

Local Relative Transfer Function for Sound Source Localization

Xiaofei Li, Radu Horaud, Laurent Girin, Sharon Gannot

▶ **To cite this version:**

Xiaofei Li, Radu Horaud, Laurent Girin, Sharon Gannot. Local Relative Transfer Function for Sound Source Localization. EUSIPCO 2015 - 23th European Signal Processing Conference, Aug 2015, Nice, France. pp.399-403, 10.1109/EUSIPCO.2015.7362413 . hal-01163675

HAL Id: hal-01163675

https://inria.hal.science/hal-01163675

Submitted on 15 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LOCAL RELATIVE TRANSFER FUNCTION FOR SOUND SOURCE LOCALIZATION

Xiaofei Li[1], *Radu Horaud*[1], *Laurent Girin*[1,2]

[1]INRIA Grenoble Rhône-Alpes
[2]GIPSA-Lab & Univ. Grenoble Alpes

Sharon Gannot

Faculty of Engineering
Bar-Ilan University

ABSTRACT

The relative transfer function (RTF), i.e. the ratio of acoustic transfer functions between two sensors, can be used for sound source localization / beamforming based on a microphone array. The RTF is usually defined with respect to a unique reference sensor. Choosing the reference sensor may be a difficult task, especially for dynamic acoustic environment and setup. In this paper we propose to use a locally normalized RTF, in short *local-RTF*, as an acoustic feature to characterize the source direction. Local-RTF takes a neighbor sensor as the reference channel for a given sensor. The estimated local-RTF vector can thus avoid the bad effects of a noisy unique reference and have smaller estimation error than conventional RTF estimators. We propose two estimators for the local-RTF and concatenate the values across sensors and frequencies to form a high-dimensional vector which is utilized for source localization. Experiments with real-world signals show the interest of this approach.

Index Terms— microphone array, relative transfer function, sound source localization.

1. INTRODUCTION

Sound source localization (SSL) is important for many applications, e.g., robot audition, video conferencing, hearing aids, etc. This paper addresses the problem of estimating the 2D (azimuth and elevation) direction of arrival (DOA) of a sound source using a microphone array. This problem has been largely addressed in the literature, and we focus here on the framework of methods based on *relative transfer function* (RTF) estimation.

For a given spatially-narrow static source, an *acoustic transfer function* (ATF) can be defined for each sensor, that characterizes the frequency-dependent effects of both environment (e.g. room reverberations) and sensor setup (e.g. dummy head with ear microphones) on the source signal. For a given environment and sensor setup, the ATF depends on the source direction, generally in an intricate manner, and so does the RTF, which is the ratio between the ATF of two sensors [1]. For an array with more than two microphones, a specific channel is generally chosen as the unique reference. The RTF vector thus concatenates the ATF ratios between each microphone and the reference. Normalized (unit-norm) RTF vectors are sometimes used, especially to facilitate clustering processes [2].

For low sensor and environment noise level, the RTF can be estimated from measured cross-spectrum of sensor signals. The estimated RTF can then be used in beamforming [1], or to directly recover the time difference of arrival (TDOA) [3] and source direction [4][1]. In such applications, the quality of RTF estimation is a critical issue [1, 8]. However, the presence of noise in the reference channel can significantly corrupt the RTF estimate [9]. Therefore, selecting the channel with the lower noise as the reference channel is beneficial in improving the robustness of RTF estimate [10, 11], but this is not an easy task for real-world acoustic environments and recording setups. In the present paper, we propose an alternative solution, that focuses on the definition of RTF itself. We propose to take the neighbor of each channel as a *local reference* channel, hence leading to so-called *local RTF*, and the corresponding local-RTF feature vector. This avoids taking a channel with intense noise as the unique reference channel. In other words, with such definition, a channel with intense noise at most affects the RTF of its direct neighbor, but not all of RTF vector entries.

The remainder of the paper is organized as follows. Section 2 recalls the usual definition and estimation of the RTF. Section 3 presents the definition of the proposed local-RTF and provides two local-RTF estimators. Section 4 presents an SSL method based on local-RTF. Experiments are presented in Section 5. Section 6 concludes the paper.

2. PROBLEM FORMULATION AND USUAL RTF

Let us consider a single static sound source and an array of M microphones. In the STFT domain, the signals received by the M microphones are approximated as:

$$\mathbf{x}(\omega, l) \approx \mathbf{h}(\omega)s(\omega, l) + \mathbf{n}(\omega, l), \quad (1)$$

This research has received funding from the EU-FP7 STREP project EARS (#609465).

[1]When multiple sources are emitting simultaneously, the problem becomes more complex. Besides beamforming, solutions based on source sparsity and source clustering in the TF domain have been proposed, especially for two-sensor configurations where the RTF is replaced with equivalent binaural cues, namely interaural level and phase differences [5, 6, 7].

where $\omega \in [0, \Omega - 1]$ and $l \in [1, L]$ are the frequency-bin and time-frame indices, $\mathbf{x}(\omega, l) = [x_1(\omega, l), \ldots, x_M(\omega, l)]^T$ is the sensor signal vector, $s(\omega, l)$ is the source signal, and $\mathbf{n}(\omega, l) = [n_1(\omega, l), \ldots, n_M(\omega, l)]^T$ is the sensor noise vector. The source and noise signals are assumed to be uncorrelated. $\mathbf{h}(\omega) = [h_1(\omega), \ldots, h_M(\omega)]^T$ is the ATF vector, which is assumed frequency-dependent and time-invariant. As stated in the introduction, the ATF indicates the relative positions between sound source and microphones, and is affected by sound reflections and sensor array configuration. The RTF for the m-th sensor is defined as the ratio $r_m(\omega) = h_m(\omega)/h_1(\omega)$. Without loss of generality, the first channel is taken as the reference, which is here unique for all RTFs. The RTF vector is $\mathbf{r}(\omega) = [r_1(\omega), \ldots, r_M(\omega)]^T$.

The RTF can be estimated using cross-spectral methods. Let us define the (empirical time-average) cross-spectrum of microphone signals between the i-th and j-th channels as:

$$\hat{\Phi}_{x_i x_j}(\omega) = \frac{1}{L}\sum_{l=1}^{L} x_i(\omega, l) x_j^*(\omega, l) \tag{2}$$

$$\approx \frac{1}{L} h_i(\omega) h_j^*(\omega) \sum_{l=1}^{L} |s(\omega, l)|^2 + \frac{1}{L}\sum_{l=1}^{L} n_i(\omega, l) n_j^*(\omega, l),$$

where * denotes the complex conjugate. The above approximation stands since all signal/noise cross-terms are small compared to the other terms. Moreover, if the noise is spatially uncorrelated, the cross-channel noise power will also be small. Since the source signal STFT does not depend on the ATFs, the RTF can be estimated by:

$$\hat{r}_m(\omega) = \frac{\hat{\Phi}_{x_m x_1}(\omega)}{\hat{\Phi}_{x_1 x_1}(\omega)}. \tag{3}$$

In [1] [9], this RTF estimator is shown to be biased, and both the bias and variance are inversely proportional to the channel average signal-to-noise ratio (SNR). In [1] an unbiased RTF estimator is also proposed based on a least squares criterion. Its variance is also inversely proportional to the average SNR. Therefore, as noise in the reference channel increases, the RTF estimation error will increase for both the biased and unbiased estimators. Consequently, choosing a high SNR channel (ideally the highest SNR channel) as the reference is beneficial in reducing the estimation error. In [10] a reference channel selection method is proposed, based on the input (or output) SNR. Its performance depends on the accuracy of the frequency-dependent SNR estimation, which is not easy in a practical (nonstationary) acoustic environment. If the acoustic environment is similar for all microphones, the reference channel can be chosen arbitrarily. But for some configurations, e.g. the microphone array is embedded in a robot head, the noise signal at each microphone can be quite different. Moreover, the variation of the microphone array position and background noise can make the acoustic environment of each channel vary significantly in time. Therefore, selecting the channel with the lower noise may not be an easy task.

3. LOCAL RELATIVE TRANSFER FUNCTION

3.1. Definition

Based on the above discussion, to avoid a potential bad unique reference we propose a *local-RTF* constructed not from a unique reference channel but rather from a *local reference*, for instance (one of) the sensor's closest neighbor sensor:

$$a_m(\omega) = \frac{|h_m(\omega)|}{\| \mathbf{h}(\omega) \|} e^{j(\arg[h_m(\omega)] - \arg[h_{m-1}(\omega)])}, \tag{4}$$

where $\arg[\cdot]$ is the phase of complex number, $\| \cdot \|$ is the l_2-norm. The corresponding *local-RTF* vector is $\mathbf{a}(\omega) = [a_1(\omega), \ldots, a_M(\omega)]^T$. Assume that the sensors indexes are ordered according to sensor proximity. For phase difference, the $(m-1)$-th channel is taken as the reference of the m-th channel (exceptionally, take the M-th channel as the reference of the first channel). The proximity of sensor pair ensures general minimization of spatial aliasing effects. As for the amplitude, we chose to normalize the local-RTF vector to unit-norm, as in [2, 12]. Compared with local amplitude ratio $\frac{|h_m(\omega)|}{|h_{m-1}(\omega)|}$, this is much more robust to estimation errors. Indeed, local amplitude ratios would be estimated using ratios of sensor signal power, which are very sensitive to the noise of the local reference when source power is small.

In summary, the local-RTF vector $\mathbf{a}(\omega)$ is the complex form of M normalized levels and M local phase differences. Note that it is not an actual transfer function vector that can be directly used for beamforming. It is rather a robust feature expected to be appropriate for SSL due to its lower sensitivity to noise (compared to usual RTF vector).

3.2. Estimation of local-RTF

We provide here two estimators to compute local-RTF vectors $\mathbf{a}(\omega)$ from microphone signals.

Estimator 1: By using the cross- and auto spectrum (2), the local-RTF of the m-th channel can be estimated as:

$$\hat{a}_m(\omega) = \frac{\sqrt{\hat{\Phi}_{x_m x_m}(\omega)}}{\sqrt{\sum_{m=1}^{M} \hat{\Phi}_{x_m x_m}(\omega)}} e^{j\arg[\hat{\Phi}_{x_m x_{m-1}}(\omega)]}. \tag{5}$$

As expected from definition and confirmed by simulations, this estimator is biased. It is however suitable for high SNR due to its small bias in this case and low computation cost.

Estimator 2: The second estimator of the local-RTF that we propose is based on the unbiased RTF estimator proposed in [8]. For each channel m, we basically replace the reference channel 1 by channel $m-1$. In a few details, the noise power spectral density (PSD) estimate $\hat{\Phi}_{n_{m-1} n_{m-1}}(\omega, l)$ of the local reference channel is first calculated by recursively averaging past spectral power values of the observed signal using a time-varying smoothing parameter adjusted by the speech presence probability [8]. The

same principle is applied to estimate the noise cross-PSD between channels m and $m-1$, namely $\hat{\Phi}_{n_m n_{m-1}}(\omega, l)$. The cross-PSD of the noisy signal $\hat{\Phi}_{x_m x_{m-1}}(\omega, l)$ is estimated from observations. The PSD estimate $\hat{\Phi}_{s_{m-1} s_{m-1}}(\omega, l)$ of the image source signal $h_{m-1}(\omega)s(\omega, l)$ in the reference channel is calculated using the *optimally modified log-spectral amplitude* (OM-LSA) technique [13]. An estimate $\hat{\rho}_m(\omega)$ of the ATF ratio $\rho_m(\omega) = \frac{h_m(\omega)}{h_{m-1}(\omega)}$ is then obtained from $\hat{\Phi}_{x_m x_{m-1}}(\omega, l)$, $\hat{\Phi}_{n_m n_{m-1}}(\omega, l)$ and $\hat{\Phi}_{s_{m-1} s_{m-1}}(\omega, l)$, by combining weighted spectral subtraction, frame averaging, and ratio (see [8], Eq. (28)). The above process is repeated for each channel. Finally, the local-RTF estimator is defined by:

$$\hat{a}_m(\omega) = \frac{\sqrt{\hat{\Phi}_{s_m s_m}(\omega)}}{\sqrt{\sum_{m=1}^{M} \hat{\Phi}_{s_m s_m}(\omega)}} e^{j \arg[\hat{\rho}_m(\omega)]}, \qquad (6)$$

where $\hat{\Phi}_{s_m s_m}(\omega) = \frac{1}{L} \sum_{l=1}^{L} \hat{\Phi}_{s_m s_m}(\omega, l)$. This estimator is more suitable than Estimator 1 for low SNRs, since spectral subtraction can (partly) remove the bias.

4. SOUND SOURCE LOCALIZATION USING LOCAL-RTF VECTOR

The local-RTF values for frequency bin ω, estimated by one of the two above estimators, are used to form the (frequency-dependent) local-RTF feature vector $\hat{\mathbf{a}}(\omega) = [\hat{a}_1(\omega), \ldots, \hat{a}_M(\omega)]^T$. Then by concatenating the local-RTF vectors across frequencies, we obtain a global feature vector in $\mathbb{C}^{M \times \Omega}$: $\hat{\mathbf{a}} = [\hat{\mathbf{a}}^T(0), \ldots, \hat{\mathbf{a}}^T(\omega), \ldots, \hat{\mathbf{a}}^T(\Omega - 1)]^T$.

In order to perform SSL based on the global local-RTF vector $\hat{\mathbf{a}}$, we adopt here a supervised approach. A large number K of local-RTF feature vectors $\mathbf{a}_k$ associated with corresponding 2D source direction vectors $\mathbf{d}_k$ (azimuth and elevation) is first collected. A regression model trained on this dataset can be used to map the high-dimensional local-RTF space to the low-dimensional source direction space [14, 15, 16]. In this paper we rather use a simple lookup table followed by interpolation technique that compares a new observed feature vector $\hat{\mathbf{a}}$ with all the K feature vectors in the dataset $\{\mathbf{a}_k\}_{k=1}^{K}$, finds the I closest ones $\{\mathbf{a}_{k_i}\}_{i=1}^{I}$, and provides the associated estimated source direction as the weighted mean:

$$\hat{\mathbf{d}} = \frac{1}{\sum_{i=1}^{I} \| \hat{\mathbf{a}} - \mathbf{a}_{k_i} \|^{-1}} \sum_{i=1}^{I} \| \hat{\mathbf{a}} - \mathbf{a}_{k_i} \|^{-1} \mathbf{d}_{k_i}. \qquad (7)$$

In all presented experiments, I was fixed to 4, significantly improving the localization compared to $I = 1$. Larger neighborhood did not work significantly better.

If the average power of the ω-th frequency bin (represented by $\sum_{m=1}^{M} \hat{\Phi}_{x_m x_m}(\omega)$ for Estimator 1, and by $\sum_{m=1}^{M} \hat{\Phi}_{s_m s_m}(\omega)$ for Estimator 2) is small (in practice lower than a small fixed threshold), due to the frequency sparsity of speech signals, the corresponding estimated local-RTF vector $\hat{\mathbf{a}}(\omega)$ is prone to a large estimation error. In that case, $\hat{\mathbf{a}}(\omega)$ is set to a zero vector. By doing so, the contribution of the ω-th frequency is discarded in the lookup procedure. Indeed, the subvectors $\mathbf{a}_k(\omega)$ in the lookup dataset are all unit vectors. Therefore, the zero subvector of $\hat{\mathbf{a}}$ has the same distance to all of these unit subvectors $\mathbf{a}_k(\omega)$, and this distance in non informative in the overall distance calculation. This contributes to make the proposed localization based on local-RTF particularly robust to the sparsity of speech signals.

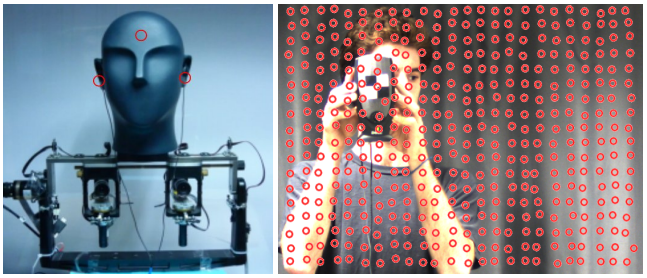


Fig. 1: Acoustic dummy head with microphones (marked with red circles) and cameras (left). Training dataset (right).

5. EXPERIMENTS

5.1. Experimental setup and data

The microphone array used in the presented experiments is composed of four microphones mounted onto a Sennheiser MKE 2002 acoustic dummy head. The microphones are plugged into the left and right ears and fixed on the forehead and on the back of the head, see Fig. 1(left). We used the audio-visual data acquisition method described in [7]: Sounds are emitted by a loudspeaker on which a visual marker is fixed ; a camera is rigidly attached to the dummy head, and the ground-truth source direction is obtained by localizing the visual marker in the image provided by the camera, see Fig. 1(right). The image resolution is 640×480 pixels, spanning a field-of-view of 28°-azimuth × 21°-elevation. Hence, 1° corresponds approximately to 23 pixels.

All data are recorded in a quiet office environment, with soft background noise (e.g., computer fans, air conditioning, etc.) with an overall SNR of about 18dB. The loudspeaker was placed at approximately 2.5m away from the dummy head. The training data which are used for generating the lookup dataset consist of 1s-duration white-noise signals emitted from 432 source directions, spanning an approximate field-of-view of 24° × 18°, see Fig. 1(right). The test data which are used to evaluate the localization method consist of 108 speech utterances of variable duration extracted from the TIMIT dataset [17], and emitted by the loudspeaker from 108 directions within the camera field-of-view. The sampling rate

Setup	Estimator 1		Estimator 2	
	Azim.	Elev.	Azim.	Elev.
Binaural	0.93	0.91	0.91	0.97
4-microphone array	0.87	0.49	0.86	0.53

Table 1: Average localization error (in degrees) for two types of microphone arrays, with no additive noise.

SNR	Estim. 1		Estim. 2		RGR		HIS	
(dB)	Azi.	Ele.	Azi.	Ele.	Azi.	Ele.	Azi.	Ele.
10	0.83	0.51	0.85	0.47	0.93	0.78	0.96	0.76
5	0.83	0.56	0.86	0.47	0.96	0.82	0.95	0.82
0	0.85	0.62	0.89	0.46	1.05	0.83	1.02	0.74
-5	1.00	0.76	1.02	0.51	1.33	1.04	1.20	1.05
-10	1.53	1.22	1.51	0.75	1.98	1.62	1.79	1.30

Table 2: Average localization error (in degrees) for the environmental noise, for both proposed local-RTF estimators, and for the RGR and HIS RTF estimators.

is 16kHz and the window length of the STFT is 32ms with 16ms overlap. One power spectrum estimate (2) was calculated for each entire test sentence (hence L depending on sentence duration), resulting in one local-RTF value and one source direction estimate for each test sentence. The performance metric is the absolute angle error (in degrees) in azimuth and elevation, respectively, averaged over the 108 test values. Note that the training data and test data have the same recording setup (room, position of microphone array, distance between source and microphone array). Reverberations are not explicitly considered but are implicitly embedded in the local-RTF features and in the look-up table. The T_{60} reverberation time of the room is about 0.37s.

In order to test the efficiency of local-RTF features for SSL in noisy environment, two types of noise signals were recorded and added to the speech test signals at various SNRs: 1) an *environmental noise* is recorded in a noisy office environment with opened door and windows. This noise comprises more diverse and nonstationary components, produced by e.g. people movements, devices, outside environment (passing cars, street noise), etc. Noise sources are neither strictly directional nor entirely diffuse either; 2) a *directional white Gaussian noise (WGN)* is emitted by the loudspeaker with a direction beyond the camera field-of-view. Note that the SNR is an average SNR because either the noise, the speech signals, or both, are nonstationary. Actual frame-wise SNR may significantly vary for a given average SNR.

5.2. 4-microphone setup vs. binaural setup experiment

As a preliminary experiment, we have tested the efficiency of using the 4-microphone array setup vs. using a *binaural* setup with only the two ear microphones, as largely considered in the SSL literature, e.g. [5, 6, 7]. No additive noise is considered here. Table 1 shows the localization results. Both

SNR	Estim. 1		Estim. 2		RGR		HIS	
(dB)	Azi.	Ele.	Azi.	Ele.	Azi.	Ele.	Azi.	Ele.
10	0.80	0.49	0.82	0.49	0.95	0.70	0.80	0.87
5	1.24	0.65	0.80	0.54	1.01	0.83	0.87	0.80
0	3.39	1.31	0.91	0.56	1.33	0.73	1.11	0.64
-5	8.33	2.74	1.40	0.77	2.70	1.17	1.31	0.75
-10	11.2	3.87	3.82	1.48	3.61	1.42	1.64	1.00

Table 3: Average localization error (in degrees) for the directional WGN, for both proposed local-RTF estimators, and for the RGR and HIS RTF estimators.

local-RTF estimators are tested. It can be seen that the localization error for the 4-microphone array setup is significantly lower than for the binaural setup, especially for the elevation, where the average error is reduced by about 45%. This is because the two additional microphones on the dummy head are located above the ear microphones, and therefore they significantly improve the discrimination for the elevation. The performance of both local-RTF estimators are here similar because of the high SNR of the recordings.

5.3. SSL in noisy conditions

Table 2 shows the localization results for the environmental noise at different SNRs. SSL using the two proposed local-RTF estimators is compared with SSL using two unbiased RTF estimators derived in [8]: the unit-RTF with a random global reference (RGR), which uses a unique reference channel selected randomly, and the highest input SNR (HIS) reference [10] based on SNR estimation [8] (see Section 2).

It can be seen that, for $0-10$dB SNR range, the two local-RTF estimators have close performance measures. Elevation estimation is more accurate than azimuth estimation. Both RGR and HIS reference methods also have similar performance, but the error is significantly larger than the error for the proposed method. The relative difference is larger for elevation (e.g. 0.82° for both RGR and HIS, vs. 0.56° and 0.47° for Estimator 1 and 2, respectively, at 5dB SNR) than for azimuth (e.g. 0.96° and 0.95° for RGR and HIS, respectively, vs. 0.83° and 0.86° for Estimator 1 and 2, respectively, at 5dB SNR). As expected, all methods exhibit degraded performance when noise power increases, but the proposed method (for any of the two estimators) remains more efficient than the reference methods. At -10 and -5dB SNR, the proposed method with Estimator 2 outperforms all other methods, since it efficiently exploits both local reference channel and noise spectral subtraction. Such results show that the proposed method is able to circumvent the problem of choosing a good reference channel. In these experiments, it works even better than the HIS method which depends on a correct estimation of the SNR at each channel (note that HIS generally performs better than RGR at low SNR).

Table 3 shows the localization results for the directional WGN. Here, the necessity of carefully taking the noise into

account is evident, either by using spectral subtraction (Estim. 2 vs. Estim. 1) or by using appropriate channel selection (HIS vs. RGR). Performance measures of Estimator 1 and RGR drop abruptly for SNR equal to and lower than 0dB and -5dB, respectively. In contrast, Estimator 2 obtains the best results in both azimuth and elevation at 5 and 0dB, and remains competitive with the HIS method at -5dB. This can be explained by the fact that when the SNR is low, the noise directivity induces a large noise power difference among channels, and the proposed method with Estimator 2 correctly exploits the information diversity. The HIS method performs well at low SNRs because the input SNR estimation is relatively accurate due to the stationarity of the directional WGN. HIS correctly estimates the highest SNR channel and uses it as an appropriate global reference. The fact that the proposed method can compete with the HIS method up to -5dB SNR is remarkable given that no channel selection is made.

6. CONCLUSION

A local-RTF acoustic feature vector has been proposed for sound source localization. This feature vector has been shown to be more robust than RTF with a unique (possibly selected) reference channel for SSL in several tested conditions. Only single-source localization in noise has been considered in the present paper. Future work will address the use of the local-RTF vector for multiple-source localization in more adverse environments. Due to the lower bias and variance of the observed local-RTF vector, this feature is expected to be a robust feature for source separation and multiple speakers localization based on clustering.

7. REFERENCES

[1] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Proc.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[3] T. G. Dvorkind and S. 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] B. Laufer, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," in *IEEE WASPAA*, (New Paltz, NY), pp. 1–4, 2013.

[5] M. Mandel, R. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 18, no. 2, pp. 382–394, 2010.

[6] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 20, no. 5, pp. 1503–1512, 2012.

[7] A. Deleforge, V. Drouard, L. Girin, and R. Horaud, "Mapping sounds onto images using binaural spectrograms," in *EUSIPCO*, (Lisbon, Portugal), 2014.

[8] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech and Audio Proc.*, vol. 12, no. 5, pp. 451–459, 2004.

[9] S. Gannot, D. Burshtein, and E. Weinstein, "Analysis of the power spectral deviation of the general transfer function GSC," *IEEE Trans. Signal Proc.*, vol. 52, no. 4, pp. 1115–1120, 2004.

[10] T. C. Lawin-Ore and S. Doclo, "Reference microphone selection for MWF-based noise reduction using distributed microphone arrays," in *ITG Conf. Speech Communication*, (Braunschweig, Germany), 2012.

[11] S. Stenzel, J. Freudenberger, and G. Schmidt, "A minimum variance beamformer for spatially distributed microphones using a soft reference selection," in *IEEE HSCMA Workshop*, (Nancy, France), 2014.

[12] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l_1-norm minimization," *EURASIP J. Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.

[13] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[14] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *Int. J. Neural Systems*, vol. 25, no. 1, 2015.

[15] A. Deleforge, R. Horaud, Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE Trans. Audio, Speech, Lang. Proc.,* accepted, 2015.

[16] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Gaussian process models for HRTF-based 3D sound localization," in *IEEE ICASSP*, (Florence, Italy), 2014.

[17] J. Garofolo, L. Lamel, W. Fisher, and coll., "TIMIT acoustic-phonetic continuous speech corpus," tech. rep., Linguistic Data Consortium, Philadelphia, 1993.