_{\Phi}^{S,(m)}$. As proposed in [34], we learn the speech filters $\boldsymbol{W}_{\Phi}^{S,(m)}$ from the clean speech signals of a speaker. The noise sources spectral components are factorized as in classical NMF:

$$\boldsymbol{V}^{N,(p)} = \boldsymbol{W}^{N,(p)}\boldsymbol{H}^{N,(p)} \tag{14}$$

A summary of the status (set, fixed or free) of the matrices involved in the proposed NTF both in the training step and the testing step is given in table I.

TABLE I
STATUS OF THE NMF/NTF MATRICES AND TENSORS FOR THE m^{th} SPEECH SOURCE AND THE p^{th} NOISE SOURCE. THE COMPONENTS ARE EITHER **SET** TO A STATIC MATRIX OR **FREE** (I.E. ESTIMATED) OR **FIXED**.

	train	**test**
$\boldsymbol{W}_{\Phi}^{S,(m)}$	free	fixed
$\boldsymbol{H}_{\Phi}^{S,(m)}$	free	free
$\boldsymbol{W}_{\text{ex}}^{S}$	set	fixed
$\boldsymbol{H}_{\text{ex}}^{S,(m)}$	-	free
$\boldsymbol{W}^{N,(p)}$	-	free
$\boldsymbol{H}^{N,(p)}$	-	free
$\boldsymbol{Q}^{S,(m)}$	-	free
$\boldsymbol{Q}^{N,(p)}$	-	free

The MU for the proposed NTF with the β-divergence can be derived straightforwardly from equations (12) , (13) and (14) and the definition of the cost function in (11) . For simplicity, the matrix based notation is used, i.e. with the matrices $\boldsymbol{X}_c$, $\boldsymbol{V}_c$ and vectors $\boldsymbol{Q}_c^{S,(m)}$, $\boldsymbol{Q}_c^{N,(p)}$ with $c \in \{l, r\}$ rather than the tensors $\mathcal{X}$, $\mathcal{V}$. Accordingly, the multiplicative updates (MU) for the free parameters $\boldsymbol{H}_{\text{ex}}^{S}, \boldsymbol{H}_{\Phi}^{S}$, $\boldsymbol{W}^{N}, \boldsymbol{H}^{N}$, $\boldsymbol{Q}_c^{S}$, and $\boldsymbol{Q}_c^{N}$ can be computed as follows.

Speech components :

$$\boldsymbol{H}_{\text{ex}}^{S,(m)} \leftarrow \boldsymbol{H}_{\text{ex}}^{S,(m)} \odot \frac{(\boldsymbol{W}_{\text{ex}}^{S})^{\top}[\sum_c \widetilde{\boldsymbol{Q}}_c^{S,(m)}(\boldsymbol{X}_c \odot \boldsymbol{V}_c^{\beta-2} \odot \boldsymbol{V}_{\Phi}^{S,(m)})]}{(\boldsymbol{W}_{\text{ex}}^{S})^{\top}[\sum_c \widetilde{\boldsymbol{Q}}_c^{S,(m)}(\boldsymbol{V}_c^{\beta-1} \odot \boldsymbol{V}_{\Phi}^{S,(m)})]}$$

$$\boldsymbol{H}_{\Phi}^{S,(m)} \leftarrow \boldsymbol{H}_{\Phi}^{S,(m)} \odot \frac{(\boldsymbol{W}_{\Phi}^{S,(m)})^{\top}[\sum_c \widetilde{\boldsymbol{Q}}_c^{S,(m)}(\boldsymbol{X}_c \odot \boldsymbol{V}_c^{\beta-2} \odot \boldsymbol{V}_{\text{ex}}^{S,(m)})]}{(\boldsymbol{W}_{\Phi}^{S,(m)})^{\top}[\sum_c \widetilde{\boldsymbol{Q}}_c^{S,(m)}(\boldsymbol{V}_c^{\beta-1} \odot \boldsymbol{V}_{\text{ex}}^{S,(m)})]}$$

Noise components :

$$\boldsymbol{W}^{N,(p)} \leftarrow \boldsymbol{W}^{N,(p)} \odot \frac{[\sum_{c\in\{l,r\}} \widetilde{\boldsymbol{Q}}_c^{N,(p)}(\boldsymbol{X}_c \odot \boldsymbol{V}_c^{\beta-2})](\boldsymbol{H}^{N,(p)})^{\top}}{[\sum_{c\in\{l,r\}} \widetilde{\boldsymbol{Q}}_c^{N,(p)}\boldsymbol{V}_c^{\beta-1}](\boldsymbol{H}^{N,(p)})^{\top}}$$

$$\boldsymbol{H}^{N,(p)} \leftarrow \boldsymbol{H}^{N,(p)} \odot \frac{(\boldsymbol{W}^{N,(p)})^{\top}[\sum_{c\in\{l,r\}} \widetilde{\boldsymbol{Q}}_c^{N,(p)}(\boldsymbol{X}_c \odot \boldsymbol{V}_c^{\beta-2})]}{(\boldsymbol{W}^{N,(p)})^{\top}[\sum_{c\in\{l,r\}} \widetilde{\boldsymbol{Q}}_c^{N,(p)}\boldsymbol{V}_c^{\beta-1}]}$$

Mixing matrices :

$$\boldsymbol{Q}_c^{S,(m)} \leftarrow \boldsymbol{Q}_c^{S,(m)} \odot \frac{[\boldsymbol{X}_c \odot \boldsymbol{V}_c^{\beta-2} \odot \boldsymbol{V}^{S,(m)}]1_{T\times 1}}{[\boldsymbol{V}_c^{\beta-1} \odot \boldsymbol{V}^{S,(m)}]1_{T\times 1}}$$

$$\boldsymbol{Q}_c^{N,(p)} \leftarrow \boldsymbol{Q}_c^{N,(p)} \odot \frac{[\boldsymbol{X}_c \odot \boldsymbol{V}_c^{\beta-2} \odot \boldsymbol{V}^{N,(p)}]1_{T\times 1}}{[\boldsymbol{V}_c^{\beta-1} \odot \boldsymbol{V}^{N,(p)}]1_{T\times 1}}$$

where $1_{T\times 1}$ is a T-vector with ones.

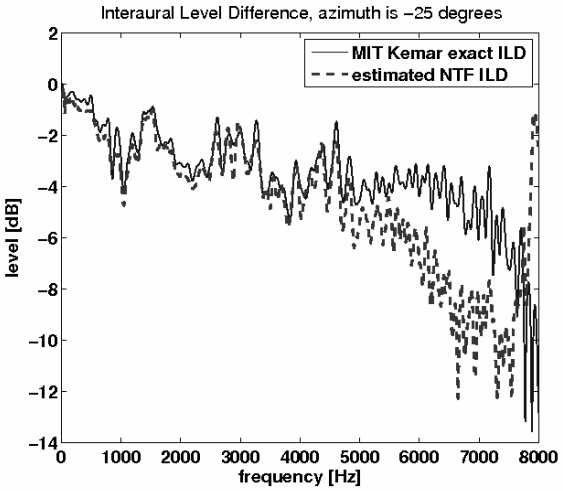


Fig. 4. Comparison of the ILD from the Kemar dummy-head magnitude HRTFs from MIT database (plain) and the estimated ILD from the $\boldsymbol{Q}^{S}$ matrix in our NTF algorithm (doted). The data is a sentence of the speaker FCJF0 at the azimuth -25^{o} without noise in an anechoic room.

C. Proposed sound source localization from NTF

The proposed NTF framework can be used to localize the sound sources, either by using the estimated binaural mixing matrix or the estimated source images. In the former case, the localization is processed internally in the NTF. In the latter case, the NTF is used as sound source separation, which is followed by an external sound source localization. As in the previous subsection, the sound sources are speech sources, though it can be applied to any kind of source.

1) Estimated binaural mixing matrix: The first idea is to perform the localization internally to the NTF. Indeed, the mixing matrix $\boldsymbol{Q}^{S,(m)}$ explicitly encodes the *magnitude* of the HRTF, which will be noted MHRTF, of the left and right ears of the binaural head. Consequently, it can be directly used to compute the localization of each speech sources based on the comparison of the MHRTFs estimated by the mixing matrices and the true MHRTFs measured on the binaural head. First, we compute the estimated ILD of the m^{th} speech source :

$$\text{ILD}^{S,(m)}(k) = 20\log_{10}\frac{Q_{l,k}^{S,(m)}}{Q_{r,k}^{S,(m)}} \tag{15}$$

The computation of the estimated ILD in equation (2) is based on the binaural spectrogram and is a function of the frequency index k and the frame index n. The estimated ILD in equation (15) is only a function of the frequency index k since it comes from the mixing matrix $\boldsymbol{Q}^{S,(m)}$ which does not depend on the frame index n. This estimation of the ILD is a global estimation for the entire audio signal.

The estimated ILD is then compared to a set of ILDs from a reference HRTF database in order to provide the corresponding azimuth. The estimation of the azimuth is given by the formula :

$$\hat{\theta}^{S,(m)} = \underset{\theta}{\text{argmin}} \sum_{k=1}^{F} g\left(\text{ILD}^{S,(m)}(k) - \text{ILD}_{\theta}^{hrtf}(k)\right) \tag{16}$$

where $g(.)$ is a cost function. In general, $g(x) = x^2$ is chosen as the cost function (Mean Square Error).

With respect to the Duplex theory, the frequency range on which the ILDs are computed in equation (16) is restricted to a high-frequency bandwidth (1500 to 4500 Hz). Figure 4 presents a comparison between the ILD estimated by NTF and the corresponding ground truth ILD from an HRTF database. Figure 5 illustrates the localization of two speakers at same level with a diffuse noise at $0dB$. It is important to notice that due to the nature of the non-negative framework which assumes non-negative real values, only the ILD can be estimated from the decomposition. This limitation could be removed by using Complex-valued NMF [36] in order to exploit all of the binaural cues including the IPD.

2) Fixed binaural mixing matrix: A variant of this idea can be obtained by using prior knowledge about the expected MHRTFs, as measured in the HRTF database. Considering the m^{th} speech source and each possible MHRTFs associated with azimuth θ, a binaural factorization is computed in equation (12) with the corresponding mixing matrix $\mathbf{Q}^{S,(m)}_{\theta,hrtf}$ being fixed as one of the expected pairs of MHRTFs. Then, the localization of the speech source is chosen as the one minimizing the cost function, as follows:

$$\hat{\theta}^{S,(m)} = \underset{\theta^{S,(m)}}{\operatorname{argmin}} \ \mathcal{D}_\beta(\mathcal{X}|\mathcal{V}_{\theta^{S,(m)}}) \tag{17}$$

where $\mathcal{V}_{\theta^{S,(m)}}$ is the decomposition obtained with a fixed mixing matrix $\mathbf{Q}^{S,(m)}_{\theta,hrtf}$ and all other mixing matrices free. This implies that all expected MHRTF mixing matrices must be tested in the NTF algorithm for the considered speech source, requiring as many decompositions as possible MHRTF mixing matrices. This variant can be used to localize multiple speech sources in a similar manner, but it is more computationally expensive, since its complexity increases as the power of the number of sound sources.

3) Speech source images: The third idea is to use the individual speech source images as separated from the environment noise sources to process the localization. The speech source images $\hat{\mathbf{S}}^{\mathrm{S},(m)}_c$ are separated using the standard generalized Wiener filter [37]:

$$\hat{\mathbf{S}}^{\mathrm{S},(m)}_c = \mathbf{Y}_c \odot \frac{\boldsymbol{Q}^{S,(m)}_c \odot \boldsymbol{V}^{S,(m)}}{\boldsymbol{V}_c} \tag{18}$$

where $\mathcal{Y} \in \mathbb{C}^{2\times F\times T}$ is the STFT of the stereo image of the *mixture*, $\mathbf{Y}_c \in \mathbb{C}^{\times F\times T}$ is the corresponding matrix on channel c and $\boldsymbol{V}_c$ is defined in equation (12). The separation can be followed externally by a SoA binaural localization algorithm in order to compute the azimuth from the estimated source images. In particular, the ILD can be computed directly from the sources images from (18), as:

$$\mathrm{ILD}^{S,(m)}(k,n) = 20\log_{10}\left(\frac{Q^{S,(m)}_{l,k}}{Q^{S,(m)}_{r,k}} \cdot \frac{|Y_{l,k,n}|}{|Y_{r,k,n}|} \cdot \frac{V_{r,k,n}}{V_{l,k,n}}\right) \tag{19}$$

This is equal to (15) in the case where the amplitude spectrograms $|Y_{c,k,n}|$ and their approximations by the NTF factorizations $V_{c,k,n}$ as defined in (12) are equal, for instance if the factorizations are full rank.

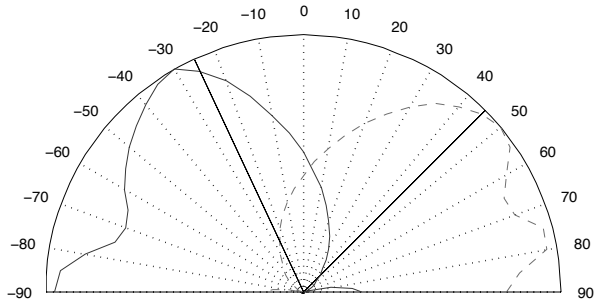


Fig. 5. Radar plot of the error function with respect to the azimuth for multiple sound sources localization in a noisy sound environment. This illustrates the NTF-based localization in the case of two speakers from TIMIT located at -25° and +40° in the presence of a diffuse noise from QUT-NOISE at 0 dB SNR. The black solid lines indicate the true position of the speakers, the solid and the dashed curves indicate the opposite of the error function estimated for the two speakers, whose maximum is the estimated position of the speaker.

IV. Experiments

A. Experimental setup

Two experiments were conducted in order to evaluate the NTF-based binaural localization for the localization of one and multiple speakers within realistic environmental noise.

1) Benchmark: The benchmark for the localization was composed of three methods: SoA binaural localization algorithms, including GCC-PHAT [11], DUET [12], and a binaural localization algorithm based on ILD and described in equation (6) (referred to as ILD); SoA binaural localization from the source images as separated by NTF (as described in Section III-C3 and referred to as NTF+sep), and binaural localization from the binaural mixing matrix of the NTF, estimated or fixed (as described respectively in Sections III-C1 and III-C2, and referred as *NTF* $+\boldsymbol{Q}^S$).

For all SoA algorithms, the localization was performed for each frame of the Short-Term Fourier Transform (STFT) of the signal windowed by a 25 ms Hamming window, with 50% overlapping, the estimated binaural cues were converted into the corresponding azimuth by using the HRTF mapping as described in equations (6) and (7), and a single azimuth was determined for each simulation as the one maximizing the azimuth histogram over the complete signal. For the NTF algorithm, the chosen cost function in (16) is $g(x) = \sqrt{|x|}$ and the range of the ILDs is restricted between 1500 and 4500 Hz accordingly to the Duplex theory. The activation matrices $\boldsymbol{H}_{\mathrm{ex}}$ and $\boldsymbol{H}_{\Phi}$ are initialized by the outputs of an NMF on the mono downmix of the mixture, and the noise matrices are randomly initialized and the mixing filters are initialized in front of the head. All simulations were conducted by using the Ircam Spat software [38] dedicated to sound spatialization and artificial reverberation, within the Max real-time audio environment [1],

[1] *Max* is a Visual Programming Language for Media, available at the web page https://cycling74.com/products/max

and by using the Kemar dummy-head HRTFs from the MIT database [28]. In this database, the HRTFs were measured at constant distance (1.4 meters) in an anechoic chamber and the distance between the two ears of the Kemar dummy-head is 18 centimeters.

2) Sound localization database: A large set of binaural sound scenes was created by mixing and spatializing speaker recordings from the TIMIT English-American speaker database [39] and environmental sounds collected from the QUT-NOISE-TIMIT database [40]. The TIMIT speaker database is composed of 630 speakers pronouncing 10 sentences each, in which 2 are shared among all speakers and the remaining 8 are different across all speakers, from which 5 males and 5 females speakers were randomly selected (MPDF0, FHEW0, FCJF0, FDAW0, FDML0, FECD0, MGRL0, MJEB1, MKLS0, MMRP0). The 2 shared sentences were used to train the speech dictionaries and the 8 remaining sentences for the localization simulations. Each speaker has been spatialized with azimuth ranging from -90 degrees to 90 degrees, with a 5 degrees increment in front of the binaural head. The other spatialization setups were fixed as follows: the yaw of the speaker was oriented to the center of the binaural head, and the aperture of the speaker was set to 40 degrees. The background sounds used were the CAFE_CAFE-1, car_WINDOWB-1, HOME_KITCHEN-1, and STREET_CITY-1 collected from QUT-NOISE-TIMIT database, which are long recordings (about 1 hour) of real sound backgrounds.

The background sounds, originally mono-channel, were used to simulate a diffuse background sound environment. To do so, sound extracts were randomly selected at different time position from the original background sounds in order to create uncorrelated sound sources. These sound sources were then located at different positions on a sphere surrounding the binaural head as specified in [41]. Finally, speech and background sound were mixed at four signal-to-noise ratio (SNR): infinite, $+6$, 0, and -6 dB. The SNR was measured as described in [41]: for each spatialized source (source of interest and background sound sources), the level is measured at the center of the virtual head in absence of the head - i.e., in absence of any HRTF effect. These measurements are used to compute the SNR between the sources of interest and the background sound sources, and then to adjust the background sound level to the desired SNR.

Two experiments were conducted: one with a single speaker mixed within a diffuse noise, and one with two speakers in the same conditions. In the single speaker case, the evaluation is done on all combinations : the ten speakers spatialized from -90 deg to +90 deg with a step of 10 degrees with the four diffuse noises, which represents 760 combinations at each SNR. In the two speakers case, the evaluation is done on a subset : the 25 female or male pairs are taken; the speakers position and the noise type are jointly and randomly chosen on a subset of 40 combinations among all possible combinations. The subset is randomly drawn for each speakers pair. Finally, there were $1,000$ combinations at each SNR.

3) Localization metrics: The main performance metric is derived from the "gross accuracy" [19] defined as the proportion of sources correctly localized within a ± 5 degrees interval. Based on studies on the human accuracy in sound source localization, a "perceptual gross accuracy" is proposed, which accounts for the fact that humans are more accurate for sound sources localized in front of them (a few degrees at 0 degree) and less accurate for sound sources localized on their side (about 10/20 degrees at $\pm$ 90 degree) [42], [43]. In consequence, the proposed "perceptual gross accuracy" (referred as GA%) has a threshold which varies linearly from 5 degrees at 0 degree to 15 degrees at $\pm$ 90 degree. Also, the mean absolute error (referred as MAE) is computed as the mean absolute difference between the real and the estimated azimuth. A single azimuth is determined for each simulation file, and the two performance measures are calculated by comparison of the real and estimated azimuths.

B. Experiment 1: Single Speaker

1) Comparison of localization algorithms: Table II reports the binaural sound source localization performance obtained with the previously described algorithms for the single speaker localization within a diffuse noise, as a function of the SNR.

TABLE II
GROSS ACCURACY (%), DIFFUSE NOISE, ONE SPEAKER. COMPARISON OF THE BINAURAL SOUND SOURCE LOCALIZATION ALGORITHMS.

SNR	-6	0	+6	$+\infty$
binLoc				
GCC-PHAT	6.6	15.9	45.9	100
DUET	6.8	11.0	29.7	98.6
ILD	16.2	45.1	78.4	100
NTF + sep				
NTF + sep + GCC-PHAT	10.0	33.9	67.6	100
NTF + sep + DUET	10.4	37.7	68.1	98.6
NTF + sep + ILD	70.1	94.3	99.6	100
NTF + $\boldsymbol{Q}^S$				
NMF + est $\boldsymbol{Q}^S$	**79.9**	**96.2**	**99.7**	100

The SoA binaural localization algorithms (*binLoc*) are dramatically affected by the presence of noise. The first set of NTF-based algorithms (*NTF + sep*), for which the binaural localization is computed from the source images as separated by NTF, clearly improves the localization performance. In particular, the NTF + sep + ILD algorithm presents a good performance at all SNR, with 99.6% GA at $+6$ dB and 94.3% GA at 0 dB SNR and 70.1% GA at -6 dB. The NTF-based algorithms (*NTF est*, NMF + est $\boldsymbol{Q}^S$) for which the localization is computed from the estimated binaural mixing matrix present the best localization performance, at all SNRs. In particular, the localization obtained from the binaural mixing matrix is clearly better than the one obtained from the source images (NTF+sep+ILD) , and with less complexity: the localization is directly computed from the binaural matrix and does not require to compute the source images.

The localization error distribution has been examined, Figure 6 presents the MAE (Mean Absolute Error) obtained as a function of the true azimuth, with different SNR for the NTF + est $\boldsymbol{Q}^S$ architecture. The error distribution is asymmetric, which is in concordance to the asymmetric nature of the HRTFs. Also, the error is increasing from the front (0 degree) to the sides (+/- 90 degree), which is a well-known issue in sound source localization and in agreement with the human localization ability as reported in [42], [43]. Finally, the figure shows that the localization error is under 5 degrees in average at $+6$ dB and 0 dB, and then increase substantially at -6 dB, especially on the sides.

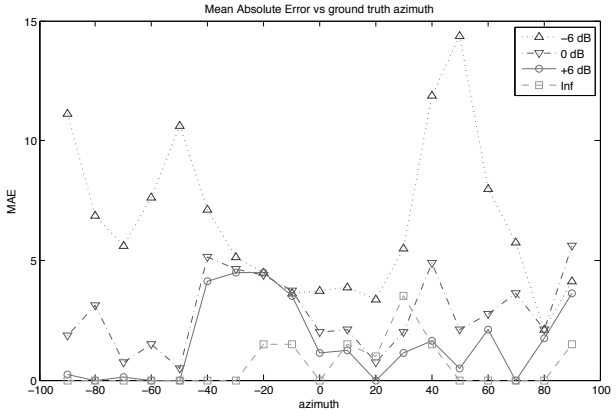


Fig. 6. Mean Absolute Error (deg.). One Speaker. Estimated vs. Real azimuth as a function of the SNR.

2) Comparison with SoA source separation algorithms: The proposed NTF binaural sound source localization algorithm, has been compared by using SoA sound source separation algorithms as presented in [21]. The SoA source separation algorithms comprises: the standard decomposition based on the Multiplicative Updates (MU), and a probabilistic model based on an Expectation-Maximization algorithm (EM) (see *Flexible Audio Source Separation Toolbox* [35]). In Table III, the two approaches are compared on the sound source localization task : from the estimated binaural mixing matrix (MU or EM + $\boldsymbol{Q}^S$) and from the source images as described in Section III-C3 (MU or EM + sep + ILD). For a fair comparison, all algorithms are based on standard NMF decomposition and free mixing matrices for the sound and noise sources. Please note that the MU + sep + ILD algorithm is strictly equivalent as the one previously referred as NTF + sep + ILD. The localization based on the estimated binaural mixing matrix yields the best performance (EM and MU), and once again is better than the localization based on the source images. The localization performance is significantly lower for the EM approach, which confirms previous report that the MU is more robust that the EM to non punctual sources (see [35]).

Table IV compares the source separation algorithms on the sound source separation task. This is done in order to assess whether a better separation of the sound sources leads to a better localization. The performance is measured in terms of SDR (Signal to Distortion Ratio), ISR (source Image to Spatial distortion Ratio), SIR (Source to Interference Ratio) and SAR (Source to Artifacts Ratio) as defined in [44]. On the one hand, the EM version has a better SIR, raising less interference in sound source separation with an acceptable SAR, thus indicating a better separation in the spectral/temporal dimensions. On the other hand, the ISR is higher with the MU approach, indicating a better separation with the MU version in the spatial domain, and consequently a better localization.

TABLE III
GROSS ACCURACY (%), DIFFUSE NOISE, ONE SPEAKER. COMPARISON OF SOA SOURCE SEPARATION ALGORITHMS FOR SOUND SOURCE LOCALIZATION.

SNR	-6	0	+6	$+\infty$
MU + $\boldsymbol{Q}^S$	**79.9**	**96.2**	**99.7**	100
MU + sep + ILD	70.1	94.3	99.6	100
EM + $\boldsymbol{Q}^S$	39.2	76.2	96.7	100
EM + sep + ILD	39.2	76.2	96.7	100

TABLE IV
SDR, ISR, SIR, SAR, DIFFUSE NOISE, ONE SPEAKER. COMPARISON OF SOA SOURCE SEPARATION ALGORITHMS FOR SOUND SOURCE SEPARATION.

SNR	SDR	ISR	SIR	SAR
MU				
-6	-1.0	8.3	-2.8	13.4
0	0.7	9.3	4.1	15.1
+6	0.5	3.3	11.1	17.3
EM				
-6	-1.0	2.5	3.6	5.2
0	0.9	5.2	9.4	8.7
+6	0.4	2.7	16.8	13.1

3) Comparison of NTF variants: Finally, several variants of the *NTF* + $\boldsymbol{Q}^S$ algorithm have been examined in table V, comprising: the nature of the speech dictionary: a standard NMF (NMF), or a source/filter NMF (SF-NMF); the nature of the speakers used for learning the dictionary and for testing: the same speaker in both steps (w. speak.), or two different speakers (w./o speak.) of the *same* genre (male/male, female/female); the nature of the noise source: none (w./o $\boldsymbol{Q}^N$), a noise source present but not localized (cst $\boldsymbol{Q}^N$ = 1, i.e., the noise source is supposed to be in front), and noise source present and localized (est $\boldsymbol{Q}^N$).

TABLE V
GROSS ACCURACY (%), DIFFUSE NOISE, ONE SPEAKER. COMPARISON OF THE NTF-BASED BINAURAL SOUND SOURCE LOCALIZATION VARIANTS.

SNR	-6	0	+6	$+\infty$
NTF + $\boldsymbol{Q}^S$				
NMF w. speak. w. est $\boldsymbol{Q}^N$	79.9	96.2	**99.7**	100
SF-NMF w. speak. w/o $\boldsymbol{Q}^N$	10.1	17.6	39.3	100
SF-NMF w. speak. w. cst $\boldsymbol{Q}^N$	52.1	71.6	85.5	80.5
SF-NMF w. speak. w. est $\boldsymbol{Q}^N$	**86.3**	**96.6**	99.5	100
SF-NMF w/o speak. w. est $\boldsymbol{Q}^N$	85.7	95.9	98.9	100

First, the use of a source/filter model specific to the speech dictionaries improves the localization as compared to a standard NMF. Second, the localization performance is not really affected by the speaker used to construct the speech

dictionaries (SF-NMF w. speak. w. est $\boldsymbol{Q}^N$ vs. SF-NMF w/o speak. w. est $\boldsymbol{Q}^N$). Not surprisingly, prior knowledge of the speaker to be localized improves his localization, but the difference is relatively slight compared to the use of an arbitrary speaker in the training step, with a loss smaller than 1% at all SNRs.

Thirdly, the integration of a noise source and its localization greatly improves the localization in unknown noisy environment. The absence of a noise source model provides a poor localization in the presence of noise. The addition of a noise source model at a fixed position substantially improves the sound source localization. Finally, the free estimation of the noise localization again substantially improves the sound source localization. A close look on the effect of the estimated noise source on the sound source localization performance as a function of the SNR shows that the lower the SNR is, the more the noise source localization helps the speaker localization ($+14.0\%$ GA at $+6$ dB, $+25.0\%$ GA at 0 dB, and $+34.2\%$ GA at -6 dB).

4) Exploiting MHRTF priors: Finally, Table VI compares the localization obtained with fixed binaural mixing matrices (SF-NMF + fixed $\boldsymbol{Q}^S$), as described in Section III-C2, with the localization obtained from the estimated one (SF-NMF + est $\boldsymbol{Q}^S$). In the former, a sparsity constraint on the noise activation matrix $\mathbf{H}^N$ has been added in order to avoid that most of the binaural signal is explained by the free noise sources, thus conducting to wrong localization. Once again, the freely estimated sound source mixing matrix yields to the best performance, while the sparsity constraint on the noise activations substantially improves the localization performance with the fixed MHRTF binaural mixing matrices.

TABLE VI
GROSS ACCURACY (%), DIFFUSE NOISE, ONE SPEAKER. COMPARISON OF FREE VERSUS FIXED BINAURAL MIXING MATRIX.

SNR	-6	0	+6	$+\infty$
SF-NMF + est $\boldsymbol{Q}^S$	**86.3**	**96.6**	**99.5**	**100**
SF-NMF + fixed $\boldsymbol{Q}^S$	70.7	70.9	65.2	50.0
SF-NMF + fixed $\boldsymbol{Q}^S$ + sparse $\mathbf{H}^N$	77.6	88.6	89.5	80.5

The main conclusions of this first experiment are: 1) The proposed NTF-based sound source localization is more robust by far to unknown background noise than SoA binaural sound source localization algorithms; 2) The localization computed from the binaural mixing matrix is better than the one from the binaural source images while requiring less computational cost. 3) The proposed NTF binaural architecture with the source/filter decomposition is better than the localization obtained with SoA sound source separation algorithms (MU/EM) with standard NMF decomposition.

C. Experiment 2: Multiple Speakers

Table VII reports the results obtained for the localization of two speakers within a diffuse noise, as a function of the SNR. In this experiment, one of the two speakers is a male and the other being a female. Only the *NTF* + $\boldsymbol{Q}^S$ algorithm and its variants were retained from the previous experiment for the localization of multiple speakers. As for the previous experiment the same behavior is observed for the localization of multiple speakers: the best localization is obtained with a source/filter model and the estimation and localization of the noise source.

TABLE VII
GROSS ACCURACY (%), DIFFUSE NOISE, TWO SPEAKERS. THE ALGORITHMS ARE THE NTF BASED ALGORITHMS. FOR THE SPEAKERS: THE TIME-FREQUENCY BASES ARE BASED ON THE STANDARD NMF AND THE SOURCE/FILTER MODEL SF-NMF. FOR THE NOISE SOURCES: THE TIME FREQUENCY BASES ARE BASED ON THE STANDARD NMF, THE MIXING MATRIX IS IGNORED (W/O $\boldsymbol{Q}^N$), FIXED TO ONE (W. CST $\boldsymbol{Q}^N$), AND ESTIMATED (W. EST $\boldsymbol{Q}^N$)

SNR	-6	0	+6	$+\infty$
one speaker				
NMF w. est $\boldsymbol{Q}^N$	79.9	96.2	**99.7**	100
SF-NMF w. est $\boldsymbol{Q}^N$	**86.3**	**96.6**	99.5	100
two speakers				
NMF w. est $\boldsymbol{Q}^N$	63.1	**74.9**	76.1	75.2
SF-NMF w/o $\boldsymbol{Q}^N$	47.2	58.4	69.8	87.5
SF-NMF w. cst $\boldsymbol{Q}^N$	59.7	69.8	74.1	74.2
SF-NMF w. est $\boldsymbol{Q}^N$	64.7	72.0	75.3	79.4
SF-NMF w. est $\boldsymbol{Q}^N$ w perm.	**68.7**	74.5	**76.5**	**80.0**

The localization performance is around 60-75% for the *NTF* + $\boldsymbol{Q}^S$ (SF-NMF w. est $\boldsymbol{Q}^N$) and increases to 70-80% by ignoring localization errors due to a speaker identification error (SF-NMF w. est $\boldsymbol{Q}^N$ w. perm), i.e. permutation of the male and the female speakers. This constitutes encouraging results for the NTF-based binaural localization of multiple sound sources, as compared to the results reported in [25], [26], especially considering the complexity of the localization of two speakers in the presence of realistic background sounds. However, there is a clear loss in localization performance for the localization of two speakers (around 30%) as compared to the one of a single speaker. This loss can be explained by three causes: 1) a mismatch between the speakers, 2) a degradation due to the presence of a noise, and 3) a wrong separation between the speakers which would in its turn affects the localization, by compensating the wrong speaker reconstruction by the mixing filters. The effect of the speaker identification error remains relatively small (from 1.2% GA at $+6$dB to 4% GA at -6dB). The effect of the noise is certainly present, but does not explain a large part of the loss, since the localization performance drops by 20-25% even in ideal conditions (with no noise, SNR=$+\infty$). In consequence, the most likely cause of the localization error is a wrong separation between the speakers. These observations show the current limitations and raises the challenges for further research on the binaural localization of multiple sound sources in realistic environments.

V. CONCLUSION

A non-negative factorization of audio signals for the binaural localization of multiple sound sources has been presented in this paper. In this proposed formulation, each sound source is represented by mean of spectral dictionaries, temporal activation, and its distribution within each channel (here, left and right ears). The binaural mixing matrix can be used directly to

estimate the ILD corresponding to each source and thus their localization. Also, spectral dictionaries can be constructed in advance for some sources of interest to be localized, while some other background sources can remain unknown and estimated on the fly. A couple of experiments conducted on simulated but realistic sound environments consisting of one or two speakers mixed within a diffuse noise shows the efficiency of the proposed method and its robustness to unknown background sounds. Further research will focus on the exploitation of the phase information in the mixing matrices in the NTF algorithm in order to exploit all binaural cues (ILD and IPD), the localization of more than two sound sources, and the experimentation within real robotic and sound environments.

REFERENCES

[1] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic Event Detection in Real-Life Recordings," in *European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, 2010, p. 1267–1271.

[2] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound Event Detection in Multisource Environments using Source Separation," in *Workshop on Machine Listening in Multisource Environments (CHiME)*, Florence, Italy, 2011, pp. 69–72.

[3] C. Cotton and D. Ellis, "Spectral vs. Spectro-Temporal Features for Acoustic Event Classification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Lyon, France, 2011, pp. 69–72.

[4] K. J. Piczak, "Environmental Sound Classification with Convolutional Neural Networks," in *International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, USA, 2015, pp. 1–6.

[5] P. Paatero, "Least squares formulation of robust non-negative factor analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 37, no. 1, pp. 23–35, 1997.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[7] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic Scene Classification with Matrix Factorization for Unsupervised Feature Learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai , China, 2016, pp. 6445–6449.

[8] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic Event Detection Method using Semi-Supervised Non-Negative Matrix Factorization with a Mixture of Local Dictionaries," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, Budapest , Hungary, 2016.

[9] C. Knapp and G. Carter, "The Generalized Cross-Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, p. 320–327, 1976.

[10] P. Aarabi, "Self-localizing Dynamic Microphone Arrays," *IEEE transactions on systems, man, and cybernetics*, vol. 32, no. 4, pp. 474–484, 2002.

[11] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Elsevier Signal Processing*, vol. 92, pp. 1950–1960, 2012.

[12] A. Jourjine, S. Rickard, , and Ö. Yilmaz, "Blind Separation of Disjoint Orthogonal Signals: Demixing n Sources From 2 Mixtures," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000, p. 2985–2988.

[13] S. Rickard and F. Dietrich, "DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET," in *IEEE Workshop on Statistical Signal and Array Processing*, Pocono Manor , USA, 2000.

[14] M. I. Mandel and D. P. W. Ellis, "EM Localization and Separation Using Interaural Level and Phase Cues," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, USA, 2007, p. 275–278.

[15] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based Expectation-Maximization Source Separation and Localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[16] H. Viste and G. Evangelista, "Binaural Source Localization," in *Digital Audio Effects (DAFx) Conference*, Naples, Italy, 2004, p. 145–150.

[17] M. Raspaud, H. Viste, and G. Evangelista, "Binaural Source Localization by Joint Estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2010.

[18] A. Deleforge, F. Forbes, and R. Horaud, "Variational EM for Binaural Sound- Source Separation and Localization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 76–80.

[19] J. Woodruff and D. Wang, "Binaural Localization of Multiple Sources in Reverberant and Noisy Environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503 – 1512, 2012.

[20] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative Tensor Factorisation for Sound Source Separation," in *Irish Signals and Systems Conference (ISSC)*, Dublin, Ireland, September 2005.

[21] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550 – 563, 2010.

[22] P. R. Mitchell and I. A. Essa, "Estimating the spatial position of spectral components in audio," in *Independent Component Analysis and Blind Signal Separation (ICA)*, Dublin, Ireland, 2006, pp. 666–673.

[23] S. Lee, S. H. Park, and K.-M. Sung, "Beamspace-Domain Multichannel Nonnegative Matrix Factorization for Audio Source Separation," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 43–46, 2012.

[24] J. Traa, P. Smaragdis, N. D. Stein, and D. Wingate, "Directional NMF for Joint Source Localization and Separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2015.

[25] F. Keyrouz, W. Maier, and K. Diepold, "Robotic Binaural Localization and Separation of More than Two Concurrent Sound Sources," in *International Symposium on Signal Processing and Its Applications (ISSPA)*, Feb 2007, pp. 1–4.

[26] K. Youssef, K. Itoyama, and K. Yoshii, "Identification and Localization of One or Two Concurrent Speakers in a Binaural Robotic Context," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Hong Kong, China, Oct 2015, pp. 407–412.

[27] H. G. Okuno and K. Nakadai, "Robot Audition: Its Rise and Perspectives," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, pp. 5610–5614.

[28] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *Journal of the Acoustic Society of America*, vol. 97, no. 6, p. 3907–3908, 1995.

[29] N. Roman, D. Wang, and G. J. Brown, "Speech Segregation based on Sound Localization," *Journal of the Acoustic Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[30] C. Viña, S. Argentieri, and M. Rébillat, "A spherical cross-channel algorithm for binaural sound localization," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 2921–2926.

[31] M. Raspaud, H. Viste, and G. Evangelista, "Binaural Source Localization by Joint Estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, Jan 2010.

[32] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[33] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[34] D. Bouvier, N. Obin, M. Liuni, and A. Roebel, "A Source/Filter Model with Adaptive Constraints for NMF-based Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 131–135.

[35] A. Ozerov, E. Vincent, and F. Bimbot, "A General Flexible Framework for the Handling of Prior Information in Audio Source Separation," Research Report, 2010.

[36] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel Extensions of Non-negative Matrix Factorization with Complex-valued Data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971 – 982, 2013.

[37] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[38] T. Carpentier, M. Noisternig, and O. Warusfel, "Twenty Years of Ircam Spat: Looking Back, Looking Forward," in *International Computer Music Conference (ICMC)*, Denton, USA, 2015, pp. 270–277.

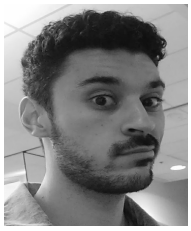
[39] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.
[40] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Interspeech*, 2010, pp. 3110–3113.
[41] C. Viña, S. Argentieri, and M. Rébillat, "A Spherical Cross-channel Algorithm for Binaural Sound Localization ," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Tokyo, Japan, 2011, pp. 2921–2926.
[42] S. S. Stevens and E. B. Newman, "The Localization of Actual Sources of Sound," *American Journal of Psychology*, vol. 21, p. 297–306, 1936.
[43] R. A. Butler, "The Bandwidth Effect on Monaural and Binaural Localization," *Journal of Hearing Research*, vol. 21, p. 67–73, 1986.
[44] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, London, United Kingdom, September 2007, pp. 552–559.

Elie Laurent Benaroya is a Research Fellow in the Image, Data, Signal department at TélécomParisTech, Paris, France. He received his engineer degree from Ecole Centrale Paris in 1997 and his M.Sc. in signal processing and machine learning from Ecole Normale Supérieure (ENS Cachan, France) in 1999. In 2003, he received his Ph.D. in the field of monaural audio source separation from IRISA (INRIA/CNRS/Université Rennes I). He worked from 2004 to 2007 in a French start-up, Audionamix, specialized in audio source separation products. He has been working as a Research Fellow since 2008 in several prestigious French institutes, in various research areas such as silent speech recognition, sound source separation, music information retrieval and machine learning. His research interests concern audio scene analysis in general and audio/speech source separation, probabilistic modeling and machine learning in particular.

Nicolas Obin (M'13) is Associate Professor at Sorbonne Université and the Institute for Research and Coordination in Acoustics & Music (IRCAM). He received a MSc degree in Acoustics, Signal Processing, and Computer science applied to Music and a Ph.D. in Computer Sciences from the University of Pierre and Marie Curie (UPMC) in 2006 and 2011. In 2006, he was a visiting researcher at CNMAT (Center for New Music and Audio Technologies) at the University of California, Berkeley. He received his Ph.D. on the modeling of speech prosody and speaking style for text-to-speech synthesis from IRCAM/UPMC, for which he was awarded the best French Ph.D. thesis in computational sciences from "La Fondation Des Treilles" in 2011. His primary research interests include speech/music processing and machine learning with application to voice conversion, speech synthesis, singing voice, sound synthesis and machine listening. He is also deeply involved into the promotion of audio technologies for creation, arts, and culture.

Marco Liuni received the Ph.D. degree in mathematics from Florence University and the Ph.D. degree in signal processing from UPMC Paris 6 University in 2012. He is currently postdoctoral fellow and a computer music designer at IRCAM. His interests lie in adaptive/intelligent sound processing and its link with human cognition, source separation, and real-time computer music.

Axel Roebel (M'08) received the Diploma in electrical engineering from Hanover University in 1990 and the Ph.D. degree (summa cum laude) in computer science from the Technical University of Berlin in 1993. In 1994 he joined the German National Research Center for Information Technology (GMD-First) in Berlin where he continued his research on adaptive modeling of time series of nonlinear dynamical systems. In 1996 he became assistant professor for digital signal processing in the communication science department of the Technical University of Berlin. In 2000 he was visiting researcher at CCRMA Standford University, where he worked on adaptive sinusoidal modeling. In the same year he joined the IRCAM to work on sound analysis, synthesis and transformation algorithms. In summer 2006 he was Edgar-Varese guest professor for computer music at the Electronic studio of the Technical University of Berlin and currently he is head of the Sound Analysis and Synthesis team at IRCAM. His current research interests are related to music and speech signal analysis and transformation.



Wilson Raumel received a Master's Degree in engineering sciences in 2016 with a specialization in sound and image processing, intelligent systems, and robotics at the University of Pierre and Marie Curie - Sorbonne Universités. He's is currently attending a Master's Degree at the University of Pierre and Marie Curie in innovation management in order to create and/or participate to a startup in information technologies. His primary interests cover audio, music, and entrepreneurship.

Sylvain Argentieri obtained the highest teaching diploma in France (Agrégation externe) in Electronical Science in the Ecole Normale Supérieure (Cachan) in 2002. He then received his Ph.D. in Computer Science from the Paul Sabatier University (Toulouse) in 2006. He is now Associate Professor in the "AMAC" group in the Institute for Intelligent Systems and Robotics (ISIR), Sorbonne Université (formerly known as the Pierre et Marie Curie University) since 2008. His research interests relate to artificial audition in robotics, especially in the binaural context, for sound source localization, speaker recognition and human–robot interaction. He is also interested in active approaches to multimodal perception and sensorimotor integration, in relation with the sensorimotor contingencies theory.