Three-dimensional sound source localization by distributed microphone arrays

Ali Dehghan Firoozabadi[1], Pablo Irarrazaval[2], Pablo Adasme[3], David Zabala-Blanco[4], Pablo Palacios-Játiva[5], Hugo Durney[1], Miguel Sanhueza[1], Cesar Azurdia-Meza[5]

[1]Department of Electricity, Universidad Tecnológica Metropolitana, Av. Jose Pedro Alessandri 1242, 7800002, Santiago, Chile

[2]Electrical Engineering Department, Pontificia Universidad Católica de Chile, 7820436, Santiago, Chile

[3]Electrical Engineering Department, Universidad de Santiago de Chile, Av. Ecuador 3519, Santiago 9170124, Chile

[4]Centro de investigación de estudios avanzados del Maule (CIEAM), Vicerrectoría de investigación y postgrado, Universidad Católica del Maule, Talca 3466706, Chile

[5]Department of Electrical Engineering, Universidad de Chile, Santiago 8370451, Chile

E-mail: adehghanfirouzabadi@utem.cl

***Abstract*— Multiple sound source localization (SSL) is one of the applicable and important areas in the speech signal processing. In this paper, a two-step method is proposed for multiple 3D SSL based on the time delay estimation (TDE) in combination with distributed microphone arrays (DMA). In the first step, the direction of speakers are estimated by the use of a circular microphone array (CMA) in the center of the room and implementing the generalized cross-correlation (GCC) function. In the second step, the distributed T-shaped microphone arrays on the walls are considered for 3D SSL. The two most closed T-shaped array to each speaker are selected, where one of them is used for horizontal and the other one for vertical direction of arrival (DOA) estimation by the use of generalized eigenvalue decomposition (GEVD) algorithm. The experiments on the simulated data for 2 and 3 simultaneous speakers show the superiority of the proposed distributed microphone array-direction of arrival estimators (DMA-DOAE) method in comparison with other previous works in noisy and reverberant environments.**

Keywords*— *Sound source localization, direction of arrival, cross-correlation, eigenvalue decomposition, microphone array.

I. Introduction

Sound source localization (SSL) is one of the important areas in the speech signal processing [1,2]. The combination between microphone arrays and SSL techniques are considered for improving the signal quality and estimating the speaker location. Steering beampattern to the direction of speakers is necessarily for improving the performance of the speech enhancement algorithms. In addition, the SSL in robotic applications is implemented for indoor and outdoor scenarios [3].

Various localization methods have been proposed in the recent decades, which are divided into the one-step and two-step algorithms. In two-step methods, the time difference of arrival (TDOA) is estimated between a pair of microphones [4]. The precision of the localization in these methods depends on the TDOA estimations, where the accuracy decreases in noisy and reverberant scenarios. The two-step methods have low computational complexity but the accuracy is low in undesirable environments. The one-step methods are based on the optimization of a cost function for some candidate points, where the location of speakers are estimated by searching the environments for finding the optimal points [5]. These categories of the methods localize the speakers with more accuracy but the computational complexity is higher in comparison with two-step methods. The steered response power-phase transform (SRP-PHAT) method is a proper algorithm for 3D SSL but the computational complexity is high because of searching the 3D candidate points. Maximo et al. proposed a proper strategy for decreasing the complexity of the SRP algorithm [6]. In the presented method, a practical implementation of the SRP-PHAT algorithm is proposed, which it uses the areas around of separated positions in the search space for source localization. Nikolas et al. proposed a perpendicular cross-spectra fusion (PCSF) algorithm as a novel method for DOA estimating, where it considers the analytic formulas for the estimations in the time-frequency (TF) domain [7]. In the presented method, the subsystems are proposed for DOA estimation, where they are implemented in a parallel structure for preparing the candidate DOAs in each TF points. Ning et. al. presented a method for binaural SSL due to the combination between the model-based information of speech spectral characteristics of sound sources and deep neural network (DNN) [8].

The main idea in this paper is the use of distributed microphone arrays (DMA) in combination with time delay estimation (TDE) algorithms for SSL. In the first step, a circular microphone array (CMA) in the center of the room in

combination with generalized cross-correlation phase transform (GCC-PHAT) method is considered for DOA estimation. Also, the i-vector probabilistic linear discriminant analysis (i-vector PLDA) algorithm [9] is selected for estimating the number of speakers. In addition, the number of peak positions in the GCC-PHAT function are extracted due to the number of speakers. Then, the two closest T-shaped microphone array to each speaker are selected for the 3D SSL. These two T-shaped microphone arrays in combination with the generalized eigenvalue decomposition (GEVD) algorithm are selected for horizontal (DOA_H) and vertical (DOA_V) direction estimation. The uncertainty area $\pm\alpha_H$ and $\pm\alpha_V$ are considered as an area for DOA estimations of T-shaped microphone arrays. This process is repeated for all speakers to find the central, horizontal, and vertical DOAs. Finally, three DOAs are intersected and the closest point to the DOA planes in the overlapped area is selected as the 3D location of a speaker.

Section 2 explains the real microphone signal model for simulations in noisy and reverberant environments. Section 3 shows the proposed idea for 3D SSL based on the DMA in combination with GCC-PHAT and GEVD algorithms. Section 4 represents the results of the evaluations on the simulated data for noisy and reverberant scenarios. Some conclusions are reported in Section 5.

II. THE MICROPHONE SIGNAL MODEL

Noise and reverberation are two undesirable environmental factors in real scenarios. The real model is proposed for considering the environmental effects, which is expressed as:

$$x_m(t) = \sum_{q=1}^{Q} x_{m,q}(t) = \sum_{q=1}^{Q} s_q(t) * \gamma_{m,q}\left(d^{(s)},t\right) + v_m(t) \quad \text{where} \quad \begin{cases} q = 1,\ldots,Q \\ m = 1,\ldots,M \end{cases}, \tag{1}$$

where $s_q(t)$ is the q-*th* sound source signal, $x_{m,q}(t)$ is the m-*th* microphone signal regarding to the q-*th* sound source, $\gamma_{m,q}\left(d^{(s)},t\right)$ is the impulse response between m-*th* microphone and q-*th* sound source, $v_m(t)$ is the Gaussian additive noise in the m-*th* microphone place, Q is the number of sound sources, M is the number of microphones, and * denotes to convolution operator.

III. THE 3D SSL BASED ON DMA, GCC-PHAT, AND GEVD ALGORITHMS

In this paper, a novel SSL system is proposed based on the DMA in combination with GCC-PHAT algorithm to avoid the complexity and GEVD method for preparing the high precision in reverberant environments. Fig. 1 shows the block diagram of the proposed system, where each part will be explained in the following.

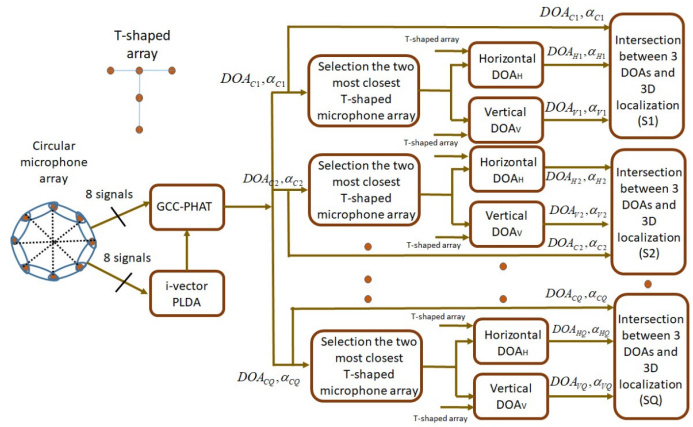

Fig. 1. The block diagram of the proposed 3D SSL algorithm based on DMA, GCC-PHAT, and GEVD methods.

A. Distributed microphone array

The structure of the microphone array highly affects the accuracy of the localization algorithms. In this paper, the DMA are proposed in combination with some localization algorithms. As shown in Fig. 2(a), in the first step a uniform CMA is considered at the center of the room. The speakers' directions are estimated by the combination between CMA and GCC-PHAT algorithm. This process has low computational complexity because of the use a few number of microphone signals in the CMA. Fig. 2 (a) shows the related microphone pairs of the CMA for the GCC-PHAT algorithm in DOA estimation. In the second step, the allocated T-shaped microphones on the walls are considered in combination with GEVD algorithm and the CMA in the center of the room with GCC-PHAT method for completing the 3D SSL process. The two closest T-shaped microphone array to the direction of each speaker are selected for estimating the horizontal (DOA_H) and vertical (DOA_V) directions. Fig. 2 (b and c) show the selected microphone pairs for horizontal and vertical DOA estimations, respectively.

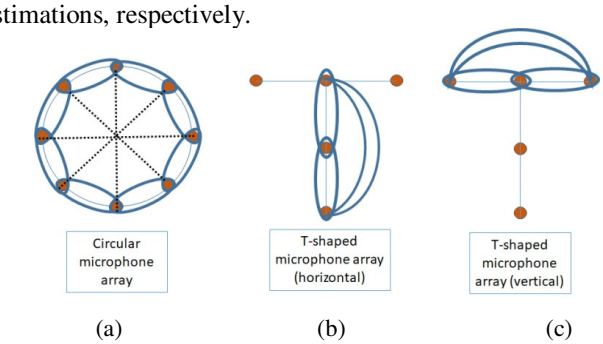

Fig. 2. The DMA for 3D SSL: a) the circular microphone array, b) the T-shaped microphone array for horizontal DOA_H estimation, and c) the T-shaped microphone array for vertical DOA_V estimation.

B. The CMA in combination with GCC-PHAT for DOA estimation

The generalized cross-correlation (GCC) is a proper function for TDE by the use of speech signals of a microphone pair. The speakers' DOAs are calculated by estimating the

TDOAs of the microphone signals [10]. The distances between microphone and speakers are shown by $r_{mq}\left(m=1,...,M\ \&\ q=1,...,Q\right)$. The TDOA between l-*th* and p-*th* microphones is shown as τ_{lp}.

The generalized cross-correlation (GCC) function ($R_{lp}(\tau)$) is the CC between the filtered version of microphone signals $x_l(t)$ and $x_p(t)$, where the Fourier transform of these filters are $G_l(\omega)$ and $G_p(\omega)$. $X_l(\omega)$ is the Fourier transform of signal $x_l(t)$ and $X'_p(\omega)$ is the complex conjugate of the Fourier transform from microphone signal $x_p(t)$. If the weighed function is defined as $\psi_{lp}(\omega)=G_l(\omega)G'_p(\omega)$, therefore the GCC function is simplified as:

$$R_{lp}(\tau)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\psi_{lp}(\omega)X_l(\omega)X'_p(\omega)e^{j\omega\tau}d\omega \tag{2}$$

The phase transform (PHAT) is considered for weighted function in GCC algorithm, which is defined as:

$$\psi_{lp}^{PHAT}(\omega)=\frac{1}{\left|X_l(\omega)X'_p(\omega)\right|} \tag{3}$$

Finally, the direction of the speakers is estimated by calculating the peaks of the GCC-PHAT function. We propose to average the DOAs of all microphone pairs for calculating the final DOAs more accurately.

$$\begin{aligned}
\hat{\theta}_{C1}&=\frac{1}{V}\sum_{v=1}^{V}\underset{0\le\theta\le2\pi}{\mathbf{arg\,max}}\,R_{lp}(\tau)\quad for\ \ l,p=1,...,Q_C\rightarrow \mathrm{DOA}_{C1}\\
\hat{\theta}_{C2}&=\frac{1}{V}\sum_{v=1}^{V}\underset{\substack{0\le\theta\le2\pi\\ \theta\neq\hat{\theta}_{C1}}}{\mathbf{arg\,max}}\,R_{lp}(\tau)\quad for\ \ l,p=1,...,Q_C\rightarrow \mathrm{DOA}_{C2}\\
&\ \ \ .\qquad\qquad .\qquad\qquad .\\
&\ \ \ .\qquad\qquad .\qquad\qquad .\\
\hat{\theta}_{CQ}&=\frac{1}{V}\sum_{v=1}^{V}\underset{\substack{0\le\theta\le2\pi\\ \theta\neq\hat{\theta}_{C1},...,\hat{\theta}_{CQ-1}}}{\mathbf{arg\,max}}\,R_{lp}(\tau)\quad for\ \ l,p=1,...,Q_C\rightarrow \mathrm{DOA}_{CQ}
\end{aligned} \tag{4}$$

where $\hat{\theta}_{C1},\hat{\theta}_{C2},...,\hat{\theta}_{CQ}$ are the estimated DOAs for the speakers by the use of CMA, Q_C is the number of microphones in the CMA, V is the number of microphone pairs in Fig. 2(a), which is considered as V =8 based on the selected CMA for the proposed algorithm. We consider the uncertainty area of estimated DOAs for each speaker as $\pm\alpha_C$, which is necessarily for calculating the final 3D location by the intersection between these areas of DMA. Therefore, the standard deviation (SD) parameter is selected for calculating the uncertainty area for each speaker as:

$$\alpha_{C,q}=\pm\sqrt{\frac{1}{V}\sum_{v=1}^{V}\left(\hat{\theta}_{Cq,v}-\hat{\theta}_{Cq}\right)^2}\quad for\ \ q=1,...,Q \tag{5}$$

where $\alpha_{C,q}$ is the uncertainty area of $\mathrm{DOA_C}$ for q-*th* speaker, $\hat{\theta}_{Cq,v}$ is the estimated DOA for q-*th* speaker by the use of v-*th* microphone pair, and $\hat{\theta}_{Cq}$ is the averaged DOAs of all microphone pairs in CMA for q-*th* speaker.

In the next step, the two closest T-shaped microphone arrays to each speaker are selected for the rest of the localization process. One of the T-shaped microphone arrays is considered for horizontal $\mathrm{DOA_H}$ estimation (see Fig. 2 (b)) and the other one for vertical $\mathrm{DOA_V}$ estimation (see Fig. 2 (c)), where 3 microphone pairs are considered for calculating the DOAs in GEVD algorithm based on the selected CMA. If the room is assumed as a linear time-invariant (LTI) system, the microphone signals ($\underline{x}_i(n)$) are written as:

$$\underline{x}_i^T(n)\underline{g}_j=\underline{x}_j^T(n)\underline{g}_i \tag{6}$$

The impulse response with length L is considered as:

$$\underline{g}_i=\left[g_{i,0},g_{i,1},.....,g_{i,L-1}\right]^T\quad,\quad i=1,2,3 \tag{7}$$

The covariance matrix R for three microphone signals is defined as following.

$$R=\begin{pmatrix} R_{x_1x_1} & R_{x_1x_2} & R_{x_1x_3}\\ R_{x_2x_1} & R_{x_2x_2} & R_{x_2x_3}\\ R_{x_3x_1} & R_{x_3x_2} & R_{x_3x_3}\end{pmatrix} \tag{8}$$

where the components of covariance matrix R are $R_{x_ix_j}=E\left\{\underline{x}_i(n)\underline{x}_j^T(n)\right\}$,$(i,j=1,2,3)$ and the vector $\underline{u}$ with length $3L$ is defined as:

$$\underline{u}=\begin{bmatrix}\underline{g}_3\\ -\underline{g}_2\\ -\underline{g}_1\end{bmatrix} \tag{9}$$

Vector $\underline{u}$ contains the eigenvector of matrix R related to the eigenvalue 0 (three impulse responses). Minimizing the cost function $\underline{u}^T R\underline{u}$ produces the optimal filter coefficients, where the error function is defined as:

$$e(n)=\frac{\underline{u}^T(n)\underline{x}(n)}{\left\|\underline{u}(n)\right\|} \tag{10}$$

where $\underline{x}(n)=\left[\underline{x}_1^T(n)\ \ \underline{x}_2^T(n)\ \ \underline{x}_3^T(n)\right]^T$ and the gradient of error function $e(n)$ based on vector $\underline{u}$ is defined as:

$$\nabla e(n) = \frac{1}{\|\underline{u}(n)\|}\left[\underline{x}(n) - e(n)\frac{\underline{u}(n)}{\|\underline{u}(n)\|}\right] \tag{11}$$

The constraint LMS algorithm is implemented as following for calculating vector $\underline{u}$.

$$\underline{u}(n+1) = \underline{u}(n) - \frac{\mu}{\|\underline{u}(n)\|}\left[\underline{x}(n)\underline{x}^T(n)\frac{\underline{u}(n)}{\|\underline{u}(n)\|} - e^2(n)\frac{\underline{u}(n)}{\|\underline{u}(n)\|}\right] \tag{12}$$

where μ is the adaptation step, small and positive value. Finally, we obtain the following expression by calculating the expected value of Eq. (11) after convergence as:

$$R\frac{\underline{u}(\infty)}{\|\underline{u}(\infty)\|} = E\left\{e^2(n)\right\}\frac{\underline{u}(\infty)}{\|\underline{u}(\infty)\|} \tag{13}$$

where $\underline{u}(\infty)$ is the eigenvector belongs to the smallest eigenvalue of covariance matrix R. Since the microphones in Fig. 2(a) are selected for horizontal direction estimation, the $\mathrm{DOA_H}$ value is calculated as:

$$\hat{\theta}_{\mathrm{T,H},q} = \frac{1}{3}\sum_{k=1}^{3}\hat{\theta}_{k,i,j,q} \quad for \quad \begin{cases} i = 1,...,3 \\ j = 1,...,3 \\ q = 1,...,Q \end{cases} \tag{14}$$

where $\hat{\theta}_{\mathrm{T,H},q}$ is the estimated $\mathrm{DOA_H}$ of the T-shaped microphone array for q-*th* speakers and the uncertainty area is estimated as:

$$\alpha_{\mathrm{T,H},q} = \pm\sqrt{\frac{1}{3}\sum_{k=1}^{3}\left(\hat{\theta}_{k,i,j,q} - \hat{\theta}_{\mathrm{T,H},q}\right)^2} \tag{15}$$

This process is repeated for calculating the vertical direction $\hat{\theta}_{\mathrm{T,V,q}}$ and uncertainty area $\alpha_{\mathrm{T,V,q}}$ for q-*th* speaker and the microphone pairs in Fig. 2(c) for obtaining the all directions $\alpha_{C,q}, \alpha_{\mathrm{T,H},q}, \alpha_{\mathrm{T,V},q}$ for final 3D location estimation. The final 3D location of q-*th* speaker is estimated by the intersection between three uncertainty areas and the closest point to all three DOA planes in the overlapped area is considered as the 3D location for a speaker. This process is repeated for estimating the 3D location for all Q speakers.

IV. Simulation and Results

The evaluations are implemented on the simulated data obtained by the TIMIT dataset [11]. The simulations are implemented on the scenarios for 2 and 3 simultaneous speakers. Two male and one female speakers are selected for data recording, Then, one male (S1) and one female (S2) speakers are considered for 2 simultaneous speakers and all speakers for 3 simultaneous speakers conditions. Fig. 3 shows a view of the simulated room with dimensions (592,475,420)cm, where the CMA is located in the center of the room, the T-shaped microphone arrays locate on the walls and the three speakers positions are S1=(108,264,174)cm, S2=(478,418,161)cm and S1=(496,94,180)cm, respectively.

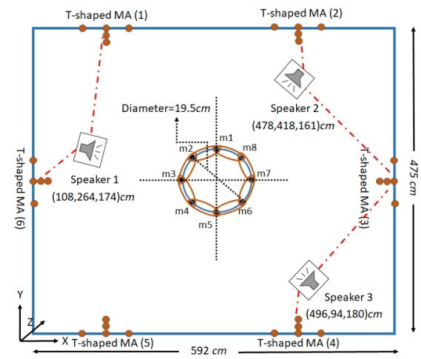

Fig. 3. A view of the simulated room with the location of speakers, CMA, and T-shaped microphone arrays.

The CMA contains of 8 microphones and the T-shaped microphone array (6 T-shaped arrays) on the walls contains of 5 microphones. In addition, the Hamming window with 60ms length and 50% overlap is selected for data segmentation.

In this paper, the additive white Gaussian noise in the microphone places is considered for simulating the environmental noises. In addition, the Image algorithm is selected for simulating the reverberation effects in the environments [12].

The proposed DMA-DOAE method is compared with MSRP-PHAT [6], PCSF [7] and SSM-DNN [8] algorithms by the use of mean absolute estimation error (MAEE) criteria in noisy and reverberant environments.

Fig. 4(a) shows the MAEE results in SNR=5dB, variable reverberation time $0 < RT_{60} < 700\text{ms}$ for 2 simultaneous speakers. As seen, the proposed DMA-DOAE method localizes the speakers more accurately in comparison with previous works. Also, the precision of all methods increases by decreasing the RT_{60} value, specially for the proposed DMA-DOAE method. Fig. 4(b) shows the results in $RT_{60} = 650\text{ ms}$ and variable SNR ($-10\text{dB} < \text{SNR} < 20\text{dB}$) for 2 simultaneous speakers. Also, this figure shows the higher precision of the proposed method in comparison with other works. The results of all methods are similar in high SNR values, but the proposed method localizes the speakers more accurately in low SNRs in comparison with previous works.

Fig. 5(a) shows the results for the proposed method in comparison with MSRP-PHAT, PCSF, and SSM-DNN algorithms by MAEE criteria for 3 simultaneous speakers for SNR=5dB and reverberation time $0 < RT_{60} < 700\text{ms}$. As shown, the proposed method has better precision in comparison with other previous algorithms. Fig. 5(b) shows the results of the proposed method for 3 simultaneous speakers in comparison with MSRP-PHAT, PCSF, and SSM-DNN algorithms for $RT_{60} = 650\text{ms}$ and variable SNRs $(-10\text{dB} < \text{SNR} < 20\text{dB})$. As shown the proposed method has better accuracy in the low SNRs in comparison with other previous works. Also, the location of the speakers are estimated with high precision by the proposed method in high SNR values. In general, the

results for 2 simultaneous speakers are better than the 3 simultaneous speakers, which is based on the reverberation effects due to the high number of speakers.

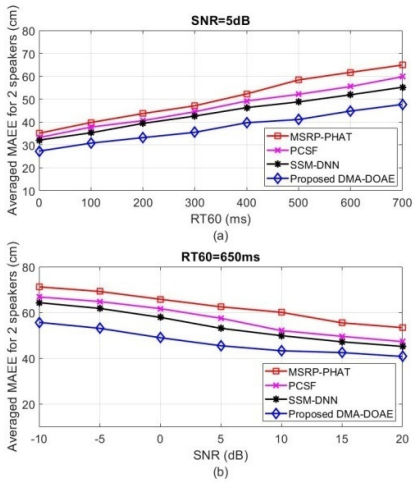

Fig. 4. The comparison between proposed DMA-DOAE, MSRP-PHAT, PCSF, and SSM-DNN algorithms by the use of MAEE (cm) criteria for 2 simultaneous speakers: a) for SNR=5dB and $0 < RT_{60} < 700$ms, b) for $RT_{60} = 650$ms and -10dB < SNR < 20dB.

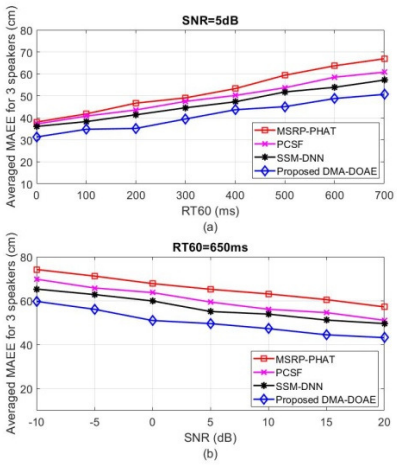

Fig. 5. The comparison between proposed DMA-DOAE, MSRP-PHAT, PCSF, and SSM-DNN algorithms by the use of MAEE (cm) criteria for 3 simultaneous speakers: a) for SNR=5dB and $0 < RT_{60} < 700$ms, b) for $RT_{60} = 650$ms and -10dB < SNR < 20dB.

V. Conclusions

In this paper, a two-step method is proposed for the SSL based on the DMA. In the first step, the direction of the speakers are estimated by the GCC-PHAT and i-vector PLDA for estimating the number of speakers. In the next step, two most closest T-shaped microphone arrays to each speaker are selected for 3D SSL. The T-shaped microphone arrays in combination with GEVD algorithm are considered for horizontal and vertical DOA estimations. Finally, the 3D speakers' locations are estimated by the intersection between the DOA of CMA and two horizontal and vertical DOAs of T-shaped microphone arrays. The intersection prepares an area, where the closest point to all surfaces is belong to the speaker location. The simulations are implemented on proposed DMA-DOAE method in comparison with MSRP-PHAT, PCSF, and SSM-DNN algorithms for 2 and 3 simultaneous speakers in noisy and reverberant environments. The results for variable SNRs and RT_{60} show the superiority of the proposed method specially in low SNRs and high RT_{60}.

Acknowledgment

The authors acknowledge financial support from: ANID/FONDECYT Postdoctorado No. 3190147 and ANID/FONDECYT No. 11180107.

References

[1] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancementand Array Signal Processing*. Singapore: Wiley-IEEE, 2018.

[2] L. Bianchi, F. Antonacci, A. Sarti, and S. Tubaro, "The ray space transform:A new framework for wave field processing," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5696-5706, Nov. 2016.

[3] D. Su, T. Vidal-Calleja, and J. V. Miro, "Towards real-time 3D soundsources mapping with linear microphone arrays," in *Proceedings IEEE International Conference Robotics and Automation*, Singapore, May/Jun. 2017, pp. 1662-1668.

[4] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploitingredundancy among multiple microphones," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 549-557, Nov. 2003.

[5] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT sourcelocation implementation using coarse-to-fine region contraction (CFRC)," in *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2007, pp. 295-298.

[6] M. Cobos, A. Marti and J. J. Lopez, "A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71-74, Jan. 2011.

[7] N. Stefanakis, D. Pavlidi and A. Mouchtaris, "Perpendicular Cross-Spectra Fusion for Sound Source Localization With a Planar Microphone Array," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1821-1835, Sept. 2017.

[8] N. Ma, J. A. Gonzalez and G. J. Brown, "Robust Binaural Localization of a Target Sound Source by Combining Spectral Source Models and Deep Neural Networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122-2131, Nov. 2018.

[9] I. Vinals, P. Gimeno, A. Ortega, A. Miguel and E. Lleida, "Estimation of the Number of Speakers with Variational Bayesian PLDA in the DIHARD Diarization Challenge," in *Proceedings Interspeech 2018,* Hyderabad, India, pp. 2803-2807, 2018.

[10] R. Lee, M. Kang, B. Kim, K. Park, S. Q. Lee and H. Park, "Sound Source Localization Based on GCC-PHAT With Diffuseness Mask in Noisy and Reverberant Environments," in *IEEE Access*, vol. 8, pp. 7373-7382, 2020.

[11] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1", Web Download. Philadelphia: Linguistic Data Consortium (1993). Available from: https://catalog.ldc.upenn.edu/LDC93S1. Last accessed May 2019.

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