TOWARDS A GENERALIZATION OF RELATIVE TRANSFER FUNCTIONS TO MORE THAN ONE SOURCE

Antoine Deleforge[*], *Sharon Gannot*[†], *Walter Kellermann*[*]

[*]University of Erlangen-Nuremberg, Germany
[†]Bar-Ilan University, Israel

ABSTRACT

We propose a natural way to generalize relative transfer functions (RTFs) to more than one source. We first prove that such a generalization is not possible using a single multichannel spectro-temporal observation, regardless of the number of microphones. We then introduce a new transform for multichannel multi-frame spectrograms, *i.e.*, containing several channels and time frames in each time-frequency bin. This transform allows a natural generalization which satisfies the three key properties of RTFs, namely, they can be directly estimated from observed signals, they capture spatial properties of the sources and they do not depend on emitted signals. Through simulated experiments, we show how this new method can localize multiple simultaneously active sound sources using short spectro-temporal windows, without relying on source separation.

***Index Terms*—** Relative Transfer Function, Grassmannian manifolds, Plücker Embedding, Multiple sound sources localization

1. INTRODUCTION

When sound propagates from an emitter to a receiver in a natural environment, objects along its path (*e.g.*, a human or robot head, walls...) lead to reflections and reverberation. This is commonly modeled as a linear filtering and described by the convolution of the emitted signal with a so called *room impulse response* (RIR). For a given room, the latter only depends on the source's spatial properties (position, orientation, directivity, diffuseness, etc.) and not on the emitted signal. The frequency domain counterparts of RIRs are *acoustic transfer functions* (ATFs). Knowledge of the ATFs involved in an acoustic setup is useful in many audio signal processing applications, *e.g.*, blind source separation [1], beamforming [2], sound source localization [3–5], acoustic echo cancellation [6].

Most existing methods to estimate ATFs rely on the synchronized emitted and received signals. However, the emitted signals are often not available, rendering the estimation of ATFs impossible without additional restrictive assumptions. For this reason, *relative transfer functions* are often considered [7]. These also capture source spatial properties and do not depend on the emitted signal, with the advantage that they can be reliably and robustly estimated directly from an observed multichannel signal [8, 9]. They are defined as a *normalized* version of ATFs, *i.e.*, the ATF at a given microphone is divided by a linear combination of the ATFs to other microphones, *e.g.*, the ATF of a reference microphone. In the case of $M = 2$ microphones, the log-magnitude and phase of RTFs are referred to as *interaural level* and *phase differences*, respectively, in the binaural hearing literature [10, 11]. Recently, supervised sound source localization methods making use of a training set of interaural cues [5] or of RTFs [4] have been proposed.

In this paper, we theoretically investigate the possibility of generalizing RTFs to more than one source. Such generalizations should preserve the three key properties of RTFs, namely, they can be directly estimated from observed signals, they capture spatial properties of the sources and they do not depend on the emitted signals. We first state and prove a theorem showing that such a generalization is not possible if a single multichannel spectro-temporal observation is used. We then consider the case of multiple time observations, and propose a new transformation for multichannel, multi-frame spectrograms, *i.e.*, containing several multichannel time frames in each time-frequency bin. This transformation builds on the Plücker embedding method for Grassmannian manifolds. We show that it yields a natural generalization of RTFs to multiple sources, when there are less sources than microphones. Through simulated experiments, we show how this method could be applied to the localization of multiple simultaneously active sound sources using short spectro-temporal windows, without having to separate them.

2. GENERALIZING RTFS

2.1. Single-source case and RTF properties

Let us consider a sound source emitting the spectrogram $\{s_{ft}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F\times T}$ recorded by an M-microphone ar-

The research leading to these results has received funding from the European Unions Seventh Framework Programme (FP7/2007-2 013) under grant agreement n° 609465 (project EARS).

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

ray, where F and T are the number of frequency bands and the number of time frames, respectively. Under noise-free, finite convolutive filtering assumptions and for long enough time frames, the multichannel observation $\boldsymbol{x}_{ft} = [x_{ft,1}, \dots, x_{ft,M}]^\top \in \mathbb{C}^M$ received by M microphones at frequency-time (f,t) is given by

$$\boldsymbol{x}_{ft} = \boldsymbol{a}_f s_{ft} \tag{1}$$

where $\boldsymbol{a}_f = [a_{f,1}, \dots, a_{f,M}] \in \mathbb{C}^M$ comprises the acoustic transfer functions from the source to the M microphones at frequency f. For a given microphone setup in a given room, $\boldsymbol{a}_f$ solely depends on the source's *spatial properties*. Therefore, (1) nicely decomposes the recorded signal into a component $\boldsymbol{a}_f$ that only captures spatial properties and a component s_{ft} that only captures the source content at (f,t).

If the emitted signal s_{ft} is unknown, unambiguously recovering $\boldsymbol{a}_f$ from observation $\boldsymbol{x}_{ft}$ is impossible, without further assumptions. However, the specific structure of Eq. 1 offers an attractive way to circumvent this. Let ν be a *normalizing function*, which divides an input vector by a linear combination of its entries, *e.g.*, the first entry. It is then easy to check that $\nu(\boldsymbol{x}_{ft}) = \nu(\boldsymbol{a}_f)$ for all $\boldsymbol{x}_{ft} \in \mathcal{I}$, where $\mathcal{I} \subseteq \mathbb{C}^M$ is the nonzero locus of the linear combination. In other words, the signal term cancels out and $\boldsymbol{r}_{ft} = \nu(\boldsymbol{x}_{ft})$, when defined, captures only the spatial properties of the source. In the signal processing literature, $\boldsymbol{r}_f$ is referred to as a *relative transfer function* (RTF) [7]. In summary, relative transfer functions possess three key desirable properties:

(I) They can be directly estimated from observed signals

(II) They capture spatial properties of the sound source

(III) They do not depend on the emitted signal

Mathematically, these three properties are verified if and only if there exists a *non-constant* function $g\colon \mathcal{I} \to \Omega$ and a function h such that $(1) \implies g(\boldsymbol{x}_{ft}) = h(\boldsymbol{a}_f)$ for all $\boldsymbol{x}_{ft} \in \mathcal{I}$, where Ω is an arbitrary set and $\mathcal{I} \subseteq \mathbb{C}^M/\{\boldsymbol{0}\}$.

2.2. Instantaneous multiple-source case

In the case of K sound sources emitting spectrograms $\{s_{ft,k}\}_{f=1,t=1}^{F,T}$ for $k = 1 \dots K$, model (1) becomes:

$$\boldsymbol{x}_{ft} = \sum_{k=1}^{K} \boldsymbol{a}_{f,k} s_{ft,k} = \mathbf{A}_{f,K} \boldsymbol{s}_{ft} \tag{2}$$

where $\boldsymbol{s}_{ft} = [s_{ft,1}, \dots, s_{ft,K}]^\top \in \mathbb{C}^K$ is the vector of emitted signals and $\mathbf{A}_{f,K} = [\boldsymbol{a}_{f,1}, \dots, \boldsymbol{a}_{f,K}] \in \mathbb{C}^{M\times K}$ comprises the K acoustic transfer functions capturing the sources' spatial properties. An interesting question is: *can we generalize relative transfer functions to more than one source, while preserving properties (I), (II) and (III)*? In other words, is there a non-constant function $g\colon \mathcal{I} \to \Omega$ and a function h such that $g(\boldsymbol{x}_{ft}) = h(\mathbf{A}_{f,K})$ for all $\boldsymbol{x}_{ft} \in \mathcal{I}$? In this section, we prove that the answer is "no" through the following theorem:

Theorem 1 *Let $\mathcal{I}$ be a subset of $\mathbb{C}^M/\{\boldsymbol{0}\}$, Ω an arbitrary set, $g\colon \mathcal{I} \to \Omega$ and $h\colon \mathbb{C}^{M\times K} \to \Omega$ two functions and $K > 1$. If for all $\mathbf{A} \in \mathbb{C}^{M\times K}$ and for all $\boldsymbol{s} \in \mathbb{C}^K$ with $\mathbf{A}\boldsymbol{s} \in \mathcal{I}$ we have $g(\mathbf{A}\boldsymbol{s}) = h(\mathbf{A})$, then g is constant.*

In other words, the only possible multiple-source instantaneous generalizations of RTFs are constant, which violates property *(II)*.

Proof of Theorem 1:
Let $g\colon \mathcal{I} \to \Omega$ and $h\colon \mathbb{C}^{M\times K} \to \Omega$ be two functions such that for all $\mathbf{A} \in \mathbb{C}^{M\times K}$ and for all $\boldsymbol{s} \in \mathbb{C}^K$ with $\mathbf{A}\boldsymbol{s} \in \mathcal{I}$ we have $g(\mathbf{A}\boldsymbol{s}) = h(\mathbf{A})$.

• Case $K \geq M$: Let $\mathbf{A} \in \mathbb{C}^{M\times K}$ be a fixed matrix with M linearly independent columns. Then, for all $\boldsymbol{x} \in \mathcal{I}$, we have $\boldsymbol{x} = \mathbf{A}\boldsymbol{s}$ with $\boldsymbol{s} = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}\boldsymbol{x}$. By definition of g and h, we thus have $g(\boldsymbol{x}) = g(\mathbf{A}\boldsymbol{s}) = h(\mathbf{A})$ for all $\boldsymbol{x} \in \mathcal{I}$. $h(\mathbf{A})$ does not depend on $\boldsymbol{x}$. Therefore, g is constant.

• Case $K < M$: Let $\mathbf{A} \in \mathbb{C}^{M\times K}$ be a fixed matrix with K linearly independent columns. Let $E_{\mathbf{A}}$ be the column space of $\mathbf{A}$, *i.e.*, the K-dimensional vector subspace of $\mathbb{C}^M$ defined by $E_{\mathbf{A}} = \{\mathbf{A}\boldsymbol{s}; \boldsymbol{s} \in \mathbb{C}^K\}$. We now prove that $g(\boldsymbol{x}) = h(\mathbf{A})$ for all $\boldsymbol{x} \in \mathcal{I}$:

- If $\boldsymbol{x} \in E_{\mathbf{A}}$, then by definition of $E_{\mathbf{A}}$ there is $\boldsymbol{s}$ such that $\boldsymbol{x} = \mathbf{A}\boldsymbol{s}$, and thus $g(\boldsymbol{x}) = g(\mathbf{A}\boldsymbol{s}) = h(\mathbf{A})$.
- If $\boldsymbol{x} \notin E_{\mathbf{A}}$, let $\boldsymbol{x}' \in E_{\mathbf{A}} \cap \mathcal{I}$. Then $\boldsymbol{x}$ and $\boldsymbol{x}'$ are linearly independent. Let $\mathbf{A}' = [\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{a}'_3, \dots, \boldsymbol{a}'_K] \in \mathbb{C}^{M\times K}$ have K linearly independent columns (note that this is only possible because $K > 1$). Let $\boldsymbol{s} = [1, 0, \dots, 0]^\top$ and $\boldsymbol{s}' = [0, 1, 0, \dots, 0]^\top$, so that $\boldsymbol{x} = \mathbf{A}'\boldsymbol{s}$ and $\boldsymbol{x}' = \mathbf{A}'\boldsymbol{s}'$. By definition of g and h, we have $g(\boldsymbol{x}) = g(\mathbf{A}'\boldsymbol{s}) = h(\mathbf{A}') = g(\mathbf{A}'\boldsymbol{s}') = g(\boldsymbol{x}')$. Since $\boldsymbol{x}' \in E_{\mathbf{A}}$, we have $g(\boldsymbol{x}') = h(\mathbf{A})$ and thus $g(\boldsymbol{x}) = h(\mathbf{A})$.

Thus, $g(\boldsymbol{x}) = h(\mathbf{A})$ for all $\boldsymbol{x} \in \mathcal{I}$, and $h(\mathbf{A})$ does not depend on $\boldsymbol{x}$. Therefore, g is constant. ■

2.3. Multiple-frame, multiple-source case

In this section we overcome the non-existence of an instantaneous generalization of RTFs by proposing a *multi-frame generalization*. More precisely, we consider the case where K rather than one observations are available along the time axis. Using the following notations:

$$\mathbf{X}_{ft,K} = [\boldsymbol{x}_{ft}, \dots, \boldsymbol{x}_{ft+K-1}] \in \mathbb{C}^{M\times K}, \tag{3}$$

$$\mathbf{S}_{ft,K} = [\boldsymbol{s}_{ft}, \dots, \boldsymbol{s}_{ft+K-1}] \in \mathbb{C}^{K\times K}, \tag{4}$$

we obtain a multiframe version of (2) for the time segment $[t \dots t+K-1]$:

$$\mathbf{X}_{ft,K} = \mathbf{A}_{f,K}\mathbf{S}_{ft,K}. \tag{5}$$

We will refer to $\{\mathbf{X}_{ft,K}\}_{f=1,t=1}^{F,T}$ as a *multichannel, K-frame spectrogram*. Each time-frequency bin contains an $M \times K$ complex matrix. The question then becomes: *is there a non-constant function g and a function h such that $g(\mathbf{X}_{ft,K}) = h(\mathbf{A}_{f,K})$ for all $\mathbf{A}_{f,K} \in \mathbb{C}^{M\times K}$ and $\mathbf{S}_{ft,K} \in \mathbb{C}^{K\times K}$?* From now and until the end of this paper, we will assume that the number of sources is strictly lower than the number of microphones, *i.e.* $K < M$. Under this assumption, an interesting candidate solution is $g = h = \mathrm{span}$, where $\mathrm{span} : \mathbb{C}^{M\times K} \to \mathrm{Gr}(K, \mathbb{C}^M)$ is the function associating a matrix to its column space. $\mathrm{Gr}(K, \mathbb{C}^M)$ is called a *Grassmannian manifold*: elements of this set are K-dimensional linear subspaces of $\mathbb{C}^M$ [12, 13]. Assuming that the square matrix $\mathbf{S}_{ft,K}$ has linearly independent columns (this assumption is further discussed in Section 2.6), it acts as a change of basis from the column space of $\mathbf{A}_{f,K}$ to the column space of $\mathbf{X}_{ft,K}$ in equation (5). Therefore, $\mathrm{span}(\mathbf{X}_{ft,K}) = \mathrm{span}(\mathbf{A}_{f,K})$ does not depend on $\mathbf{S}_{ft,K}$, and span possesses the desired properties to generalize RTFs.

However, the output values of span are not vectors but vector subspaces. These cannot be manipulated numerically. We thus need a way to map the Grassmannian manifold $\mathrm{Gr}(K, \mathbb{C}^M)$ to a numerical space. This is possible using a method known as *Plücker embedding* [12]. The method was first introduced in the case $K = 2$ and $M = 4$ by Julius Plücker in 1865, and later generalized to any K and M values by Hermann Grassmann. Building on this, we propose a new transform for multichannel, multi-frame spectrograms. This transform applied to equation (5) will yield an equation of the form (1), allowing a generalization of RTFs to multiple sources. We shall name it the *Plücker spectrogram transform* after the work of Julius Plücker.

2.4. The Plücker spectrogram transform

Let $\{\mathbf{X}_{ft,K}\}_{f=1,t=1}^{F,T}$ be an M-channel K-frame spectrogram. We denote by $\mathbf{X}_{ft,K|i_1,i_2,\dots,i_K}$ the $K \times K$ matrix formed by the K rows of $\mathbf{X}_{ft,K}$ with indexes $i_1, i_2, \dots, i_K$. Let $\xi(1), \dots, \xi(L)$ be the lexicographically-ordered list of cardinal-K sublists of $\{1, \dots, M\}$ with $L = \binom{M}{K}$. We define the *Plücker spectrogram transform of order K* as follows:

$$\mathfrak{p}_K(\mathbf{X}_{ft,K}) = \frac{1}{K!}\begin{pmatrix} \det(\mathbf{X}_{ft,K|\xi(1)}) \\ \det(\mathbf{X}_{ft,K|\xi(2)}) \\ \vdots \\ \det(\mathbf{X}_{ft,K|\xi(L)}) \end{pmatrix} \in \mathbb{C}^L. \tag{6}$$

This transform applied to (5) yields the following remarkable identity:

$$\mathfrak{p}_K(\mathbf{X}_{ft,K}) = \mathfrak{p}_K(\mathbf{A}_{f,K})\det(\mathbf{S}_{ft,K}). \tag{7}$$

This follows from the determinant property $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$ for square matrices $\mathbf{A}$ and $\mathbf{B}$ of equal sizes. Interestingly, (7) has the same form as equation (1). In other words, the Plücker spectrogram transform changes an M-microphone observation of K sources into an $\binom{M}{K}$-microphone observation of a single (compound) source. As a consequence, we have:

$$\boldsymbol{r}_{f,K} = \nu(\mathfrak{p}_K(\mathbf{X}_{ft,K})) = \nu(\mathfrak{p}_K(\mathbf{A}_{f,K})). \tag{8}$$

Therefore, $\boldsymbol{r}_{f,K}$ is a suitable generalization of RTFs to K sources and M microphones ($K < M$) using multiframe spectrograms. Namely, it verifies properties *(I)*, (*II*) and (*III*), and for $K = 1$, the RTF definition given in Section 2.1 is exactly recovered.

2.5. Relation to subspace methods

The proposed approach shares a lot of similarities with the so-called *subspace methods* for sound source localization. A well-known example is the method MUSIC, which stands for MUltiple SIgnal Classification, [14, 15]. MUSIC starts by computing the covariance matrix of a multichannel signal in a given frequency band. An eigenvalue decomposition of this matrix is then performed, allowing to identify the *signal subspace*, spanned by the principal eigenvectors, and the orthogonal *noise subspace*, spanned by the remaining eigenvectors. As showed in Section 2.3, the signal subspace corresponds to the space spanned by the ATF, or equivalently the RTF vectors associated to the emitting sources, *i.e.*, $\mathrm{span}(\mathbf{A}_{f,K})$. In contrast, RTF vectors are orthogonal to the noise subspace. Therefore, sound source directions are those whose associated RTF vectors have minimal projections onto the noise subspace. They are usually estimated by finding the smallest projections of a predefined set of RTF vectors.

Alternatively, in equation (8), we introduce a new vector $\boldsymbol{r}_{f,K}$ which *uniquely characterizes* the signal subspace $\mathrm{span}(\mathbf{A}_{f,K})$, using a minimal number of observations. This vector can thus be directly mapped to the spatial properties of all sources, provided that the associated mapping function is known. This mapping may either be directly obtained from a sound propagation model or learned from a predefined set of RTF vectors, as demonstrated in Section 3. An intrinsic difference between this approach and MUSIC is that it does not require the estimation and decomposition of covariance matrices. On the other hand, it requires a mapping from generalized RTFs to multiple-source spatial characteristics, while MUSIC only requires single-source mappings.

2.6. Conditions of applicability and properties

Assuming that the normalizing function ν divides a vector by, *e.g.*, its first entry, (8) is only valid if $\det(\mathbf{X}_{ft,K|\xi(1)}) \neq 0$.

Using (6), (7) and properties of the determinant, it follows that such singularity only occurs in the following situations:

- If one or more sources are completely silent in all K time frames $(t \dots t+K-1)$ at frequency f.
- If two or more sources are perfectly correlated over the segment, *i.e.*, their absolute normalized cross-correlation is 1.
- If two or more sources have *similar* spatial properties, *i.e.* $\boldsymbol{a}_{f,k} = \alpha \boldsymbol{a}_{f,l}$ for some $\alpha \in \mathbb{C}, k \neq l$. This may occur if, *e.g.*, they have identical directions in the free-field case.
- If the K transfer functions and emitted signals are such that observations are linearly dependent, by coincidence.

Let us define audio sources as objects emitting distinguishable sounds from distinguishable locations. Then, the first three cases may be interpreted as a violation of the assumption that there are K sources. The fourth case is harder to interpret, but it has a zero probability of occurrence assuming that distinguishable transfer functions and signals are mutually statistically independent. In other words, the proposed generalization of RTF is sound if the assumed number of sources K is correct. If the actual number of sources P at (f,t) is less than K then $\mathfrak{p}_K(\mathbf{X}_{ft,K}) = 0$. If $P > K$, the desirable properties are no longer preserved. A straightforward way to determine P is to note that:

$$P = \operatorname{rank}(\mathbf{X}_{ft,K}) \text{ for } K > P. \tag{9}$$

If $P < M$, P can thus be deduced by successively calculating $\operatorname{rank}(\mathbf{X}_{ft,K})$ for $K = 1 \dots M-1$.

3. SIMULATED EXPERIMENTS

We test the potential of the proposed generalization of RTF for multiple sound-source localization (SSL). In what follows, spectrograms are computed on signals sampled at 8,000 kHz using 32 ms sliding windows with 50% overlap. This results in $F = 128$ positive frequencies and $T = 64$ time frames per second of signal. We use a dataset of *head-related transfer functions* (HRTFs) for the humanoid robot NAO. These HRTFs are simulated using a 3D model of the head in an anechoic environment and the boundary element method, as done in [16]. Corresponding impulse responses have a maximal length of 10ms. The subset $\mathcal{H}$ used contains $N = 21$ HRTFs $\{\boldsymbol{a}_f(\boldsymbol{\theta}_n)\}_{f=1,n=1}^{F,N} \subset \mathbb{C}^M$ for the $M = 4$ microphones placed on the head. Here $\Theta = \{\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_N\}$ is a set of source directions with azimuth and elevations randomly picked in $[-180°, 180°]$ and $[-10°, 10°]$ respectively. From this dataset, the following *generalized RTF* (GRTF) training sets are generated, for K=1 to 3:

$$\begin{aligned}\mathcal{R}_K = \big\{&\nu(\mathfrak{p}_K([\boldsymbol{a}_f(\boldsymbol{\theta}_1),\dots,\boldsymbol{a}_f(\boldsymbol{\theta}_K)]));\\ &\boldsymbol{\theta}_1 < \dots < \boldsymbol{\theta}_K \in \Theta, f = 1 \dots F\big\}\end{aligned}$$

where the cardinality of $\mathcal{R}_K$ is $F\binom{N}{K}$. We then simulate all possible M-microphone mixtures of one to three white-noise sources coming from distinct directions in Θ, by convolving random signals of one second duration with the HRTFs in $\mathcal{H}$. The minimum distance between distinct sources is 1° in azimuth and 3° in elevation. These mixtures are perturbed by additive Gaussian noise with 10 dB or 50 dB *signal-to-noise ratios* (SNRs). The Plücker spectrogram transform of order K (6) is then applied to all individual K-frame time segments of all these mixtures, where K is the number of sources, assumed known. The F GRTFs associated with the F frequency bins at each segment are concatenated and compared to those of the corresponding training set $\mathcal{R}_K$, in terms of Euclidean distance. The set of K directions minimizing this distance gives the estimated sound source directions. For $K = 1$, 2 and 3, this respectively corresponds to approximately $1,300$, $26,000$ and $250,000$ localization tasks using time segments of length 32ms, 48ms and 64ms. The mean computational times per source per second of signal where respectively 81ms, 87ms and 436ms using our Matlab implementation on a conventional PC. Mean absolute azimuth localization errors obtained with this procedure are summarized in Table 1 (GRTF).

Table 1. Mean absolute azimuth localization error using generalized RTFs on mixtures of 1 to 3 sources, with 10 or 50 dB signal-to-noise ratios.

Number of sources	1	2	3
GRTF (SNR=50 dB)	0.04°	0.68°	1.45°
GRTF (SNR=10 dB)	10.9°	17.5°	27.4°

The results confirm that the proposed generalization of RTF captures spatial properties of sources under low noise level (50 dB SNR). However, performance is severely degraded for higher noise levels (10 dB SNR). While these results are only preliminary, they reveal two intrinsic benefits of the proposed approach. First, it can localize K simultaneous sound sources using only K spectrogram time frames. For $K = 3$ and 50 dB SNR, 91% of the 250,000 individual sources were perfectly localized using GRTFs on 64ms segments. This is impossible using methods such as MUSIC [15], where at least M and typically more time frames are required to reliably estimate spatial covariance matrices. Second, the K sound sources are jointly localized without using source separation, even though their spectra are strongly overlapping (white noise). This makes the method intrinsically efficient computationally, and contrasts with many existing multiple sound source localization methods, which rely on source separation [5, 17, 18]. These two features put forward GRTFs as a promising tool to efficiently localize multiple sound sources using short time windows. This ability may turn out to be critical, *e.g.*, in realistic human-robot interaction scenarios where sound sources may be fast moving and computational resources are limited.

4. CONCLUSION

We proposed a natural way of generalizing relative transfer functions to K sources using K spectro-temporal observations, where K is lower than the number of microphones. To the best of the authors' knowledge, this is the first study of this kind in signal processing. This work is mostly preliminary and theoretical. In the future, we plan an in-depth theoretical and empirical study of the noisy case, and an extension to natural sounds with sparse spectrograms such as speech. Moreover, several leads will be investigated to improve robustness to noise, *e.g.*, estimating the number of sources, combining Plücker transforms of different orders and weighting time-frequency observations. Finally, the possibility of learning the mapping function from GRTFs to source directions will be investigated, following [5]. This would bypass the need for a comprehensive training set containing all possible combination of source positions.

REFERENCES

[1] Lucas Parra and Clay Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[2] Sofiene Affes and Yves Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 425–437, 1997.

[3] Tsvi G. Dvorkind and Sharon Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[4] Bracha Laufer, Ronen Talmon, and Sharon Gannot, "Relative transfer function modeling for supervised source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013*. IEEE, 2013.

[5] Antoine Deleforge, Florence Forbes, and Radu Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International journal of neural systems*, vol. 25, no. 1, pp. 1–21, 2015.

[6] Jacob Benesty, Tomas Gänsler, Dennis R Morgan, M Mohan Sondhi, Steven L Gay, et al., *Advances in network and acoustic echo cancellation*, Springer, 2001.

[7] Sharon Gannot, David Burshtein, and Ehud Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[8] Shmulik Markovich, Sharon Gannot, and Israel Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[9] Klaus Reindl, S Markovich-Golan, Hendrik Barfuss, Sharon Gannot, and Walter Kellermann, "Geometrically constrained trinicon-based relative transfer function estimation in underdetermined scenarios," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013*. IEEE, 2013.

[10] Richard O Duda, "Elevation dependence of the interaural transfer function," *Binaural and spatial hearing in real and virtual environments*, pp. 49–75, 1997.

[11] Jens Blauert, *The technology of binaural listening*, Springer, 2013.

[12] Phillip Griffiths and Joseph Harris, *Principles of algebraic geometry*, John Wiley & Sons, 1994.

[13] Maja Taseska and Emanuel AP Habets, "A subspace-based perspective on spatial filtering performance with distributed and co-located microphone arrays," in *ITG Fachtagung Sprachkommunikation*. VDE, 2014.

[14] Ralph O Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276–280, 1986.

[15] Sylvain Argentieri and Patrick Danes, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2007*. IEEE, 2007, pp. 2009–2014.

[16] Vladimir Tourbabin and Boaz Rafaely, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1803–1814, 2014.

[17] Anthony Lombard, Tobias Rosenkranz, Herbert Buchner, and Walter Kellermann, "Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2009*. IEEE, 2009, pp. 233–236.

[18] Michael I. Mandel, Ron J. Weiss, and Daniel 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.