Journal of ELECTRICAL ENGINEERING, VOL 71(2020), NO3, 150–164

sciendo

PAPERS

Evaluation of localization precision by proposed quasi-spherical nested microphone array in combination with multiresolution adaptive steered response power

Ali Dehghan Firoozabadi[1*], Pablo Irarrazaval[2,3,4], Pablo Adasme[5], David Zabala-Blanco[6], Cesar Azurdia-Meza[7]

Multiple sound source localization in noisy and reverberant conditions is one of the important challenges in the speech signal processing. The aim of this article is three-dimensional sound source localization in undesirable scenarios. For the localization algorithms, the spatial aliasing is one of the destructive factors in reducing the accuracy. Firstly, a 3D quasi-spherical nested microphone array (QSNMA) is proposed for eliminating the spatial aliasing. Since the speech signal has the windowed-disjoint orthogonality property, the speech information differs in terms of the frequency bands. Then, the Gammatone filter bank is introduced for the speech subband processing. In the following, the multiresolution steered response power (SRP) algorithm is adaptively implemented on subbands with the phase transform (PHAT)/maximum likelihood (ML) weighted functions based on the levels of the noise and reverberation. The peaks of the multiresolution adaptive SRP (MASRP) algorithm are extracted in each subband based on the number of speakers for continuous time frames. Finally, the distribution of these peaks are calculated in each subband and they are merged by the use of weighted averaging method. The final 3D speakers locations are estimated by extracting the peaks in the final distribution. The proposed QSNMA-MASRP(PHAT/ML) algorithm is evaluated on real and simulated data for 2 and 3 simultaneous speakers in noisy and reverberant conditions. The proposed method is compared with SRP-PHAT, spectral source model-deep neural network, and spherical harmonic temporal extension of multiple response model sparse Bayesian learning algorithms on different range of signal-to-noise ratio and reverberation time. The mean absolute estimation error, averaged standard deviation for absolute estimation error, and computational complexity results show the superiority of the proposed method.

K e y w o r d s: sound source localization, nested microphone array, subband processing, time delay estimation, filter bank

1 Introduction

In the recent years, many researches have been done on the signal processing related to smart meeting rooms and robotics, which are considered as important fields for the source localization algorithms. Localization and tracking [1] are two important categories of these processes, which are obtained directly for sound source localization (SSL) implementations, and indirectly for speech enhancement applications by steering the beampattern to the speaker direction for preparing the better quality of recorded signals. Also, other applications are be for robotic systems, where the robots follow the speakers in order to realize some specific tasks. Therefore, they should first localize the speakers' locations and after that, they are able to record the voice instructions and commands. In the simultaneous multi-speakers condition, it is important the robots find the true speaker and follow the instructions. The localization algorithms are used for sound sources as human speakers or noise sources such as fan, air condition, car, etc. Also, the localization algorithms are implemented for one or multiple speakers scenarios, where unknown number of speakers is another challenges in this area. The accuracy of the localization algorithms is highly dependent on the noise and reverberation, which make uncertainly in the final results. In addition, the detection of the speech signal overlapped area and the implementation of the localization methods on single or multiple speakers regions are other important factors for localization algorithms. Most of the localization methods use the microphone array for digital audio signal recording [2]. The advantages of microphone array are pattern control in the desired direction and information redundancy because of the number of microphones. But spatial aliasing is the disadvantage to use the microphone arrays because of inter-microphone distances.

In many applications, SSL is a key part of signal processing systems, where still there are many challenges in this area. SSL is used in many recent researches such as: speaker separation [3], steering cameras to speaker direction in smart rooms [4,5], source localization in low light areas where video systems cannot be used [6], and

Department of Electricity, Universidad Tecnológica Metropolitana, Av. José Pedro Alessandri 1242, Santiago 7800002, Chile, [2] Electrical Engineering Department, [3] Biomedical Imaging Center, [4] Institute for Biological and Medical Engineering, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile, [5] Electrical Engineering Department, Universidad de Santiago de Chile, Av. Ecuador 3519, Santiago 9170124, Chile, [6] Department of Computing and Industries, Universidad Católica del Maule, Talca 3466706, Chile, [7] Department of Electrical Engineering, Universidad de Chile, Santiago 8370451, Chile, * Corresponding author: adehghanfirouzabadi@utem.cl, ORCID: 0000-0002-6391-6863

DOI:10.2478/jee-2020–0022, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

source identification in robotic systems [7,8]. Although many methods have been proposed in the recent years, but multiple SSL in noisy and reverberant environments is still an important challenge in the speech signal processing. There are two main categories of SSL algorithms due to the received signals by microphone array. Both categories use the generalized cross-correlation (GCC) calculation for the pairs of microphone array signals [9]. The first category of methods are one-step algorithms based on the steered response power (SRP), where the aim is maximizing the output of this function due to a series of candidate source locations [10,11]. The SRP function is calculated for all candidate locations of the search space and its global maxima is considered as the estimated source location, which contains high computational complexity process. One of the important keys in the use of SRP algorithm is the synchronization of recorded signals from microphone array to prepare the accurate results. The SRP method has high computational complexity in the implementations but it works accurately in the noisy and reverberant conditions. In contrast, there are two-step methods based on the GCC function for SSL [12]. In the first step, the time difference of arrival (TDOA) values are calculated due to the peak positions of GCC function for each microphone pairs [13]. Then, the direction or 3D location of speakers are estimated by the use of these TDOAs and the geometry of the microphone array, which considers such criteria as hypercone fitting problem [14]. The TDOA-based methods for SSL are very sensitive against of the noise and reverberation and their accuracy is greatly diminished, but they can be implemented faster due to the complexity. Also, the SSL methods are divided into the parametric and non-parametric algorithms. The parametric methods [15] are the beamformer and maximum likelihood (ML) functions but the non-parametric methods [16] are based on the subband signals and eigenvalue analysis, where two important algorithms are multiple signal classification (MUSIC) [17] and estimating signal parameters via rotational invariance technique (ESPRIT) [18], which have higher resolution in comparison with parametric methods.

In the last decade, many algorithms have been proposed for single and multiple speaker localization in indoor environments. Some methods use the two-microphone structure and the others consider the microphone array for audio recording. In addition, the localization algorithms are proposed for direction of arrival estimation (DOA) estimation or 3D SSL. In the following, the most important researches for SSL are explained. Byoungho *et al*]; proposed a multiple speaker localization method based on a spatial mapped GCC function [19]. This method uses the GCC function for SSL but the source locations are not estimated directly based on the calculated TDOAs and a spatial mapping is considered in the localization procedure. In the presented method, the source locations are estimated by the use of spatial mapped function. The spatial functions are transferred to other coordinate and a mathematical expression is obtained for the source location by the summation of mapped GCC functions. The maximums of the summed GCC function are selected as the source locations.

Yusuke *et al* proposed the histogram mapping method for multiple SSL based on the TDOA calculations [20]. Firstly, the TDOAs are estimated by the use of each pair of microphone signals. Then, the DOAs are calculated by averaging the estimated histograms. This method does not need any initial estimation of speaker DOAs. Finally, the mapping stage is implemented on the estimated DOAs for positioning the speakers.

Mojtaba *et al* proposed a method based on the relative transfer function (RTF) for SSL in hearing aid system applications [21]. This article explains the binaural aid systems for target DOA estimation in the noise-free conditions. A system based on ML function is proposed for DOA estimation, which models the shadowing effects of user's head on microphone signals as a RTF between hearing aid system of microphone signals. Three various RTF methods are presented with different precisions and resolutions. Also, the presented DOA estimators are shown based on the inverse discrete Fourier transform to be able for calculation the ML computational complexity.

Nikolaos *et al* proposed a perpendicular cross-spectra fusion function for multiple simultaneous SSL by the use of microphone array [22]. In the presented work, the perpendicular cross-spectra fusion method is introduced as a novel algorithm for DOA estimation, which uses the analytic formulas in time-frequency (TF) domain for estimating the speakers directions. The proposed method prepares the various candidate DOAs in each TF point in addition to estimating the parallel speakers locations. Also, a coherence criteria based on the divergence property of DOA estimators is presented for evaluating the reliability of different parts of speech signal, which permits to consider just the TF bins with high quality of information. This criteria prepares the high precision for multiple SSL in undesirable noisy and reverberant conditions.

Ning *et al* presented a binaural SSL method by combination the spectral source model and deep neural network (SSM-DNN) [23]. A few model-based methods have been proposed for simultaneous SSL in adverse environmental conditions. In this presented work, a new framework is proposed for simultaneous binaural SSL, which considers a combination of model-based information of speech spectral features and DNN structure. Firstly, a model for target and another for background sources are selected in the phase training step by the use of extracted spectral features. If the background source identity is unknown, a universal model can be considered for this condition. In the next step, the source models are considered jointly for mixed observations and improving the localization procedure by the use of weighed azimuth selection with a DNN-based localization system. Finally, the proposed method uses the combination of model-based and data-driven information for introducing a single computational framework for SSL. The presented method works accurately

in reverberant scenarios with the presence of interfering noise sources.

Wei and Huawei proposed a simultaneous SSL method for reverberant conditions by the use of spherical harmonic Bayesian learning [24]. In the recent years, the spherical localization methods have been proposed based on three-dimensional microphone array. The performance of sparse methods is decreased for multiple SSL in indoor conditions due to the reverberation. In this presented work, a sparse-based multiple localization method is proposed for reverberant environments, which considers the spherical harmonic temporal extension of multiple response model sparse Bayesian learning (SH-TMSBL) for SSL. The precision of SSL algorithms are increased by the use of 3D microphone structure for array signal processing in spherical harmonic domain. The results show the accuracy of the presented method in indoor reverberant conditions. In this article, a novel method for three-dimensional multiple simultaneous SSL is presented by use of the proposed 3D quasi-spherical nested microphone array (QSNMA) in combination with the multiresolution adaptive SRP algorithm based on the Gammatone filter bank, which is adaptively implemented with phase transform (PHAT)/ML weighted functions (MASRP(PHAT/ML)). The microphone array advantage is information redundancy because of the number of microphones, and its disadvantage is the spatial aliasing due to the inter-microphone distances. The spatial aliasing decreases the precision of the localization algorithms. Firstly, a 3D QSNMA is proposed as a solution for eliminating the spatial aliasing. The inter-microphone distance is adjusted in the proposed QSNMA to consider the specific microphone pairs for each subarray, which avoids to have spatial aliasing. Also, this nested microphone array (NMA) decreases highly the computational complexity by selecting the specific microphone pairs for each subband. The speech is a non-stationary and wideband signal with windowed-disjoint orthogonality (W-DO) property [25], which means the lower frequency bands have more spectral information. In the case of several simultaneous speakers, each microphone receives a mixed signal from the speakers. As long as the time representation (waveform) of the each microphone signal shows a mixed and overlapped version of the speech signals from individual speakers, there is no overlap between the signals of the speakers in many of the TF points. Then, each TF point can be considered with high probability just for one speaker based on the W-DO property. Then, the Gammatone filter bank, as a human hearing based filter, is proposed for subband processing of microphone array signals. In the following, the multiresolution SRP algorithm is implemented on the microphone signals for each subband, and adaptively, by the use of PHAT/ML weighted functions. The largest peaks for the MASRP(PHAT/ML) algorithm are extracted in each subband based on the number of speakers, and this process is repeated for continues time frames. Finally, the MASRP(PHAT/ML) peaks histogram is calculated for each subband, and the final histogram is obtained by the fusion between subband histograms with the weighted averaging method. The first N-peaks in the final histogram are selected as the speakers locations. As initial work [26], we introduced this idea briefly on the simulated data for 2 simultaneous speakers on the specific environmental conditions. In this article, the proposed algorithm is explained comprehensively by the extended formulas, and the proposed method is implemented on the real data for 2 and 3 simultaneous speakers. Also, the accuracy of the proposed method is evaluated by the mean absolute estimation error (MAEE) criteria on the fixed and variable range of signal-to-noise ratio (SNR) and reverberation time. In addition, the SSM-DNN method is included in the results section, and the computational complexity of the proposed algorithm is compared with other previous works, which shows the superiority of the proposed QSNMA-MASRP(PHAT/ML) algorithm based on the high accuracy and acceptable computational complexity.

We present the microphone signal model and the propose a quasi-spherical nested microphone array. After showing the proposed localization algorithm based on the Gammatone filter bank, multiresolution SRP, and adaptive use of PHAT/ML weighted functions the simulations and the real data are discussed.

2 Microphone signal model and proposed quasi-spherical NMA

The microphone signal modeling is a principal part of the implementation of the localization algorithms. The microphone signal models are explained below and proposed 3D quasi-spherical NMA as a proper tool for elimination the spatial aliasing in the localization algorithms is introduced.

2.1 Microphone signal model

In simulations for the speech signal processing, specially localization algorithms, the microphone signal model is one of the important factors for the evaluation of the proposed algorithms. The selected model should prepares the conditions similar to real environments. Then, the ideal and real models are considered for the microphone signals in the simulations. In the ideal model, it is assumed that the signal received by microphone is a weakened and delayed version of the speech source signal. This model is mostly selected for outdoor scenarios, which is expressed as follows

$$x_m(t) = \frac{1}{r_m} s(t - \tau_m) + \breve{v}_m(t)\,, \tag{1}$$

where $x_m(t)$ is the signal received by m-th microphone, τ_m is the time delay, r_m is the distance between the source and m-th microphone, $s(t)$ is the speech source signal, and $\breve{v}_m(t)$ is the additive noise in m-th microphone. This model is named ideal because the reverberation and reflective surfaces effects are not considered.

Therefore, it cannot model the recorded speech signals in indoor conditions. In contrast, the real model is presented for microphone signals, which contains the room reverberation effects. The real model for microphone signals is expressed as follows

$$x_m(t) = \frac{1}{r_m} s(t - \tau_m) * \gamma_m(t) + \breve{v}_m(t)\,, \qquad (2)$$

where $\gamma_m(t)$ is the impulse response of m-th microphone to particular source. Symbol $*$ denotes the convolution operator, which contains the reverberation and all reflective effects in the indoor conditions. The signal received by m-th microphone is expressed as the convolution of this "room impulse response" and the source signal, including the noise. In addition, the near-field and far-field assumptions are defined for microphone signals. In the near-field assumption, the speech signal arrives to the microphone array spherically due to the short distance between the source and microphone array in comparison with array dimensions. But in the far-field assumption, the speech signals reach to the microphone as a flat shape because of the greater distance. The near field assumption is considered in the evaluations based on the environmental conditions, shown in Fig. 1.

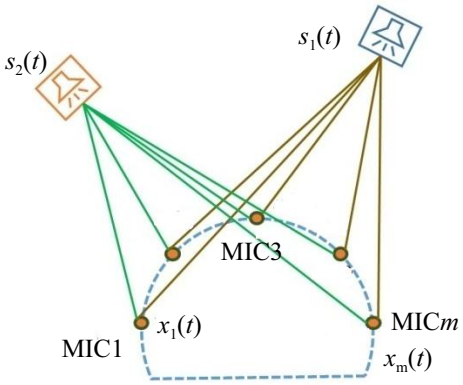

Fig. 1. The near-field assumption for microphone signals in the proposed localization algorithms

2.2 The proposed quasi-spherical NMA

The microphone array provides the suitable information due to the many numbers of microphones, but the spatial aliasing appears as a disadvantage because of the inter-microphone distances. The spatial aliasing decreases the accuracy of the localization algorithms due to the destruction of speech spectral information. The nested microphone array is introduced to solve this problem. The nested microphone array is mostly used linearly in speech enhancement algorithms for noise elimination of a speech signal [27], but the linear array is not useful for 3D SSL. In this section, a 3D quasi-spherical NMA is proposed for three-dimensional SSL algorithms. This array increases the localization accuracy in combination with multiresolution adaptive SRP(PHAT/ML) algorithm. Fig. 2 shows the block diagram of the proposed QSNMA-MASRP(PHAT/ML) algorithm for 3D multiple simultaneous SSL, where the NMA part is specified in the left side.

The inter-microphone distance (d) for each subarray should be adjusted to the condition $d < \lambda/2$ (where λ is the speech signal wavelength for the maximum frequency component in each subband) to avoid the spatial aliasing. In the proposed QSNMA, the inter-microphone distances are adjusted in a symmetrical way to have the same accuracy for speakers in different directions. The speech signal is considered for the following frequency range $B = 0-8$ kHz. The proposed QSNMA is structured as 4 subarrays, each one to cover a part of speech spectral components. Also, the sampling frequency is selected as based on the maximum speech signal frequency to comply the Nyquist condition. The first subarray is designed to cover the frequency range of $B_1 = 4-8$ kHz. Then, the central frequency for the analysis filter $H_1(z)$ and the inter-microphone distances are estimated as $f_{c1} = 6$ kHz and $d_1 < 2.3$ cm respectively. The second subarray is designed for the frequency range $B_2 = 2-4$ kHz, to avoid the spatial aliasing. Therefore, the microphones in the proposed nested array are adjusted to provide the central frequency $f_{c2} = 3$ kHz for the filter $H_2(z)$ and the inter microphone distance is estimated as $d_2 = 2d_1 < 4.6$ cm. The third subarray for frequency range $B_3 = 1-2$ kHz with the filter $H_3(z)$ has the central frequency $f_{c3} = 1.5$ kHz. The inter-microphone distance is selected as $d_3 = 4d_1 < 9.2$ cm. Finally, the parameters of the forth $H_4(z)$ subarray are $d_4 = 8d_1 < 18.4$ cm and $f_{c4} = 0.5$ kHz. Figure 3 shows the proposed 3D QSNMA with $M = 18$ microphones, and its diameter is 18.4 cm.

The proposed QSNMA is designed such a way to prepare a proper microphone pairs for each subband. Figure 4 shows the designed subarrays for each subband related to the analysis filters. The first subarray is designed for the analysis filter $H_1(z)$, which contains the microphone pairs $(m1, m2)$, $(m2, m3)$, $(m3, m4)$, $(m4, m5)$, $(m5, m6)$, $(m6, m7)$, $(m7, m8)$, and $(m8, m1)$. The microphone pairs $(m9, m13)$, $(m10, m14)$, $(m11, m15)$, and $(m12, m16)$ are structured for the second subarray and analysis filter The third subarray contains the microphone pairs $(m13, m18)$, $(m14, m18)$, $(m15, m18)$, $(m16, m18)$, $(m9, m17)$, $(m10, m17)$, $(m11, m17)$, and $(m12, m17)$, which are designed for the analysis filter Finally, the microphone pairs $(m9, m15)$, $(m12, m14)$, $(m11, m13)$, $(m10, m16)$, and $(m17, m18)$ are used for the forth subarray.

Analysis filters are considered for each subarray to prepare the proper frequency components for the related microphone pairs to eliminate the spatial aliasing. A multirate sampling with down-samplers are required to design the analysis filters $H_i(z)$ [27]. Figure 5 shows the filters together with down-samplers D_i as a tree structure. Each level of the tree contains a low-pass filter (LPF) high-pass filter (HPF) and down-sampler $D_i(\downarrow 2)$.

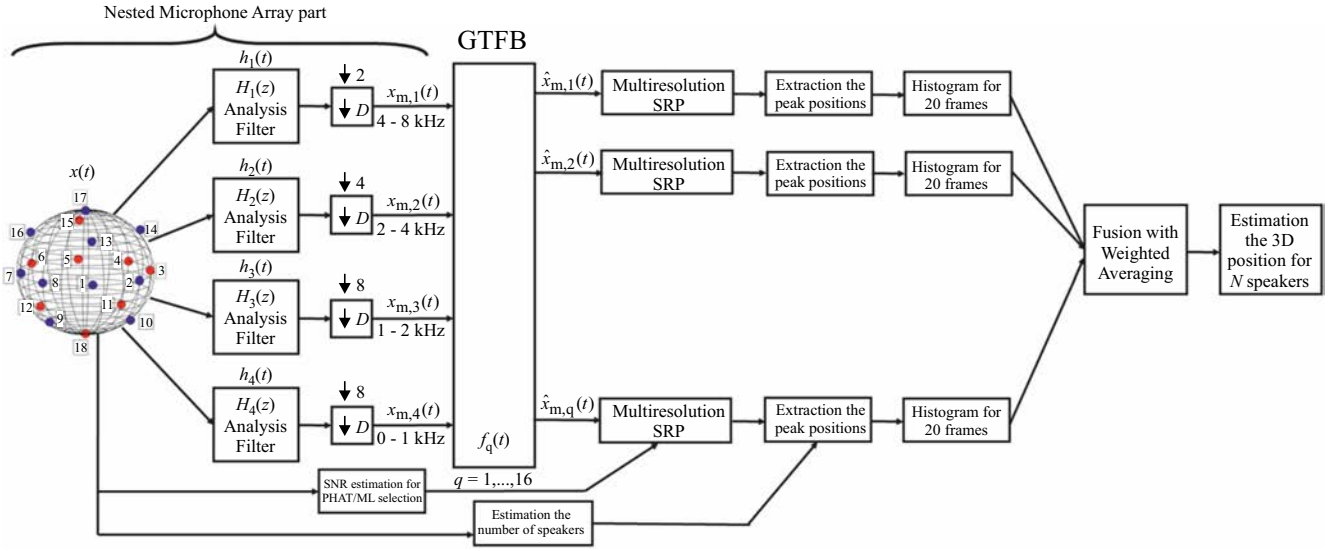

Fig. 2. The block diagram of the proposed sound source localization algorithm with QS-NMA, Gammatone filter bank (GTFB) and multiresolution SRP with adaptive PHAT and ML weighed functions

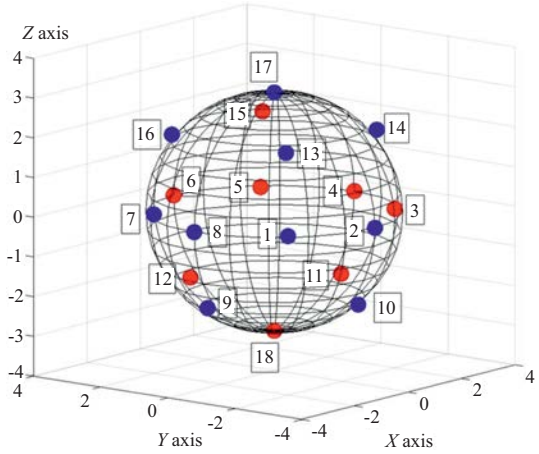

Fig. 3. The proposed 3D QSNMA for SSL in combination with MASRP(PHAT/ML) algorithm

The relation between LPFs $LP_i(z)$, HPFs $HP_i(z)$ and down-samplers D_i is

$$\begin{aligned} H_1(z) &= HP_1(z), \\ H_2(z) &= LP_1(z) \times HP_2(z^2), \\ H_3(z) &= LP_1(z) \times LP_2(z^2) \times HP_3(z^4), \\ H_4(z) &= LP_1(z) \times LP_2(z^2) \times LP_3(z^4)\,. \end{aligned} \tag{3}$$

As seen, the analysis filters $H_i(z)$ are implemented in a multi-level structure based on a series of LPFs and HPFs. Therefore, there are possibilities to develop more frequency bands, if more computational complexity is allowed. Figure 6 shows the frequency response for the analysis filters related to the QSNMA. As shown, the filter $H_1(z)$ is related to the highest frequency band and filter $H_4(z)$ covers the lowest frequency range. The analysis filter $H_1(z)$ is for the nearest microphone pairs in the first subarray and the analysis filter $H_4(z)$ is for the farthest microphone pairs in the forth subarray.

Each level of the tree contains a LPF and a HPF, which are designed with the Remez method. Therefore, the transition band is 0.0625 and the stop band is selected as 50 dB. These analysis filters not only eliminate the spatial aliasing, but also prepare the proper frequency bands for each pair of microphones in the proposed QSNMA.

3 The proposed multiresolution adaptive SRP(PHAT/ML) method based on the Gammatone filter bank

The speech is a wideband and non-stationary signal with a non-uniform spectral distribution, where the lower frequency bands of the speech signal have more spectral information. Therefore, more paying attention to these frequency components prepares the proper information and increases the localization accuracy. Then, the multiresolution adaptive SRP(PHAT/ML) is proposed for implementing on different frequency bands of the speech signal, which are recorded by the use of proposed QSNMA. The use of GammaTone filter bank is proposed for subband processing in the implementation of the MASRP method. The GammaTone filter bank is based on the human auditory system, where it has a high frequency resolution in low frequency components. Since the speech signal has the W-DO property, where each TF point with high probability is related just to one speaker, the Gammatone filter bank increases the precision of the localization algorithms by increasing the frequency resolution. In the following, the multiresolution SRP algorithm is proposed with the adaptive use of PHAT/ML weighted functions on the overlapped speech signal for 3D SSL.

3.1 The Gammatone filter bank for speech subband processing

The Gammatone filter bank is a proper tool for speech signal subband processing, which is defined based on the

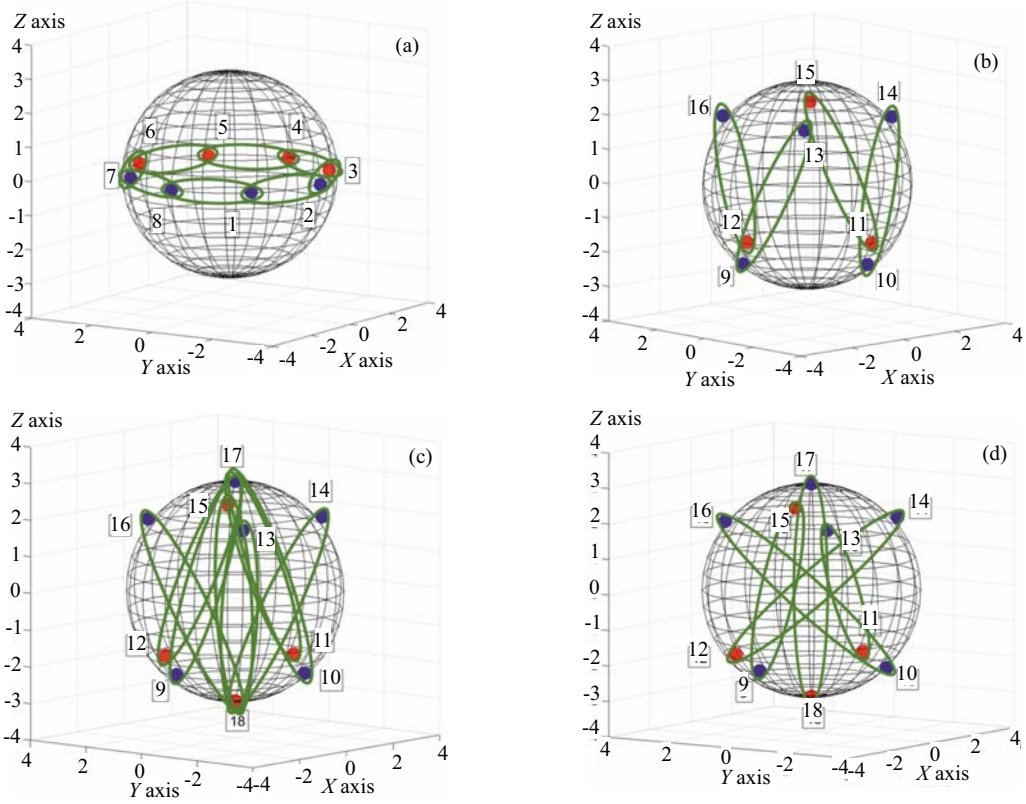

Fig. 4. The four subarrays related to the proposed QSNMA

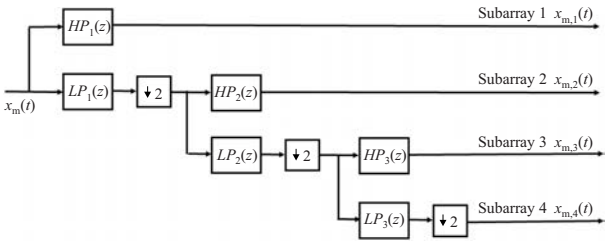

Fig. 5. A multirate tree structure fo the analysis filter bank related to the proposed QSNMA

speech spectral components. This filter was introduced by Johannesma in 1972 for physiological modeling of the impulse response associated with the human auditory system [28]. This filter was named as Gammatone in 1980 by Aersten and Johannesma due to the use a carrier tone with amplitude modulation, which has the exponential shape similar to the gamma distribution in probabilistic topics [29]. In 1987, more studies were done on the Gammatone filter bank, which shows its complete adaptation with the human auditory system, where it has been widely used for auditory system modeling [30]. The time domain impulse response for the Gammatone filter bank is

$$h_g(t) = a_n t^{n-1} e^{-\Omega t} \cos(\omega_0 t + \varphi), \quad t \geq 0, \tag{4}$$

where $\Omega = 2\pi b$, $\omega_0 = 2\pi f_0$ with f_0 being the filter central frequency, while n is the filter order, and b is the scaling parameter. For a fixed n value, the filter bandwidth is raised by increasing the variable b. The variable n controls the envelope of the Gammatone filter bank in the time domain, where the ascending and descending areas are skewed by increasing the variable n. The variable φ is considered as 0 because its effect is not considerable on the Gammatone filter spectral features. The Gammatone filter bank in frequency domain is, [28]

$$H_g(f) = \left[1 + j\frac{f - f_0}{b}\right]^{-n} + \left[1 + j\frac{f + f_0}{b}\right]^{-n}, \quad (-\infty < f < \infty). \tag{5}$$

The second part in (5) can be neglected, based on the human auditory system model. Then, the Gammatone filters of order n can be implemented with a cascade structure of n Gammatone filters of order 1 with frequency response $\left[1 + j\frac{f-f_0}{b}\right]^{-1}$ where each one is a shifted LPF in the frequency domain, which can be structured with recursive implementation method as follows.

Vector $x(i)$ is considered as the input signal with a sampling period Δt. First, the signal $x[i]$ is shifted in the frequency domain by for f_0. Then the complex vector will be

$$V[i] = e^{-j\omega_0 i \Delta t} x[i]. \tag{6}$$

Signal $V[i]$ passing through the first-order recursive filter is changed to $W[i]$. The filter output is passed $n-1$ times through the recursive filter to get the output

$$W[i] = W[i-1] + \left(1 - e^{-\Omega \Delta t}\right)\left(V[i-1] - W[i-1]\right). \tag{7}$$

Finally, the output of Gammatone filter bank is produced with a frequency shift $+f_0$ and by considering the real part of (7)

$$y[i] = \Re\{e^{j\omega_0 i\Delta t} W[i]\}. \tag{8}$$

The output of the Gammatone filter bank, from Fig. 2, is

$$\hat{x}_{m,q}(t) = x_m(t) * f_q(t) \quad \text{for} \quad \begin{cases} m = 1, \ldots, \ (M = 18) \\ q = 1, \ldots, \ (Q = 16), \end{cases} \tag{9}$$

where M is number of the microphones, Q is number of the subbands, and $f_q(t)$ is the frequency response of the Gammatone filter. The subband signals $\hat{x}_{m,q}(t)$, generated by the QSNMA and Gammatone filter bank, are considered as the input signals for the proposed MASRP(PHAT/ML) localization algorithm. Figure 7 shows the frequency response of the Gammatone filter bank. As seen, the frequency resolution in higher in lower frequencies.

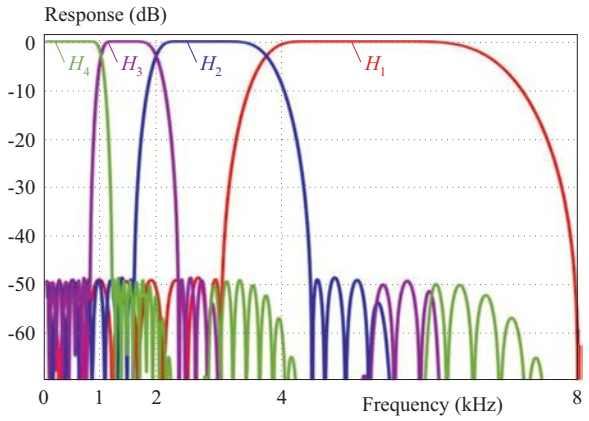

Fig. 6. Frequency response of analysis filter bank related to the proposed QSNMA

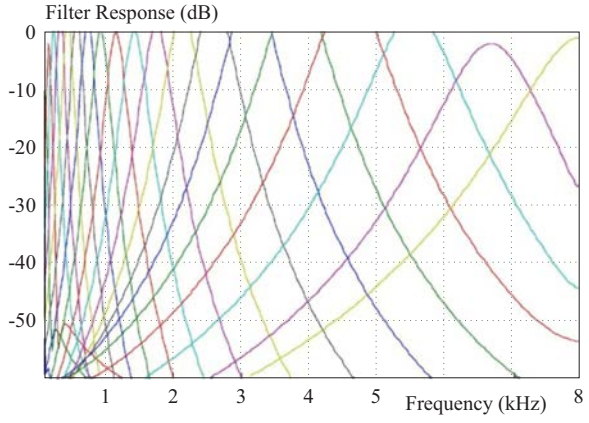

Fig. 7. Frequency response of the Gammatone filter bank in the proposed MASRP(PHAT/ML) localization algorithm

3.2 The multiresolution adaptive SRP algorithm with PHAT/ML weighted functions for 3D SSL

The 3D simultaneous SSL in noisy and reverberant conditions with high accuracy and low computational complexity is always a challenge in the speech signal processing. The one-step methods for localization are more applicable because of their precision. In this section, a multiresolution SRP method is proposed in combination with the adaptive use of PHAT/ML weighted functions and the proposed QSNMA for multiple simultaneous SSL in noisy and reverberant conditions. As mentioned, the output of Gammatone filter bank is considered as $\hat{x}_{m,q}(t)$, where the output of delay and sum beamformer (DSB) is, [10]

$$y_q(t) = \sum_{m=1}^{M_q} \hat{x}_{m,q}(t - \Delta_m), \tag{10}$$

where, q is the subband index, M_q is the number of microphones in q-th subband, and Δ_m is the delay selected for steering the array to the speaker direction. The filter and sum beamformer (FSB) is proposed to improve the robustness of the DSB at presence of noise to decrease its effect and the reverberation of the input signal. The output of the FSB block with microphones in q-th subband in the frequency domain is

$$Y_q(\omega) = \sum_{m=1}^{M_q} G_m(\omega)\hat{X}_{m,q}(\omega)e^{-j\omega\Delta_m}, \tag{11}$$

where $G_m(\omega)$ is the Fourier transform of a filter transfer functions associated with FSB and $\hat{X}_{m,q}(\omega)$ are the Fourier transform of the microphone signals in the MASRP algorithm for the q-th subband and M_q microphones related to the subband. The MASRP is a function of steered delays. The beam-pattern of the microphone array is adjusted to the specific direction in the 3D environment by changing the steered delays. The steered response is calculated by sweeping the controlled area of the beamformer. The steered response power is maximized, when the searching area by the beamformer is the same as the source location. The SRP has multiple peaks in the case of existence multiple simultaneous speakers, but the reverberation generates many incorrect peaks, which prepares the wrong source location estimations. The MASRP is expressed by the output power of the FSB is

$$P_q(\Delta_1 \ldots \Delta_{M_q}) = \int_{-\infty}^{\infty} Y_q(\omega)Y_q'(\omega)\mathrm{d}\omega, \tag{12}$$

where $Y_q(\omega, \Delta_1 \ldots \Delta_{M_q})$ is the output of the FSB for the q-th subband, and the prime denotes its complex conjugate. The MASRP of the FSB is calculated by the combination of (11) and (12). After some manipulation this will lead to

$$P_q(\Delta_1, \ldots, \Delta_{M_q}) = \sum_{l=1}^{M_q}\sum_{r=1}^{M_q}\int_{-\infty}^{\infty} \psi_{lr,q}(\omega)\hat{X}_{l,q}(\omega)\hat{X}'_{r,q}(\omega)e^{j\omega(\Delta_l - \Delta_r)}\mathrm{d}\omega, \tag{13}$$

where we have introduced the weighting function $\psi_{lr,q}(\omega) = G_{l,q}(\omega)G'_{r,q}(\omega)$.

The weighed function is an important factor in the performance of the MASRP algorithm. This function has a effect on the precision of the localization algorithms under the noisy and reverberant conditions. Two important weighted functions are PHAT and ML for placement in the MASRP algorithm. It has been shown in [31] that the PHAT weighed function is more efficient for reverberant and less noisy conditions ($SNR > 10$ dB). Therefore, a noise estimation block is proposed in the algorithm for the adaptive use of PHAT and ML weighted functions in different environmental conditions. The PHAT weighted function is defined as

$$\psi_{lr,q}^{PHAT}(\omega) = \frac{1}{\left|\hat{X}_{l,q}(\omega)\hat{X}'_{r,q}(\omega)\right|}. \tag{14}$$

This function works ideally for the free-reverberant conditions, and has a high efficiency in the reverberant scenarios. This weighted function does whitening on the microphone signals by normalizing the Fourier transform for the signals amplitude. The MASRP with PHAT weighted function is

$$P_q^{PHAT}(\Delta_1,\ldots,\Delta_{M_q}) = \sum_{l=1}^{M_q}\sum_{r=1}^{M_q}\int_{-\infty}^{\infty}\psi_{lr,q}^{PHAT}(\omega)\hat{X}_{l,q}(\omega)\hat{X}'_{r,q}(\omega)e^{j\omega(\Delta_l-\Delta_r)}\mathrm{d}\omega. \tag{15}$$

Based on the assumption in the use of PHAT filter, this function is considered for the reverberant scenarios with $SNR > 10$ dB.

If the environment is noisy and the speech and noise signals are uncorrelated, another ML weighed function is more efficient

$$\psi_{lr,q}^{ML}(\omega) = \frac{|\hat{X}_{l,q}(\omega)|\,|\hat{X}_{r,q}(\omega)|}{|V_{l,q}(\omega)|^2|\hat{X}_{r,q}(\omega)|^2 + |V_{r,q}(\omega)|^2|\hat{X}_{l,q}(\omega)|^2}, \tag{16}$$

where $V_{l,q}(\omega)$ and $V_{r,q}(\omega)$ are spectra of the additive noise in q-th subband of l-th and r-th microphones, respectively. The noise spectra are estimated on the silence part of the microphone signals. Therefore, the MASRP/ML function is expressed as

$$P_q^{ML}(\Delta_1,\ldots,\Delta_{M_q}) = \sum_{l=1}^{M_q}\sum_{r=1}^{M_q}\int_{-\infty}^{\infty}\psi_{lr,q}^{ML}(\omega)\hat{X}_{l,q}(\omega)\hat{X}'_{r,q}(\omega)e^{j\omega(\Delta_l-\Delta_r)}\mathrm{d}\omega. \tag{17}$$

The MASRP algorithm is implemented adaptively by the use of PHAT/ML weighted functions MASRP (PHAT/ML) in all subbands associated with the Gammatone filter bank and microphone pairs related to the proposed QSNMA. The number of extracted peaks in the MASRP algorithm is the same as the number of speakers, where the i-vector probabilistic linear discriminant analysis (i-vector PLDA, [32]) is considered for estimating the number of speakers. Then, the N-first peaks of the MASRP(PHAT/ML) algorithm for q-th subband denoting $\hat{T}_n \equiv \hat{T}_{n,(x,y,z)(b,q)}$ on $(x,y,z) \in \mathcal{R}$ are extracted as

$$\begin{aligned}
\hat{T}_1 &= \arg\max \quad P_q^{PHAT/ML}(\Delta_1,\ldots,\Delta_{M_q}),\\
\hat{T}_2 &= \arg\max_{\hat{T}_2\neq\hat{T}_1} \quad P_q^{PHAT/ML}(\Delta_1,\ldots,\Delta_{M_q}),\\
\hat{T}_N &= \arg\max_{\hat{T}_N\neq\hat{T}_1,\ldots,\hat{T}_{N-1}} \quad P_q^{PHAT/ML}(\Delta_1,\ldots,\Delta_{M_q}),
\end{aligned} \tag{18}$$

where $\hat{T}_1,\ldots,\hat{T}_N$ are the N-first peaks of MASRP(PHAT/ML) algorithm for the q-th subband and microphones related to this subband. The MASRP(PHAT/ML) is a function based on the steered delays $\Delta_1,\ldots,\Delta_N$ for any candidate point. The MASRP function is calculated for all 3D candidate points each having a specific steered delay in indoor condition. Therefore, the MASRP(PHAT/ML) function is maximized for all 3D candidate room points with position $r \equiv r(x,y,z)$. This process is iterated for continuous frames of overlapped speech signals and separately for each subband to extract the histogram of MASRP(PHAT/ML) peaks as

$$D_q = H\left\{\hat{T}_{1,r,q},\hat{T}_{2,r,q},\ldots,\hat{T}_{N,r,q},\ \ \forall m \in M_q\right\}, \tag{19}$$

where D_q is the histogram of the N-first peaks of MASRP(PHAT/ML) method for q-th subband. This process is repeated for all Q subbands to provide a separate histogram for each subband. Finally, the weighted averaging (WA) method is selected for combining the histograms for all subband to generate the final histogram

$$H_{\mathrm{WA}}(\hat{T}) = \frac{1}{Q}\sum_{q=1}^{Q}\frac{S_{1,q}}{\sum_{i=2}^{N}S_{i,q}}D_q, \tag{20}$$

where

$$S_{i,q} = \max_{S_{i,q}\neq S_{1,q}\neq S_{2,q}\cdots\neq S_{i-1,q}} D_q. \tag{21}$$

The effect of each speaker in the related subband is highlighted by the weighted averaging. For example, if the first speaker has more contents in the frequency band 2-2.5 kHz, then it has more effect on the combination between histograms. The final histogram is calculated on $\hat{T} \in \mathcal{R}$ by the fusion of histograms for each subband, and the N-first peaks are the 3D location of N speakers as

$$\begin{aligned}
\hat{T}_{1,r} &= \arg\max H_{\mathrm{WA}}(\hat{T}),\\
\hat{T}_{2,r} &= \arg\max_{\hat{T}\neq\hat{T}_{1,r}} H_{\mathrm{WA}}(\hat{T}),\\
&\vdots\\
\hat{T}_{N,r} &= \arg\max_{\hat{T}\neq\hat{T}_{1,r}\ldots\hat{T}_{N-1,r}} H_{\mathrm{WA}}(\hat{T}),
\end{aligned} \tag{22}$$

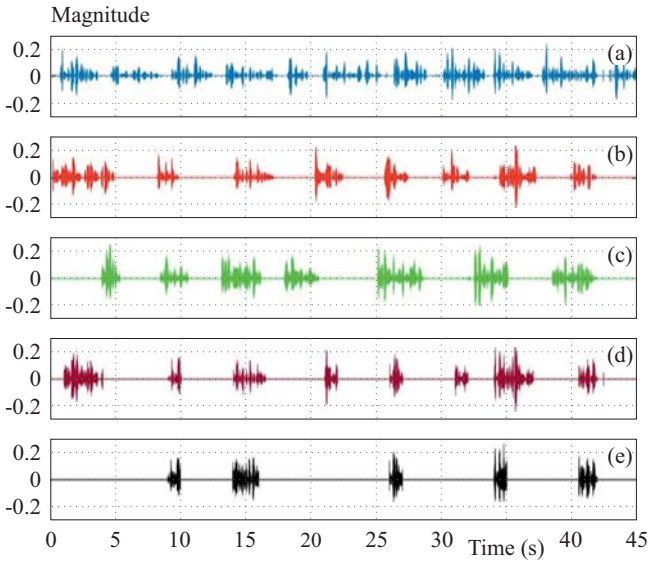

Fig. 8. The time-domain speech signal for: (a) – speaker S1, (b) – speaker S2, (c) – speaker S3, (d) – overlapped speech of speaker S1 and speaker S2, and (e) – overlapped speech between three speakers

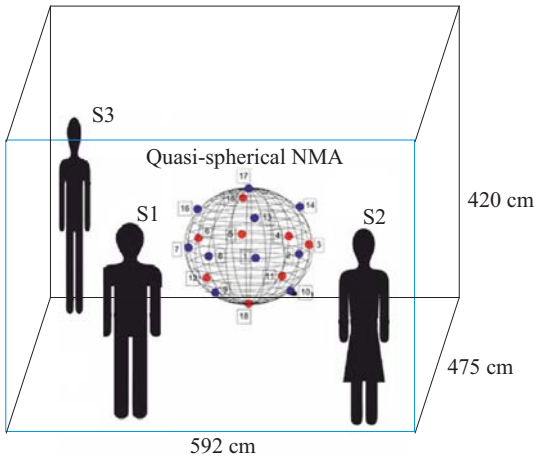

Fig. 9. A view of the simulated recording room with the proposed QSNMA and three speakers

where $\hat{T}_{1,r} \dots \hat{T}_{N,r}$ are the estimated 3D locations for N simultaneous speakers by the proposed QSNMA-MASRP(PHAT/ML) method. The estimated locations by the proposed method are very accurate because of use the QSNMA, subband processing with GammaTone filter bank, and adaptive PHAT/ML weighted functions in combination with SRP algorithm.

4 Results and discussions

The evaluations for the proposed method are implemented on real and simulated data. Texas Instruments and Massachusetts Institute of Technology (TIMIT) dataset is selected for simulated data, where we considered some of the speech signals in this dataset [33]. The simulation scenarios are considered in a way to prepare the conditions as same as the real environments. Also, the real data was recorded with the proposed QSNMA in the speech processing laboratory, at the Universidad Tecnológica Metropolitana, Santiago, Chile. The experiments in multi-speaker scenarios show about 9% of overlapped speech is for two simultaneous speakers, 8% for three overlapped speakers, and 2% for four or more speakers [34]. Therefore, the simulations are implemented on the scenarios with two and three simultaneous speakers to cover the most environmental conditions. Then, two males (24 and 35 years old) and one female (31 years old) speakers are selected for data recording. The Omni-directional microphones are considered for data recording with a sampling frequency 16000Hz in a room under the conditions of temperature 23 °Cand 25% humidity. 45 s speech signal is recorded for each speaker, where 17.3 s is for two simultaneous speakers (speaker S1 and speaker S2), and 6.8 s is for three simultaneous speakers. Figure 8 shows the time-domain speech signal for each speakers, and overlapped between two and three simultaneous speaker, respectively.

The proposed QSNMA with 18 microphones is located in the middle of the room, where the room dimensions are (475,592,420) cm. In addition, the three speakers are located at (55,130,180)cm (S1), (110,520,170) cm (S2), and (460,45,175) cm (S3), respectively. Figure 9 shows a view of the simulated recording room, where the speakers are located in different directions to show the symmetry effects of the proposed QSNMA

Robustness and precision of the localization algorithms in noisy and reverberant environments are two important factors in the evaluations of the proposed method. Noise and reverberation are the important challenges, which decrease the accuracy of the localization algorithms. In the simulations, a Gaussian white noise is considered additively with the speech signal in the microphone place. In the real environment, the effect of the

noise is considered by playing a Gaussian noise with a speaker. The reverberation appears in the indoor environments due to the existence of reflective surfaces such as walls and tables, which are recorded by the microphones as same as the original signals. Therefore, Image model is selected for simulating the reverberation effect in the environments [35]. This model simulates the reverberation similar to the real conditions. The Image algorithm produce the room impulse response between the source and microphone by considering the room dimensions, speaker location, microphone position, room reverberation time, sampling frequency, impulse response length, and reflective coefficients of the surfaces. The received signal to the microphone is produced by the convolution between the source signal and generated room impulse response by the Image method. Two categories of experiments are selected for the evaluations. In the first category, the proposed algorithm is evaluated for the fixed SNR, variable RT_{60} and vice versa for 2 and 3 simultaneous speakers to show the effect of the noise and reverberation changes on the robustness and accuracy. In the second category, some scenarios are defined, which happens commonly in the real environments to evaluate the precision of the proposed algorithm. Three scenarios are defined for the second category of evaluations. The first scenario is named the reverberant environment, where $RT_{60} = 650$ ms and $SNR = 20$ dB. The second scenario is a noisy environment with $SNR = 20$ dB and $RT_{60} = 250$ ms. The most challenging scenario is noisy-reverberant environment by selecting and The precision and accuracy of the proposed method is evaluated by the comparison between the proposed method with other previous works on these scenarios. A Hamming window with length 60ms and 50% overlap is selected for data windowing in the evaluations. These length and overlap values prepare the best stationary for the data in the localization algorithm. In addition, the experiments are implemented on a PC with CPU Intel(R) core i7-7700 (4.2 GHz), ×64-based processor, 32 GB RAM, and WINDOWS 64-bit operating system by the use of MATLAB software version 2018b for the simulations. The MATLAB software is considered for the implementations because is user friendly and more applicable for the use of existing functions and preparing the figures in the results section. Otherwise, the other software such as Python and C, or implementing with digital signal processor hardware are the options to be closer to real-time implementation.

The proposed QSNMA-MASRP(PHAT/ML) method is compared with SRP-PHAT [11], SSM-DNN [23], and SH-TMSBL [24] algorithms based on the precision and accuracy on the real and simulated data for 2 and 3 simultaneous speakers. The MAEE criteria in frame b, of estimated ($\hat{T}_{N,(x,y,z)[b]}$ and true ($T_{N,(x,y,z)}[b]$) locations is considered for comparison on L time frames

$$MAEE_N = \frac{1}{L}\sum_{b=1}^{L}\left|\hat{T}_{N,(x,y,z)}[b] - T_{N,(x,y,z)}[b]\right|. \quad (23)$$

The methods with smaller MAEE values for one or averages multiple of speakers are considered as the algorithms with better accuracy.

Figure 10 and 11 show the energy distribution of the proposed QSNMA-MASRP (PHAT/ML) method for the subband 1-1.2 kHz in noisy, reverberant, and noisy-reverberant scenarios for 2 and 3 simultaneous speakers, respectively. These 3D figures shows the information for x and y axis (the two most important axis since the z values for all speakers are very similar). By considering all axis (x, y, z) with the energy as the output of QSNMA-MASRP (PHAT/ML) function, then it would be a 4D shape, which is not possible to show as a clear energy distribution. Then, we decided to drop the z axis just in the plot, which is similar for all speakers, and a clear 3D energy distribution is plotted for the proposed algorithm. As seen, the noisy-reverberant condition is the worst scenario, where there are many extra peaks in the energy distribution diagram, which conduce the wrong estimation for the speaker location. After that, noisy scenarios still has many extra peaks but the reverberant scenario has less peaks in comparison with the other scenarios. In addition, each speaker has different information in the sub-figures for the energy distribution based on the subbands. These information differences are due to the use of GammaTone filter bank, which provides the good separation between the speakers based on the frequency components in different subbands. In addition, each speaker is dominant in different subband by the effect of Gammatone filer bank.

Table 1 shows the results for two simultaneous speakers (speaker S1 and speaker S2) for real and simulated data. The simulations are implemented on 3 scenarios and the comparison with SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms are realized. The mean absolute estimation error (MAEE) and averaged standard deviation (SD) for absolute estimation error, in cm, are considered to compare the simulation results with real speakers positions. The MAEE results are calculated by averaging of 10 time frames to be robust in different scenarios. The algorithm was tested by various number of frames to find with how many frames the accuracy and the computational complexity of the algorithm are in an acceptable range. We found that by increasing the number of overlapped frames, the accuracy is increased by the use of around 20 frames, but by having more frames the accuracy does not change and just more complexity is added to the algorithm. Then, we decided to keep the number of frames as 20 in order to have acceptable accuracy as well as the complexity. Then, n=20 frames means (floor =660 ms of the data from the overlapped speech between two speakers (speaker S1 and speaker S2) for the scenario with 2 simultaneous speakers. Also, in the scenario with 3 simultaneous speaker, the data are taken from are of the overlapped speech between three speakers in Fig. 8. As seen, the proposed method has the smaller MAEE and SD in comparison with other methods, which means the speakers are localized more accurately. Also, the difference of accuracy in all scenarios in the proposed method

Table 1. The MAEE and averaged SD of absolute estimation error results for the proposed QSNMA-MASRP(PHAT/ML) method in comparison with SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms for 2 simultaneous speakers in different scenarios for real and simulated data

MAEE (cm)	SRP-PHAT, [11]			SSM-DNN [23]			SH-TMSBL [24]			Proposed*		
					Real data							
Speaker	S1	S2	SD	S1	S2	SD	S1	S2	SD	S1	S2	SD
Scenario 1	54	61	7.7	49	52	7.3	33	44	6.7	22	34	5.9
Scenario 2	55	63	7.9	51	54	7.6	36	42	6.9	24	31	6.1
Scenario 3	67	68	8.3	58	62	7.7	52	56	7.1	35	38	6.3
					Simulated data							
Speaker	S1	S2	SD	S1	S2	SD	S1	S2	SD	S1	S2	SD
Scenario 1	53	55	7.2	42	46	6.8	32	34	6.4	17	22	5.7
Scenario 2	56	51	7.4	44	47	7.0	35	37	6.5	21	27	5.8
Scenario 3	59	66	8.1	50	56	7.5	44	47	6.8	24	28	6.0

* QSNMA-MASRP (PHAT/ML)

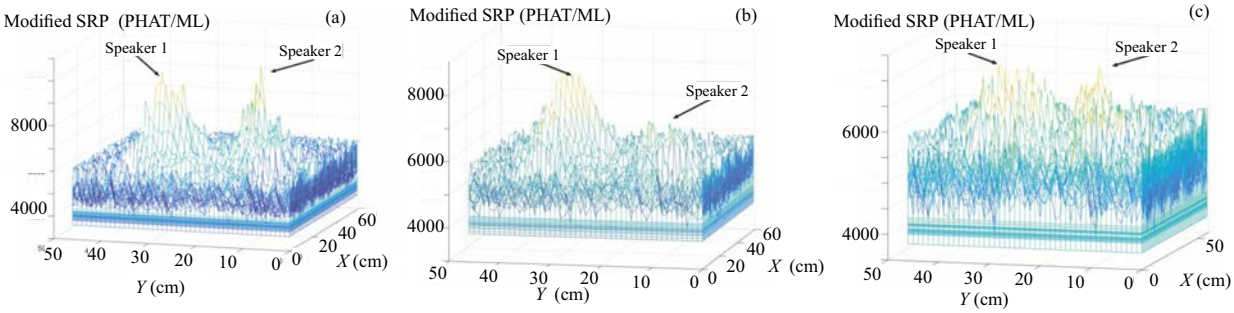

Fig. 10. Energy distribution curves for proposed multi-resolution SRP-PHAT/ML method by use of Gammatone filter bank and QS-NMA for 3 simultaneous speakers (speaker 1 and speaker 2 and speaker 3) and for: (a) – reverberant, (b) – noisy, and (c) – noisy-reverberant scenarios

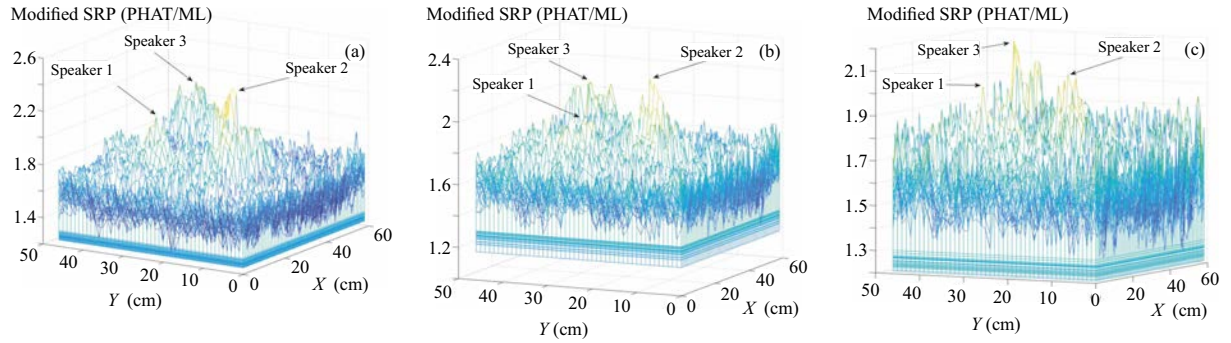

Fig. 11. The averaged MAEE (cm) for the proposed multi-resolution SRP-PHAT/ML method by use of Gammatone filter bank and QS-NMA in comparison with traditional SRP-PHAT [17] and SH-TMSBL [13] methods for 3 simultaneous speakers (real and simulated data) for: (a) – different RT_{60} values for $SNR = 5$ dB, and (b) – different SNR values and $RT_{60} = 650$ ms

is less than the SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms. The reason is due to the use of PHAT and ML filters adaptively, which considers the best weighted filter based on the different acoustical environments to increase the accuracy of the proposed method. The MAEE and SD results for the real data are larger than the simulated data. Because measuring the noise and reverberation in the real environment is not completely accurate.

Figure 12 shows the averaged MAEE criteria, in cm, for the proposed QSNMA-MASRP(PHAT/ML) method in comparison with SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms for 2 simultaneous speakers on real and simulated data. Figure 12(a) shows the results for and different range RT_{60} of values. As seen, our proposed method has better results in comparison with other previous works in both real and simulated data. For example, the averaged MAEE value for the proposed method in $SNR = 5$ dB and $RT_{60} = 200$ ms for real data is 23 cm in comparison with 56 cm for SRP-PHAT, 48 cm for SSM-DNN, and 35 cm for SH-TMSBL algorithm. In addi-

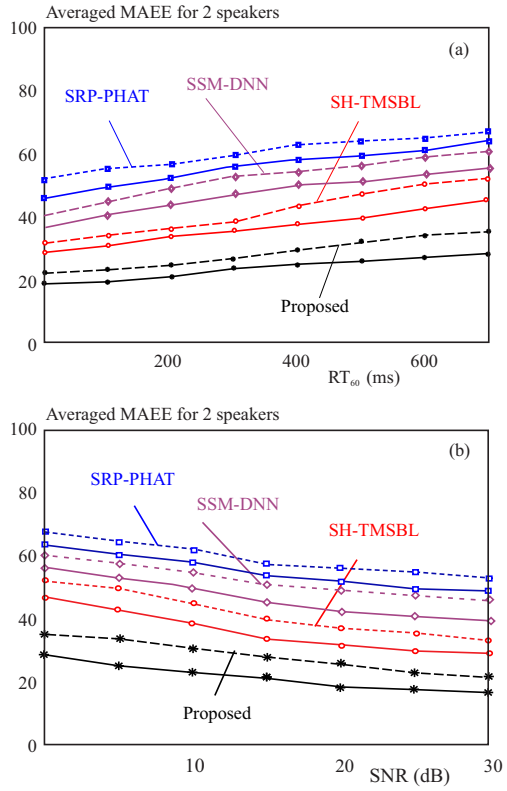

Fig. 12. The averaged MAEE (cm) for the proposed multi-resolution SRP-PHAT/ML method by use of Gammatone filter bank and QS-NMA in comparison with traditional SRP-PHAT [17] and SH-TMSBL [13] methods for 2 simultaneous speakers, real (dashed lines) and simulated (solid lines) data for: (a) – different RT_{60} values for $SNR = 5$ dB, and (b) – different SNR values and $RT_{60} = 650$ ms

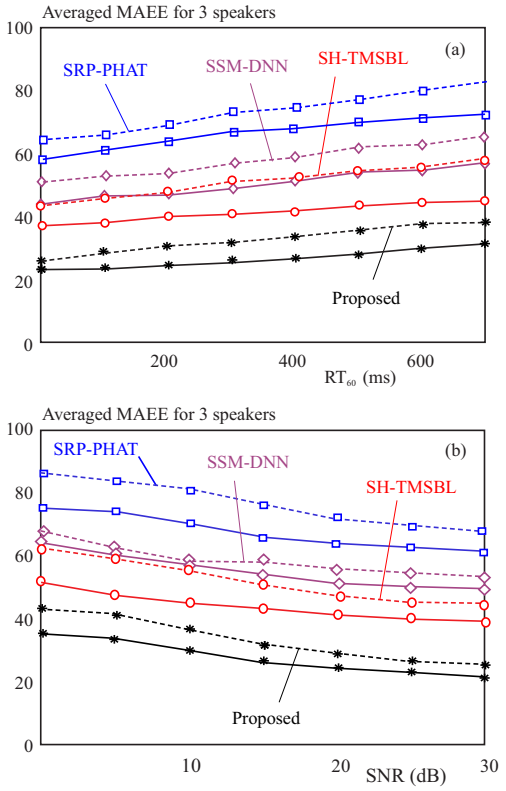

Fig. 13. The averaged MAEE (cm) for the proposed multi-resolution SRP-PHAT/ML method by use of Gammatone filter bank and QS-NMA in comparison with traditional SRP-PHAT [17] and SH-TMSBL [13] methods for 3 simultaneous speakers real (dashed lines) and simulated (solid lines) data for: (a) – different RT_{60} values for $SNR = 5$ dB, and (b) – different SNR values and $RT_{60} = 650$ ms

Table 2. The MAEE and averaged SD of absolute estimation error results for the proposed QSNMA-MASRP(PHAT/ML) method in comparison with SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms for 3 simultaneous speakers in different scenarios for real and simulated data

MAEE (cm)	SRP-PHAT, [11]				SSM-DNN [23]				SH-TMSBL [24]				Proposed*			
Real data																
Speaker	S1	S2	S3	SD	S1	S2	S3	SD	S1	S2	S3	SD	S1	S2	S3	SD
Scenario 1	67	74	75	7.9	56	55	58	7.7	42	45	48	6.8	21	27	33	6.2
Scenario 2	70	75	78	8.1	57	59	60	7.9	47	52	56	7.3	29	32	37	6.4
Scenario 3	80	84	87	8.7	61	64	66	8.0	53	61	64	7.4	35	37	41	6.5
Simulated data																
Speaker	S1	S2	S3	SD	S1	S2	S3	SD	S1	S2	S3	SD	S1	S2	S3	SD
Scenario 1	57	63	67	7.6	48	54	51	7.5	34	41	43	6.6	21	22	26	5.9
Scenario 2	64	65	72	7.8	50	56	55	7.6	43	42	46	7.0	23	26	28	6.1
Scenario 3	67	72	79	8.3	57	63	59	7.9	44	48	49	7.2	25	36	37	6.3

* QSNMA-MASRP (PHAT/ML)

tion, the averaged MAEE value for the proposed method in $SNR = 10$ dB and $RT_{60} = 650$ ms for real data is 31 cm in comparison with 63 cm for SRP-PHAT, 56 cm for SSM-DNN, and 45 cm for SH-TMSBL algorithm. Also, the accuracy of the proposed method does not decrease highly by increasing the value. For example, the variation in the MAEE criteria for the proposed method in and is just 8cm for simulated and 12 cm for real data, in comparison with 18 cm for simulated and 14 cm for real data in SRP-PHAT, 19 cm for simulated and 21 cm for real data

Table 3. The computational complexity for the proposed QSNMA-MASRP(PHAT/ML) method in comparison with SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms for 2 and 3 simultaneous speakers on real data

Run time (s)	SRP -PHAT [11]	SSM -DNN [23]	SH- -TMSBL [24]	Pro- posed*
Two simultaneous speakers				
Scenario 1	675	547	252	248
Scenario 2	631	529	226	235
Scenario 3	692	578	284	276
Three simultaneous speakers				
Scenario 1	709	566	292	302
Scenario 2	683	547	273	261
Scenario 3	724	603	332	325

* QSNMA-MASRP (PHAT/ML)

in SSM-DNN, and 17 cm for simulated and 20 cm for real data in SH-TMSBL algorithms. Figure 12(b) shows the results for $RT_{60} = 650$ ms and a different range of SNR values. As shown, our proposed method has better accuracy in comparison with the SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms for 2 simultaneous speakers. The variation in the MAEE values for the proposed method in $SNR = 0$ dB and $SNR = 30$ dBis just 12 cm for simulated data and 13 cm for real data, in comparison with 14 cm for simulated and 17 cm for real data in SRP-PHAT, 18 cm for simulated and 17 cm for real data in SSM-DNN, and 18cm for simulated and 19 cm for real data in SH-TMSBL algorithms.

Table 2 shows the averaged MAEE and averaged standard deviation (SD) for absolute estimation error, for 3 simultaneous speakers in reverberant, noisy, and noisy-reverberant scenarios for real and simulated data. Also, these results are calculated for the proposed QSNMAMASRP(PHAT/ML) method in comparison with SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms. As shown, the proposed method has smaller MAEE and SD values in comparison with other previous works. Even by adding one more speaker, the result for the proposed method does not change highly and is similar to the results for two simultaneous speakers (by comparison of Tab. 2 and Tab. 1). These MAEE and SD results show the superiority of the proposed method in comparison with SRP-PHAT, SSM-DNN, and SHTMSBL algorithms for two and three simultaneous speakers in different undesirable scenarios.

Figure 13 shows the results for 3 simultaneous speakers for the proposed QSNMA-MASRP(PHAT/ML) method in comparison with SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms for real and simulated data. In both subfigures ((a) – different RT_{60} values for $SNR = 5$ dB and (b) – different SNR values for $RT_{60} = 650ms$), where the proposed method has better results in comparison with other previous works. For example, the averaged MAEE value for the proposed method in $SNR = 5$ dB and $RT_{60} = 200$ ms for real data is 34 cm in comparison with 71 cm for SRP-PHAT, 57 cm for SSM-DNN, and 50 cm for SH-TMSBL algorithm. In addition, the averaged MAEE value for the proposed method in $SNR = 10$ dB and $RT_{60} = 650$ ms for real data is 36 cm in comparison with 81 cm for SRP-PHAT, 58 cm for SSM-DNN, and 54 cm for SH-TMSBL algorithm. In Fig. 13(a), the variation in the MAEE parameter for the proposed method in $RT_{60} = 0$ ms and $RT_{60} = 700$ ms is just cm for simulated and 12 cm for real data in comparison with 14cm for simulated and 19 cm for real data in SRP-PHAT, 13 cm for simulated and 15cm for real data in SSM-DNN, and 10 cm for simulated and 14 cm for real data in SH-TMSBL algorithms. In Fig.13(b), the variation in the MAEE for the proposed method in $SNR = 0$ dB and $SNR = 30$ dB is just 13 cm for simulated and 16 cm for real data in comparison with 16 cm for simulated and 20 cm for real data in SRP-PHAT, 16cm for simulated and 18cm for real data in SSM-DNN, and 14cm for simulated and 19cm for real data in SH-TMSBL algorithms. Therefore, based on the results in Fig. 12 and Fig. 13, our proposed method is more robust in different range of SNR and RT_{60}. Also, the proposed method has better accuracy and less error in comparison with other previous works.

Table 3 shows the computational complexity of the proposed QSNMA-MASRP(PHAT/ML) method in comparison with SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms. The run-time of the MATLAB software, in second, for 2 and 3 simultaneous speakers on all environmental scenarios for real data is considered for this comparison. As seen, the SRP-PHAT method has the most computational complexity in comparison with other works due to the space search of candidate locations. After that, the complexity of the SSM-DNN algorithm is high because of use the neural networks in the training step. The proposed QSNMAMASRP(PHAT/ML) algorithm has lower complexity in comparison with the SRP-PHAT and SSM-DNN methods due to the nested microphone array and allocating some specific microphone pairs to each subarray. Also, the computational complexity of the proposed method is similar to the SHTMSBL algorithm in most of the scenarios. In some cases, the SH-TMSBL method has less complexity in comparison with the proposed work, which is mentioned in the table. Therefore, the complexity of the proposed method is acceptable for implementing as a localization algorithm in comparison with other previous works.

5 Conclusions

The multiple SSL from overlapped speech signal in noisy and reverberant environments is one of the important challenges in the speech signal processing. Some one-step and two-step methods were proposed for SSL, where they have high accuracy and low computational complexity, respectively. Also, the spatial aliasing is one of the destructive factors in the precision of the localization algorithms due to the intermicrophone distances. Firstly, a

quasi-spherical nested microphone array is proposed for eliminating the spatial aliasing. The proposed QSNMA is structured in a way to prepare enough microphone pairs for each subarray, and to have the spatial symmetry for speakers in all directions. In addition, it prepares the capacity for the three-dimensional SSL because of its 3D shape and distribution among the microphones in all dimensions. Speech signal has the W-DO property, which means each TF point of the overlapped speech signal with high probability is related to one single speaker. Therefore, the GammaTone filter bank is selected for signal subband processing. This filter has a high frequency resolution in low frequency components of the speech signal due to its design based on the human auditory system. In following, the multiresolution SRP algorithm is implemented adaptively with PHAT/ML weighted functions on QSNMA signals. The PHAT and ML weighted functions are considered adaptively for low and high noisy part of the speech signal, respectively. The MASRP peaks are extracted based on the number of speakers in each subband and this process is iterated for continuous time frames. Then, the distribution of the MASRP peaks is calculated for each subband and they are combined by the weighted averaging method. The maximums of the final peaks distribution are selected based on the number of speakers as the speakers locations. The proposed QSNMAMASRP(PHAT/ML) method is compared by the MAEE criteria and computational complexity with SRP-PHAT, SSM-DNN, and SH-TMSBL algorithms. The experiments are implemented on noisy, reverberant, and noisy-reverberant scenarios for 2 and 3 simultaneous speakers in different range of SNR and RT_{60}. In all scenarios, the proposed method has higher accuracy and lower computational complexity in comparison with the other previous works, which shows the superiority of the proposed algorithm.

Acknowledgment

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

References

[1] X. Sheng and Y.-H. Hu, "Maximum Likelihood Multiple-Source Localization Using Acoustic Energy Measurements with Wireless Sensor Networks", *IEEE Transactions on Signal Processing*, vol. 53, pp. 44-53, 2005.

[2] A. Ikeda, H. Mizoguchi, Y. Sasaki, T. Enomoto, and S. Kagami, "2D Sound Source Localization in Azimuth & Elevation from Microphone Array by Using a Directional Pattern of Element", *IEEE SENSORS,* Atlanta, GA, pp. 1213-1216, 2007.

[3] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation maximization source separation and localization", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, p. 382-394, 2010.

[4] F. Antonacci, M. Matteucci, D. Migliore, D. Riva, A. Sarti, M. Tagliasacchi, and S. Tubaro, "Tracking multiple acoustic sources in reverberant environments using regularized particle filter", *In Proceedings IEEE International Conference on Digital Signal Processing*, Cardiff, UK, pp. 99-102, 2017.

[5] Q. Yan, J. Chen, G. Ottoy, and L. D. Strycker, "Robust AOA based acoustic source localization method with unreliable measurements", *Signal Processing*, vol. 152, pp. 13-21, 2018.

[6] K. Na, Y. Kim, and H. Cha, *Acoustic sensor network-based parking lot surveillance system*, Berlin, Heidelberg: Springer Berlin Heidelberg, p. 247-262, 2009.

[7] D. Su, K. Nakamura, K. Nakadai, and J. V. Miro, "Robust sound source mapping using three-layered selective audio rays for mobile robots", *In Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, Daejeon, Korea, pp. 2771-2777, 2016.

[8] D. Su, T. V. Calleja, and J. V. Miro, "Towards real-time 3D sound sources mapping with linear microphone arrays", *In Proceedings IEEE International Conference on Robotics and Automation, Singapore, Singapore*, pp. 1662-1668, 2017.

[9] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, ""*Robust Localization in Reverberant Rooms*, Springer, Berlin, Heidelberg, Ch. 8, pp. 157-180, 2001.

[10] 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, 2011.

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

[12] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, p. 439-443, 2013.

[13] T. 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.

[14] A. Canclini, P. Bestagini, F. Antonacci, M. Compagnoni, A. Sarti, and S. Tubaro, "A robust and low-complexity source localization algorithm for asynchronous distributed microphone networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1563-1575, 2015.

[15] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach", *IEEE SP Magazine*, vol. 13, pp. 67-94, 1996.

[16] P. Stoica and R. Mose, ""*Introduction to Spectral Analysis*, Prentice-Hall, 1997.

[17] R. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation", *IEEE Transactions on Antennas and Propagation*, vol. AP-34, pp. 276-280, 1986.

[18] R. Roy and K. Kailath, "ESPRIT-Estimation of Signal Parameters via Rotational Invariance Techniques", *IEEE Transactions on ASSP*, vol. 37, no. 7, pp. 984-995, 1989.

[19] B. Kwon, Y. Park, and Y. S. Park, "Multiple sound source localization using the spatially mapped GGC function", In *ICROS-SICE International Conference*, Japan, pp. 1773-1776, 2009.

[20] Y. Hikoa, M. Matsuo, and N. Hamada, "Multiple-speech-source -localization using advanced histogram mapping method", *Acoustical Science and Technology*, vol. 30, no. 2, pp. 143-146, 2009.

[21] M. Farmani, M. S. Pedersen, Z. Tan, and J. Jensen, "Informed Sound Source Localization Using Relative Transfer Functions for Hearing Aid Applications", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611-623, 2017.

[22] 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, 2017.

[23] 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, 2018.

[24] W. Dai and H. Chen, "Multiple Speech Sources Localization in Room Reverberant Environment Using Spherical Harmonic Sparse Bayesian Learning", *IEEE Sensors Letters*, vol. 3, no. 2, pp. 1-4, 2019.

[25] S. Rickard and F. Dietrich, "DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET", *Proceedings of the Tenth IEEE Workshop on Statistical Signal and Array Processing* (Cat. No. 00TH8496), Pocono Manor, PA, USA, pp. 311-314, 2000.

[26] A. D. Firoozabadi, P. Irarrazaval, P. Adasme, H. Durney, and M. S. Olave, "A Novel Quasi-Spherical Nested Microphone Array and Multiresolution Modified SRP by GammaTone Filterbank for Multiple Speakers Localization", *In Proceedings Signal Processing: Algorithms, Architectures, Arrangements, and Applications* (SPA), Poznan, Poland, pp. 208-213, 2019.

[27] Y. R. Zheng, R. A. Goubran, and M. E. Tanany, "Experimental Evaluation of a Nested Microphone Array With Adaptive Noise Cancellers", *IEEE Transactions on Instrumentation and Measurement Journal*, vol. 53, no. 3, pp. 777-786, 2004.

[28] P. I. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus", *Symposium on Hearing Theory*, IPO Eindhoven, Holland, pp. 58-69, 1972.

[29] A. Aertsen, P. Johannesma, and D. Hermes, "Spectro-temporal receptive fields of auditory neurons in the grass frog", *Biological Cybernetics*, vol. 38, no. 4, pp. 235-248, 1980.

[30] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function", *In a meeting of the IOC Speech Group on Auditory Modeling at RSRE*, vol. 2, no. 7, 1987.

[31] A. D. Firoozabadi and H. R. Abutalebi, "SRP-ML: A Robust SRP-based speech source localization method for Noisy environments," *18-th Iranian Conference on Electrical Engineering* (ICEE), Isfahan, Iran, pp. 2950-2955, 2010.

[32] 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", *Proceeding Interspeech*, pp. 2803-2807, 2018.

[33] 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.

[34] O. Cetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition", *Proceeding Interspeech*, pp. 293-296, 2006.

[35] J. Allen and D. Berkley, "Image method for efficiently simulating small room acoustics", *The Journal of the Acoustical Society of America*, vol. 65, pp. 943-950, 1979.

Received 9 May 2020

Ali Dehghan Firoozabadi (Prof, Dr, Ing), was born in 1985 in Yazd, Iran, graduated in BSc and MSc from the Yazd University in 2007 and 2009 respectively, where he passed his doctoral study in Electrical Engineering-Telecommunications in 2015 at the same institute. During 2015 until 2017 he has worked as a postdoctoral researcher at Universidad de Chile, Universidad de Santiago de Chile, and Pontificia Universidad Católica de Chile. He was appointed as an associated professor in 2017 at the Departamento de Electricidad, Facultad de Ingeniera, Universidad Tecnológica Metropolitana. His primary interest includes the areas of microphone array signal processing, speaker localization and tracking, speech enhancement, light source localization.

Pablo Irarrázaval received his BSc in electrical engineering in 1988 from the Pontificia Universidad Católica de Chile (PUC), and his MSc (1991) and PhD (1995) in Electrical Engineering from Stanford University. He is a Professor with the Electrical Engineering Department and the Institute for Biological and Medical Engineering in the Pontificia Universidad Católica de Chile. He was chairman of the Department and Vice-dean of Academic Affairs of the College of Engineering at PUC. His research interest is in Medical Imaging, particularly with Magnetic Resonance Imaging acquisition and reconstruction, and with Image Perception.

Pablo Adasme is an associate researcher and full professor in computer science at the Electrical Engineering Department of the Universidad de Santiago de Chile. He was born in 1972 in Santiago, Chile. He obtained the title of Industrial Engineer together with Bachelor and Master degrees from Universidad de Santiago de Chile in 2000 and 2003, respectively. In 2010, he received a PhD degree in computer science from the Universit de Paris Sud 11, in Paris, France. Currently, his main research interests are related with deterministic and stochastic combinatorial optimization problems applied to a diverse range of engineering domains including wireless communications, signal processing, network design and energy problems.

David Zabala-Blanco received the BSc degree in electronic systems engineering from Escuela Militar de Ingenieria, La Paz, Bolivia in 2011, and the MSc and PhD degrees in telecommunication engineering from Tecnologico de Monterrey, Monterrey, Mexico in 2014 and 2018, respectively. During 2017, he was in a research stay by working in the project FONDECYT Iniciación No. 11160517 (Analysis, design, and implementation of Nyquist pulses in OFDM next generation wireless communication systems) at the University of Chile, Santiago, Chile. He is currently a Postdoc at the Universidad Catóica del Maule. His research interest includes OFDM-based systems, optical communications, Nyquist-I pulses, and extreme learning machines.

Cesar A. Azurdia-Meza (BSc, MSc, PhD) was born in 1981 in Antigua Guatemala, Sacatepquez, Guatemala. He received the BSc degree in electronics engineering from Universidad del Valle de Guatemala, Guatemala in 2005;and the MSc degree in electrical engineering from Linnaeus University, Sweden in 2009. In 2013 he obtained the PhD degree in Electronics and Radio Engineering, Kyung Hee University, Republic of Korea. He joined the Department of Electrical Engineering, University of Chile as an Assistant Professor in August 2013, where he is currently lecturing on wireless and mobile communication systems. He has served as Technical Program Committee (TPC) member for multiple conferences, as well as a reviewer in journals such as IEEE Communications Letters, IEEE Transactions on Wireless Communications, Wireless Personal Communications, IEEE ACCESS, IET Communications, EURASIP Journal on Advances in Signal Processing, among others. Dr. Azurdia is an IEEE Communications Society Member, as well as Member of the IEICE Communications Society. His research interests include topics such as Nyquist's ISI criterion, OFDM-based systems, SC-FDMA, visible light communication systems, vehicular communications, 5G and beyond enabling technologies, and signal processing techniques for communication systems. He is a co-recipient of the 2019 IEEE LATINCOM Best Paper Award, as well as the 2016 IEEE CONESCAPAN Best Paper Award.