Speakers counting by proposed nested microphone array in combination with limited space SRP

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, 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*— In this paper, a novel method is presented for estimating the number of speakers based on the microphone arrays. Firstly, a 3D snowflake nested microphone array (SNMA) is proposed for recording the speech signals. In the following, the steered response power (SRP) algorithm is implemented on subbands in limited spaces conditions for all microphone pairs related to the subarrays. Therefore, a weighted averaging method is implemented on subband limited spaces SRPs (LSRP), and the final energy map is compared with the histogram of the maximums of the SRP function on different subbands for various time frames. The passed candidate points are categorized by unsupervised K-means clustering and the number of speakers is estimated by the silhouette criteria. The accuracy of the proposed method is compared with PENS, i-vector PLDA, and wavelet-GEVD algorithms. The results show the superiority of the proposed method in comparison with other previous research.**

Keywords*— *Speakers counting, nested microphone array, subband processing, classification, filtering.

I. Introduction

Knowing the number of speakers is one of the requirements in such multisensory data processing systems as direction of arrival (DOA) estimation [1], sound source localization [2], speaker tracking [3], etc. Noise, reverberation, and spatial aliasing are the most well-known undesirable factors to decrease the accuracy of speakers counting algorithms. Therefore, the use of microphone array is an appropriate solution for increasing the performance of speakers counting algorithms.

In the recent years, various speakers counting methods have been proposed, where their main weakness is low accuracy in the high number of speakers and undesirable environmental situations. A group of methods are based on the eigenvalue calculation of covariance matrix from recorded data of the microphone arrays [4]. These methods work properly for up to 3 simultaneous speakers but the accuracy is decreased in the high number of speakers condition.

Mati and Thomas proposed the minimum description length (MDL) method for estimating the number of speakers [5]. Kumara et al. proposed a speakers counting method by the use of two-microphone structure [6]. For a specific speaker, the relative distance of certain excitation in vocal track is unchangeable for the recorded speech signal in two-microphone structures. Therefore, the time delays, in other words the number of speakers, are calculated according to the cross-correlation for the Hilbert envelop from linear estimation residual in the recorded signals. Ignacio et al. presented a speakers counting method based on the i-vector probabilistic linear discriminant analysis (PLDA) in diarization applications [7]. Firstly, the i-vectors are calculated for the segments of the recorded speech. In the following, these values are classified by the use of fully Bayesian PLDA algorithm. The number of speakers are counted based on the comparison between some hypothesis according to the differences of such information criteria. Halim and Siham proposed an estimating the number of speakers method according to the statistical calculation of the 7th Mel coefficients form speech spectrum components [8]. In our previous work, we proposed a speakers counting method based on a circular microphone array in combination with adaptive generalized eigenvalue decomposition (GEVD) algorithm [9].

In this paper, a novel method for speakers counting is proposed by the use of microphone array in undesirable environmental conditions. The spatial aliasing is one of the disadvantages in the use of microphone arrays. Firstly, we propose a 3D snowflake nested microphone array (SNMA) with the best inter-microphone distances for eliminating the

spatial aliasing and preparing better spectral components of microphone signals. In the following, the obtained speech signals of the microphone pairs related to each subband enter to the limited space subband SRP (LSRP) function. The computational complexity of the SRP method is decreased by the use of limited space area. Then, the outputs of the LSRP algorithm is weighted according to the SRP's maximums and the subband order. In addition, the LSRP algorithm is implemented repeatedly on each subband for preparing the histograms of the peak positions of the proposed function. The histograms are compared with weighted LSRP energy map, and the peaks with less distance of a threshold are selected and the rest values are denied. The selected peaks are categorized by the K-means clustering and the number of speakers is estimated by the silhouette criteria.

Section 2 shows the microphone signal model. Section 3 introduces the 3D SNMA for the speech data. In addition, the proposed LSRP algorithm in combination with K-means clustering and silhouette criteria are explained in this part. Section 4 shows the results of the evaluations and simulations. Some conclusions are included in section 5.

II. THE MICROPHONE SIGNAL MODEL FOR THE NESTED ARRAY

The real model for microphone signals is designed to consider the reflective effects in the environments as following:

$$x_m(t) = \sum_{q=1}^{N}\sum_{l=0}^{D-1} h_{mq}(l)\, s_q(t-l) + n_m(t), \quad \text{where} \quad (m = 1,\ldots,M) \quad (1)$$

where s_q is the speech signal in q-th source, h_{mq} is the room impulse response between q-th source and m-th microphone, n_m is the additive white Gaussian noise in m-th microphone, N is the number of speakers, M is the number of microphones, and D is the impulse response length.

III. THE PROPOSED NESTED ARRAY AND SPEAKERS COUNTING ALGORITHM

In this section, the proposed 3D NMA is proposed as a proper array for eliminating the spatial aliasing. In the following, the proposed speakers counting method is presented based on the LSRP algorithm, weighted averaging on energy maps, and K-means clustering with silhouette criteria. Fig. 1 shows the diagram of the proposed speakers counting method.

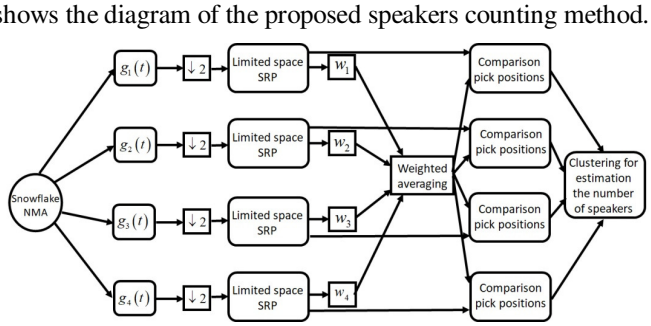

Fig. 1. The diagram of the proposed speakers counting methodof 3D SNMA.

A. The proposed 3D snowflake nested microphone array

The microphone array increases the accuracy of the speech processing algorithms due to preparing the more information. But the spatial aliasing according to the inter-microphone distances decreases the precision of the algorithms in the same time. Therefore, the nested microphone array were proposed for eliminating the spatial aliasing. We propose a 3D SNMA, which is implemented in combination with suggested method for estimating the number of speakers. Fig. 2 shows the proposed SNMA in the speaker counting applications.

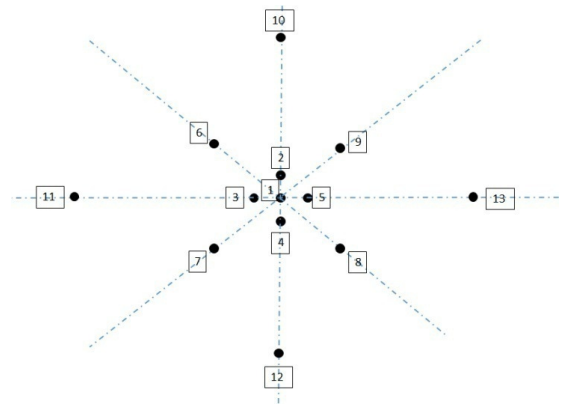

Fig. 2. The proposed snowflake nested microphone array for speakers counting application (Microphone 1 is in the z plane).

In this paper, the speech signal in frequency range [0-7800]Hz is selected in telephony applications with sampling frequency $F_s = 16000\text{Hz}$. The proposed SNMA is structured of 4 sub-arrays. The first sub-array is designed to cover the frequency range B1=[3900-7800]Hz. The relation between inter-microphone distance (d) and the wavelength for the maximum frequency component (λ) is $d \le \lambda/2$ to avoid the spatial aliasing. Therefore, the inter-microphone distance for the first sub-array is $d_1 \le 2.2cm$ and the central frequency for analysis filter is selected as $F_{c1} = 5850\text{Hz}$. The second sub-array covers the frequency range B2=[1950-3900]Hz, where the inter-microphone distance and the central frequency are selected as $d_2 \le 4.4cm$ and $F_{c2} = 2925\text{Hz}$, respectively. The third sub-array is structured for the frequency range B3=[975-1950]Hz for avoiding the spatial aliasing. Therefore, the inter-microphone distance is $d_3 \le 8.8cm$ and the central frequency is selected as $F_{c3} = 1462.5\text{Hz}$. Finally, the forth sub-array is designed for the frequency range B4=[0-975]Hz with the inter-microphone distance $d_4 \le 17.6cm$ and the central frequency $F_{c4} = 487.5\text{Hz}$. The spatial aliasing is eliminated completely by the use of designed SNMA. Fig. 3 shows the four sub-arrays related to the proposed SNMA.

The physical 3D array in the real environment is implemented by considering $d_1 \le 2.2cm$, $d2 \le 4.4cm$, $d_3 \le 8.8cm$, and $d_4 \le 16.9cm$, based on the available spaces in the environment. The analysis filter bank is required for avoiding the spatial aliasing and preparing the proper

frequency range for each sub-array. A multirate sampling with down samplers are selected to design this analysis filter bank.

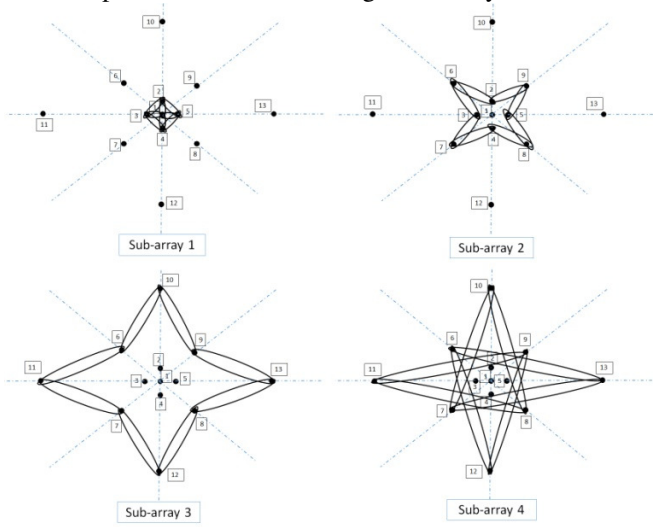

Fig. 3. Four designed sub-arrays for the proposed 3D snowflake nested microphone array.

B. *The proposed LSRP algorithm for speakers counting*

The proposed method for speakers counting is based on the source localization verification of the estimations according to a decision criteria, weighting of the information in different subbands, and the final classification for estimating the number of speakers.

If the microphone signal in Eq. (1) is considered for subband i (i=1,...,4) as $x_{m,i}(t)$, the output of the filter-and-sum beamformer (FSB) is defined as:

$$Y_i\left(\omega,\Delta_1...\Delta_{M_i}\right)\equiv\sum_{m=1}^{M_i}H_m(\omega)X_{m,i}(\omega)e^{-j\omega\Delta_m} \tag{2}$$

where $\Delta_1...\Delta_M$ are M_i (the number of microphones is subband i) steered delays, $X_1(\omega),....,X_{M_i}(\omega)$ are the Fourier transforms of the microphone signals and $H_1(\omega),....,H_{M_i}(\omega)$ are the Fourier transforms of filters for the i-th subband. We select a limited space (LS) area in the Z axis instead of considering whole 3D space to decrease the computational complexity of the proposed method. The steered response power for limited space area in each subband is defined as [10]:

$$P_i^{LS}\left(\Delta_1...\Delta_{M_i}\right)\equiv\int_{-\infty}^{+\infty}Y_i\left(\omega,\Delta_1...\Delta_{M_i}\right)Y_i'\left(\omega,\Delta_1...\Delta_{M_i}\right)d\omega \tag{3}$$

where $Y_i\left(\omega,\Delta_1...\Delta_{M_i}\right)$ is the output of FSB in frequency domain for i-th subband and $Y_i'\left(\omega,\Delta_1...\Delta_{M_i}\right)$ is its complex conjugate. The limited space SRP function for i-th subband is calculated by the combination of Eqs. (2) and (5) as:

$$P_i^{LS}\left(\Delta_1...\Delta_{M_i}\right)=\int_{-\infty}^{+\infty}\left(\sum_{u=1}^{M_i}H_u(\omega)X_{u,i}(\omega)e^{-j\omega\Delta_u}\right)\left(\sum_{v=1}^{M_i}H_v'(\omega)X_{v,i}'(\omega)e^{j\omega\Delta_v}\right)d\omega \tag{4}$$

The phase transform (PHAT) filter is a proper weighting function in reverberant conditions for the speech signals. This function decreases the effect of reverberation by whitening the speech spectrum components, where its combination with subband LSRP is defined as:

$$P_i^{LS(\mathrm{PHAT})}\left(\Delta_1...\Delta_{M_i}\right)=\sum_{u=1}^{M_i}\sum_{v=1}^{M_i}\int_{-\infty}^{+\infty}\frac{1}{\left|X_{u,i}(\omega)X_{v,i}'(\omega)\right|}X_{u,i}(\omega)X_{v,i}'(\omega)e^{j\omega(\Delta_v-\Delta_u)}d\omega \tag{5}$$

The peak positions in the LSRP function for each subband are related to the source locations, which can be considered as the number of speakers but it is not trustable because of noisy and reverberant conditions, where in following we propose some processes to improve its performance. Since the spectral components of various subbands are selected for the process, the weighted averaging is able to decrease the effect of undesirable environmental factors. In addition, the weighting increases the effect of low frequency components of the speech signal, which includes more information. Therefore, the weights are defined as:

$$w_i=\frac{i}{I}\cdot\frac{S_1\left(P_i^{LS(\mathrm{PHAT})}\left(\Delta_1...\Delta_{M_i}\right)\right)}{\sum_{j=2}^{R}S_j\left(P_i^{LS(\mathrm{PHAT})}\left(\Delta_1...\Delta_{M_i}\right)\right)},\quad (i=1,...,4) \tag{6}$$

where S_1 is the largest peak position in subband LSRP function, R is the number of microphone pairs in each sub-array (R=8 for the proposed SNMA), and I is the number of subbands (I=4). Vector S_j and weighted averaging of subband LSRP function are defined as:

$$S_j\left(P_i^{LS(\mathrm{PHAT})}\left(\Delta_1...\Delta_{M_i}\right)\right)=\underset{S_j\left(P_i^{LS(\mathrm{PHAT})}\right)\neq S_1\left(P_i^{LS(\mathrm{PHAT})}\right)\neq....\neq S_{j-1}\left(P_i^{LS(\mathrm{PHAT})}\right)}{\mathbf{argmax}}P_i^{LS(\mathrm{PHAT})}\left(\Delta_1...\Delta_{M_i}\right) \tag{7}$$

and,

$$\overline{P}_{weighted}^{LS(\mathrm{PHAT})}\left(\Delta_1...\Delta_{M_i}\right)=\frac{1}{I}\cdot\sum_{i=1}^{I}w_i.P_i^{LS(\mathrm{PHAT})}\left(\Delta_1...\Delta_{M_i}\right) \tag{8}$$

This weighted averaging is implemented in just one frame. In addition, the subband LSRP(PHAT) function is calculated

for various time frames and all subbands for preparing the histogram of the first R peaks as:

$$T_i = \left\{ S_1\left(P_i^{LS(\mathrm{PHAT})}\right), S_2\left(P_i^{LS(\mathrm{PHAT})}\right), ..., S_r\left(P_i^{LS(\mathrm{PHAT})}\right) \middle| \forall i \in I, r \in R \right\} \quad (9)$$

The histogram's peak positions with less than e cm (e=10 cm in this paper) in comparison with peaks positions of the subband LSRP(PHAT) function are selected and they enter to the classification unit. The other peaks with distance bigger than e value are denied.

$$\bar{A} = \left\{ \left| \mathbf{argmax}(T_i) - \mathbf{argmax}\, \bar{P}_{weighted}^{LS(\mathrm{PHAT})}\left(\Delta_1 ... \Delta_{M_i}\right) \right| \prec e, \; for \;\; i = 1, ..., I \right\} \quad (10)$$

In the following, the vector $\bar{A}$, which includes the enhanced peaks of all subbands enters to the K-means clustering with silhouette criteria for estimating the number of speakers. Since K-means is an unsupervised clustering method, the cost function is defined as:

$$Z = \sum_{h=1}^{N} \sum_{c=1}^{n} \left\| A_c^{(h)} - C_h \right\|^2 \quad (11)$$

where $\left\| A_c^{(h)} - C_h \right\|^2$ is the Euclidean distance between data $A_c^{(h)}$ and centroid C_h in each cluster. Finally, the silhouette criteria is implemented as a method to find the best number of clusters, where in this paper is the number of speakers [11].

IV. Simulation and Results

The proposed SNMA-LSRP algorithm for speakers counting is evaluated on various environmental conditions. The TIMIT dataset is selected for implementing the algorithms on simulated data [12]. The evaluations are implemented in the scenarios up to five simultaneous speakers for covering a wide range of real situations. Two males and three females speech signals are selected for the simulations. The simulations are implemented in a room with dimension (350,250, 230) cm, where the first to fifth speakers are located at S1=(85,218,175) cm, S2=(107,115,165) cm, S3=(90,46,161) cm, S4=(302,215,182) cm, and S5=(325,44,179) cm, respectively. Fig. 4 shows a view of the simulated room with the positions of SNMA and speakers.

A Hamming window with 50% overlap is selected for the simulations to prepare the best efficiency of the speech signal. The Image algorithm is selected for simulating the reverberation effects in the indoor environments [13]. In the first step, the proposed algorithm is implemented on three different scenarios and for up to five simultaneous speakers. The first scenario is called the reverberant conditions, where the effect of the reverberation is dominant in front of the noise with $RT_{60} = 650$ ms and SNR=20 dB. The second scenario is noisy environment by SNR=5 dB and $RT_{60} = 250$ ms where the noise is dominant. The last scenario is named noisy-reverberant condition, where both factors highly affect the speech signal as SNR=5 dB and $RT_{60} = 650$ ms The proposed SNMA-LSRP algorithm is compared with PENS [8], i-vector PLDA [7], and wavelet-GEVD [9] methods in different scenarios.

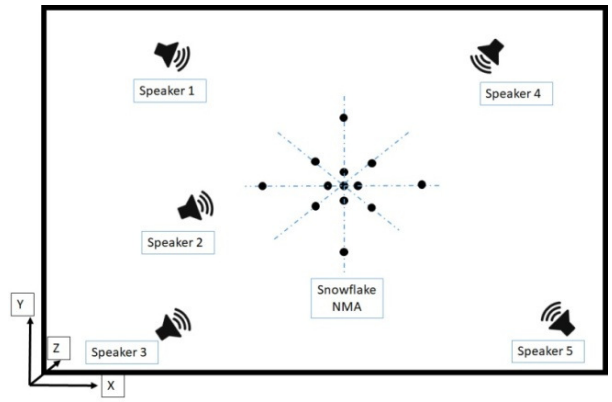

Fig. 4. A view of the simulated room with the speakers and SNMA.

Fig. 5 shows the results for noisy and reverberant environments from two to five simultaneous speakers. As seen, all methods have similar results in range 95-98% correct estimating for two simultaneous speakers. But the proposed method is more accurate for three, four and five simultaneous speakers in comparison with other previous works specially in five overlapped speech, where the proposed method provides 79% correct estimating the number of speakers in comparison with 54% in PENS, 55% in i-vector PLDA, and 71% in wavelet-GEVD methods.

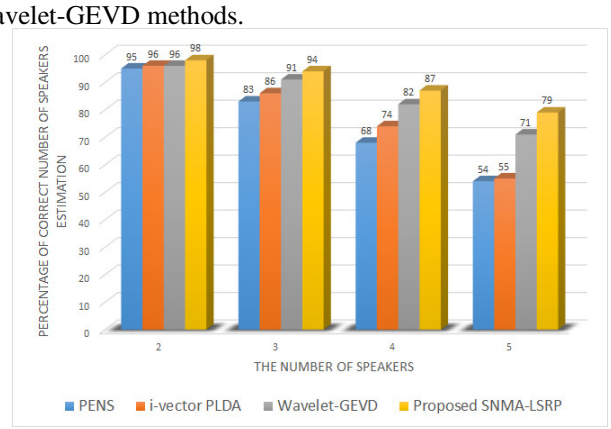

Fig. 5. The results of the proposed SNMA-LSRP method in comparison with PENS, i-vector PLDA, and wavelet-GEVD in estimating the number of speakers on simulated data for the noisy and reverberant environments (SNR=5 dB and $RT_{60} = 650$ ms)

In addition, the second category of experiments is for evaluating the proposed SNMA-LSRP method in comparison with PENS, i-vector PLDA, and wavelet-GEVD in the effect of the noise and reverberation variations. Fig. 6(a) shows the results of speakers counting in fixed noise (SNR=20 dB) and variable reverberation ($0 < RT_{60} < 700$ ms) five simultaneous speakers. As seen, all methods have proper accuracy in the

low RT_{60} values, but the accuracy is decreased by increasing the RT_{60}. The proposed method estimates the number of speakers more accurately in comparison with other previous works in all reverberation times, especially in bigger RT_{60}. Fig. 6(b) shows the results for fixed $RT_{60} = 250$ ms and variable SNR (-10 dB<SNR<20 dB) of the proposed method in comparison with other researches. As seen, the accuracy of the methods decrease in low SNRs but the proposed SNMA-LSRP method has better accuracy in all SNRs in comparison with PENS, i-vector PLDA, and wavelet-GEVD algorithms.

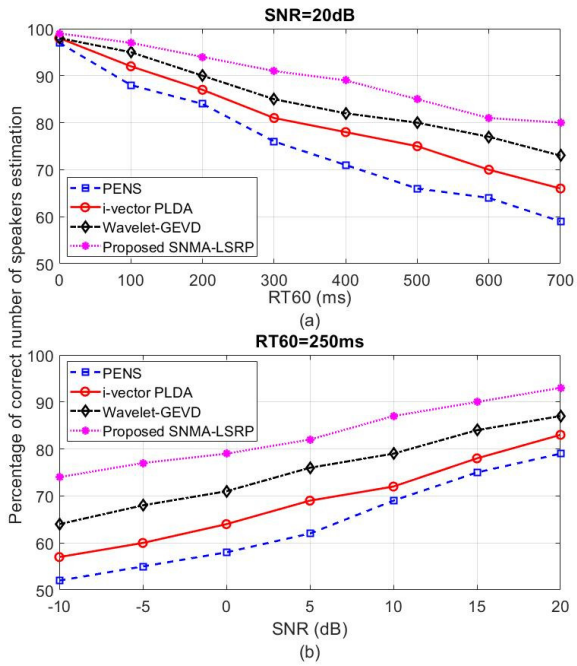

Fig. 6. The results of the proposed SNMA-LSRP method in comparison with PENS, i-vector PLDA and wavelet-GEVD for five simultaneous speakers in the next scenarios: a) fixed noise (SNR=20 dB) and variable reverberation ($0 < RT_{60} < 700$ ms), and b) fixed reverberation time $RT_{60} = 250$ ms and variable SNR (-10 dB<SNR<20 dB).

V. Conclusions

In this paper, a 3D snowflake nested microphone array was proposed for eliminating the spatial aliasing. The inter-microphone distance is adjusted in a way to prepare the proper information for speakers counting algorithm. In addition, the analysis filter in SNMA provides the possibility for subband processing. The SRP algorithm is implemented on limited space and subband format for preparing the energy map in each subband. A proposed weighted averaging method is considered for combining the LSRP-based information of all subbands to obtain the final LSRP energy map. In parallel, the LSRP algorithm is implemented on various time frames for each subband for calculating the histogram of peak positions. These histograms in different subbands are compared with combined LSRP energy map and the peaks with distance less than a threshold are selected for clustering section and the other peaks are denied. Finally, the K-means clustering algorithm with silhouette criteria is considered for estimating the number of speakers based on the selected peaks. The proposed SNMA-LSRP method is compared with PENS, i-vector PLDA, and wavelet-GEVD algorithms in the noisy and reverberant scenarios for up to five simultaneous speakers. The results in all conditions show the superiority of the proposed method in comparison with other previous works.

Acknowledgment

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

References

[1] Y. Zhou, Y. Li, L. Wang, C. Wen and W. Nie, “The Compressed Nested Array for Underdetermined DOA Estimation by Fourth-order Difference Coarrays,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, Barcelona, Spain, pp. 4617-4621, 2020.

[2] 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,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122-2131, Nov. 2018.

[3] K. Omiya and K. Suyama, “Multiple sound source tracking using low complexity directional estimation,” in *Proc. 17th International Symposium on Communications and Information Technologies (ISCIT)*, Cairns, QLD, pp. 1-6, 2017.

[4] M. S. Bartlett, “A note on the multiplying factors for various x approximations,” *Journal of the Royal Statistical Society*, vol. 16, no. ser B, pp. 296-298, 1954.

[5] M. Wax and T. Kailath, “Detection of signals by information theoretic criteria,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 2, pp. 387-392, Apr. 1985.

[6] R. K. Swamy, K. S. R. Murty and B. Yegnanarayana, "Determining Number of Speakers From Multispeaker Speech Signals Using Excitation Source Information," *IEEE Signal Processing Letters*, vol. 14, no. 7, pp. 481-484, July 2007.

[7] 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 *Proc. Interspeech 2018,* Hyderabad, India, pp. 2803-2807, 2018.

[8] H. Sayoud and S. Ouamour, “Proposal of a new confidence parameter estimating the number of speakers-An experimental investigation,” *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, no. 2, pp. 101-109, 2010.

[9] A. Dehghan Firoozabadi, P. Irarrazaval, P. Adasme, D. Zabala-Blanco and C. Azurdia-Meza, “A novel method for estimating the number of speakers based on generalized eigenvalue decomposition and adaptive wavelet transform by using K-means clustering,” *Signal, Image and Video Processing (SIVP)*, Vol. 14, pp. 1017-1025, July 2020.

[10] S. Tervo and T. Lokki, “Interpolation methods for the SRP-PHAT algorithm,” in *Proc. 11th International Workshop on Acoustic Echo and Noise Control* (IWAENC), 2008.

[11] J. R. Peter, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Computational and applied mathematics*, vol. 20, 53-65, 1987.

[12] 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.

[13] 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.