Frequency-Sliding Generalized Cross-Correlation: A Sub-band Time Delay Estimation Approach

Maximo Cobos, Fabio Antonacci, Luca Comanducci, and Augusto Sarti

M. Cobos is with the Departament d'Informàtica, Universitat de València, Burjassot, 46100 Spain e-mail: (maximo.cobos@uv.es).F. Antonacci, A. Sarti and L. Comanducci are with the Dipartimento di Elettronica, Informazione e Bioingegneria at the Politecnico di Milano, Italy. E-mail: (firstname.lastname@polimi.it)This work has been partially supported by FEDER and the Spanish Ministry of Science, Innovation and Universities under Grants RTI2018-097045-B-C21 and PRX19/00075.

Abstract

The generalized cross-correlation (GCC) is regarded as the most popular approach for estimating the time difference of arrival (TDOA) between the signals received at two sensors. Time delay estimates are obtained by maximizing the GCC output, where the direct-path delay is usually observed as a prominent peak. Moreover, GCCs play also an important role in steered response power (SRP) localization algorithms, where the SRP functional can be written as an accumulation of the GCCs computed from multiple sensor pairs. Unfortunately, the accuracy of TDOA estimates is affected by multiple factors, including noise, reverberation and signal bandwidth. In this paper, a sub-band approach for time delay estimation aimed at improving the performance of the conventional GCC is presented. The proposed method is based on the extraction of multiple GCCs corresponding to different frequency bands of the cross-power spectrum phase in a sliding-window fashion. The major contributions of this paper include: 1) a sub-band GCC representation of the cross-power spectrum phase that, despite having a reduced temporal resolution, provides a more suitable representation for estimating the true TDOA; 2) such matrix representation is shown to be rank one in the ideal noiseless case, a property that is exploited in more adverse scenarios to obtain a more robust and accurate GCC; 3) we propose a set of low-rank approximation alternatives for processing the sub-band GCC matrix, leading to better TDOA estimates and source localization performance. An extensive set of experiments is presented to demonstrate the validity of the proposed approach.

Index Terms: Time delay estimation, GCC, SVD, weighted SVD, sub-band processing, SRP-PHAT.

I   INTRODUCTION

Time delay estimation (TDE) refers to finding the time differences-of-arrival (TDOAs) between signals received at an array of sensors. Traditionally, TDE has played an important role in many location-aware systems, including radar, sonar, wireless systems, sensor calibration or seismology. In acoustic signal processing, TDE is essential for localizing and tracking acoustic sources [1, 2]. Automatic camera steering [3], speaker localization [4] or speech enhancement systems are application examples strongly relying on TDE [5].

The generalized cross-correlation (GCC), originally proposed by Knapp and Carter in their 1976 seminal paper [6], is still today the most popular technique for TDE. By using GCCs, the TDOA between two signals is estimated as the time lag that maximizes the cross-correlation between filtered versions of such signals. In this context, the GCC may consider different weighting functions (filters) characterized by a particular behavior [7, 8, 9]. Roth, smoothed coherence transform, Eckart or Hannan-Thomson (maximum likelihood) are examples of such weighting schemes [6, 10]. The GCC with phase transform (GCC-PHAT) has been repeatedly shown to be a suitable alternative for TDE in real reverberant scenarios [11, 12]. However, it has also been demonstrated that its performance can only be considered optimal under uncorrelated noise conditions and a high signal-to-noise ratio (SNR) [13]. The present work is particularly focused on PHAT weighting, thus, the terms GCC and GCC-PHAT will be used indistinctly throughout this paper. Since the GCC was first proposed, many approaches have appeared to ameliorate the robustness of TDE techniques, with improvements that are mostly achieved by exploiting the spatial diversity provided by more than two microphones [14, 15, 16], blind channel estimation [17, 18], modified GCC weightings [19], or the incorporation of some a priori information [20, 21]. Further improvements to deal with reverberation and correlated noise fields have also been proposed [22], addressing some of the limitations of the GCC. In addition, other methods exploiting temporal diversity by means of Bayesian filtering have shown great potential [23, 24].

In contrast to the above methods, this paper proposes a novel sub-band approach to TDE with two sensors which is not directly categorized within the above processing improvements. Concretely, the method is based on the exploration of the cross-power spectrum phase by following a sliding window approach, obtaining a set of sub-band GCCs that encode the contribution of different frequency bands to the estimated TDOA. The resulting

sub-band GCC matrix is shown to be rank-one for a full-band signal in the noiseless single-path case. This fact is exploited to obtain a robust GCC in adverse scenarios, proposing low-rank approximations of the sub-band GCC matrix that ultimately lead to better estimation accuracy, reduced level of spurious peaks and lower probability of anomalous estimates.

The rest of the paper is structured as follows. Section II summarizes the background concerning GCC-based TDOA estimation. Section III presents the frequency-sliding GCC (FS-GCC) representation. Section IV discusses our proposed methods for TDE based on the FS-GCC. The experimental evaluation is in Section V. Finally, the conclusions of this work are summarized in Section VI.

II   TIME DELAY ESTIMATION

This section summarizes the conventional GCC approach for TDE. To this end, the ideal anechoic model is first presented, discussing the main problems arising in realistic acoustic conditions.

II-A   *Anechoic Signal Model*

Let us consider a pair of sensors with spatial coordinates given by column vectors $\mathbf{m}_1$, $\mathbf{m}_2$ $\in \mathbb{R}^3$ and an emitting acoustic source located at $\mathbf{s} \in \mathbb{R}^3$. The time difference-of-arrival (TDOA) measured in samples is defined as

$$\tau_0 \triangleq \lfloor \frac{\|\mathbf{s} - \mathbf{m}_1\| - \|\mathbf{s} - \mathbf{m}_2\|}{c} f_s \rceil = \eta_1 - \eta_2, \tag{1}$$

where c is the wave propagation speed, $\lfloor \cdot \rceil$ denotes the rounding operator and f_s is the sampling frequency. The terms η_1 and η_2 represent the time of flight (TOF) of the sound to the sensors in samples.

Assuming an anechoic scenario, the signals received by the two sensors can be modeled as

$$x_m[n] = \beta_m s[n - \eta_m] + w_m[n], \quad m = 1, 2, \tag{2}$$

where $\beta_m \in \mathbb{R}_+$ is a positive amplitude decay factor, $s[n]$ is the source signal and $w_m[n]$ is an additive noise term. In the discrete-time Fourier transform (DTFT) domain, the sensor signals can be written as

$$X_m(\omega) = \beta_m S(\omega) e^{-j\omega\eta_m} + W_m(\omega), \quad m = 1, 2, \tag{3}$$

where $S(\omega)$, $W_m(\omega) \in \mathbb{C}$, are the DTFTs of the source signal and the noise signal, respectively, and $j = -1$.

II-B *Generalized Cross-Correlation*

The GCC of a pair of sensor signals is defined as the inverse Fourier transform of the weighted cross-power spectrum, i.e.

$$R[\tau] \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} \Psi(\omega) e^{j\omega\tau} \, d\omega = \mathcal{F}^{-1}\{\Psi(\omega)\}, \tag{4}$$

where $\tau \in \mathbb{Z}$ represents time delay and $\Psi(\omega) \in \mathbb{C}$ represents the phase transform (PHAT) cross-power spectrum:

$$\Psi(\omega) \triangleq \frac{X_1(\omega)X_2^*(\omega)}{|X_1(\omega)X_2^*(\omega)|}, \tag{5}$$

where $(\cdot)^*$ denotes complex conjugation. Note that, in the ideal anechoic and noiseless case,

$$\Psi(\omega) = \frac{\beta_1\beta_2|S(\omega)|^2}{\left|\beta_1\beta_2|S(\omega)|^2\right|} e^{-j\omega\tau_0} = e^{-j\omega\tau_0}, \tag{6}$$

leading to

$$R[\tau] = \mathcal{F}^{-1}\left\{e^{-j\omega\tau_0}\right\} = \delta[\tau - \tau_0], \tag{7}$$

where $\delta[\tau]$ is the Kronecker delta function. Therefore, the ideal GCC-PHAT for a full-band source signal will show a unit impulse located at the true TDOA. This motivates the use of the following estimator for time-delay estimation over signals acquired by a pair of sensors:

$$\hat{\tau}_0 = \arg\max_{\tau} R[\tau]. \tag{8}$$

The GCC-PHAT inherently discards the magnitude information of the signals and provides a time-delay estimate based only on the phase of the cross-power spectrum, i.e. the phase of the frequency domain analysis of the cross-correlation between the two signals. It is important to remark that the measured time delay is an integer multiple of the sampling period. Note, however, that a finer resolution can be achieved by interpolating between consecutive samples of the GCC function if necessary.

II-C *Problems in Realistic Scenarios*

It is well-known that several problems arise when using GCC-PHAT for estimating the TDOA in realistic scenarios. Indeed, TDOA measurements are very sensitive to reverberation, noise, and the presence of potential interferers:

- In reverberant environments, for certain locations and orientations of the source signal, the peak of the GCC related to a reflective path could overcome that of the direct

path.

- In noisy scenarios, for some time instants, the noise level could exceed that of the signal, making the estimated TDOA unreliable.

- Peaks corresponding to the direct path, reflections or combinations of interfering signals may also lead to errors when estimating the TDOA of a target source.

The above issues can be even more problematic when the spectral characteristics of the target source or the additive noise lead to a reduced signal-to-noise ratio (SNR) at some frequency bands [25]. The phase information can be completely lost at those frequency bands where either there is no signal information or where its content is especially affected by the noise. Consider, for example, the (normalized) GCCs shown in Figure 1. The top row corresponds to the ideal case of a noiseless full-band signal. The other two rows correspond to the same signal but affected by different spectral noise profiles. As it can be observed, the noisy phase introduces many spurious peaks that can cause anomalous TDOA estimates. Although noisy frequency bands could be filtered out to obtain a cleaner GCC (last column of Fig. 1), knowing beforehand the best frequency range accommodating the time delay information is not always straightforward. Moreover, the rippling GCCs resulting from pass-band-like signals (third row) complicate substantially source localization tasks in adverse conditions [25].
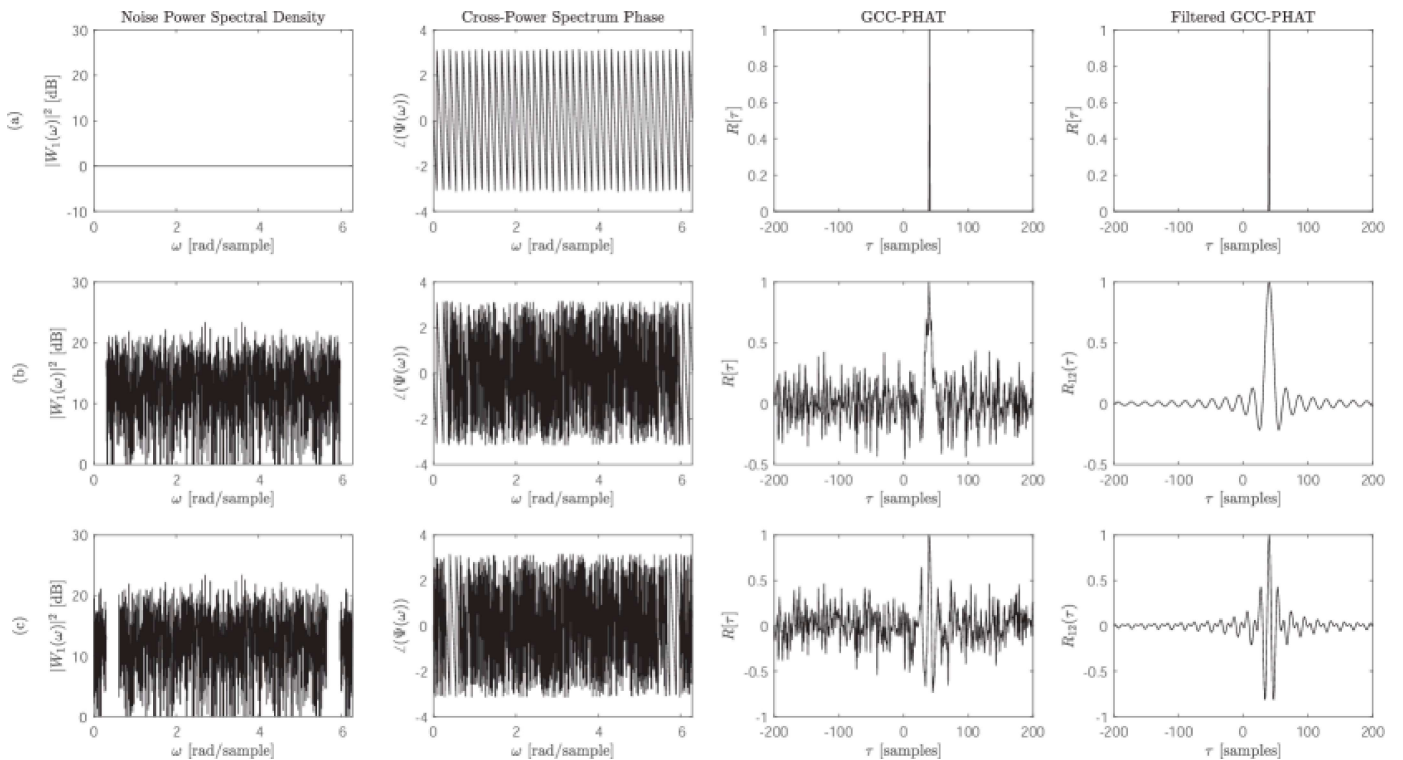

Figure 1: GCCs obtained from two signals corrupted by additive white noise (SNR $= -15$ dB) affecting different frequency bands: (a) No noise. (b) Noise affecting high frequencies. (c) Noise affecting high and low frequencies. The true TDOA is $\tau_0 = 40$ samples.

III   FREQUENCY-SLIDING GENERALIZED CROSS-CORRELATION

This section proposes a frequency-sliding GCC-PHAT method that generalizes the filtering approach discussed in the last section with the aim of obtaining a robust GCC representation. This representation will be useful to explore the likelihood of different frequency bands contributing to a direct-path delay estimate.

III-A   *Sub-band GCCs*

Let us define the sub-band GCC for an arbitrary frequency band l as

$$
\begin{aligned}
R[\tau, l] &\triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} \Psi(\omega + \omega_l)\Phi(\omega)e^{j\omega\tau} \, d\omega \\
&= \mathcal{F}^{-1}\big\{\Psi(\omega + \omega_l)\Phi(\omega)\big\},
\end{aligned}
$$

where ω_l is the frequency offset corresponding to band l. The function $\Phi(\omega) \in \mathbb{R}$ is a symmetric frequency-domain window, centered at $\omega = 0$ with frequency support $B_\Phi \in [0, \pi]$, i.e. $\Phi(\omega) = 0$ for $|\omega| \geq B_\Phi$.

The effect of the window and the cross-power spectrum can be separated as follows:

$$
\begin{aligned}
R[\tau, l] &= \mathcal{F}^{-1}\big\{\Psi(\omega + \omega_l)\big\} * \mathcal{F}^{-1}\big\{\Phi(\omega)\big\} \\
&= \Big(e^{-j\omega_l\tau}\mathcal{F}^{-1}\big\{\Psi(\omega)\big\}\Big) * \phi[\tau] \\
&= \Big(e^{-j\omega_l\tau}R[\tau]\Big) * \phi[\tau],
\end{aligned}
\tag{10}
$$

or equivalently,

$$
\begin{aligned}
R[\tau, l] &= \mathcal{F}^{-1}\big\{\Psi(\omega)\Phi(\omega - \omega_l)\big\}e^{-j\omega_l\tau}, \\
&= \Big(R[\tau] * \big(\phi[\tau]e^{j\omega_l\tau}\big)\Big)e^{-j\omega_l\tau},
\end{aligned}
\tag{11}
$$

where $\phi[\tau] \in \mathbb{R}$ represents the inverse Fourier transform of the spectral window $\Phi(\omega)$ and $*$ denotes convolution.

A frequency-sliding sub-band GCC can be obtained by sweeping the cross-power spectrum phase over (possibly) overlapping frequency bands:

$$
\omega_l = lM_\Phi, \quad l = 0, ..., L - 1,
\tag{12}
$$

where M_Φ is the frequency hop. The number of bands L can be conveniently chosen to cover those frequencies up to the Nyquist limit:

$$
L = \lfloor \frac{\pi - B_\Phi + M_\Phi}{M_\Phi} \rfloor.
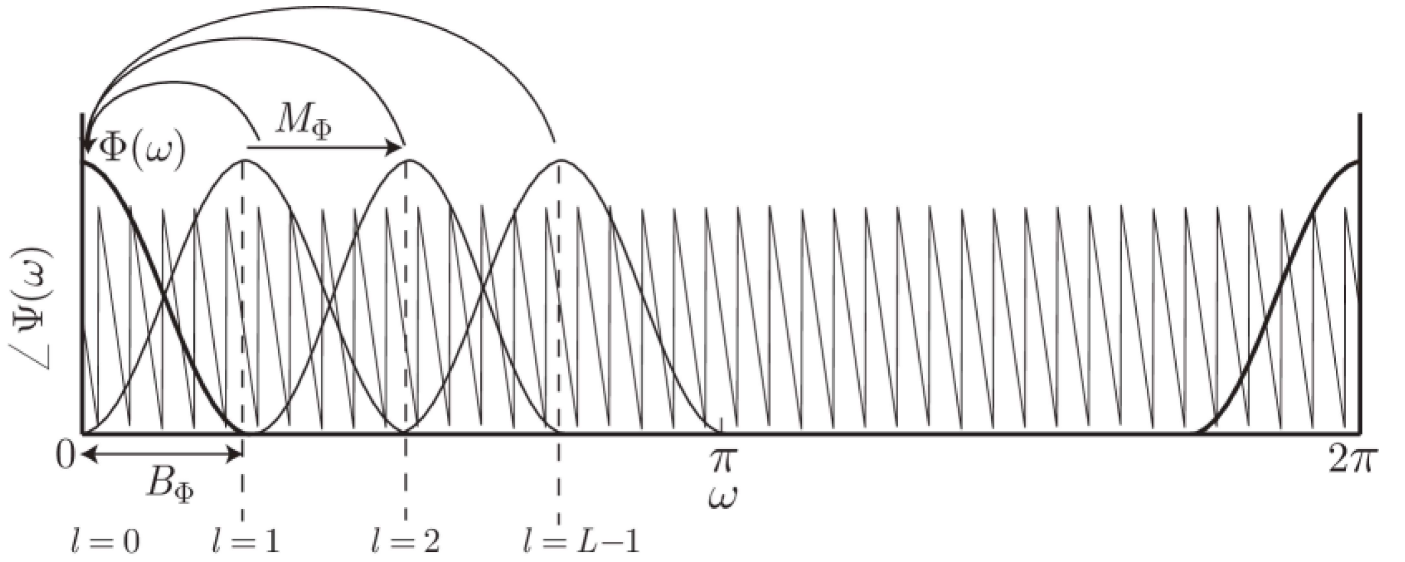\tag{13}
$$

Figure 2: Interpretation of the frequency-sliding GCC.

An interpretation of such sliding operation is shown in Fig. 2. As given by Eq. (11), the sub-band GCCs can be interpreted as the product of $\Psi(\omega)$ with a shifted version of $\Phi(\omega)$ to ω_l, shifted back in frequency to zero before taking the inverse Fourier transform.

III-B *Frequency-Sliding GCC Matrix*

In practice, sub-band GCCs are extracted considering the discrete Fourier transform (DFT) of the microphone signals $x_m[n]$:

$$\mathbf{X}_m = \left[X_m[0], X_m[1], ..., X_m[N-1] \right]^T, \quad m = 1, 2, \tag{14}$$

where the elements of $\mathbf{X}_m \in \mathbb{C}^N$ are the coefficients $X_m[k]$ corresponding to the discrete frequencies $\omega_k = k\dfrac{2\pi}{N}$. Similarly, consider the vector $\mathbf{\Phi}$ containing the discrete-frequency samples $\Phi[k] = \Phi(\omega_k)$ of the selected spectral window

$$\mathbf{\Phi} = \left[\Phi[0], \Phi[1], ..., 0, 0, ..., \Phi[N-1] \right]^T \in \mathbb{R}^N, \tag{15}$$

symmetrically padded with zeros to contain only $B = \lfloor 2B_\Phi \dfrac{N}{2\pi} \rfloor$ non-zero elements.

The sub-band GCC vectors $\mathbf{r}_l \in \mathbb{C}^N$ are obtained by taking the inverse DFT of the windowed PHAT spectrum, i.e.

$$\mathbf{r}_l = \left[r_l[0], r_l[1], ..., r_l[N-1] \right]^T, \quad l = 0, ..., L-1, \tag{16}$$

with

$$r_l[n] = \frac{1}{N} \sum_{k=0}^{N-1} \frac{X_1[k+lM]X_2^*[k+lM]}{|X_1[k+lM]X_2^*[k+lM]|} \Phi[k] e^{j\frac{2\pi}{N}kn}, \tag{17}$$

where $M = \lfloor M_\Phi \frac{N}{2\pi} \rfloor$ is the discrete frequency hop.

The frequency-sliding GCC (FS-GCC) matrix is constructed by stacking all the sub-band GCC vectors together, i.e.

$$\mathbf{R} = [\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_{L-1}] \in \mathbb{C}^{N \times L}, \tag{18}$$

noting that

$$\mathbf{R}[n, l] = \begin{cases} R[n, l], & n = 0, ..., N/2 - 1. \\ R[n - N, l], & n = N/2, ..., N - 1. \end{cases} \tag{19}$$

III-C *Ideal Sub-band GCC Model*

To get an insight into the properties underlying the presented FS-GCC representation, let us start by considering the ideal full-band cross-power spectrum of Eq. (6), $\Psi(\omega) = e^{-j\omega\tau_0}$. By inserting Eq. (7) into Eq. (10), the ideal sub-band GCCs (denoted with a tilde) are given by

$$\begin{aligned} \tilde{R}[\tau, l] &= e^{-j\omega_l\tau}\delta[\tau - \tau_0] * \phi[\tau] \\ &= e^{-j\omega_l\tau_0}\phi[\tau - \tau_0], \quad l = 1, ..., L. \end{aligned}$$

By analyzing the above result, some interesting observations arise:

- The magnitudes of the sub-band GCCs are independent of the selected frequency offset ω_l, and correspond to the inverse Fourier transform of the sub-band analysis window centered at the true TDOA:

$$|\tilde{R}[\tau, l]| = |\phi[\tau - \tau_0]| \quad \forall l. \tag{21}$$

- The maximum value of $|\tilde{R}[\tau, l]|$ is located at $\tau = \tau_0$ for every sub-band l, $l = 0, ..., L - 1$.

- The peak width at the true TDOA ($\mathcal{W}$) as well as the level of its side lobes, depends on the selected type of window Φ and its frequency support B (e.g. $\mathcal{W} = 2N/B$ for a rectangular window, while $\mathcal{W} = 4N/B$ for a Hann window).

- Since the spectral window is real and even, $\phi[\tau]$ is also real and even. Then, the phase pattern of the set of sub-band GCCs corresponds to a downsampled version of $\angle(\Psi(\omega))$, modulated by the sign of $\phi[\tau]$.

- The full-band conventional GCC can be recovered from the sub-band GCCs if the constant overlap-add (COLA) property is fulfilled.

Consequently, the FS-GCC matrix will present the following features:

- The columns of **R** correspond to the window response shifted to the true TDOA and multiplied by a different complex number, i.e.

$$\tilde{\mathbf{r}}_l = e^{-j\omega_l \tau_0}\boldsymbol{\phi}_0, \tag{22}$$

where $\boldsymbol{\phi}_0 \in \mathbb{R}^N$ is a vector containing N samples of the shifted window response:

$$\begin{aligned}
\boldsymbol{\phi}_0 &\triangleq \phi[n - \tau_0] = \left[\phi[0 - \tau_0] \ldots \phi[N \ / \ 2 - 1 - \tau_0] \right. \\
&\quad \left. \phi[-N \ / \ 2 - \tau_0] \ldots \phi[-1 - \tau_0] \right]^T.
\end{aligned} \tag{23}$$

- The ideal FS-GCC matrix, $\tilde{\mathbf{R}}$, can be expressed as an outer product, resulting in a rank-one matrix:

$$\tilde{\mathbf{R}} = \boldsymbol{\phi}_0 \mathbf{e}^H, \tag{24}$$

where $\mathbf{e} \triangleq [e^{j\omega_0 \tau_0}, \ldots, e^{j\omega_{L-1} \tau_0}]^T \in \mathbb{C}^L$ and $(\cdot)^H$ denotes the conjugate transpose operator.

This last observation will be used to estimate robustly the time-delay in realistic cases (noisy and/or reverberant scenarios with reduced/varying bandwidth signals).

III-D *Noisy Sub-band GCC Model*

In a more general case, the columns of the FS-GCC matrix will contain different amounts of noise depending on the SNR characterizing the different sub-bands, i.e.

$$\mathbf{r}_l = \alpha_l \tilde{\mathbf{r}}_l + (1 - \alpha_l)\mathbf{n}_l, \quad l = 0, \ldots, L - 1, \tag{25}$$

where $\mathbf{n}_l \in \mathbb{C}^N$ models the GCC noise component of the l-th sub-band, and $\alpha_l \in [0, 1]$ is a scalar that balances the contribution of the noise in the l-th sub-band. Since the noise vectors in the frequency domain have a normalized magnitude spectrum given by the spectral window $\Phi(\omega)$, it follows from Parseval's theorem that

$$\mathbf{n}_l^H \mathbf{n}_l = \|\mathbf{n}_l\|^2 = \|\phi\|^2 = \int_{-\pi}^{\pi} \Phi^2(\omega) \, d\omega. \tag{26}$$

Similarly, the **R** matrix can be expressed as:

$$\mathbf{R} = \tilde{\mathbf{R}}\mathbf{G} + \mathbf{N}(\mathbf{I} - \mathbf{G}), \tag{27}$$

where $\mathbf{I} \in \mathbb{R}^{L \times L}$ is the identity matrix, $\mathbf{N} \in \mathbb{C}^{N \times L}$ is the matrix containing all the noise vectors, i.e.

$$\mathbf{N} = [\mathbf{n}_0, \mathbf{n}_1, \ldots, \mathbf{n}_{L-1}], \tag{28}$$

and $\mathbf{G} \in \mathbb{R}^{L \times L}$ is the diagonal matrix

$$\mathbf{G} = \begin{bmatrix} \alpha_0 & 0 & \cdots & 0 \\ 0 & \alpha_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_{L-1} \end{bmatrix}. \tag{29}$$

In contrast to the ideal case, the non-ideal FS-GCC matrix $\mathbf{R}$ will have rank L unless more than one of the diagonal entries α_l are equal to 1.
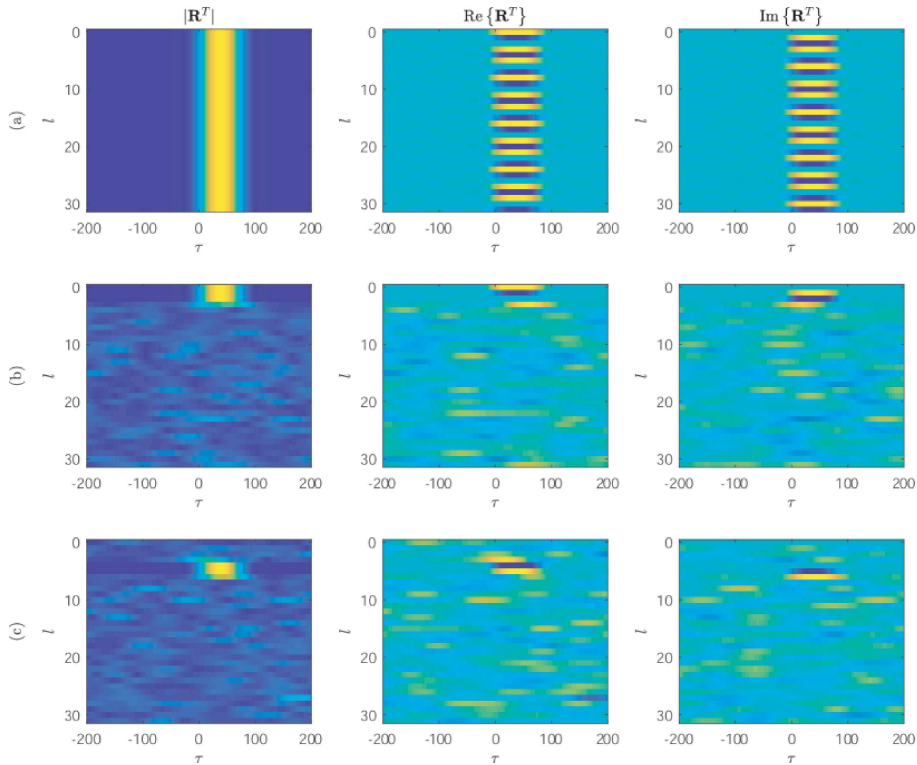

Figure 3: Sub-band GCCs obatined for the synthetic examples shown in Fig.1 (SNR $= -15$ dB, $\tau_0 = 40$): magnitude (left), real part (center) and imaginary part (right).

III-E *Examples*

Fig. 3(a-c) show the resulting FS-GCCs for the three signals analyzed in Fig. 1. A Hann spectral window with a frequency support of $B = 64$ bins ($B_\Phi = 0.1$) and 50% overlap was used, resulting in $L = 32$ sub-bands. As it can be observed, the first row corresponding to the noiseless full-band signal shows a peak perfectly aligned across all the analyzed bands and centered at the true TDOA ($\tau_0 = 40$ samples). Due to the use of the Hann window, the side-lobes are considerably small. The other rows corresponding to the signal affected by noise show also a similar pattern, but only at those sub-bands where the SNR is sufficient to preserve the phase information of the cross-power spectrum. The noise over the rest of sub-bands can be also identified, showing that no useful delay information can be extracted from other frequencies.

An example considering a more realistic signal can be observed in the first column of Fig. 4, corresponding to three different time frames of a male speech signal ($N = 2048$ samples, $f_s = 44.1$ kHz). In this case, B has been set to 128 bins, resulting in a narrower peak. As observed, the useful bandwidth of common audio signals as speech may change considerably with time and the FS-GCC representation shows clearly which frequency bands are actively contributing to a reliable time delay estimate. Note also that the last speech frame (c) shows a bandwidth that is considerably narrower than the other two, which makes TDE more difficult in this case.

While the full-band ideal case results in a high peak continuously present over all the frequency-bands leading to a rank one matrix, in a general case, noise and reflections will interact with the linear phase component corresponding to the direct-path delay, producing a matrix with full column rank. TDE from the FS-GCC matrix should be therefore oriented towards the extraction of the reliable components of $\mathbf{R}$, discarding properly the information from noisy bands.

Finally, a rule of thumb to select an appropriate value for the involved parameters can be established by allowing the main lobe width ($\mathcal{W}$) of the GCC to fit within the expected range of possible time delays, i.e. $\mathcal{W} \leq f_s \frac{d}{c}$, where d is the inter-sensor distance. For example, for a Hann window ($\mathcal{W} = 4N / B$), $d = 0.5$ m and $f_s = 44100$ Hz, it follows that $B \approx 128$ and $M = 32$ (75% overlap).

IV   LOW-RANK APPROXIMATIONS OF THE FS-GCC MATRIX

This section exploits the properties of the presented sub-band FS-GCC representation by proposing a low-rank approximation framework based on singular value decompositions. Such low-rank representations will be shown to provide a robust GCC for TDE in adverse conditions.

IV-A   *Singular Value Decomposition*

A low-rank approximation of $\mathbf{R}$ can be obtained by solving

$$\min_{\hat{\mathbf{R}}} \quad \|\mathbf{R} - \hat{\mathbf{R}}\|_F, \quad \text{subject to} \quad \text{rank}(\hat{\mathbf{R}}) \leq r \tag{30}$$

where r is the rank of the approximating matrix $\hat{\mathbf{R}}$, and $\|\cdot\|_F$ denotes the Frobenius norm. The problem has analytic solution in terms of the singular value decomposition (SVD) of $\mathbf{R}$, as given by the Eckart-Young-Mirsky theorem. Let us factorize $\mathbf{R}$ as

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H, \tag{31}$$

where $\mathbf{U} \in \mathbb{C}^{N \times L}$ is the matrix containing the left singular vectors, $\boldsymbol{\Sigma} \in \mathbb{R}^{L \times L}$ is the diagonal matrix containing the ordered singular values and $\mathbf{V} \in \mathbb{C}^{L \times L}$ is the matrix containing

the right singular vectors. The particular rank-r matrix that best approximates $\mathbf{R}$ is given by

$$\mathbf{R^r} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \tag{32}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the i-th columns of the corresponding SVD matrices and σ_i are the ordered singular values [26].

IV-B *Separation of GCC Components*

Let us assume that the sub-band GCC matrix can be expressed by two separable components: a target-path delay component and a noise component. Moreover, as it will be later discussed in Section IV-C, we assume that the direct-path component is the one contributing to the greatest singular value, while the noise component is obtained by the addition of the rest of separable matrices, i.e.

$$\mathbf{R}^{\text{target}} = \mathbf{R^1} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T, \tag{33}$$

$$\mathbf{R}^{\text{noise}} = \sum_{i=2}^{L} \sigma_i \mathbf{u}_i \mathbf{v}_i^T. \tag{34}$$

An example decomposition applied over the three speech frames discussed in the previous section is shown in Fig. 4. Note that the true delay is substantially enhanced by the target delay component of the SVD. The third speech frame (row 3) is the one having the noisiest FS-GCC, due to its reduced SNR.

IV-C *Target GCC Extraction*

The singular values and the left singular vectors correspond, respectively, to the square-root of the eigenvalues and the orthonormal eigenvectors of $\mathbf{RR}^H$. In the ideal case, taking into account Eq. (24):

$$\tilde{\mathbf{R}}\tilde{\mathbf{R}}^H = \phi_0 \mathbf{e}^H \mathbf{e} \phi_0^H = L \phi_0 \phi_0^H \tag{35}$$

is singular, with $(L-1)$ zero eigenvalues and one non-zero eigenvalue $\lambda = L\phi_0^H \phi_0 = L\|\phi\|^2$ associated to the eigenvector ϕ_0. This means that such eigenvector should contain the information of the TDOA-shifted window of Eq. (23).

In the general noisy case, the left singular vectors will be eigenvectors of

$$\begin{aligned} \mathbf{RR}^H &= \left(\tilde{\mathbf{R}}\mathbf{G} + \mathbf{N}(\mathbf{I} - \mathbf{G}) \right) \left(\tilde{\mathbf{R}}\mathbf{G} + \mathbf{N}(\mathbf{I} - \mathbf{G}) \right)^H \tag{36} \\ &= \left(\sum_{l=0}^{L-1} \alpha_l^2 \right) \phi_0 \phi_0^H + \sum_{l=0}^{L-1} (1 - \alpha_l)^2 \mathbf{n}_l \mathbf{n}_l^H \end{aligned}$$

$$+ \sum_{l=0}^{L-1} (\alpha_l - \alpha_l^2) e^{j\omega_l \tau_0} \mathbf{n}_l \phi_0^H +$$

$$+ \sum_{l=0}^{L-1} (\alpha_l - \alpha_l^2) e^{-j\omega_l \tau_0} \phi_0 \mathbf{n}_l^H.$$

It is readily seen that when $\mathbf{G} = \mathbf{I}$, i.e. when no noise is present in any of the sub-bands, the result is the same as that of Eq. (35), given by the first term of Eq. (36). The second term, which depends on the noise components of the GCC, is a rank L hermitian matrix with L non-zero real eigenvalues. The third term is a rank-one matrix with a non-zero eigenvalue associated to a noise-dependent complex eigenvector. Finally, the fourth term is another rank-one matrix, with one non-zero eigenvalue related to the eigenvector ϕ_0.

As a simplified example, a matrix $\mathbf{G}$ with L_g ones and $(L - L_g)$ zeros in its diagonal, i.e. either with perfect sub-bands or completely noisy sub-bands, will only contain the first (target) rank-one term and the second (noise) term with rank $(L - L_g)$. The eigenvalue corresponding to the target term will be $L_g \|\phi\|^2$, while the sum of the eigenvalues of the second term will be $(L - L_g)\|\phi\|^2$. In most practical cases, the greatest singular value is expected to be dominated by the target rank-one component, which justifies the use of a rank one approximation in terms of $\mathbf{u}_1$.

Taking the above discussion into account, under high SNR conditions, the first left singular vector should be dominated by the first term of Eq. (36) corresponding to the eigenvector ϕ_0. Since such eigenvector should be ideally real, we take the real part of $\mathbf{u}_1$, $\Re\{\mathbf{u}_1\}$:

$$\hat{\phi}_0[n] = \Re\{\mathbf{u}_1\} \operatorname{sign}\left(\Re\{u_1[\gamma]\}\right), \tag{37}$$

where $\Re\{u_1[\gamma]\}$ is the element of $\Re\{\mathbf{u}_1\}$ having the maximum absolute value, i.e.

$$\gamma = \arg\max_n |\Re\{u_1[n]\}|. \tag{38}$$

Note that this last equation and the $\operatorname{sign}(\cdot)$ function in Eq. (37) are used as a criterion for solving the singular vector sign ambiguity (both $\mathbf{u}_1$ and $-\mathbf{u}_1$ are eigenvectors of $\mathbf{RR}^H$), forcing a positive peak in $\hat{\phi}_0$. The estimated TDOA is given by the position of such peak:

$$\hat{\tau}_0 = \arg\max_n \hat{\phi}_0[n]. \tag{39}$$

The fourth column of Fig. 5 shows the recovered GCCs from the target SVD (SVD FS-GCC), $\hat{\phi}_0$, for the three speech frames shown in the previous examples. For comparison purposes, the results for conventional GCC-PHAT are shown in the third column.
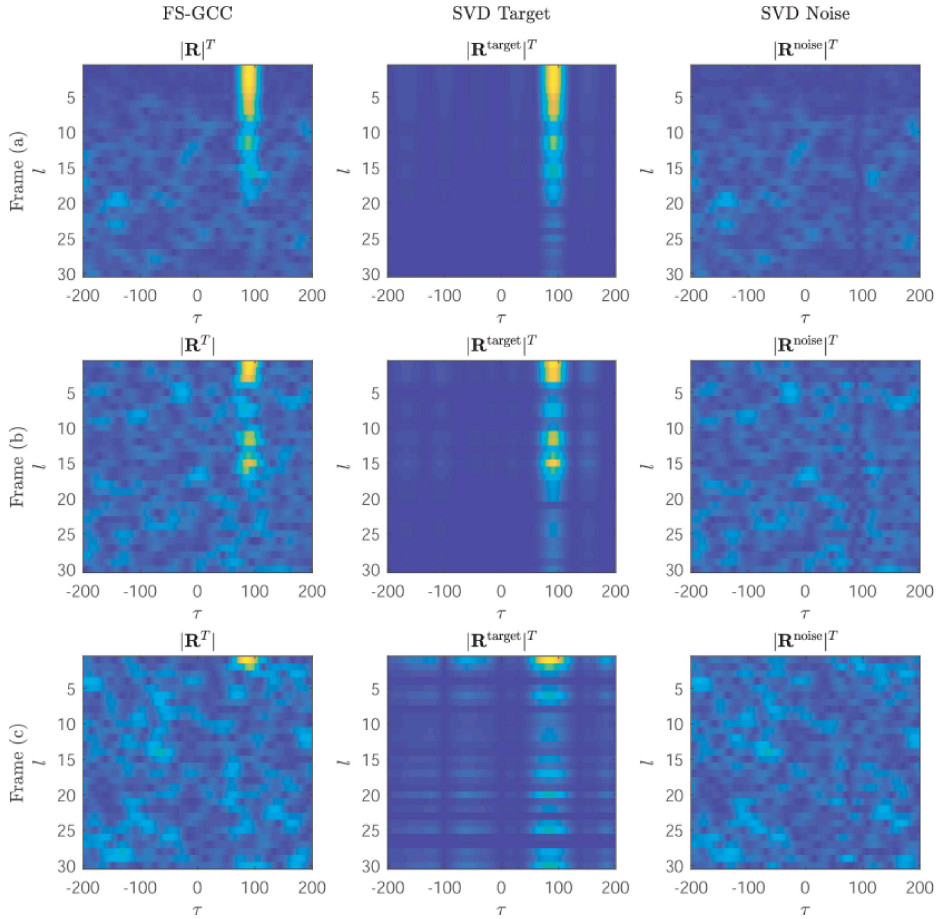
Figure 4: Extracted SVD components from three different speech frames (SNR = 25 dB) with true TDOA $\tau_0 = 92$. Left column: FS-GCC matrix. Middle column: target time-delay component. Right column: noise component.

IV-D *Weighted Low-Rank Approximation*

The Frobenius norm weights uniformly all the elements of the approximation error $(\mathbf{R} - \hat{\mathbf{R}})$. However, it can be very useful in practice to assign particular weights to the different sub-band GCCs, based on some defined confidence measure. Therefore, weighted low-rank approximations are of interest here, where the problem is now to solve [27]

$$\min_{\hat{\mathbf{R}}} \quad \left\| \left(\mathbf{R} - \hat{\mathbf{R}} \right) \odot \mathbf{W} \right\|_F, \quad \text{subject to} \quad \text{rank}(\hat{\mathbf{R}}) \leq r, \tag{40}$$

where $\mathbf{W} \in \mathbb{R}_+^{N \times L}$ is a weight matrix with non-negative weights and $\odot$ denotes the Hadamard product operator. The above equation is equivalent to solving a weighted Frobenius norm, where a given weight is assigned to each entry of the matrix $\mathbf{R}$, i.e.

$$\left\| \left(\mathbf{R} - \hat{\mathbf{R}} \right) \odot \mathbf{W} \right\|_F = \sum_{n=1}^{N} \sum_{l=1}^{L} w_{n,l}^2 \left(\mathbf{R}[n,l] - \hat{\mathbf{R}}[n,l] \right)^2, \tag{41}$$

where $w_{n,l}$ are the weight entries of the matrix $\mathbf{W}$. To address the problem, it is useful to consider the decomposition $\mathbf{R} = \mathbf{A}\mathbf{B}^H$, where $\mathbf{A} \in \mathbb{C}^{N \times r}$ and $\mathbf{B} \in \mathbb{C}^{L \times r}$. Any pair of matrices $(\mathbf{A}, \mathbf{B})$ providing such a decomposition of a rank-r matrix could be a potential solution, so

the problem can be formulated as an unconstrained minimization over pairs of matrices $(\mathbf{A}, \mathbf{B})$ [27]:

$$\min_{\mathbf{A}, \mathbf{B}} \quad \|\left(\mathbf{R} - \mathbf{A}\mathbf{B}^H\right) \odot \mathbf{W}\|_F. \tag{42}$$

In our particular application, the weight matrix can be restricted to follow the specific structure:

$$\mathbf{W} = \begin{bmatrix} w_0 \, w_1 \, \dots \, w_{L-1} \\ w_0 \, w_1 \, \dots \, w_{L-1} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ w_0 \, w_1 \, \dots \, w_{L-1} \end{bmatrix} = \mathbf{1}\mathbf{w}^T, \tag{43}$$

where $\mathbf{1}$ is a column vector of length N and $\mathbf{w} = [w_0, w_1, \dots, w_{L-1}]^T$ is a vector containing confidence weights $w_l \in [0, 1]$ assigned to each of the L sub-bands. Under such case, due to the structure of $\mathbf{W}$, the problem of Eq. (40) can be rewritten as the factorization of a modified matrix $\mathbf{R}_w$ [28]:

$$\min_{\mathbf{A}, \mathbf{B}_w} \quad \|\left(\mathbf{R}_w - \mathbf{A}\mathbf{B}_w^H\right)\|_F, \tag{44}$$

where

$$\mathbf{R}_w = \mathbf{R}\widetilde{\mathbf{W}}, \tag{45}$$

$$\mathbf{B}_w = \widetilde{\mathbf{W}}\mathbf{B}, \tag{46}$$

$$\widetilde{\mathbf{W}} = \operatorname{diag}(w_1, w_2, \dots, w_L). \tag{47}$$

By applying SVD factorization to $\mathbf{R}_w$ and conveniently truncating to the largest r singular values, the estimates of $\mathbf{A}$ and $\mathbf{B}_w$ are obtained as $\hat{\mathbf{A}} = \mathbf{U}\hat{\mathbf{\Sigma}}^{1/2}$ and $\hat{\mathbf{B}}_w = \mathbf{V}\hat{\mathbf{\Sigma}}^{1/2}$, where $\hat{\mathbf{\Sigma}}$ is the truncated singular value matrix. The rank-r approximation is therefore recovered as:

$$\mathbf{R}^{\mathbf{r}} = \hat{\mathbf{A}}\hat{\mathbf{B}}^H = \hat{\mathbf{A}}(\hat{\mathbf{W}}^{-1}\hat{\mathbf{B}}_w)^H. \tag{48}$$

Taking again into consideration that the noiseless sub-band GCC matrix is rank one, the target and noise components are extracted as:

$$\mathbf{R}^{\text{target}} = \hat{\mathbf{a}}_1 \hat{\mathbf{b}}_1^H, \tag{49}$$

$$\mathbf{R}^{\text{noise}} = \sum_{i=2}^{L} \hat{\mathbf{a}}_i \hat{\mathbf{b}}_i^H, \tag{50}$$

where $\hat{\mathbf{a}}_i$ and $\hat{\mathbf{b}}_i$ are, respectively, the columns of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. Similarly to the SVD case, the recovered GCC is obtained as

$$\hat{\phi}_0[n] = \Re\{\hat{\mathbf{a}}_1\} \operatorname{sign}\left(\Re\{\hat{a}_1[\gamma]\}\right), \tag{51}$$

where $\Re\{\hat{a}_1[\gamma]\}$ is the element of $\Re\{\hat{\mathbf{a}}_1\}$ having the maximum absolute value, i.e.

$$\gamma = \arg\max_n |\Re\{\hat{a}_1[n]\}|. \tag{52}$$

The target components resulting from the considered speech examples are shown in the second column of Fig. 5, while the recovered GCCs from the weighted SVD (WSVD FS-GCC) are shown in the last column.[1]
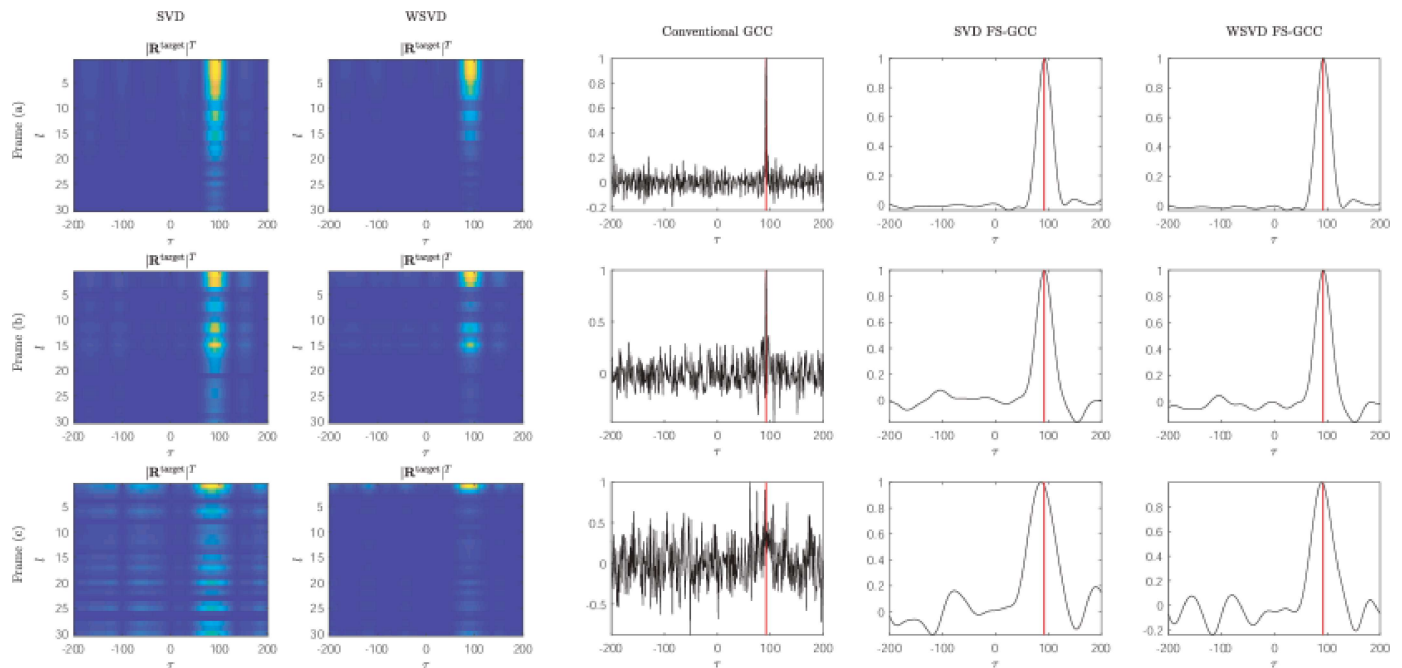
[1]Code available at https://github.com/spatUV/fs-gcc

Figure 5: Results for three speech frames using SVD FS-GCC and WSVD-FSGCC. The two first columns show the extracted target components using SVD and WSVD. The rest of columns show the recovered GCCs and the true TDOA (red line, $\tau_0 = 92$ samples). Note that the conventional GCC (third column) fails in speech frame (c).

IV-D1 Weighting

Note that the weights w_l can be directly related to the α_l coefficients of the model in Eq. (25). Weights having a value close to 1 must reflect that the corresponding band provides almost perfect delay information. In contrast, weights close to zero should reflect that the frequency band is dominated by noise and its information should be discarded. Similarly, perfect sub-bands are expected to have a magnitude GCC which is only dependent on the displaced window response ϕ_0, as expressed by Eq. (22). Therefore, the average of the magnitude of a perfect sub-band GCC ($\alpha_l = 1$), regardless the true TDOA, will be:

$$|\bar{\mathbf{r}}_l||_{\alpha=1} = \frac{1}{N} \sum_{n=0}^{N-1} |\phi[n]|. \tag{53}$$

In contrast, GCCs from completely noisy sub-bands ($\alpha_l = 0$) are given by noise realizations $\mathbf{n}_l$. The magnitude of a noisy band, which follows a Rayleigh distribution, will tend to its mean value:

$$|\bar{\mathbf{r}}_l||_{\alpha=0} = \frac{1}{2N} \|\phi\|^2 \frac{\pi}{2}. \tag{54}$$

Taking the above two extreme cases into account, the proposed weights are:

$$w_l = \begin{cases} g_l, & g_l \geq 0 \\ 0, & g_l < 0, \end{cases} \tag{55}$$

where

$$g_l = \frac{\left(|\bar{\mathbf{r}}_l||_{\alpha=0} - \frac{1}{N} \sum_{n=0}^{N-1} |r_l[n]| \right)}{\left(|\bar{\mathbf{r}}_l||_{\alpha=0} - |\bar{\mathbf{r}}_l||_{\alpha=1} \right)}. \tag{56}$$

Therefore, the weights are expected to vary between 1 (perfect sub-bands) and 0 (completely noisy sub-bands). Note that the above equations were derived by assuming that "noisy" sub-bands only comprise frequency bins where either the SNR is very low or there is no signal at all. These bands are not expected to incorporate early echoes or reverberation effects, at least, as long as the frame size is longer than the length of the acoustic channel. Under this assumption, the cross-power spectrum phase will be uniformly distributed in the range $[-\pi, \pi]$, and so the sub-band GCC coefficients wil correspond to a filtered complex Gaussian process [29] with Rayleigh-distributed magnitude.

V EXPERIMENTS

This section describes the experiments conducted to show the advantages of the proposed FS-GCC method in terms of TDE performance. The criteria used to assess such performance and the complete experimental set-up are next described.

V-A Performance Criteria

We classify a time delay estimate as either an anomaly or a nonanomaly according to its absolute error $e_i = |\tau - \hat{\tau}_i|$, where τ is the true time delay and $\hat{\tau}_i$ is the i-th time delay estimate. If $e_i > T_c / 2$, the estimate is assumed to be anomalous, where T_c is the signal correlation time [30]. For our particular source signal (speech), T_c was computed as the width of the main lobe of its autocorrelation function (taken between the -3 dB points), which is equal to 24 samples. The TDE performance is evaluated in terms of the percentage of

anomalous estimates over the total estimates ($P_{\hat{\tau}}$), the average GCC first-to-second peak ratio (FSPR), the mean and the standard deviation of the absolute error (all for the subset of nonanomalous estimates: $\text{FSPR}_{\hat{\tau},\text{na}}$, $\text{MAE}_{\hat{\tau},\text{na}}$, $\text{SDAE}_{\hat{\tau},\text{na}}$). These measures are defined as

$$P_{\hat{\tau}} = \frac{N_\text{a}}{N_\text{T}}, \tag{57}$$

$$\text{MAE}_{\hat{\tau},\text{na}} = \frac{1}{N_\text{na}} \sum_{i \in \chi_\text{na}} e_i, \tag{58}$$

$$\text{SDAE}_{\hat{\tau},\text{na}} = \frac{1}{N_\text{na}} \sum_{i \in \chi_\text{na}} \left(e_i - \text{MAE}_{\hat{\tau},\text{na}} \right)^2, \tag{59}$$

where N_T denotes the total number of estimates, N_a is the number of estimates that are identified as anomalies, N_na is the number of nonanomalous estimates, and χ_na represents the subset of nonanomalous estimates. The FSPR is defined as the average gain (over the subset of nonanomalous estimates) of the maximum GCC peak with respect to the second larger peak, i.e.

$$\text{FSPR}_{\hat{\tau},\text{na}} = \frac{1}{N_\text{na}} \sum_{i \in \chi_\text{na}} 20 \log_{10} \left(\frac{\hat{\phi}_0^{(i)}[\hat{\tau}_0]}{\hat{\phi}_0^{(i)}[\hat{\tau}_1]} \right), \tag{60}$$

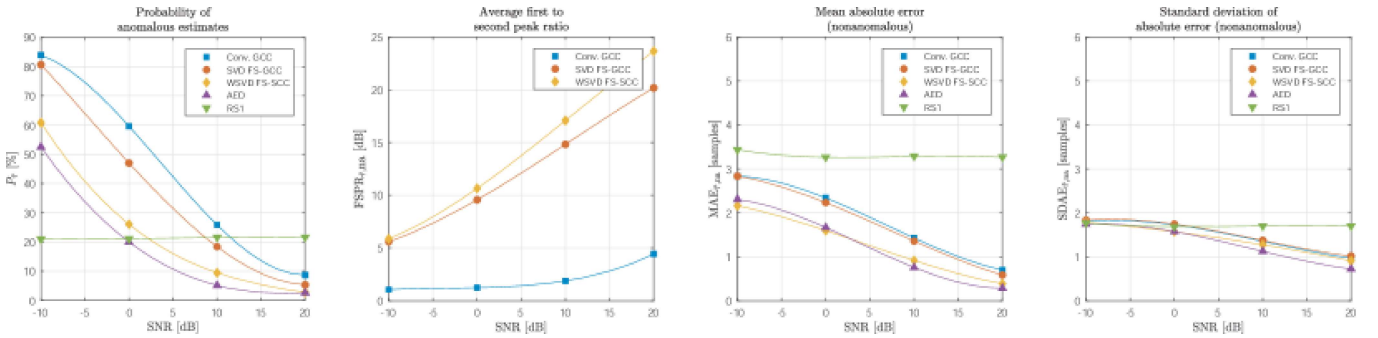where $\hat{\tau}_1$ denotes the lag of the second larger peak in the GCC.

Figure 6: Performance evaluation in anechoic scenario ($r_i = 0$). The fitting curve is a third order polynomial.
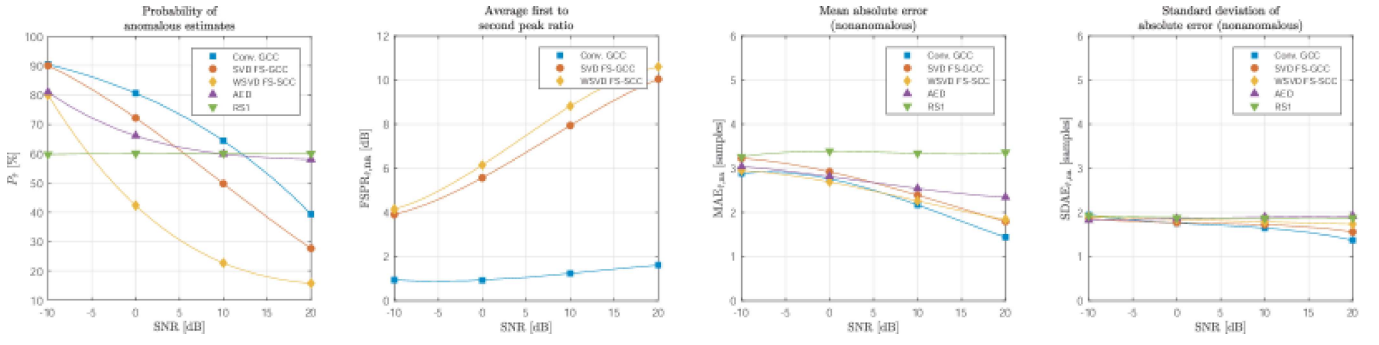
Figure 7: Performance evaluation in reverberant scenario ($r_i = 0.8$, $T_{60} \approx 0.3$ s). The fitting curve is a third order polynomial.

V-B *Simulation Set-up and Algorithm Parameters*

We consider a rectangular room simulated by the image-source method in a single source scenario [31]. Synthetic impulse responses were generated for a pair of sensors separated 0.5 meters considering 10 random array positions and orientations within the room, as well as 10 random source locations for each microphone configuration. The simulations were repeated for each reverberant condition. The following parameters were used:

- Room dimensions: $6 \times 7 \times 3$ meters ($x \times y \times z$).

- Uniform reflection coefficients: $r_i \in \{0.0, 0.8\}$.

- Source positions: 10 random positions on the plane ($x, y, z = 1.25$).

- Microphone positions: two-microphone array with inter-sensor spacing 0.5 m, with 10 random positions and orientations on the $x - y$ plane ($z = 1.25$).

- SNR: Varying between -10 dB and 20 dB. For each array set-up and source position, 10 different noise realizations were generated for each SNR condition. To control the SNR, mutually independent white Gaussian noise was properly scaled and added to each microphone signal.

- Source signal: A male speech signal of 2 seconds duration, digitized with 16-bit resolution at 44.1 kHz. The signal was processed to eliminate non-activity segments.

The synthetic microphone signals obtained by convolving the source signal with the generated impulse responses were processed by the different methods in the Short-Time Fourier Transform (STFT) domain. We used a frame length of 2048 samples and Hann windowing with 75% overlap, leading to 177 frames per audio example. The total number of estimates used to evaluate each method at each SNR and reverberant condition is therefore $N_T = 10 \times 10 \times 10 \times 177 = 177,000$.

The parameters used for evauating the FS-GCC approach were $B = 128$ (spectral window length) and $M = 32$ (hop) frequency bins. Results for a blind channel identification method, Adaptive Eigenvalue Decomposition (AED) [17], and a method based on ratios of acoustical transfer functions (RTFs) (the algorithm denoted as RS1 in [22]) are also given for comparison purposes. The AED method was configured to use rectangular windows of the same size (2048), a filter length of 512 samples and an adaptation step of $\mu = 0.003$. The RS1 method was configured as in a tracking scenario, considering as frames rectangular windows of 2048 samples with 75% overlap. The power spectral density (PSD) estimation needed for RS1 was performed using the Welch [32] method with Hann windows of 256 samples and 50% overlap, resulting in 15 periodograms for each PSD estimate. The Recursive Least Square (RLS) method, applied in RS1, was configured using a forgetting factor of $\alpha = 0.8222$, following [22]. Since AED and RS1 are not GCC-based methods, the FSPR is not provided for such algorithms.

V-C *TDE Results*

The results for the anechoic condition with a varying SNR are shown in Fig. 6. The percentage of anomalous estimates is clearly reduced for all the methods with respect to the conventional GCC-PHAT, with significant improvements achieved by WSVD FS-GCC, RS1 and, especially, AED. At the lowest SNR, RS1 is the most robust method, but its performance seems to be bounded for higher SNRs. Note that the different behavior of RS1 may be due to the aforementioned tracking scenario configuration. AED and WSVD FS-GCC follow a similar behavior. The biggest improvement is observed for SNR $= 0$ dB, where the difference between WSVD FS-GCC and conventional GCC is close to 35 percentage points (40 in the case of AED and RS1). The FSPR is significantly better for the FS-GCC methods than for the conventional GCC, especially at higher SNRs. All the methods outperform as well the conventional GCC-PHAT in terms of mean and standard deviation of time-delay errors, following a similar behavior (except RS1, which has a higher average error). This is an interesting result, since it means that having a lower temporal resolution due to the windowing effect does not affect negatively the TDE accuracy of FS-GCC.

The results for the reverberant case $r_i = 0.8$ (reverberation time, $T_{60} \approx 0.3$ s) are shown in Fig. 7. The percentage of anomalous estimates are now worse for all the methods, but in this case, both AED and RS1 are significantly affected by reverberation. In contrast, WSVD FS-GCC still shows very good robustness, with the most significant difference (more than 40 percentage points) at SNR $= 10$ dB. The FSPR is again considerably better for both FS-GCC methods, with SVD providing slightly better results than WSVD. Regarding nonanomalous TDOA errors, the FS-GCC methods provide slightly better results at low SNRs. At higher SNRs the conventional GCC-PHAT seems to be slightly more accurate, although in all cases the differences are very small (below 1 sample).

V-D *Impact on Source Localization Performance*

Although accurate time delay estimates are assumed to lead to better localization results, the advantages in terms of FSPR are also expected to contribute to better localization in SRP-based approaches, due to the mitigation of noise in the GCCs. To support such claim, experiments were conducted considering the same acoustic conditions and source signals but with six distributed microphones placed at the walls and corners of the room. The modified SRP algorithm (M-SRP) [33] was applied considering a grid resolution of 0.15 m. Fig. (8) shows an example of the resulting SRP maps when using conventional GCCs (a) and the proposed WSVD FS-GCCs (b) for the same signal frame (SNR = 0 dB). The improvement in terms of robustness to noise is clearly observed. The results for the mean and median absolute location errors are specified in Tables I and II, respectively, for anechoic and reverberant conditions. For comparison purposes, localization performance for AED is also provided by estimating the source location as the point of the grid having the lowest mean squared error considering all the available TDOAs. It can be observed that FS-GCC provides always more accurate location estimates and less anomalous detections (lower median) both in the anechoic case and in the reverberant case. Note that in the very worst case of SNR $= -10$ dB and reverberation, all the methods provide unreliable location estimates.
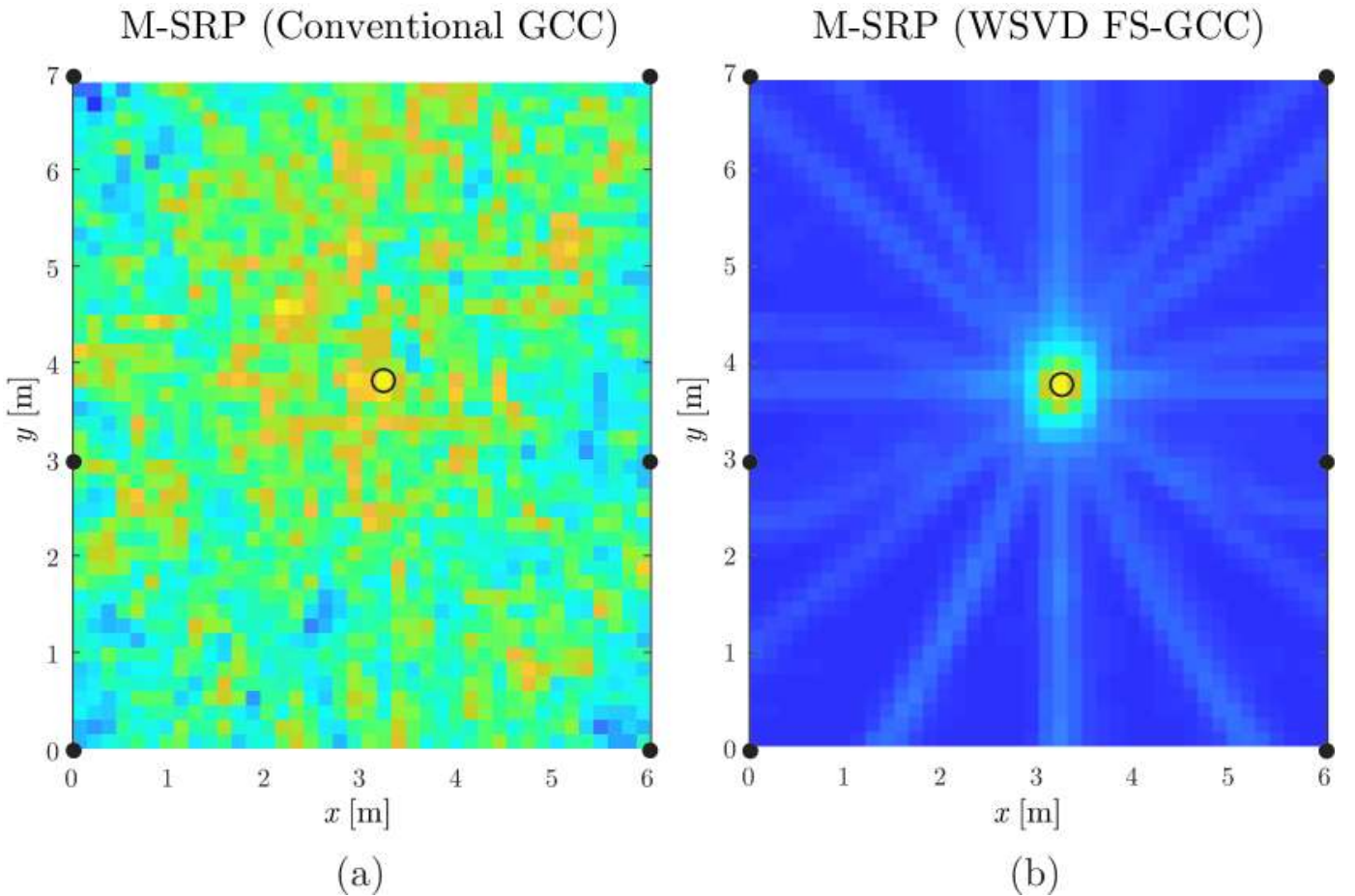

(a)

(b)

Figure 8: Example M-SRP maps (SNR = 0 dB). Microphones are represented with black dots. The circle indicates the source location. (a) Using conventional GCCs. (b) Using WSVD FS-GCCs.

TABLE I: Localization error [meters] for $r_i = 0$

Method	SNR = 20 dB		SNR = 10 dB		SNR = 0 dB		SNR = -10 dB	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GCC	0.2510	0.0839	0.6786	0.0844	1.9216	1.8230	2.7655	2.7102
FS-GCC	0.0716	0.0595	0.0819	0.0702	0.3685	0.0839	1.5298	0.4895
AED	0.9542	0.0702	0.9802	0.0702	1.3457	0.5862	1.9669	1.7781

TABLE II: Localization error [meters] for $r_i = 0.8$, $T_{60} \approx 0.3$ s

Method	SNR = 20 dB		SNR = 10 dB		SNR = 0 dB		SNR = -10 dB	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GCC	1.5722	0.9079	2.4151	2.4198	2.7016	2.6541	2.7025	2.6055
FS-GCC	0.2825	0.0839	0.6675	0.0926	1.5757	0.9293	2.7053	2.5615
AED	1.8427	1.6497	1.8589	1.6620	1.9365	1.8054	2.6129	2.5434

V-E *Direction of Arrival Estimation in Real Recordings*

In order to prove the effectiveness of the FS-GCC method also in a real environment, we tested Direction of Arrival (DoA) estimation performances over the LOCATA challenge dataset [2]. Specifically, we considered Task 1, consisting in a single static source and a single static microphone array, from the evaluation version of the dataset, for a total of 13 different recordings and 43122 samples. Among available setups, we chose the DICIT planar array, and considered as sub-arrays, microphones separated by a distance of 0.032 m, since this is the closest setup to the ones considered in the rest of the paper, thus facilitating the comparison. DoA estimation is performed using the M-SRP algorithm with a uniform spherical grid of regular triangles formed by 5 icosahedron subdivisions. Voice-Active Periods (VAP) were extracted considering the Voice-Activity Detection (VAD) labels provided with the dataset.

Azimuth and elevation average errors during VAP for Task 1 are shown in Tab. III. FS-GCC based M-SRP outperforms GCC both with respect to azimuth and elevation estimation, although only by a small amount. This behavior is expected, since the LOCATA dataset is not

excessively noisy and, as shown in Fig. 7, FS-GCC and GCC TDE performances become progressively similar as the SNR increases. Results related to elevation are significantly worse than the ones related to the azimuth, due to the fact that the DICIT array has a lower diversity on the z-axis (of the 15 microphones considered only two are positioned at a different height with the respect to the remaining ones). Although [2] provides only mean azimuth average errors, the values obtained using both FS-GCC and GCC are comparable with the other methods submitted to the LOCATA challenge and considering Task 1. It should be also emphasized that no tracking mechanism was used, as opposed to some of the methods evaluated in LOCATA.

TABLE III: Average azimuth and elevation errors [degrees] for task 1 of the LOCATA challenge

Method	Azimuth	Elevation
GCC	2.514	9.333
FS-GCC	2.391	8.536

VI Conclusion

This paper presented an improved GCC-based technique for TDE based on a sub-band analysis of the cross-power spectrum phase. The properties resulting from the so-called frequency-sliding GCC (FS-GCC) allows the recovering of denoised correlation signals by means of low-rank approximations of the FS-GCC matrix. The use of SVD and weighted SVD for obtaining both robust GCCs and accurate time delay estimates has been validated and compared to other well-known approaches, showing the relevant impact that the proposed technique can have in TDE and source localization performance.

References

[1]   M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

[2] C. Evers, H. Loellmann, H. Mellmann, A. Schmidt, H. Barfuss, P. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *arXiv*, vol. abs/1909.01008, 2019.

[3] A. Marti, M. Cobos, and J. J. Lopez, "Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2592–2595, May 2011.

[4] J. A. Belloch, A. Gonzalez, A. M. Vidal, and M. Cobos, "On the performance of multi-gpu-based expert systems for acoustic localization involving massive microphone arrays," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5607 – 5620, 2015.

[5] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, p. 026503, Dec 2006.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, August 1976.

[7] J. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, pp. 998–1003, December 1982.

[8] A. Weiss and E. Weinstein, "Fundamental limitations in passive time delay estimation–part i: Narrow-band systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, pp. 472–486, April 1983.

[9] E. Weinstein and A. Weiss, "Fundamental limitations in passive time-delay estimation–part ii: Wide-band systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 1064–1078, October 1984.

[10] G. C. Carter, "Coherence and time delay estimation," *Proceedings of the IEEE*, vol. 75, pp. 236–255, Feb 1987.

[11] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 791–803, Nov 2003.

[12] J. Perez-Lorenzo, R. Viciana-Abad, P. Reche-Lopez, F. Rivas, and J. Escolano, "Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments," *Applied Acoustics*, vol. 73, no. 8, pp. 698 – 712, 2012.

[13] Cha Zhang, D. Florencio, and Zhengyou Zhang, "Why does phat work well in lownoise, reverberative environments?," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2565–2568, March 2008.

[14] Jingdong Chen, J. Benesty, and Yiteng Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 549–557, Nov 2003.

[15] X. Alameda-Pineda and R. Horaud, "Geometrically-constrained robust time delay estimation using non-coplanar microphone arrays," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp. 1309–1313, Aug 2012.

[16] J. Benesty, Jingdong Chen, and Yiteng Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 509–519, Sep. 2004.

[17] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, pp. 384–91, 02 2000.

[18] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, p. 495250, Oct 2003.

[19] T. Padois, O. Doutres, and S. Sgard, "On the use of modified phase transform weighting functions for acoustic imaging with the generalized cross correlation," *The Journal of the Acoustical Society of America*, vol. 145, pp. 1546–1555, March 2019.

[20]   M. S. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," in *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, p. 4, Oct 1997.

[21]   B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 1110–1118, Nov 2005.

[22]   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.

[23]   J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 3021–3024 vol.5, May 2001.

[24]   K. Wu, V. G. Reju, A. W. H. Khong, and S. T. Goh, "Swarm intelligence based particle filter for alternating talker localization and tracking using microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 1384–1397, June 2017.

[25]   M. Cobos, M. García-Pineda, and M. Arevalillo-Herráez, "Steered response power localization of acoustic passband signals," *IEEE Signal Processing Letters*, vol. 24, pp. 717–721, May 2017.

[26] G. Strang, *Linear Algebra and Learning from Data.* Cambridge University Press, 1 ed., 1 2019.

[27] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 720–727, AAAI Press, 2003.

[28] P. M. Aguiar and J. M. F. Moura, "Factorization as a rank 1 problem," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 1, pp. 178–184 Vol. 1, June 1999.

[29] D. R. Brillinger, *Time Series: Data Analysis and Theory (Classics in Applied Mathematics), Ch. 4.* SIAM, 2001.

[30] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 148–152, March 1996.

[31] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[32] P. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[33] M. Cobos, A. Marti, and J. J. López, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.,* vol. 18, no. 1, pp. 71–74, 2011.

Maximo Cobos (SM'12) received the master's degree in telecommunications and the Ph.D. degree in telecommunications engineering from the Universitat Politècnica de València, Valencia, Spain, in 2009 and 2007, respectively. He completed with honors his studies under the University Faculty Training program (FPU) and was the recipient of the Ericsson Best Ph.D. Thesis Award from the Spanish National Telecommunications Engineering Association. In 2010, he received a Campus de Excelencia postdoctoral fellowship to work at the Institute of Telecommunications and Multimedia Applications. In 2011, he joined the Universitat de València, where he is currently an Associate Professor. His work is focused on the area of digital signal processing and machine learning for audio and multimedia applications, where he has authored/coauthored more than 100 technical papers in international journals and conferences. He is a member of the Audio Signal Processing Technical Committee of the European Acoustics Association and serves as Associate Editor for IEEE Signal Processing Letters.

Fabio Antonacci (M'14) was born in Bari, Italy, on July 26, 1979. He received the Laurea degree in 2004 in telecommunication engineering and the Ph.D. degree in information engineering in 2008, both from the Politecnico di Milano, Milan, Italy. He is currently an Assistant Professor at the Politecnico di Milano. His research focuses on spacetime processing of audio signals, for both speaker and microphone arrays (source localization, acoustic scene analysis, rendering of spatial sound) and on modeling of acoustic propagation. He is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee and of the EURASIP SAT on Audio, Speech and Music Signal Processing.

Luca Comanducci (S'19) received the B.S. degree in Music Information Science from the University of Milan, Milan, Italy, in 2014 and the M.S. degree in Computer Science and Engineering from Politecnico di Milano, Milan, Italy, in 2018, where he is currently working toward the Ph.D. degree in information engineering within the Dipartimento di Elettronica, Informazione and Bioingegneria. His main research interests concern the application of Deep Learning techniques to space-time audio signal processing.
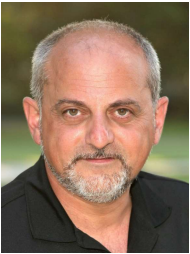
Augusto Sarti (M'04-SM'13) received his Ph.D. in Information Engineering from the University of Padova, Italy, in 1993 (joint program with University of California, Berkeley). After a few months as Post-Doc researcher at UC Berkeley, he joined the Faculty of the Politecnico di Milano (PoliMI), Italy, where he is currently a Full Professor. In 2013, he also joined the University of California, Davis, as a Full Professor (adjunct). He is the co-founder of the Image and Sound Processing Group at PoliMI, with which he promoted/coordinated and/or contributed to numerous European projects. He currently coordinates the activities of ISPG's audio labs. He is also the founder and coordinator of PoliMI's M.Sci. program in Music and Acoustic Engineering. He coauthored over 300 scientific publications on international journals and congresses and numerous patents in the multimedia signal processing area. His research interests are in the area of multimedia signal processing, with particular focus on sound analysis, synthesis, and processing; space-time audio processing; geometrical acoustics; and music information retrieval. He has been an Active Member of the IEEE Technical Committee on Audio and Acoustics Signal Processing and of the Editorial Board of the IEEE.