Estimation of the Direct-Path Relative Transfer Function for Supervised Sound-Source Localization

Xiaofei Li, Laurent Girin, Radu Horaud, Sharon Gannot

Estimation of the Direct-Path Relative Transfer Function for Supervised Sound-Source Localization

Xiaofei Li, Laurent Girin, Radu Horaud and Sharon Gannot

***Abstract*—This paper addresses the problem of binaural localization of a single speech source in noisy and reverberant environments. For a given binaural microphone setup, the binaural response corresponding to the direct-path propagation of a single source is a function of the source direction. In practice, this response is contaminated by noise and reverberations. The direct-path relative transfer function (DP-RTF) is defined as the ratio between the direct-path acoustic transfer function of the two channels. We propose a method to estimate the DP-RTF from the noisy and reverberant microphone signals in the short-time Fourier transform domain. First, the convolutive transfer function approximation is adopted to accurately represent the impulse response of the sensors in the STFT domain. Second, the DP-RTF is estimated by using the auto- and cross-power spectral densities at each frequency and over multiple frames. In the presence of stationary noise, an inter-frame spectral subtraction algorithm is proposed, which enables to achieve the estimation of noise-free auto- and cross-power spectral densities. Finally, the estimated DP-RTFs are concatenated across frequencies and used as a feature vector for the localization of speech source. Experiments with both simulated and real data show that the proposed localization method performs well, even under severe adverse acoustic conditions, and outperforms state-of-the-art localization methods under most of the acoustic conditions.**

***Index Terms*—binaural source localization, direct-path relative transfer function, inter-frame spectral subtraction.**

I. Introduction

Sound-source localization (SSL) is an important task for many applications, e.g., robot audition, video conferencing, hearing aids, to cite just a few. In the framework of human-inspired binaural hearing, two interaural cues are widely used for SSL, namely the interaural phase difference (IPD) and the interaural level difference (ILD) [1], [2], [3], [4], [5], [6], [7]. In the general case where the sensor array is not free-field, i.e. the microphones are placed inside the ears of a dummy head or on a robot head, the interaural cues are frequency-dependent due to the effects on sound propagation induced by the shape of the outer ears, head and torso [8]. This is true even for anechoic recordings, i.e. in the absence of reverberations. SSL is then based on the relationship between interaural cues and direction of arrival (DOA) of the emitting source.

X. Li and R. Horaud are with INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France. E-mail: first.last@inria.fr

L. Girin is with INRIA Grenoble Rhône-Alpes and with Univ. Grenoble Alpes, GIPSA-lab, Grenoble, France. E-mail: laurent.girin@gipsa-lab.grenoble-inp.fr

Sharon Gannot is with Bar Ilan University, Faculty of Engineering, Israel. E-mail: Sharon.Gannot@biu.ac.il

X. Li, L. Girin and R. Horaud acknowledge support from the EU FP7 STREP project EARS #609465 and from the ERC Advanced Grant VHIA #340113.

When the short-time Fourier transform (STFT) is used, the ILD and IPD correspond to the magnitude and argument, respectively, of the relative transfer function (RTF), which is the ratio between the acoustic transfer functions (ATF) of the two channels [9]. In a reverberant environment, the RTF contains both direct-path information, namely the direct wave propagation path from the source location to the microphone locations, and information representing early and late reverberations. Extracting the direct path is of crucial importance for SSL. In an anechoic and noise-free environment the source direction can be easily estimated from the RTF. However, in practice, noise and reverberations are often present and contaminate SSL estimation.

In the presence of noise, based on the stationarity of the noise and the non-stationarity of the desired signal, the RTF was estimated in [9] by solving a set of linear equations, and in [10] by solving a set of nonlinear decorrelation equations. In [10], the time difference of arrival (TDOA) was estimated based on RTF, and a TDOA tracking method was also proposed. These methods have the limitation that a significant amount of noisy frames are included in the estimation. An RTF identification method based on the probability of speech presence and on spectral subtraction was proposed in [11]: this method uses only the frames which are highly likely to contain speech. The unbiased RTF estimator proposed in [12] is based on segmental power spectral density matrix subtraction, which is a more efficient method to remove noise compared with the approaches just mentioned. The performance of these spectral subtraction techniques was analyzed and compared with eigenvalues decomposition techniques in [13].

The RTF estimators mentioned above assume a multiplicative transfer function (MTF) approximation [14], i.e., the source-to-microphone filtering process is assumed to be represented by a multiplicative process in the STFT domain. Unfortunately, this is only justified when the length of the filter impulse response is shorter than the length of the STFT window, which is rarely the case in practice. Moreover, the RTF is usually estimated from the ratio between two ATFs that include reverberation, rather than from the ratio between ATFs that only correspond to the direct-path sound propagation. Therefore, currently available RTF estimators are poorly suitable for SSL in reverberant environments.

The influence of reverberation on the interaural cues is analyzed in [15]. The relative early transfer function was introduced in [16] to suppress reverberation. Several techniques were proposed to extract the RTF that corresponds to the direct-path sound propagation, e.g., based on detecting time frames with less reverberations. The precedence effect, e.g.,

[17], widely used for SSL, relies on the principle that signal onsets are dominated by the direct path. Based on band-pass filter banks, the localization cues are extracted only from reliable frames, such as the onset frames in [18], the frames preceding a notable maximum [19], the frames weighted by the precedence model [20], etc. Interaural coherence was proposed in [21] to select binaural cues not contaminated by reverberations. Based on Fourier transform, the coherence test [22], and the direct-path dominance test [23] are proposed to detect the frames dominated by one active source, from which localization cues can be estimated. However, in practice, there are always reflection components in the frames selected by these methods, due to an inaccurate model or an improper decision threshold.

Contributions and Method Overview: In this paper, we propose a direct-path RTF estimator suitable for the localization of a single speech-source in noisy and reverberant environments. We build on the cross-band filter proposed in [24] for system identification in the STFT domain. This filter represents the impulse response in the STFT domain by a cross-band convolutive transfer function instead of the multiplicative (MTF) approximation. In practice we consider the use of a simplified convolutive transfer function (CTF) approximation, as used in [25]. The first coefficient of the CTF at different frequencies represents the STFT of the first segment of the channel impulse response, which is composed of the direct-path impulse response, plus possibly few early reflections. In particular, if the time delay between the direct-path wave and the first notable reflection is large, less reflections are included. Therefore, we refer to the first coefficient of the CTF as the direct-path acoustic transfer function, and the ratio between the coefficients from two channels is referred to as the *direct-path relative transfer function* (DP-RTF).

Inspired by [26] and based on the relationship of the CTFs between the two channels, we use the auto- and cross-power spectral densities (PSD) estimated over multiple STFT frames, to construct a set of linear equations in which the DP-RTF is the unknown variable. Therefore, the DP-RTF can be estimated via standard least squares. In the presence of noise, an inter-frame spectral subtraction technique is proposed, extending our previous work [12]. The auto- and cross-PSD estimated in a frame with low speech power are subtracted from the PSDs estimated in a frame with high speech power. After subtraction, low noise power and high speech power are left due to the stationarity of the noise and the non-stationarity of the speech signal. The DP-RTF is estimated using the remaining signal's auto- and cross-PSD. This PSD subtraction process does not require an explicit estimation of the noise PSD, hence it does not suffer from noise PSD estimation errors.

Finally, the estimated DP-RTFs are concatenated over frequencies and plugged into an SSL method, e.g., [6]. Experiments with simulated and real data were conducted under various acoustic conditions, e.g., different reverberation times, source-to-sensor distances, and signal-to-noise ratios. The experimental results show that the proposed method performs well, even in adverse acoustic conditions, and outperforms the MTF-based method [12], the coherence test method [22] and the conventional SRP-PHAT method in most of the tested conditions.

The remainder of this paper is organized as follows. Section II formulates the sensor signals based on the crossband filter. Section III presents the DP-RTF estimator in a noise-free environment. The DP-RTF estimator in the presence of noise is presented in Section IV. In Section V, the SSL algorithm is described. Experimental results are presented in Section VI and VII, and Section VIII draws some conclusions.

II. Cross-band Filter and Convolutive Transfer Function

We consider first a non-stationary source signal $s(n)$, e.g., speech, emitted in a noise-free environment. The received binaural signals are

$$\begin{aligned} x(n) &= s(n) \star a(n) \\ y(n) &= s(n) \star b(n), \end{aligned} \tag{1}$$

where $\star$ denotes convolution, and $a(n)$ and $b(n)$ are the binaural room impulse responses (BRIR) from the source to the two microphones. The BRIRs combine the effects of the room acoustics (reverberations) and the effects of the sensor set-up (e.g., dummy head/ears). Applying the STFT, (1) is approximated in the time-frequency (TF) domain as

$$\begin{aligned} x_{p,k} &= s_{p,k}\, a_k \\ y_{p,k} &= s_{p,k}\, b_k, \end{aligned} \tag{2}$$

where $x_{p,k}$, $y_{p,k}$ and $s_{p,k}$ are the STFT of the corresponding signals (p is the time frame index and k is the frequency bin index), and a_k and b_k are the ATFs corresponding to the BRIRs. Let N denote the length of a time frame or, equivalently, the size of the STFT window. Eq. (2) corresponds to the MTF approximation, which is only valid when the impulse response $a(n)$ is shorter than the STFT window. In the case of non-stationary acoustic signals, such as speech, a relatively small value for N is typically chosen to assume *local* stationarity, i.e., within a frame. Therefore, the MTF approximation (2) is questionable in a reverberant environment, since the room impulse response could be much longer than the STFT window.

To address this problem cross-band filters were introduced [24] to represent more accurately a linear system with long impulse response in the STFT domain. Let L denote the frame step. The cross-band filter model consists in representing the STFT coefficient $x_{p,k}$ in (2) as a summation over multiple convolutions across frequency bins (there is an equivalent expression for $y_{p,k}$):

$$x_{p,k} = \sum_{p'=-C}^{Q_k-1} \sum_{k'=0}^{N-1} s_{p-p',k'}\; a_{p',k',k}. \tag{3}$$

From [24], if $L < N$, then $a_{p',k',k}$ is non-causal, with $C = \lceil N/L \rceil - 1$ non-causal coefficients. The number of causal filter coefficients Q_k is related to the reverberation time at the k-th frequency bin, which will be discussed in detail in Section VI.

The TF-domain impulse response $a_{p',k',k}$ is related to the time-domain impulse response $a(n)$ by:

$$a_{p',k',k} = (a(n) \star \zeta_{k,k'}(n))|_{n=p'L}, \tag{4}$$

which represents the convolution with respect to the time index n evaluated at frame steps, with

$$\zeta_{k,k'}(n) = e^{j\frac{2\pi}{N}k'n} \sum_{m=-\infty}^{+\infty} \overline{\omega}(m)\,\omega(n+m)\,e^{-j\frac{2\pi}{N}m(k-k')}, \tag{5}$$

where $\overline{\omega}(n)$ and $\omega(n)$ denote the STFT analysis and synthesis windows, respectively. A convolutive transfer function (CTF) approximation is further introduced and used in [25] to simplify the analysis, i.e., only band-to-band filters are considered, $k = k'$. Hence, (3) is rewritten as

$$x_{p,k} = \sum_{p'=0}^{Q_k-1} s_{p-p',k} a_{p',k} = s_{p,k} \star a_{p,k}, \tag{6}$$

where we assumed $L \approx N$ such that non-causal coefficients are disregarded. Note that $a_{p',k',k}$ is replaced with $a_{p',k}$ to simplify the notations. The cross-band filter and CTF formalism will now be used to extract the impulse response of the direct-path propagation.

III. Direct-Path Relative Transfer Function

From (4) and (5), with $k' = k$ and $p' = 0$, the first coefficient of $a_{p',k}$ in the CTF approximation (6) can be derived as

$$\begin{aligned} a_{0,k} = (a(n) \star \zeta_{k,k}(n))|_{n=0} &= \sum_{t=0}^{T-1} a(t)\zeta_{k,k}(-t) \\ &= \sum_{t=0}^{N-1} a(t)\nu(t)e^{-j\frac{2\pi}{N}kt}, \end{aligned} \tag{7}$$

where T is the length of the BRIR and

$$\nu(n) = \begin{cases} \sum_{m=0}^{N} \overline{\omega}(m)\omega(m-n) & \text{if } 1-N \le n \le N-1, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, $a_{0,k}$ (as well as $b_{0,k}$) can be interpreted as the k-th Fourier coefficient of the impulse response segment $a(n)|_{n=0}^{N-1}$ windowed by $\nu(n)|_{n=0}^{N-1}$. Without loss of generality, we assume that the room impulse responses $a(n)$ and $b(n)$ begin with the impulse responses of the direct-path propagation. If the frame length N is properly chosen, $a(n)|_{n=0}^{N-1}$ and $b(n)|_{n=0}^{N-1}$ are composed of the impulse responses of the direct-path and a few reflections. Particularly, if the initial time delay gap (ITDG), i.e. the time delay between the direct-path wave and the first notable reflection, is large compared to N, $a(n)|_{n=0}^{N-1}$ and $b(n)|_{n=0}^{N-1}$ mainly contain the direct-path impulse response. Therefore we refer to $a_{0,k}$ and $b_{0,k}$ as the direct-path ATFs. By definition, the DP-RTF is given by (we remind that the direct path is relevant for sound source localization):

$$d_k = \frac{b_{0,k}}{a_{0,k}}. \tag{8}$$

In summary, the CTF approximation offers a nice framework to encode the direct-path part of a room impulse response into the first CTF coefficients. Applying this to each channel of a BRIR and taking the ratio between the first CTF coefficients of each channel provides the DP-RTF. Of course, in practice, the DP-RTF must be estimated from the sensor signals.

A. Direct-Path Estimation

Since both channels are assumed to follow the CTF model, we can write:

$$x_{p,k} \star b_{p,k} = s_{p,k} \star a_{p,k} \star b_{p,k} = y_{p,k} \star a_{p,k}. \tag{9}$$

This relation was proposed in [26], [27] for the time-domain TDOA estimation and is here extended to the CTF domain. In vector form (9) can be written as

$$\mathbf{x}_{p,k}^\top \mathbf{b}_k = \mathbf{y}_{p,k}^\top \mathbf{a}_k, \tag{10}$$

where $^\top$ denotes vector or matrix transpose, and

$$\begin{aligned} \mathbf{x}_{p,k} &= [x_{p,k}, x_{p-1,k}, \dots, x_{p-Q_k+1,k}]^\top, \\ \mathbf{y}_{p,k} &= [y_{p,k}, y_{p-1,k}, \dots, y_{p-Q_k+1,k}]^\top, \\ \mathbf{b}_k &= [b_{0,k}, b_{1,k}, \dots, b_{Q_k-1,k}]^\top, \\ \mathbf{a}_k &= [a_{0,k}, a_{1,k}, \dots, a_{Q_k-1,k}]^\top. \end{aligned}$$

Dividing both sides of (10) by $a_{0,k}$ and reorganizing the terms, we can write:

$$y_{p,k} = \mathbf{z}_{p,k}^\top \mathbf{g}_k, \tag{11}$$

where

$$\begin{aligned} \mathbf{z}_{p,k} &= [x_{p,k}, \dots, x_{p-Q_k+1,k}, y_{p-1,k}, \dots, y_{p-Q_k+1,k}]^\top, \\ \mathbf{g}_k &= \left[\frac{b_{0,k}}{a_{0,k}}, \dots, \frac{b_{Q_k-1,k}}{a_{0,k}}, -\frac{a_{1,k}}{a_{0,k}}, \dots, -\frac{a_{Q_k-1,k}}{a_{0,k}}\right]^\top. \end{aligned}$$

We see that the DP-RTF appears as the first entry of $\mathbf{g}_k$. Hence, in the following, we base the estimation of the DP-RTF on the construction of $y_{p,k}$ and $\mathbf{z}_{p,k}$ statistics. More specifically, multiplying both sides of (11) by $y_{p,k}^*$ (the complex conjugate of $y_{p,k}$) and taking the expectation, $E\{\cdot\}$, we obtain:

$$\phi_{yy}(p,k) = \boldsymbol{\phi}_{zy}^\top(p,k)\,\mathbf{g}_k, \tag{12}$$

where $\phi_{yy}(p,k) = E\{y_{p,k}y_{p,k}^*\}$ is the PSD of $y(n)$ at TF bin (p,k), and

$$\begin{aligned} \boldsymbol{\phi}_{zy}(p,k) = &[E\{x_{p,k}y_{p,k}^*\}, \dots, E\{x_{p-Q_k+1,k}y_{p,k}^*\}, \\ &E\{y_{p-1,k}y_{p,k}^*\}, \dots, E\{y_{p-Q_k+1,k}y_{p,k}^*\}]^\top \end{aligned}$$

is a vector composed of cross-PSD terms between the elements of $\mathbf{z}_{p,k}$ and $y_{p,k}$.[1] In practice, these auto- and cross-PSD terms can be estimated by averaging the corresponding auto- and cross-STFT spectra over D frames:

$$\hat{\phi}_{yy}(p,k) = \frac{1}{D}\sum_{d=0}^{D-1} y_{p-d,k}\,y_{p-d,k}^*. \tag{13}$$

[1]More precisely, $\boldsymbol{\phi}_{zy}(p,k)$ is composed of y PSD 'cross-terms', i.e., y taken at frame p and previous frames, and of x, y cross-PSD terms for y taken at frame p and x taken at previous frames.

The elements in $\boldsymbol{\phi}_{zy}(p,k)$ can be estimated by using the same principle. Consequently, in practice (12) is approximated as

$$\hat{\phi}_{yy}(p,k) = \hat{\boldsymbol{\phi}}_{zy}^{\top}(p,k)\,\mathbf{g}_k. \tag{14}$$

Let P denote the total number of the STFT frames. Q_k is the minimum index of p to guarantee that the elements in $\mathbf{z}_{p,k}$ are available from the STFT coefficients of the binaural signals. For PSD estimation, the previous $D-1$ frames of the current frame are utilized as shown in (13). Therefore, $p_f = Q_k + D - 1$ is the minimum index of p to guarantee that all the frames for computing $\hat{\boldsymbol{\phi}}_{zy}(p,k)$ are available from the STFT coefficients of the binaural signals. By concatenating the frames from p_f to P, (14) can be written in matrix-vector form:

$$\hat{\boldsymbol{\phi}}_{yy}(k) = \hat{\boldsymbol{\Phi}}_{zy}(k)\,\mathbf{g}_k, \tag{15}$$

with

$$\hat{\boldsymbol{\phi}}_{yy}(k) = [\hat{\phi}_{yy}(p_f,k),\ldots,\hat{\phi}_{yy}(p,k),\ldots,\hat{\phi}_{yy}(P,k)]^{\top},$$
$$\hat{\boldsymbol{\Phi}}_{zy}(k) = [\hat{\boldsymbol{\phi}}_{zy}(p_f,k),\ldots,\hat{\boldsymbol{\phi}}_{zy}(p,k),\ldots,\hat{\boldsymbol{\phi}}_{zy}(P,k)]^{\top}.$$

Note that $\hat{\boldsymbol{\phi}}_{yy}(k)$ is a $(P-p_f+1)\times 1$ vector and $\hat{\boldsymbol{\Phi}}_{zy}(k)$ is a $(P-p_f+1)\times(2Q_k-1)$ matrix. In principle, an estimate $\hat{\mathbf{g}}_k$ of $\mathbf{g}_k$ can be found be solving this linear equation. However, in practice, the sensor signals contain noise and thus the estimated PSD contain noise power. Therefore, we have to remove this noise power before estimating $\mathbf{g}_k$.

IV. DP-RTF Estimation in the Presence of Noise

Noise always exists in real-world configurations. In the presence of noise, some frames in (15) are dominated by noise. Besides, the PSD estimate of speech signals is deteriorated by noise. In this section, an inter-frame subtraction technique enabling to improve the DP-RTF estimation in noise is described, based on a speech frame selection process.

A. Noisy Signals and PSD Estimates

In the presence of additive noise (1) becomes

$$\begin{aligned}\tilde{x}(n) &= x(n) + u(n) = a(n) \star s(n) + u(n),\\ \tilde{y}(n) &= y(n) + v(n) = b(n) \star s(n) + v(n),\end{aligned} \tag{16}$$

where $u(n)$ and $v(n)$, the noise signals, are assumed to be individually wide-sense stationary (WSS) and uncorrelated with $s(n)$. Moreover, $u(n)$ and $v(n)$ are assumed to be either uncorrelated, or correlated but jointly WSS. Applying the STFT to the binaural signals in (16) leads to

$$\begin{aligned}\tilde{x}_{p,k} &= x_{p,k} + u_{p,k}\\ \tilde{y}_{p,k} &= y_{p,k} + v_{p,k},\end{aligned}$$

in which each quantity is the STFT coefficient of its corresponding time-domain signal. Similarly to $\mathbf{z}_{p,k}$, we define

$$\begin{aligned}\tilde{\mathbf{z}}_{p,k} &= [\tilde{x}_{p,k},\ldots,\tilde{x}_{p-Q_k+1,k},\tilde{y}_{p-1,k},\ldots,\tilde{y}_{p-Q_k+1,k}]^{\top}\\ &= \mathbf{z}_{p,k} + \mathbf{w}_{p,k}\end{aligned}$$

where

$$\mathbf{w}_{p,k} = [u_{p,k},\ldots,u_{p-Q_k+1,k},v_{p-1,k},\ldots,v_{p-Q_k+1,k}]^{\top}.$$

The PSD of $\tilde{y}_{p,k}$ is $\phi_{\tilde{y}\tilde{y}}(p,k)$. We define the PSD vector $\boldsymbol{\phi}_{\tilde{z}\tilde{y}}(p,k)$ composed of the auto- and cross-PSDs between the elements of $\tilde{\mathbf{z}}_{p,k}$ and $\tilde{y}_{p,k}$. Following (13), these PSDs can be estimated as $\hat{\phi}_{\tilde{y}\tilde{y}}(p,k)$ and $\hat{\boldsymbol{\phi}}_{\tilde{z}\tilde{y}}(p,k)$ by averaging the auto- and cross-STFT spectra of input signals over D frames. Since the speech and noise signals are uncorrelated, we can write

$$\begin{aligned}\hat{\phi}_{\tilde{y}\tilde{y}}(p,k) &= \hat{\phi}_{yy}(p,k) + \hat{\phi}_{vv}(p,k),\\ \hat{\boldsymbol{\phi}}_{\tilde{z}\tilde{y}}(p,k) &= \hat{\boldsymbol{\phi}}_{zy}(p,k) + \hat{\boldsymbol{\phi}}_{wv}(p,k),\end{aligned} \tag{17}$$

where $\hat{\phi}_{vv}(p,k)$ is an estimation of the PSD of $v_{p,k}$, and $\hat{\boldsymbol{\phi}}_{wv}(p,k)$ is a vector composed of the estimated auto- or cross-PSDs between the entries of $\mathbf{w}_{p,k}$ and $v_{p,k}$.

B. Inter-Frame Spectral Subtraction

From (14) and (17), we have for any frame p:

$$\hat{\phi}_{\tilde{y}\tilde{y}}(p,k) - \hat{\phi}_{vv}(p,k) = (\hat{\boldsymbol{\phi}}_{\tilde{z}\tilde{y}}(p,k) - \hat{\boldsymbol{\phi}}_{wv}(p,k))^{\top}\mathbf{g}_k, \tag{18}$$

or alternately:

$$\hat{\phi}_{\tilde{y}\tilde{y}}(p,k) = \hat{\boldsymbol{\phi}}_{\tilde{z}\tilde{y}}(p,k)^{\top}\mathbf{g}_k + \hat{\phi}_{vv}(p,k) - \hat{\boldsymbol{\phi}}_{wv}(p,k)^{\top}\mathbf{g}_k. \tag{19}$$

By subtracting the estimated PSD $\hat{\phi}_{\tilde{y}\tilde{y}}(p,k)$ of one frame, e.g. p_2, from the estimated PSD of another frame, e.g. p_1, we obtain

$$\begin{aligned}\hat{\phi}^{s}_{\tilde{y}\tilde{y}}(p_1,k) &\triangleq \hat{\phi}_{\tilde{y}\tilde{y}}(p_1,k) - \hat{\phi}_{\tilde{y}\tilde{y}}(p_2,k)\\ &= \hat{\phi}^{s}_{yy}(p_1,k) + e_{vv}(p_1,k)\end{aligned} \tag{20}$$

with

$$\begin{aligned}\hat{\phi}^{s}_{yy}(p_1,k) &= \hat{\phi}_{yy}(p_1,k) - \hat{\phi}_{yy}(p_2,k),\\ e_{vv}(p_1,k) &= \hat{\phi}_{vv}(p_1,k) - \hat{\phi}_{vv}(p_2,k).\end{aligned}$$

Applying the same principle to $\hat{\boldsymbol{\phi}}_{\tilde{z}\tilde{y}}(p,k)$, we have:

$$\begin{aligned}\hat{\boldsymbol{\phi}}^{s}_{\tilde{z}\tilde{y}}(p_1,k) &\triangleq \hat{\boldsymbol{\phi}}_{\tilde{z}\tilde{y}}(p_1,k) - \hat{\boldsymbol{\phi}}_{\tilde{z}\tilde{y}}(p_2,k)\\ &= \hat{\boldsymbol{\phi}}^{s}_{zy}(p_1,k) + \mathbf{e}_{wv}(p_1,k),\end{aligned} \tag{21}$$

with

$$\begin{aligned}\hat{\boldsymbol{\phi}}^{s}_{zy}(p_1,k) &= \hat{\boldsymbol{\phi}}_{zy}(p_1,k) - \hat{\boldsymbol{\phi}}_{zy}(p_2,k),\\ \mathbf{e}_{wv}(p_1,k) &= \hat{\boldsymbol{\phi}}_{wv}(p_1,k) - \hat{\boldsymbol{\phi}}_{wv}(p_2,k).\end{aligned}$$

Applying (19) to frames p_1 and p_2 and subtracting the resulting equations, we obtain:

$$\hat{\phi}^{s}_{\tilde{y}\tilde{y}}(p_1,k) = \hat{\boldsymbol{\phi}}^{s}_{\tilde{z}\tilde{y}}(p_1,k)^{\top}\mathbf{g}_k + e(p_1,k), \tag{22}$$

where

$$e(p_1,k) = e_{vv}(p_1,k) - \mathbf{e}_{wv}(p_1,k)^{\top}\mathbf{g}_k. \tag{23}$$

Because $v(n)$ is stationary, $e_{vv}(p_1,k)$ is small. Conversely, the fluctuations of speech signals are much larger than the fluctuations of the noise signal because the speech signals are both non-stationarity and sparse, i.e., speech power spectrum can vary significantly over frames. Thence, by properly

choosing the frame indexes p_1 and p_2, for instance in such a way that the speech power $\hat{\phi}_{yy}(p_1,k)$ is high and the speech power $\hat{\phi}_{yy}(p_2,k)$ is low, we have $\hat{\phi}^s_{yy}(p_1,k) \gg e_{vv}(p_1,k)$, or equivalently $\hat{\phi}^s_{\tilde{y}\tilde{y}}(p_1,k) \gg e_{vv}(p_1,k)$. The same reasoning applies to $\mathbf{e}_{wv}(p_1,k)$, except that the u-v cross-terms of $\mathbf{e}_{wv}(p_1,k)$ are small compared to $\hat{\phi}^s_{\tilde{y}\tilde{y}}(p_1,k)$ either if u and v are uncorrelated, or if u and v are jointly WSS, which are our (quite reasonable) working assumptions.

The choice of the frame index necessitates to classify the frames into two sets, $\mathcal{P}_1$ and $\mathcal{P}_2$, which have high speech power and very low speech power, respectively. This is done in Subsection IV-D using the minimum and maximum statistics of noise spectrum. Before that, we finalize the estimation of the DP-RTF in the noisy case, based on (22).

C. DP-RTF Estimation

Let $P_1 = |\mathcal{P}_1|$ denote the cardinality of $\mathcal{P}_1$. The PSD subtractions (20) and (21) are applied to all the frames $p_1 \in \mathcal{P}_1$ using their corresponding frames $p_2 \in \mathcal{P}_2$, denoted as $p_2(p_1)$. In practice, $p_2(p_1)$ is the frame in $\mathcal{P}_2$ that is nearest to p_1, since the closer the two frames, the smaller the difference of their noise PSD and the difference of their transfer function. The resulting PSDs and cross-PSD vectors are gathered into a $P_1 \times 1$ vector and a $P_1 \times (2Q_k - 1)$ matrix, respectively, as:

$$\hat{\boldsymbol{\phi}}^s_{\tilde{y}\tilde{y}}(k) = [\hat{\phi}^s_{\tilde{y}\tilde{y}}(1,k), \ldots, \hat{\phi}^s_{\tilde{y}\tilde{y}}(p_1,k), \ldots, \hat{\phi}^s_{\tilde{y}\tilde{y}}(P_1,k)]^\top,$$
$$\hat{\boldsymbol{\Phi}}^s_{\tilde{z}\tilde{y}}(k) = [\hat{\boldsymbol{\phi}}^s_{\tilde{z}\tilde{y}}(1,k), \ldots, \hat{\boldsymbol{\phi}}^s_{\tilde{z}\tilde{y}}(p_1,k), \ldots, \hat{\boldsymbol{\phi}}^s_{\tilde{z}\tilde{y}}(P_1,k)]^\top.$$

Let us denote $\mathbf{e}(k) = [e(1,k), \ldots, e(p_1,k), \ldots, e(P_1,k)]^\top$ the $P_1 \times 1$ vector that concatenates the residual noise for the P_1 frames. Then, from (22) we obtain the following linear equation, which is the "noisy version" of (15):

$$\hat{\boldsymbol{\phi}}^s_{\tilde{y}\tilde{y}}(k) = \hat{\boldsymbol{\Phi}}^s_{\tilde{z}\tilde{y}}(k)\mathbf{g}_k + \mathbf{e}(k). \tag{24}$$

Assuming that the sequence of residual noise entries in $\mathbf{e}(k)$ is i.i.d.[2] and also assuming $P_1 \geq (2Q_k - 1)$, the least square solution to (24) is given by:

$$\hat{\mathbf{g}}_k = (\hat{\boldsymbol{\Phi}}^s_{\tilde{z}\tilde{y}}(k)^H \hat{\boldsymbol{\Phi}}^s_{\tilde{z}\tilde{y}}(k))^{-1} \hat{\boldsymbol{\Phi}}^s_{\tilde{z}\tilde{y}}(k)^H \hat{\boldsymbol{\phi}}^s_{\tilde{y}\tilde{y}}(k), \tag{25}$$

where ^{H} denotes matrix conjugate transpose. Finally, the estimation of the DP-RTF d_k defined in (8) is provided by the first element of $\hat{\mathbf{g}}_k$, denoted as $\hat{g}_{0,k}$.

Note that if two frames in $\mathcal{P}_1$ are close to each other, their corresponding elements in vector $\hat{\boldsymbol{\phi}}^s_{\tilde{y}\tilde{y}}(k)$ (or corresponding rows in matrix $\hat{\boldsymbol{\Phi}}^s_{\tilde{z}\tilde{y}}(k)$) will be correlated. This correlation yields some redundancy of the linear equations. However, in practice, we keep this redundancy to make full use of data and give a more robust solution to (24).

[2]This assumption is made to simplify the analysis. In practice, $e(p_1,k)$ may be a correlated sequence because of the possible correlation of $\hat{\phi}_{vv}(p,k)$ (or $\hat{\phi}_{wv}(p,k)$) across frames. Taking this correlation into account would lead to a weighted least square solution to (24), involving a weight matrix in (25). This weight matrix is not easy to estimate, and in practice, (25) delivers a good estimate of $\hat{g}_{0,k}$, as assessed in our experiments.

Still assuming that $e(p_1,k)$ is i.i.d and denoting its variance by σ^2_k, the covariance matrix of $\hat{\mathbf{g}}_k$ is given by [28]:

$$\mathbf{cov}\{\hat{\mathbf{g}}_k\} = \sigma^2_k (\hat{\boldsymbol{\Phi}}^s_{\tilde{z}\tilde{y}}(k)^H \hat{\boldsymbol{\Phi}}^s_{\tilde{z}\tilde{y}}(k))^{-1}. \tag{26}$$

The statistical analysis of the auto- and cross-PSD estimates show that σ^2_k is inversely proportional to the number of smoothing frames D [28]. Thence using a large D leads to a small error variance σ^2_k. However, increasing D decreases the fluctuation of the estimated speech PSD among frames and thus makes the elements in the matrix $\hat{\boldsymbol{\Phi}}^s_{\tilde{z}\tilde{y}}(k)^H \hat{\boldsymbol{\Phi}}^s_{\tilde{z}\tilde{y}}(k)$ smaller, which results in a larger variance of $\hat{\mathbf{g}}_k$. Therefore, an appropriate value of D should be chosen to achieve a good tradeoff between smoothing the noise spectrum and preserving the fluctuation of speech spectrum.

Finally, to improve the robustness of the DP-RTF estimation, we also calculate (25) after exchanging the roles of the two channels in the whole process. This delivers an estimate $\hat{g}'_{0,k}$ of the inverse of (8), i.e. an estimate of the inverse DP-RFT $\frac{a_{0,k}}{b_{0,k}}$. Both $\hat{g}_{0,k}$ and ${\hat{g}'_{0,k}}^{-1}$ are estimates of $\frac{b_{0,k}}{a_{0,k}}$. The final DP-RTF estimate is given by averaging these two estimates as:

$$\hat{c}_k = \frac{1}{2}(\hat{g}_{0,k} + {\hat{g}'_{0,k}}^{-1}). \tag{27}$$

D. Frame Classification

We adopt the minimum-maximum statistics for frame classification, which was first introduced in [12], and is applied to a different feature in this paper. Frame classification is based on the estimation of $\tilde{y}$ PSD, i.e., $\hat{\phi}_{\tilde{y}\tilde{y}}(p,k)$. The frame p_1 is selected such that $\hat{\phi}^s_{\tilde{y}\tilde{y}}(p_1,k)$ in (22) is large compared to $e(p_1,k)$, and thus (22) matches well the noise-free case.

As shown in (17), the PSD estimation $\hat{\phi}_{\tilde{y}\tilde{y}}(p,k)$ is composed of both speech and noise powers. A minimum statistics formulation was proposed in [29], where the minimum value of the smoothed periodograms with respect to the index p, multiplied by a bias correction factor, is used as the estimation of the noise PSD. Here we introduce an equivalent sequence length for analyzing the minimum and maximum statistics of noise spectra, and propose to use two classification thresholds (for two classes $\mathcal{P}_1$ and $\mathcal{P}_2$) defined from the ratios between the maximum and minimum statistics. In short, we classify the frames by using the minimum controlled maximum border.

Formally, the noise power in $\hat{\phi}_{\tilde{y}\tilde{y}}(p,k)$ is

$$\xi_{p,k} \triangleq \hat{\phi}_{vv}(p,k) = \frac{1}{D}\sum_{d=0}^{D-1} |v_{p-d,k}|^2. \tag{28}$$

For a stationary Gaussian signal, the probability density function (PDF) of periodogram $|v_{p,k}|^2$ obeys the exponential distribution [29]

$$f(|v_{p,k}|^2;\lambda) = \frac{1}{\lambda} e^{-|v_{p,k}|^2/\lambda} \tag{29}$$

where $\lambda = E\{|v_{p,k}|^2\}$ is the noise PSD. Assume that the sequence of $|v_{p,k}|^2$ values at different frames are i.i.d. random variables. The averaged periodogram $\xi_{p,k}$ obeys the Erlang

distribution [30] with scale parameter $\mu = \lambda/D$ and shape parameter D:

$$f(\xi_{p,k}; D, \mu) = \frac{\xi_{p,k}^{D-1} e^{-\frac{\xi_{p,k}}{\mu}}}{\mu^D (D-1)!}. \tag{30}$$

We are interested in characterizing and estimating the ratio between the maximum and minimum statistics of the sequence $\xi_{p,k}$. Since the maximum and minimum statistics are both linearly proportional to μ [29], we assume, without loss of generality, that $\mu = 1$. Consequently the mean value of $\xi_{p,k}$ is equal to D.

As mentioned in Section III-A, the frame index of the estimated PSDs $\hat{\phi}_{yy}(p,k)$ and $\xi_{p,k}$ is confined to the range p_f to P. Let R denote the increment of the frame index p of the estimated PSDs. If R is equal to or larger than D, for two adjacent estimated PSD $\xi_{p,k}$ and $\xi_{p+R,k}$, there is no frame overlap. The sequence $\xi_{p,k},\ p = p_f : R : P$ is then an independent random sequence. The length of this sequence is $\tilde{P} = \lceil \frac{P - p_f + 1}{R} \rceil$. The PDFs of the minimum and maximum of these $\tilde{P}$ independent variables are [31]:

$$\begin{aligned} f_{min}(\xi) &= \tilde{P} \cdot (1 - F(\xi))^{\tilde{P}-1} \cdot f(\xi), \\ f_{max}(\xi) &= \tilde{P} \cdot F(\xi)^{\tilde{P}-1} \cdot f(\xi), \end{aligned} \tag{31}$$

where $F(\cdot)$ denotes the cumulative distribution function (CDF) associated with the PDF (30). Conversely, if $R < D$, $\xi_{p,k}$ is a correlated sequence, and the correlation coefficient is linearly proportional to the frame overlap. For this case, (31) will not be valid anymore. Based on a large amount of simulations using white Gaussian noise (WGN),[3] it was found that the following approximate equivalent sequence length

$$\tilde{P}' = \frac{\tilde{P} R}{D} \cdot \left(1 + \log\left(\frac{D}{R}\right)\right) \tag{32}$$

can replace $\tilde{P}$ in order to make (31) valid for the correlated sequence. We observe that the ratio between the number D of frames used for spectrum averaging and the frame increment R of PSD estimates, is replaced with its logarithm. Note that this is an empirical result, for which theoretical foundation remains to be investigated.

Then, the expectation of the minimum can be approximately computed as

$$\bar{\xi}_{min} \approx \frac{\sum_{\xi_i} \xi_i \cdot f_{min}(\xi_i)}{\sum_{\xi_i} f_{min}(\xi_i)}, \tag{33}$$

where $\xi_i \in \{0, 0.1D, 0.2D, \ldots, 3D\}$ is a grid used to approximate the integral operation, which well covers the support of the Erlang distribution with shape D and scale 1. Similarly, the CDF of the maximum can be estimated as

$$F_{max}(\xi) \approx \sum\nolimits_{\xi_i} f_{max}(\xi_i). \tag{34}$$

Finally, we define two classification thresholds that are two specific values of the maximum and minimum ratios, namely

$$r_1 = \frac{\xi_{F_{max}(\xi)=0.95}}{\bar{\xi}_{min}}, \text{ and } r_2 = \frac{\xi_{F_{max}(\xi)=0.5}}{\bar{\xi}_{min}}, \tag{35}$$

where $\xi_{F_{max}(\xi)=0.95}$ and $\xi_{F_{max}(\xi)=0.5}$ are the values of ξ for which the CDF of the maximum is equal to 0.95 and 0.5, respectively. Classes $\mathcal{P}_1$ and $\mathcal{P}_2$ are then obtained with

$$\mathcal{P}_1 = \{p \mid \xi_{p,k} > r_1 \cdot \min_p\{\xi_{p,k}\}\}, \tag{36}$$

$$\mathcal{P}_2 = \{p \mid \xi_{p,k} \leq r_2 \cdot \min_p\{\xi_{p,k}\}\}. \tag{37}$$

These two thresholds are set to ensure that the frames in $\mathcal{P}_1$ contain large speech power and the frames in $\mathcal{P}_2$ contain negligible speech power. The speech power for the other frames are probabilistically uncertain, making them unsuitable for either $\mathcal{P}_1$ or $\mathcal{P}_2$. Using two different thresholds evidently separates speech region and noise-only region. In other words, there is a low probability to have a frame classified into $\mathcal{P}_1$ in the proximity of $\mathcal{P}_2$ frames, and vice versa. Therefore, in general, the PSD of a frame in $\mathcal{P}_1$ is estimated using D frames that are not included in the noise-only region, and vice versa. Note that if there are no frames with speech content, e.g., during long speech pauses, class $\mathcal{P}_1$ will be empty with a probability of 0.95 due to threshold r_1.

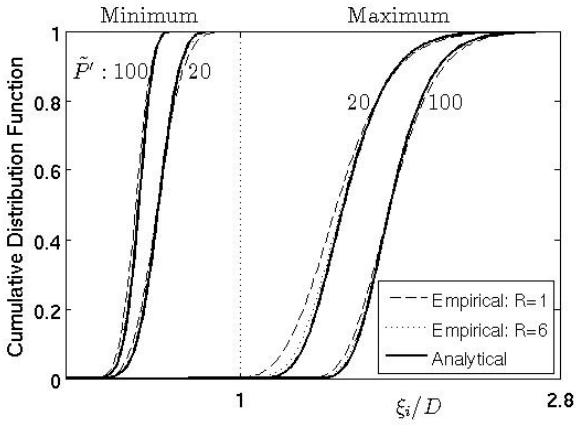

Fig. 1: Cumulative distribution function (CDF) of the minimum and maximum statistics of $\xi_{p,k}$ for $D = 12$.

As an illustration of (32), Fig. 1 shows the CDF for $D = 12$. The empirical curves are simulated using WGN, and the analytical curves are computed using the equivalent sequence length in (32). The minimum CDF and maximum CDF of two groups of simulations are shown, for which the equivalent sequence lengths $\tilde{P}'$ are fixed at 20 and 100, respectively. For each equivalent sequence length $\tilde{P}'$, two empirical curves with frame increment $R = 1$ and $R = 6$ are simulated using WGN, whose corresponding original sequence lengths are $\tilde{P} = 69$ and $\tilde{P} = 24$ for $\tilde{P}' = 20$, and $\tilde{P} = 344$ and $\tilde{P} = 118$

[3]The simulations are done with the following procedure: applying STFT to a number of WGN signals with identical long duration. For each time-frequency bin, estimate the PSD by averaging the periodograms of the past D frames. Without loss of generality, the scale parameter μ of the PSD estimation can be set to 1 by adjusting the noise PSD λ to D. A sequence of correlated PSD estimates is generated by picking PSD estimates from the complete sequence, with frame increment R (with $R < D$). The length of the correlated sequence is $\tilde{P}$. The minimum/maximum values of each correlated sequence are collected at each frequency for all the WGN signals. The PDF and CDF of the minimum/maximum statistics are simulated by the histograms of these minimum/maximum values. Fig. 1 shows some examples of this empirical CDF.

for $\tilde{P}' = 100$, respectively. This shows that the equivalent sequence length in (32) is accurate for the minimum and maximum statistics.

V. Sound Source Localization method

The amplitude and the phase of DP-RTF represent the amplitude ratio and phase difference between two source-to-microphone direct-path ATFs. In other words, in case of two microphones, the DP-RTF is equivalent to the interaural cues, ILD and IPD, associated to the direct path. More generally, we consider here J microphones. This is a slight generalization that will directly exploit the previous developments, since we consider these J microphones pair-wise. As in [32], [33], we consider the normalized version of the DP-RTF estimate (27) between microphones i and j:

$$c_{k,ij} = \frac{\hat{c}_{k,ij}}{\sqrt{1+|\hat{c}_{k,ij}|^2}}. \tag{38}$$

Compared to the amplitude ratio, the normalized DP-RTF is more robust. In particular, when the reference transfer function $a_{0,k}$ is much smaller than $b_{0,k}$, the amplitude ratio estimation is sensitive to noise present in the reference channel. By concatenating (38) across K frequencies and across $(J-1)J/2$ microphone pairs, we obtain a high-dimensional feature vector $\mathbf{c} \in \mathbb{R}^{J(J-1)K/2}$. Since speech signals have a sparse STFT representation, we denote by $\boldsymbol{h} \in \mathbb{C}^{J(J-1)K/2}$ an indicator vector whose elements are either equal to 1 if the energy at the corresponding frequency is significant, or equal to 0 if the energy is negligible. In practice, the indicator vector entries at a given frequency k are set to 0 if the corresponding matrix $\hat{\mathbf{\Phi}}^{s}_{\tilde{z}\tilde{y}}(k)$ is underdetermined, i.e. $P_1 < (2Q_k - 1)$ for that frequency. This way, we do not use any DP-RTF calculated from (25) for such "missing frequency" (see below).

The proposed DP-RTF estimation method is suitable for the most general case of microphone setup where the microphones are not necessarily placed in free-field. In other words it can be applied to any microphone pair in any microphone array setup. For instance, in the present paper, the microphones are placed in the ears of a dummy head or on the head of a robot. In these cases, there is no clear (analytical) relationship between the HRIR/HRTF/DP-RTF and the DOA of the emitting source, even after removal of the noise and reverberations. In order to perform SSL based on the feature vector $\mathbf{c}$, we adopt here a supervised framework: A training set $D_{\mathbf{c},\mathbf{q}}$ of I pairs $\{\mathbf{c}_i, \mathbf{q}_i\}_{i=1}^{I}$ is available, where $\mathbf{c}_i$ is a DP-RTF feature vector generated with an anechoic head-related impulse response (HRIR), and $\mathbf{q}_i$ is the corresponding source-direction vector. Then, for an observed (test) feature vector $\mathbf{c}$ that is extracted from the microphone signals, the corresponding direction is estimated using either (i) nearest-neighbor search in the training set (considered as a look-up table) or (ii) a regression whose parameters have been tuned from the training set. Note that the training set and the observed test features should be recorded using the same microphone set-up. This way, the HRIR of the training set (corresponding to an anechoic condition) corresponds to the direct-path of the BRIR of the test condition (recorded in reverberant condition).

Nearest-neighbor search corresponds to solving the following minimization problem ($\odot$ denotes the Hadamard product, i.e. entry-wise product):

$$\hat{\mathbf{q}} = \underset{i\in[1,I]}{\operatorname{argmin}} \parallel \boldsymbol{h} \odot (\mathbf{c} - \mathbf{c}_i) \parallel . \tag{39}$$

As mentioned above, the indicator vector $\boldsymbol{h}$ enables to select the relevant DP-RTF vector components, i.e. the ones corresponding to frequencies with (over)determined solution to (24). Because of the sparse nature of the test feature vectors, not any regression technique could be used. Indeed, one needs a regression method that allows training with full-spectrum signals and testing with sparse-spectrum signals. Moreover, the input DP-RTF vectors are high dimensional and not any regression method can handle high-dimensional input data. For these reasons we adopted the probabilistic piece-wise linear regression technique of [6].

VI. Experiments with Simulated Data

We report results with experiments carried out in order to evaluate the performance of the proposed method. We simulated various experimental conditions in terms of reverberation and additive noise.

A. The Dataset

The BRIRs are generated with the ROOMSIM simulator [34] and with the head related transfer function (HRTF) of a KEMAR dummy head [35]. The responses are simulated in a rectangular room of dimension 8 m $\times$ 5 m $\times$ 3 m. The KEMAR dummy head is located at $(4, 1, 1.5)$ m. The sound sources are placed in front of the dummy head with azimuths varying from -90° to 90°, spaced by 5°, an elevation of 0°, and distances of 1 m, 2 m, and 3 m., see Fig.2.

The absorption coefficients of the six walls are equal, and adjusted to control T_{60} at 0.22 s, 0.5 s and 0.79 s, respectively. Two other quantities, i.e. the ITDG and the direct-to-reverberation ratio (DRR), are also important to measure the intensity of the reverberation. In general, the larger the source-to-sensors distance is, the smaller the ITDG and DRR are. For example, when T_{60} is 0.5 s, the DRRs for 1, 2, 3 m are about 1.6, -4.5 and -8.1 dB, respectively. Speech signals from the TIMIT dataset [36] are used as the speech source signals, which are convolved with the simulated BRIRs to generate the sensor signals. Each BRIR is convolved with 10 different speech signals from TIMIT to achieve reliable SSL results. Note that the elevation of the speech sources is always equal to 0° in the BRIR dataset, thence in these simulated-data experiments the source direction corresponds to the azimuth only. The feature vectors in the training set $\{\mathbf{c}_i\}_{i=1}^{I}$ are generated with the anechoic HRIRs of the KEMAR dummy head from the azimuth range $[-90^\circ, 90^\circ]$, spaced by 5°, i.e. $I = 37$. In this section, the nearest-neighbor search is adopted for localization.

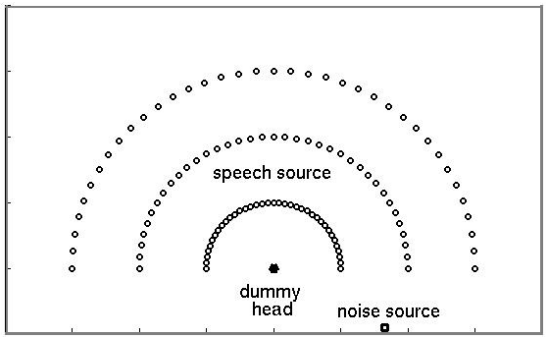

Fig. 2: Configurations of room, dummy head, speech sources and noise source for the BRIR dataset.

Two types of noise signals are generated: (i) a "directional noise" is obtained by convolving a single channel WGN signal with a BRIR corresponding to position beside the wall with azimuth of $120°$, elevation of $30°$ and distance of 2.2 m, see Fig. 2; (ii) an "uncorrelated noise" consists of an independent WGN signal on each channel. Noise signals are added to the speech sensor signals with various signal-to-noise ratios.

B. Setting the Parameters

The sampling rate is 16 kHz. Only the frequency band from 0 to 4 kHz is considered for speech source localization. The setting of all three parameters N, Q_k and D is crucial for a good estimation of the DP-RTF. Intuitively, Q_k should correspond to the value of T_{60} at the k-th frequency bin. For simplicity, we set Q_k to be the same for all frequencies and denote it as Q. In the following of this subsection, we present preliminary SSL experiments that were done in order to tune N, Q and D to an "optimal tradeoff" setting that would ensure good SSL performance for a large range of acoustic conditions. Since considering all possible joint settings of these three parameters is a hard task, when exploring the setting of one of them, we may fix the others.

In all the following, the localization error is taken as the performance metric. It is computed by averaging the absolute errors between the localized directions and their corresponding ground truth (in degrees) over the complete test dataset.

Let us first consider the setting of Q. Here we fix $N = 256$ with 50% overlap, and $D = 12$. Table I shows the localization errors for Q values corresponding to CTF length $\in [0.1T_{60}, \ldots, 0.4T_{60}]$ with $T_{60} = 0.5$ s. When the SNR is high (first four lines; SNR = 10 dB), the influence of noise is small, and the DRR plays a dominant role. Comparing the localization errors for source-to-sensors distances between 1 m and 2 m, we see that small localization errors are obtained with rather small Q values for 1 m, and with the larger Q values for 2 m. This result indicates that, for a given T_{60}, Q should be increased when the DRR is decreased. The CTF should cover most of the energy of the room impulse response. By comparing the results for the uncorrelated noise of 10 dB and -5 dB, source at 2 m (second and fifth lines), we observe that the smallest localization error is achieved by a smaller Q for the low SNR case, compared to the high SNR case. Note that a larger Q corresponds to a greater model complexity, which needs more reliable (less noisy) data to be estimated. The intense uncorrelated noise degrades the data, thence a small Q is preferred. In contrast, for the directional noise, a large Q is also suitable for the low SNR case (sixth line). The reason is possibly that the directional noise signal has a similar convolution structure as the speech signal, and the noise residual $\mathbf{e}(k)$ also has a similar convolution structure. Thence the data reliability is not degraded much. In conclusion, the optimal Q varies with the T_{60}, DRR, noise characteristics, and noise intensity. In practice, it is difficult to obtain these features automatically, thence we assume that T_{60} is known, and we set Q to correspond to $0.25T_{60}$ as a tradeoff for different acoustic conditions.

Let us now consider the setting of D. Here, we set Q to correspond to $0.25T_{60}$, and $N = 256$ with 50% overlap. The number of frames D is crucial for an efficient spectral subtraction (Section IV-B). A large D yields a small noise residual. However, the remaining speech power after spectral subtraction may also be small because of the small fluctuations of the speech PSD estimate between frames when D is large. Table II shows the localization errors for $D \in [6, \ldots, 20]$ under different conditions. Note that only the results for the low SNR case (-5 dB) are shown, for which the effect of noise suppression plays a more important role. It can be seen (first line) that a large D yields the smallest localization error, which means that removing noise power is more important than retaining speech power for this condition. The reason is that the DRR is large for source-to-sensors distance of 1 m, so that the direct-path speech power is relatively large. As D increases, the remaining direct-path speech power decreases only slightly, compared to the decrease of the noise residual. In contrast, a small D yields the smallest localization error for the directional noise at 2 m (fourth line), which means that retaining speech power is more important than removing noise power for this condition. The reasons are that (i) as described above, the data reliability is not degraded much by the directional noise in the sense of convolution, and (ii) the direct-path speech power is relatively small for a source-to-sensors distance of 2 m. The conditions of the second and third lines fall in between the first line and the fourth line, and these results do not strongly depend on D. It is difficult to choose a D value that is optimal for all the acoustic conditions. In the following, we set $D = 12$ frames (100 ms) as a fair tradeoff.

As for the setting of N, let us remind that the reflections present in $a(n)|_{n=0}^{N}$ lead to a biased definition of DP-RTF. In order to minimize the reflections contained in $a(n)|_{n=0}^{N}$, the STFT window length N should be as small as possible, while still capturing the direct-path response. However, in practice, a small N requires a large Q for the CTF to cover well the room impulse response, which increases the complexity of the DP-RTF estimate. We tested the localization performance for three STFT window sizes: 8 ms ($N = 128$ samples), 16 ms ($N = 256$ samples), and 32 ms ($N = 512$ samples), with 50% overlap. Again, Q corresponds to $0.25T_{60}$. For example, with $T_{60} = 0.79$ s and with $N = 128, 256, 512$ respectively, Q is

TABLE I: Localization errors (degrees) for different values of Q in different conditions. $T_{60} = 0.5$ s. "Distance" stands for source-to-sensors distance. The bold value is the minimum localization error for each condition.

Conditions			Q/T_{60} ($T_{60} = 0.5$ s)						
Noise type	SNR	Distance	0.1	0.15	0.2	0.25	0.3	0.35	0.4
Uncorrelated	10 dB	1 m	0.122	0.081	**0.077**	0.081	0.099	0.108	0.113
Uncorrelated	10 dB	2 m	1.338	0.847	0.716	0.649	0.629	0.608	**0.568**
Directional	10 dB	1 m	0.135	**0.113**	0.122	0.131	0.149	0.158	0.162
Directional	10 dB	2 m	1.437	0.869	0.829	0.680	0.644	0.626	**0.617**
Uncorrelated	-5 dB	2 m	7.824	6.833	6.703	**6.680**	6.802	6.964	7.149
Directional	-5 dB	2 m	13.36	12.25	11.90	11.23	10.96	10.52	**10.38**

TABLE II: Localization errors (degrees) for different values of D in different conditions. $T_{60} = 0.5$ s. "Distance" stands for source-to-sensors distance. The bold value is the minimum localization error for each condition.

Conditions			D frames							
Noise type	SNR	Distance	6	8	10	12	14	16	18	20
Uncorrelated	-5 dB	1 m	2.59	2.15	2.09	1.99	1.86	1.81	1.64	**1.59**
Uncorrelated	-5 dB	2 m	7.37	6.03	6.17	6.68	6.08	6.40	6.90	6.50
Directional	-5 dB	1 m	3.83	3.42	3.51	3.23	3.70	3.47	2.96	3.45
Directional	-5 dB	2 m	**9.80**	10.28	10.32	11.23	11.60	13.18	13.62	15.35

TABLE III: Localization errors (degrees) for three values of N. "Distance" is the sensors-to-source distance. The bold value is the minimum localization error. In this experiment, the noise signal is generated by summing the directional noise and uncorrelated noise with identical powers.

Conditions			STFT window length N		
SNR	Distance	T_{60}	128 (8 ms)	256 (16 ms)	512 (32 ms)
10 dB	1 m	0.22 s	**0.01**	**0.01**	0.02
	3 m	0.22 s	**0.58**	1.19	1.89
	3 m	0.79 s	9.60	**9.22**	9.55
-5 dB	1 m	0.22 s	1.89	1.62	**1.49**
	3 m	0.22 s	8.07	**6.30**	7.04
	3 m	0.79 s	22.66	20.81	**17.75**

equal to 50, 25, 13 frames respectively. D is set to 100 ms. For $N = 128$, 256, 512, D is 24, 12, 6 frames, respectively. Table III shows the localization errors under various acoustic conditions. We first discuss the case of high SNR (first three lines). When the source-to-sensors distance is small (1 m; first line), the ITDG is relatively large and we observe that $N = 128$ and $N = 256$ (8 ms and 16 ms windows) achieve comparable performance. This indicates that, if the ITDG is relatively large, there are not much more reflections in $a(n)|_{n=0}^{N}$ for a 16-ms window, compared with an 8-ms window. The next results (second line) show that, when T_{60} is small (0.22 s), the localization performance decreases much more for a 16-ms and a 32-ms window than for an 8-ms window, as the sensor-to-noise distance increases from 1 m to 3 m. A lower ITDG yields a larger DP-RTF estimation error due to the presence of more reflections in $a(n)|_{n=0}^{N}$. When T_{60} increases to 0.79 s, Q becomes larger, especially for $N = 128$. It can be seen (third line) that here $N = 256$ yields a better performance than other values. This is because the lack of data leads to a large DP-RTF estimation error for $N = 128$, and the reflections in $a(n)|_{n=0}^{N}$ bring a large DP-RTF estimation error for $N = 512$. When the SNR is low (-5 dB; last three lines), less reliable data are available due to noise contamination. In that case, a large N achieves the best performance. Finally, we set $N = 256$ (16-ms STFT window) as a good overall tradeoff between all tested conditions.

C. DP-RTF Estimation

We provide several representative examples showing the influence of both reverberation and noise on the DP-RTF estimates. The phase and normalized amplitude of the estimated DP-RTF for three acoustic conditions are shown in Fig. 3. The SNR is set to 30 dB in the first two examples, hence the noise is negligible. The difference between the estimated and the ground-truth phase is referred to as the phase estimation error. It can be seen that, for most frequency bins, the mean value (over ten trials) of the phase estimation error is very small (but nonzero, which indicates that the estimated DP-RTF is biased). As mentioned above, the bias is brought in by the reflections in the impulse response segment $a(n)|_{n=0}^{N}$. In addition, if the DRR gets smaller, a longer CTF is required to cover the room impulse response. However, for a given T_{60}, the CTF length Q is set as a constant, for instance $0.25T_{60}$. In this example, this improper value of Q leads to an inaccurate CTF model, which causes the DP-RTF estimate bias. When the source-to-sensors distance increases, both the ITDG and DRR become smaller. Therefore, for both phase and amplitude, the estimation bias of the second example of Fig. 3 (middle) is larger than the bias of the first example (left). Moreover, the DP-RTF $\frac{b_{0,k}}{a_{0,k}}$ in $\mathbf{g}_k$ plays a less important role relative to other elements, with decreasing DRR, which makes the variance of both the phase and amplitude estimation errors to be larger than in the first example. By comparing the first and last examples of Fig. 3, it is not surprising to observe that the estimation error increases as noise power increases. When the SNR is low, less reliable speech frames are available in the high frequency band, due to the intense noise. Therefore, there is no DP-RTF estimation for the frequency bins satisfying $P_1 < 2Q_k - 1$.

D. Baseline Methods

In our previous work [12], the proposed inter-frame spectral subtraction scheme was applied to RTF estimators (as opposed to the DP-RTF estimators proposed in the present paper). The results were compared with the RTF estimators proposed in [9] and [11] in the presence of WGN or babble noise. The

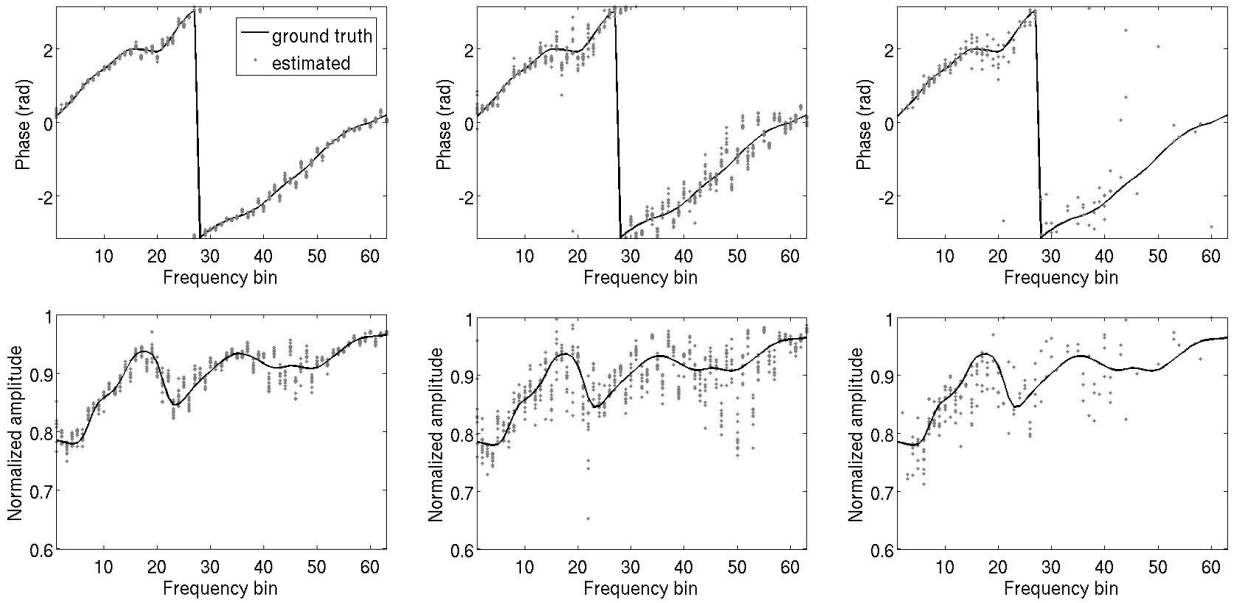

Fig. 3: The phase (top) and normalized amplitude (bottom) of the normalized estimated DP-RTF (38) as a function of frequency bins. The source direction is 30°. $T_{60} = 0.5$ s. The continuous curve corresponds to the ground-truth DP-RTF d_k computed from the anechoic HRTF. Left: 1 m source-to-sensors distance, 30 dB SNR. Middle: 2 m source-to-sensors distance, 30 dB SNR. Right: 1 m source-to-sensors distance, 0 dB SNR. For each acoustic condition, the BRIR is convolved with 10 different speech recordings as the sensor signals, whose DP-RTF estimations are all shown. In this experiment, the noise signal is generated by summing the directional noise and uncorrelated noise with identical powers.

efficiency of the inter-frame spectral subtraction to remove the noise was demonstrated. Thence, the focus of the present set of experiments is mainly aimed at (i) comparing the robustness to reverberation of the proposed DP-RTF feature with respect to other features, in a similar SSL framework, and at (ii) comparing the proposed SSL method with a conventional SSL method.

To this aim, we compare our method with three other methods: (i) an unbiased RTF identification method [12], in which a spectral subtraction procedure (similar to the one described in Section IV-B) is used to suppress noise. Since this RTF estimator is based on the MTF approximation, we refer to this method as RTF-MTF. (ii) a method based on a STFT-domain coherence test (CT) [22].[4] We refer to this method as RTF-CT. The coherence test is used in [22] to search the rank-1 time-frequency bins which are supposed to be dominated by one active source. We adopt the coherence test for single speaker localization, in which one active source denotes the direct-path source signal. The TF bins that involve notable reflections have low coherence. We first detect the maximum coherence over all the frames at each frequency bin, and then set the coherence test threshold for each frequency bin to 0.9 times its maximum coherence. In our experiments, this threshold achieves the best performance. The covariance matrix is estimated by taking a 120 ms (15 adjacent frames) averaging. The auto- and cross-PSD spectral subtraction is applied to the frames that have high speech power and a coherence larger than the threshold, and then are averaged over frames for RTF estimation. (iii) a classic one-stage algorithm: the steered-response power (SRP) utilizing the phase transform (PHAT) [39], [40]. The azimuth directions $-90^\circ : 5^\circ : 90^\circ$ are taken as the steering directions, and their HRIRs are used as the steering responses.

Note that for both RTF-MTF and RTF-CT methods, the features used in the SSL are obtained after the inter-frame spectral subtraction procedure. The SSL method presented in Section V is adopted. The training set used as a look-up table or used for training the regression is the same as for the DP-RTF.

E. Localization Results

Fig. 4 shows the localization results in terms of localization error (let us remind that this error is an average absolute error between the localized directions and their corresponding ground truth (in degrees) over the complete test dataset). Note that in real world, directional noise source, e.g. fan, refrigerator, etc., and diffuse background noise co-exist. Thence in this experiment, the noise signal was generated by summing the directional noise and uncorrelated noise with identical powers.

Let us first discuss the localization performance shown in Fig. 4-top for $T_{60} = 0.22$ s. When the DRR is high (1 m source-to-sensors distance; solid-line), compared with the proposed method, RTF-MTF has a comparable performance under high SNR conditions, and a slightly better performance

[4]Note that [21] introduces a similar technique based on interaural coherence, using features extracted from band-pass filter banks. Also, a binaural coherent-to-diffuse ratio approach was proposed in [37], [38] and applied to dereverberation but not to SSL.

under low SNR conditions (lower than 0 dB). This indicates that when the reverberation is low, the MTF approximation is valid. When less reliable data are available (under low SNR conditions), the proposed method perform slightly worse than RTF-MTF due to its greater model complexity. Note that both the RTF-MTF and the proposed DP-RTF methods achieve very good localization performance: The localization error goes from almost 0° at SNR = 10 dB to about 5° at SNR = -10 dB. RTF-CT achieves the worst performance. This indicates that when the direct-path impulse response is slightly contaminated by the reflections, employing all the data (as done by RTF-MTF and DP-RTF) obtains a smaller localization error than employing only the data selected by the coherence test. In general, for mild reverberations, the performance gap between RTF-MTF, RTF-CT and the proposed method is small and the noise level plays a decisive role for good localization.

The SRP-PHAT method achieves comparable performance measures with the three other methods when the SNR is high (10 dB). However, the performance measures of SRP-PHAT degrades immediately and dramatically when the SNR decreases. The steered-response power is severely influenced by intense noise, especially by the directional noise. This indicates that the inter-frame spectral subtraction algorithm applied to RTF-MTF, RTF-CT and the proposed method is efficient to reduce the noise.

When the DRR decreases (2 m source-to-sensors distance, grey lines; 3 m source-to-sensors distance, dashed lines), the performance measures of RTF-MTF degrades notably. For SNR = 10 dB, the localization error of RTF-MTF increases from 0.07° to 1.51° and to 6.35° for source-to-sensors distances of 1 m, 2 m and 3 m, respectively. The direct-path impulse response is severely contaminated by the reflections. At high SNRs, RTF-CT performs slightly better than RTF-MTF. Indeed, RTF-CT selects the frames that contain less reverberations for calculating the RTF estimate, which improves the performance at high SNR conditions. However, when the noise level increases, the precision of RTF-CT also degrades. The performance of RTF-CT is influenced not only by the residual noise but also by the decline of the coherence test precision, which make it fall even faster than RTF-MTF with decreasing SNR (it has a larger localization error at -5 dB and -10 dB).

The proposed method also has a larger localization error when the source-to-sensors distance increases: the DP-RTF estimation is possibly influenced by the increased amount of early reflections in the impulse response segment $a(n)|_{n=0}^{N}$, by the effect of an improper Q setting, and by the decreased importance of $\frac{b_{0,k}}{a_{0,k}}$ in vector $\mathbf{g}_k$. However, the performance of the proposed DP-RTF method degrades much slower than the ones of RTF-MTF when the source distance increases. For an SNR of 10 dB, the localization error of the proposed method increases from 0.06° to 0.16° and 1.19° as the source-to-sensors distance increases from 1 m to 2 m and 3 m. It can be seen that the performance of the proposed method also falls faster than RTF-MTF with decreasing SNR, since the available data is less reliable. The localization error of the proposed method is larger than the MTF error at -10 dB. It is observed that the proposed method prominently outperforms RTF-CT. It is shown in [23] that the coherence test is influenced by the coherent reflections (very early reflections) of the source signal. Moreover, it is difficult to automatically set a coherence test threshold that could perfectly select the desired frames.

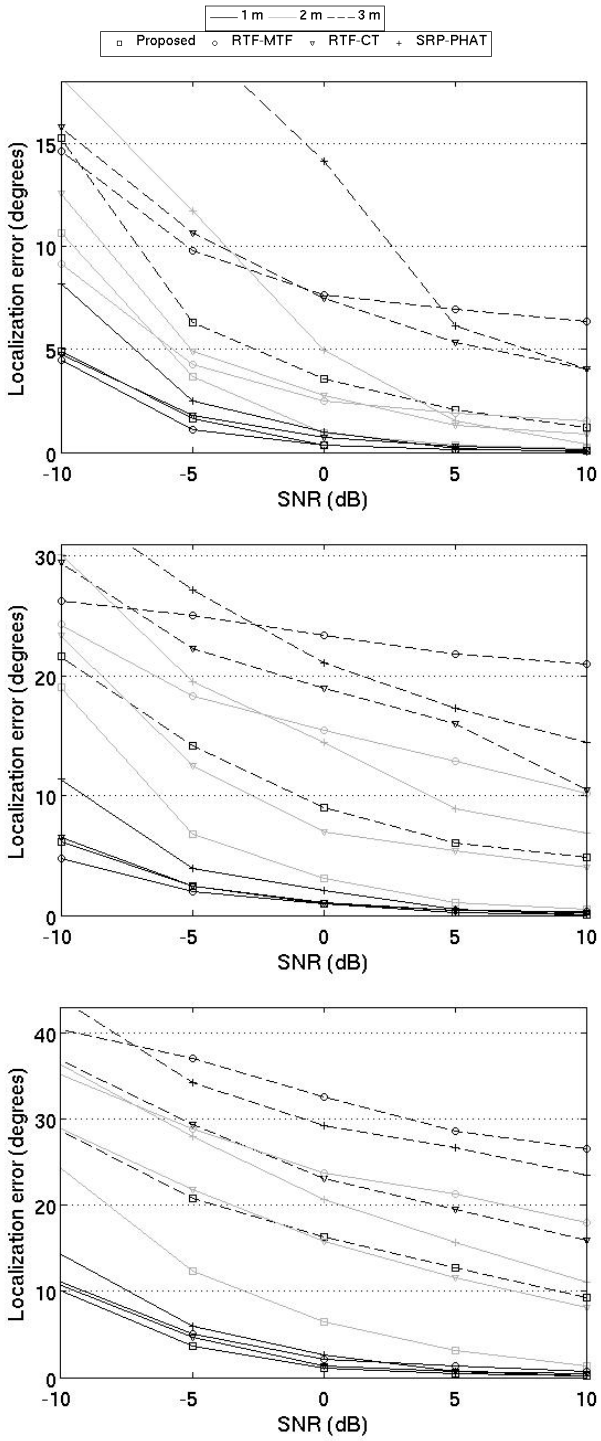

Fig. 4: Localization errors under various reverberation and noise conditions. Top: $T_{60} = 0.22$ s. Middle: $T_{60} = 0.5$ s. Bottom: $T_{60} = 0.79$ s. The localization errors are shown as a function of SNR for source-to-sensors distances of 1 m, 2 m and 3 m.

Many frames that have a coherence larger than the threshold include reflections.

The performance of SRP-PHAT also degrades with the DRR decrease. It is known that PHAT-based method are quite sensible to reverberations and noise in general. Briefly, the performance measures of SRP-PHAT are in between the performance measures of RTF-MTF and RTF-CT for high SNRs, which indicates that the PHAT weight could suppress the reverberations only to a certain extent. Below 5 dB, SRP-PHAT performs worst of the four methods.

Fig. 4 (bottom) displays the results for $T_{60} = 0.79$ s. Obviously, the performance measures of all four methods degrade as T_{60} increases. Indeed, the MTF approximation is not accurate; there are only a few time-frequency bins with a rank-1 coherence; and a large value of Q has to be utilized in the proposed method, for which there may not always be enough reliable data. Here, it can be seen that RTF-CT performs better than RTF-MTF for any SNR value and source-to-sensors distance. Even SRP-PHAT performs better than RTF-MTF (for 2 m and 3 m source-to-sensors distance). This shows that the RTF estimation error brought by the MTF approximation largely increases as T_{60} increases. For 1 m source-to-sensors distance, the proposed method performs slightly better than all other three methods. For 2 m and 3 m source-to-sensors distance, the proposed method largely outperforms the other three methods, at all SNRs. For example, at SNR = 0 dB, the proposed method achieves about 6.5° of localization error at 2 m source-to-sensors distance, while RTF-CT (the best of the three baseline methods) achieves about 15.8°, hence the gain for the proposed method over the best baseline is about 9.3°. However, the performance of the proposed method and of RTF-CT still have a faster degradation with decreasing SNR compared to RTF-MTF.

Finally, we can see from Fig. 4 (middle), that the performance of the different methods for $T_{60} = 0.5$ s falls in between the other two cases shown on the same figure, and the trends of performance evolution with T_{60} is consistent with our comments above.

In summary, the proposed method outperforms the three other methods under most acoustic conditions. In a general manner, the gain over the baseline methods increases as the source-to-sensors distance increases (or the DRR decreases) and as the reverberation time increases (but the influence of the noise level is more intricate). As a result, the proposed method achieves acceptable localization performance in quite adverse conditions. For example (among many others), with $T_{60} = 0.5$ s, source-to-sensors distance of 3 m and an SNR of 0 dB, the localization error is about 9°, and with $T_{60} = 0.79$ s, source-to-sensors distance of 2 m, and an SNR of 0 dB, the localization error is about 6.5°.

In all the above results, the duration of the signal used for localization was not considered with great attention: The localization errors were averaged over 10 sentences of TIMIT of possibly quite different duration, from 1 s to 5 s. Yet the number of available frames that are used to construct (24) depends on the speech duration, which is crucial for the least square DP-RTF estimation in (25). Here we complete the simulation results with a basic test of the influence of the speech duration on localization performance. To this aim we classified our TIMIT test sentences according to their duration (closer to 1 s, 2 s, 3 s or 4 s) and proceeded to localization evaluation for each new group (of 10 sentences), for a limited set of acoustic conditions (SNR = 10 dB and 0 dB, $T_{60} = 0.5$ s). Table IV shows the localization errors of the proposed method, the RTF-MTF, and the RTF-CT method, for the four tested approximate speech durations. We can see that, as expected, all three methods achieve a smaller localization error when increasing speech duration, for both tested SNRs. The improvement is more pronounced for the proposed method and the RTF-CT method compared to the RTF-MTF method. For example, for SNR = 10 dB, the localization error is reduced by 66% (from 1.57° to 0.54°) for the proposed method, and by 49% (from 6.24° to 3.21°) for the RTF-CT method when the speech duration rises from 1 s to 4 s. In contrast, the localization error of RTF-MTF is quite larger and is only reduced by 11% (from 12.60° to 11.16°).

TABLE IV: Localization errors (in degrees) as a function of speech duration, for $T_{60} = 0.5$ s and a source-to-sensors distance of 2 m.

SNR	Method	Speech duration (s) 1	2	3	4
10 dB	Proposed	1.57	0.88	0.79	0.54
	RTF-CT	6.24	4.43	3.86	3.21
	RTF-MTF	12.60	12.01	11.25	11.16
0 dB	Proposed	7.36	4.62	4.05	3.07
	RTF-CT	12.97	11.33	10.04	9.67
	RTF-MTF	17.56	15.29	14.94	15.01

VII. Experiments with the NAO Robot

In this section we present several experiments that were conducted using the NAO robot (Version 5) in various real-world environments. NAO is a humanoid companion robot developed and commercialized by Aldebaran Robotics.[5] NAO's head has four microphones that are nearly coplanar, see Fig. 5. The recordings contain ego-noise, i.e. noise produced by the robot. In particular, it contains a loud fan noise, which is stationary and partially interchannel correlated [41]. The spectral energy of the fan noise is notable up to 4 kHz, thence the speech signals are significantly contaminated. Note that the experiments reported in this section adopt the parameter settings discussed in Section VI-B.

A. The Datasets

The data are recorded in three environments: laboratory, office, e.g., Fig. 6-(right), and cafeteria, with reverberation times (T_{60}) that are approximately 0.52 s, 0.47 s and 0.24 s, respectively. Two **test datasets** are recorded in these environments:

1) The *audio-only* dataset: In the laboratory, speech utterances

[5]https://www.ald.softbankrobotics.com.

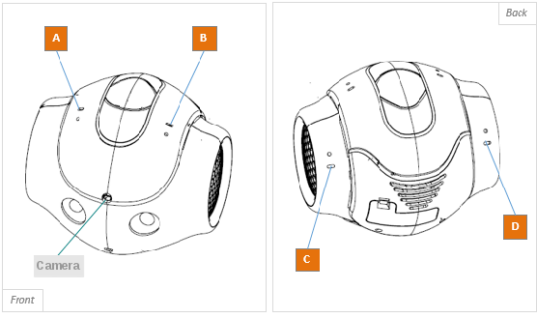

Fig. 5: NAO's head has four microphones and one camera.

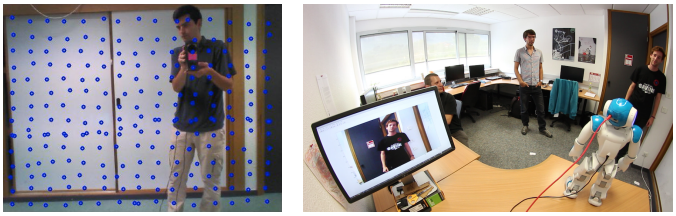

Fig. 6: The *audio-visual* training dataset (left) is obtained by moving a loudspeaker in front of a microphone/camera setup. Sounds are emitted by a loudspeaker. A LED placed on the loudspeaker enables to associate each sound direction with an image location (a blue circle). The data contain pairs of acoustic recordings and sound directions. A typical localization scenario with the NAO robot (right).

from the TIMIT dataset [36] are emitted by a loudspeaker in front of NAO. Two groups of data are recorded with a source-to-robot distance of 1.1 m and 2.1 m, respectively. For each group, 174 sounds are emitted from directions uniformly distributed in azimuth and elevation, in the range $[-120°, 120°]$ (azimuth), and $[-15°, 25°]$ (elevation).

2) The *audio-visual* dataset: Sounds are emitted by a loudspeaker lying in the field of view of NAO's camera. The image resolution is of 640×480 pixels, corresponding to approximately $60°$ ($-30°$ to $30°$) azimuth range and to approximately $48°$ ($-24°$ to $24°$) elevation range, so $1°$ of azimuth/elevation corresponds to approximately 10.5 horizontal/vertical pixels. A LED placed on the loudspeaker enables to estimate the loudspeaker location in the image, hence ground-truth localization data are available with the audio-visual dataset. Three sets of audio-visual data are recorded in three different rooms. For each set, sounds are emitted from about 230 directions uniformly distributed in the camera field-of-view. Fig. 6-(left) shows the source positions shown as blue dots in the image plane. The source-to-robot distance is about 1.5 m in this dataset.

In both datasets, ambient noise is much lower than fan noise, hence the noise of recorded signals mainly corresponds to fan noise. In the case of the audio-only dataset, the SNR is 14 dB and 11 dB for source-to-robot distances of 1.1 m and 2.1 m, respectively. For the audio-visual dataset the SNR is 2 dB.

The **training dataset** for the *audio-only* localization experiments is generated with the NAO head HRIRs of $1,002$ directions uniformly distributed over the same azimuth-elevation range as the test dataset. The training dataset for *audio-visual* experiments is generated with the NAO head HRIR of 378 directions uniformly distributed over the camera field-of-view. HRIRs are measured in the laboratory: white Gaussian noise is emitted from each direction, and the cross-correlation between the microphone and source signals yields the BRIR of each direction. In order to obtain anechoic HRIRs, the BRIRs are manually truncated before the first reflection. The regression method of [6], outlined in Section V, is used for supervised localization. The SRP-PHAT method takes the source directions in the training set as the steering directions.

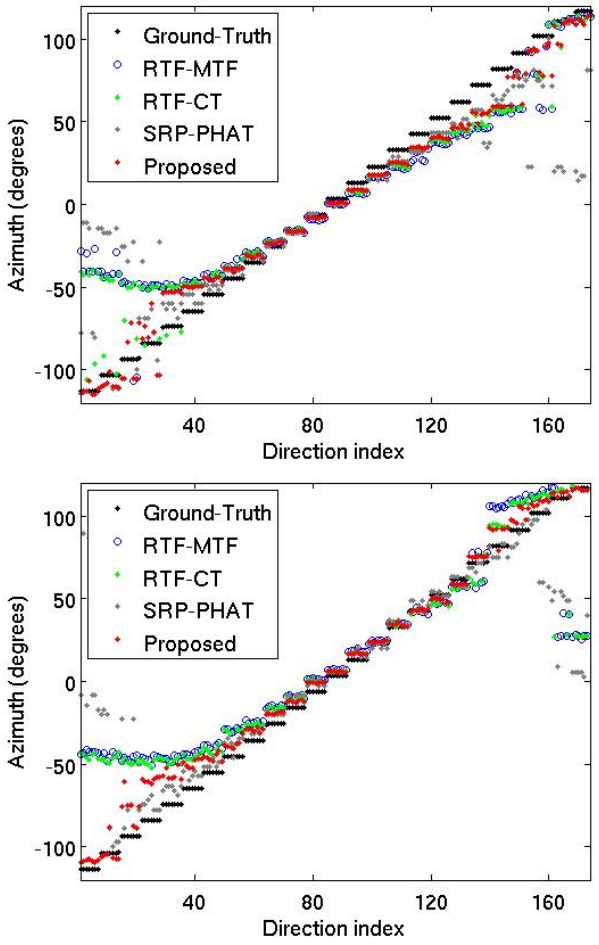

Fig. 7: Azimuth estimation for the audio-only dataset. Source-to-robot distance is 1.1 m (top) and 2.1 m (bottom).

B. Localization Results for the Audio-Only Dataset

Experiments with the audio-only dataset first show that elevation estimation in the range $[-15° \, 25°]$ is unreliable for all the four methods. This can be explained by the fact that the four microphones are coplanar. Therefore we only present the azimuth estimation results in the following.

The azimuth estimation results for the audio-only dataset are given in Fig. 7. The results are quite consistent across the two conditions, i.e. source-to-robot distance of 1.1 m

(Fig. 7-top) and 2.1 m (Fig. 7-bottom). Globally, for the azimuth range $[-50°, 50°]$ all four methods provide good localization, i.e. they follow the ground-truth line quite well, for both source-to-robot distances. In this range, the proposed method achieves slightly better results than the RTF-MTF and RTF-CT methods. The performance of all methods drops significantly for directions out of this range, but globally, the proposed method remains the closest to the ground-truth. In more details, in the approximate range $[-120°, -50°]$ and $[50°, 120°]$ it can be seen that SRP-PHAT and RTF-MTF have the largest localization error and many localization outliers caused by reverberations (SRP-PHAT performs slightly better than RTF-MTF in the zones just after $-50°$ and $50°$, possibly due to PHAT weighting). By selecting frames that involve less reverberations, RTF-CT performs slightly better than RTF-MTF. The proposed method outperforms the others by extracting the binaural cues associated with the direct-path propagation. Importantly, in the extremities of the range, the proposed method does not generate major outliers nor large deviation from the ground-truth, as opposed to the other methods.

C. Localization Results for the Audio-Visual Dataset

The azimuth and elevation in the audio-visual dataset are limited to a small range around $0°$ azimuth. As a consequence, both the azimuth and elevation localization results of this dataset are better than the results of audio-only dataset in average. Table V shows the localization errors for azimuth (Azim.) and elevation (Elev.) for the audio-visual dataset. The elevation errors are always larger than the azimuth errors, due to the low elevation resolution of the microphone array that we already mentioned (the microphone are coplanar and the microphone plane is horizontal). The cafeteria has the smaller reverberation time, $T_{60} = 0.24$ s. Consequently, the RTF-MTF and RTF-CT methods yields performance measures that are comparable with the proposed method. The office and laboratory have larger reverberation times, 0.47 s and 0.52 s, respectively, so the MTF approximation is no more accurate. A bit surprisingly RTF-MTF performs better than RTF-CT for the office (though the errors are quite close), this is probably due to the fact that the coherence test does not work well under low SNR conditions (let us remind that the SNR of the audio-visual dataset is around 2 dB). Globally, SRP-PHAT performs the worst, due to the intense noise. As a result of the presence of notable reverberations, the proposed method performs here significantly better than the three other methods. For example, in the laboratory environment, the proposed method provides $0.84°$ azimuth error and $1.84°$ elevation error, vs. $1.41°$ azimuth error and $2.30°$ elevation error for the best baseline methods (for instance SRP-PHAT and RTF-MTF respectively).

TABLE V: Localization error (in degrees) for the audio-visual dataset. The best results are shown in bold.

	Cafeteria		Office		Laboratory	
Method	Azim.	Elev.	Azim.	Elev.	Azim.	Elev.
RTF-MTF	0.47	1.58	0.62	2.14	1.46	2.30
RTF-CT	**0.43**	1.49	0.68	2.30	1.59	2.40
SRP-PHAT	0.77	1.95	1.03	2.80	1.41	3.33
Proposed	0.48	**1.46**	**0.55**	**1.86**	**0.84**	**1.84**

VIII. Conclusion

We proposed a method for the estimation of the direct-path relative transfer function (DP-RTF). Compared with the conventional RTF, the DP-RTF is defined as the ratio between two direct-path acoustic transfer functions. Therefore, the DP-RTF definition and estimation implies the removal of the reverberations, and it provides a more reliable feature, in particular for sound source localization. To estimate the DP-RTF, we adopted the convolutive transfer function (CTF) model instead of the multiplicative transfer function (MTF) approximation. By doing this, the DP-RTF can be estimated by solving a set of linear equations constructed from the reverberant sensor signals. Moreover, an inter-frame spectral subtraction method was proposed to remove noise power. This spectral subtraction process does not require explicit estimation of the noise PSD, hence it does not suffer from noise PSD estimation errors.

Based on the DP-RTF we proposed a supervised sound-source localization algorithm. The latter relies on a training dataset that is composed of pairs of DP-RTF feature vectors and their associated sound directions. The training dataset is pre-processed in such a way that it only contains anechoic head-related impulse responses. Hence the training dataset does not depend on the particular acoustic properties of the recording environment. Only the sensors set-up must be consistent between training and testing (e.g. using the same dummy/robot head). In practice we implemented two supervised methods, namely a nearest-neighbor search and a mixture of linear regressions. Experiments with both simulated data and real data recorded with four microphones embedded in a robot head, showed that the proposed method outperforms an MTF-based method and a method based on a coherence test, as well as a conventional SRP-PHAT method, in reverberant environments.

In the presented experiments the model parameters Q, D and N (Section VI-B) were set to constant values which were chosen as a tradeoff yielding good results in a variety of acoustic conditions. In the future, to improve the robustness of DP-RTF, we plan to estimate the acoustic conditions using the microphone signals, such that an optimal set of parameters can be adaptively adjusted. We also plan to extend the DP-RTF estimator and its use in SSL to the more complex case of multiple sound sources.

References

[1] H. Viste and G. Evangelista, "Binaural source localization," in *International Conference on Digital Audio Effects*, pp. 145–150, 2004.

[2] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 5, pp. 982–994, 2006.

[3] R. M. Stern, G. J. Brown, and D. Wang, "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (D. Wang and G. J. Brown, eds.), pp. 147–185, 2006.

[4] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2010.

[5] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503–1512, 2012.

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

[7] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 4, pp. 718–731, 2015.

[8] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

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

[10] 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.

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

[12] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 320–324, 2015.

[13] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 544–548, 2015.

[14] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[15] C. M. Zannini, R. Parisi, and A. Uncini, "Binaural sound source localization in the presence of reverberation," in *International Conference on Digital Signal Processing*, pp. 1–6, 2011.

[16] O. Schwartz, S. Gannot, E. Habets, *et al.*, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2015.

[17] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.

[18] D. Bechler and K. Kroschel, "Reliability criteria evaluation for TDOA estimates in a variety of real environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2005.

[19] M. Heckmann, T. Rodemann, F. Joublin, C. Goerick, and B. Scholling, "Auditory inspired binaural robust sound source localization in echoic and noisy environments," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 368–373, 2006.

[20] C. Hummersone, R. Mason, and T. Brookes, "A comparison of computational precedence models for source separation in reverberant environments," *Journal of the Audio Engineering Society*, vol. 61, no. 7/8, pp. 508–520, 2013.

[21] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, 2004.

[22] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2136–2147, 2008.

[23] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[24] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[25] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[26] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[27] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3099–3102, 1995.

[28] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*, vol. 46. Artech House Norwood, 2005.

[29] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[30] C. Forbes, M. Evans, N. Hastings, and B. Peacock, *Statistical distributions*. John Wiley & Sons, 2011.

[31] R. Martin, "Spectral subtraction based on minimum statistics," *Power*, vol. 6, p. 8, 1994.

[32] 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.

[33] X. Li, R. Horaud, L. Girin, and S. Gannot, "Local relative transfer function for sound source localization," in *The European Signal Processing Conference*, 2015.

[34] D. Campbell, "The roomsim user guide (v3. 3)," 2004.

[35] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR dummy-head microphone," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

[36] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[37] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 6, pp. 1006–1018, 2015.

[38] C. Zheng, A. Schwarz, W. Kellermann, and X. Li, "Binaural coherent-to-diffuse-ratio estimation for dereverberation using an ITD model," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pp. 1048–1052, IEEE, 2015.

[39] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180, Springer, 2001.

[40] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, pp. I–121, IEEE, 2007.

[41] H. W. Loellmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *Proceedings of Speech Communication*, pp. 1–4, VDE, 2014.

Xiaofei Li received a Bachelor degree in Electronic Information from Beijing Institute of Machinery, China in 2007 and a PhD in Electronics at Peking University, China during 2007 to 2013. Since February 2014 Xiaofei Li is with the PERCEPTION team at INRIA Grenoble Rhne-Alpes as a post-doctoral fellow. Dr. Li's interests include audio and speech signal processing, sound and speech recognition, sound source localization and audio-visual fusion.

Laurent Girin received the M.Sc. and Ph.D. degrees in Signal Processing from the Institut National Polytechnique de Grenoble (INPG), Grenoble, France, in 1994 and 1997, respectively. In 1999, he joined the Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble (ENSERG), as an Associate Professor. He is now a Professor at Phelma (Physics, Electronics, and Materials Department of Grenoble-INP), where he lectures signal processing theory and applications to audio. His research activity is carried out at GIPSA-Lab (Grenoble Laboratory of Image, Speech, Signal, and Automation). It deals with speech and audio processing (analysis, modeling, coding, transformation, synthesis), with a special interest in multimodal speech processing (e.g. audiovisual, articulatory-acoustic, etc.) and speech/audio source separation. Prof. Girin is also a regular collaborator of INRIA (French Computer Science Research Institute), as an associate member of the Perception Team.

Radu Horaud received the B.Sc. degree in Electrical Engineering, the M.Sc. degree in Control Engineering, and the Ph.D. degree in Computer Science from the Institut National Polytechnique de Grenoble, France. Currently he holds a position of director of research with INRIA Grenoble, Montbonnot Saint-Martin, France, where he is the founder and head of the PERCEPTION team. His research interests include computer vision, machine learning, audio signal processing, audiovisual analysis, and robotics. He is an area editor of the *Elsevier Computer Vision and Image Understanding*, a member of the advisory board of the *Sage International Journal of Robotics Research*, and an associate editor of the *Kluwer International Journal of Computer Vision*. In 2013 Radu Horaud was awarded an ERC Advanced Grant for his project *Vision and Hearing in Action* (VHIA).

Sharon Gannot (S'92-M'01-SM'06) received his B.Sc. degree (summa cum laude) from the Technion Israel Institute of Technology, Haifa, Israel in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Israel in 1995 and 2000 respectively, all in Electrical Engineering. In 2001 he held a post-doctoral position at the department of Electrical Engineering (ESAT-SISTA) at K.U.Leuven, Belgium. From 2002 to 2003 he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is a Full Professor at the Faculty of Engineering, Bar-Ilan University, Israel, where he is heading the Speech and Signal Processing laboratory and the Signal Processing Track. Prof. Gannot is the recipient of Bar-Ilan University outstanding lecturer award for 2010 and 2014. Prof. Gannot has served as an Associate Editor of the EURASIP Journal of Advances in Signal Processing in 2003-2012, and as an Editor of several special issues on Multi-microphone Speech Processing of the same journal. He has also served as a guest editor of ELSEVIER Speech Communication and Signal Processing journals. Prof. Gannot has served as an Associate Editor of IEEE Transactions on Speech, Audio and Language Processing in 2009-2013. Currently, he is a Senior Area Chair of the same journal. He also serves as a reviewer of many IEEE journals and conferences. Prof. Gannot is a member of the Audio and Acoustic Signal Processing (AASP) technical committee of the IEEE since Jan., 2010. Currently, he serves as the committee vice-chair. He is also a member of the Technical and Steering committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005 and was the general co-chair of IWAENC held at Tel-Aviv, Israel in August 2010. Prof. Gannot has served as the general co-chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in October 2013. Prof. Gannot was selected (with colleagues) to present a tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013 and EUSIPCO 2013. Prof. Gannot research interests include multi-microphone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation; dereverberation; single microphone speech enhancement and speaker localization and tracking.