3D Localization of Multiple Simultaneous Speakers with Discrete Wavelet Transform and Proposed 3D Nested Microphone Array

Ali Dehghan Firoozabadi[1], Hugo Durney[1], Ismael Soto[2], Miguel Sanhueza Olave[1]

[1]Department of Electricity, Universidad Tecnologica Metropolitana, Av. Jose Pedro Alessandri 1242, 7800002, Santiago, Chile
[2]Electrical Engineering Department, Universidad de Santiago de Chile, Santiago, Chile
E-mail: adehghanfirouzabadi@utem.cl

Abstract— **Multiple sound source localization is one of the important topic in speech processing. GCC function is used as a traditional algorithm for sound source localization. This function estimates DOA for multiple speakers by calculation the cross-correlation between microphone signals but its accuracy decreases in adverse conditions. The aim of proposed method in this paper is localization of multiple simultaneous speakers in undesirable condition. The proposed method is based on novel 3D nested microphone array in combination with obtained information of Discrete Wavelet Transform (DWT) and subband processing. The proposed 3D nested microphone array prepares the condition for 3D localization and eliminates the spatial aliasing between microphone signals. Also, we propose the DWT for extraction the information of speech signal. Since, the spectral information of speech signal concentrates on low frequencies, we propose a structure of filter bank based on DWT to increase the frequency resolution on low frequencies. The performed evaluation on real and simulated data shows the superiority of our proposed method in comparison with Fullband and subband processing with uniform filters and uniform microphone array.**

Keywords— Simultaneous sound source localization; Wavelet Transform; Generalized Cross-Correlation; Nested microphone array; Subband processing.

I. Introduction

Today's society is dependent to work with different systems. Optimization of these systems is one of the important part in technology improvements [1]. These systems implement the special process based on the received information of environment. One of the input data for these systems is voice instructions of users. The high quality signal of users is required to receive accurate instructions. Microphone arrays are an appropriate structures to receive these instructions of users. Therefore, we need to have the accurate position of speakers to have the high quality of speech signal [2,3].

Some previous researchers tried to localize multiple speakers. Spatial aliasing and close speakers are two important challenges in speaker localization. The traditional methods do not have an appropriate accuracy in adverse conditions. Then, it is necessary to propose a method to localize the speakers with high accuracy.

In recent decade, some research have been done on sound source localization. The most proposed algorithms are for single and non-simultaneous multiple speakers [4,5]. Also, there are some works for multiple simultaneous sound source localization. Time Difference Of Arrival (TDOA)-based localization usually performs by GCC-based methods. Claudio and Parisi [6] made an innovation on this method for localization. Also, Cross Power Spectrum Phase (CPSP) analysis is one of the common methods in speaker localization. The CPSP method is not working well when the aim is localization of multiple simultaneous speakers, because of the cross-correlation between different sound sources [7]. Moreover, in recent years, some efforts have seen to improve the localization algorithm from machine learning point of view [8].

In our previous work [9], we proposed a subband processing method for sound source localization. The weaknesses of the method are: 1- The lake of attention to different information of various subbands, 2- Inaccuracy for localization of close speakers, and 3- limitation to 2D sound source localization (Direction Of Arrival (DOA) estimation).

The proposed method in this paper is 3D simultaneous sound source localization by wavelet transform in combination with 3D nested microphone array. The accurate attention to the frequency band of speech signal is the point of view in subband processing. As known, the frequency components of speech signal are different in subbands. Also, we should solve spatial aliasing problem in microphone signals. For this purpose, firstly, 3D nested microphone array is proposed in this paper to perform the 3D localization and elimination the spatial aliasing. In the next step, we propose to design the DWT of speech signal to prepare the accurate spectral information. In following, the subband processing method is implemented on this information. The most important advantage of proposed method is 3D sound source localization for simultaneous speakers. The DOA of speakers in X, Y and Z directions are calculated and finally, the 3D location of sound sources are estimated by combination of these information. In the proposed method in this paper we assume the number of speaker is known.

In section II, the model of microphone signals and GCC function are introduced. Section III, shows the proposed 3D

ISBN 978-90-827970-1-5 © EURASIP 2018

nested microphone array and also, the proposed method based on this nested array in combination with DWT for subband processing for multiple simultaneous sound source localization. Section IV, explains the simulation conditions, evaluations and experimental results. Finally, some conclusions are presented in Section V.

II. Microphone signal model and GCC function

Two models are considered for microphone signal in speech processing: ideal and real model. In ideal model, the microphone signal is delayed and weakened of source signal. This model is unuseful because it doesn't consider the effect of reverberation in modeling. Then, the real model is selected for simulation the microphone signals. Microphone signals are expressed as:

$$x_m[n] = s[n] * \gamma_m[\vec{d}^{(s)}, n] + v_m[n] \tag{1}$$

where $x_m[n]$ is received signal in m-th microphone, $s[n]$ is speech signal of sound source (speaker), $\gamma_m[\vec{d}^{(s)}, n]$ is room impulse response for path m-th, $v_m[n]$ is additive noise in the position of m-th microphone, $\vec{d}^{(s)}$ is the distance between speaker and m-th microphone, and * denotes convolution operator. The near-field assumption is used for simulations in this paper for the close distance between speakers and microphone array.

The proposed method in this paper is based on the calculation of GCC function in subbands between microphone signals. The distance between speaker and m-th microphone is shown by r_m (for m=1,..,M). τ_{lq} is the TDOA between microphones l and q as:

$$\tau_{lq} = \tau_l - \tau_q \quad for \quad l,q = 1,.....,M \tag{2}$$

The GCC function is implemented on short-time of speech signal in real conditions. If we consider microphones l and q as a pair $\{l,q\}$, then these microphone signals ($x_l[n], x_q[n]$) are used to calculate GCC based on FFT in frame b. If we define $X_{l,b}[k]$, $X_{q,b}[k]$ as Fourier transform of microphone signals $\{l,q\}$ in block b, then GCC function for a continues value of τ can be shown as [10]:

$$\tilde{P}_{lq,b}(\tau) = \frac{1}{K}\sum_{k=0}^{K-1} \left|G_l[k]G'_q[k]\right| \left|X_{l,b}[k]\right| \left|X'_{q,b}[k]\right| e^{jk\frac{2\pi}{K}\tau} \tag{3}$$

The GCC function ($\tilde{P}_{lq,b}(\tau)$) is the cross-correlation of filtered version of $x_l[n]$ and $x_q[n]$ where the Fourier transform for these filters are $G_l[k]$ and $G_q[k]$. If we define $\psi_{lq,b}[k] = \left|G_l[k]G'_q[k]\right|$ as the frequency domain weighted function and $C_{lq,b}[k] = \left|X_{l,b}[k]X'_{q,b}[k]\right|$ as the amplitude of the k-th sample of cross-spectrum, then the GCC function can be re-written as:

$$\tilde{P}_{lq,b}(\tau) = \frac{1}{K}\sum_{k=0}^{K-1} \psi_{lq,b}[k] C_{lq,b}[k] e^{jk\frac{2\pi}{K}\tau} \tag{4}$$

The delay parameter τ can be written as $\tau = \frac{d}{c}\sin\theta$ where τ is related to parameter θ. Then we can calculate DOA instead of TDOA based on this equation. The DOA for multiple speakers can be estimated by maximization of GCC function based on parameter θ. Also, we use of PHAse Transform (PHAT) for weighted function $\psi_{lq,b}[k]$ in GCC function.

$$\psi_{lq,b}^{PHAT}[k] = \frac{1}{\left|X_{l,b}[k]X'_{q,b}[k]\right|} \tag{5}$$

This function performs the whitening of GCC integrated by the normalization of signal's amplitude. In our previous work, we used GCC-PHAT function in subband processing but the algorithm cannot localize close speakers with high accuracy due to using of uniform filters. Also, the spatial aliasing is another undesirable factor in signal processing of spectral information. In the proposed method, spatial aliasing is eliminated in microphone signals due to using of 3D nested array and also, there is a good focus on spectral information of speech signal due to using of DWT.

III. The proposed method for multiple simultaneous sound source localization based on DWT

The purpose of proposed method in this paper is to have a good attention on spectral information of speech signal and also, propose a new 3D nested microphone array. The proposed method uses of W-DO assumption of speech sources for speaker localization. As we mentioned in our previous work [9], the subband processing can increase the percentage of correct estimations for DOA of speakers but the purpose of proposed method in this paper is to have higher accuracy, elimination the spatial aliasing and 3D sound source localization. Fig. 1 shows the block diagram of proposed method.

A. Proposed 3D nested microphone array

Speech signal usually is used in frequency range [0-8000]Hz with sampling frequency Fs=16000Hz. The proposed 3D nested microphone array has been designed to cover the frequency range B=[0-8000]Hz. The 3D microphone array is divided to 4 sub-arrays. The 1st sub-array assigns the highest frequency band where is B1=[4000-8000]Hz and the central frequency is $f_{c1} = 6000Hz$. The inter-microphone distance for this sub-array is considered as $d < \lambda/2$ to eliminate the spatial aliasing where is $d_1 < 2.85cm$. The frequency band for 2nd sub-array is B2=[2000-4000]Hz and central frequency is calculated as $f_{c2} = 3000Hz$. The inter-microphone distance for this sub-array is $d_2 < 5.71cm$. For the 3rd sub-array the frequency range, central frequency and inter-microphone distance are considered B3=[1000-2000]Hz, $f_{c3} = 1500Hz$ and $d_3 < 11.4cm$ respectively. Finally, the frequency band for lowest sub-array is B4=[0-1000]Hz and central frequency $f_{c4} = 500Hz$. Then the inter-microphone distance is calculated as $d_4 < 22.8cm$. Table I shows the information to design this 3D nested microphone array.

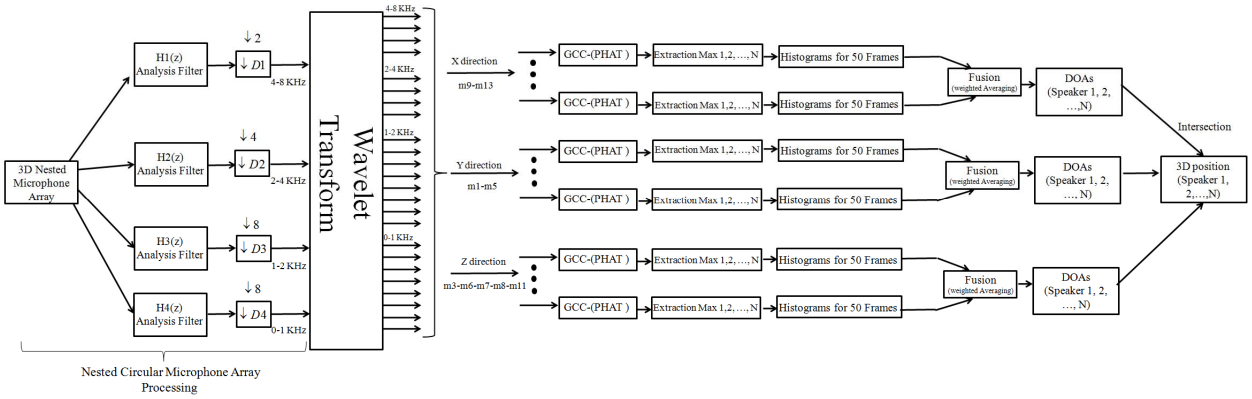


Fig. 1. Block diagram of proposed method based on 3D nested microphone array in combination with DWT.

TABLE I. THE REQUIRED INFORMATION TO DESIGN 3D NESTED MICROPHONE ARRAY.

Band	Bandwidth	f_C	d
1	B1=[4000-8000]Hz	6000 Hz	< 2.85 cm
2	B2=[2000-4000]Hz	3000 Hz	< 5.71 cm
3	B3=[1000-2000]Hz	1500 Hz	< 11.4 cm
4	B4=[0-1000]Hz	500 Hz	< 22.8 cm

Fig. 2 shows the designed 3D nested microphone array based on the information on Table I. The most important advantage of this propose microphone array is non-existence of spatial aliasing between each pair of microphone in sub-arrays. Then, all received information of microphones are used for localization.

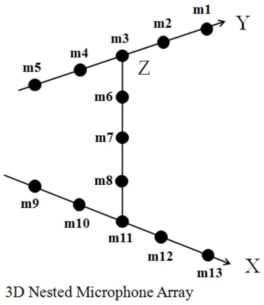


Fig. 2. The structure for proposed 3D nested linear microphone array for sound source localization.

B. *The proposed algorithm in combination with DWT and 3D nested array*

The frequency component of speech signal is higher in low frequencies. Then an appropriate filtering is required for subband processing. Discrete wavelet transform is used as an accurate filtering because it has potential to change the frequency resolution in different bands. Firstly, microphone signals are passed of analysis filters for nested array.

$$x_{ha,b}[n] = x_a[n] * h_b[n] \qquad for \;\; a = 1,.....,13 \;\; \text{and} \;\; b = 1,.....,4 \quad (6)$$

where $h_b[n]$ are the analysis filters. This 3D nested microphone array divide microphone signals to 4 categories related to 4 subbands. The frequency spectrum related to analysis filters have been shown in Fig. 3. In the next step, the output of analysis filters related to nested array are passed of wavelet transform (wavelet filters). We need to design a wavelet transform with different frequency resolution in different bands due to different frequency components of speech signal in various frequency range. Fig. 4 shows the proposed wavelet transform structure to use in this paper.

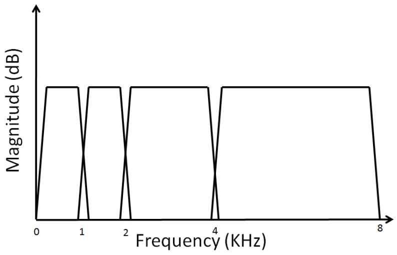


Fig. 3. The frequency spectrum for analysis filters of 3D nested microphone array.

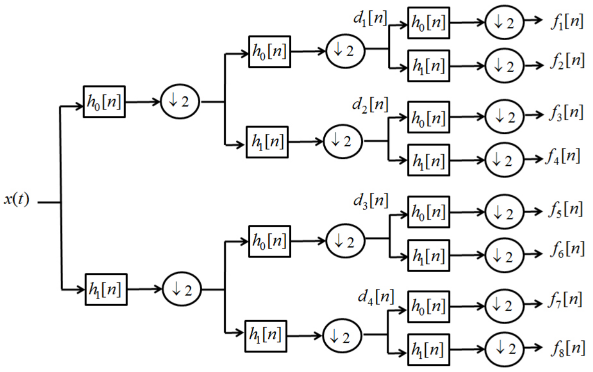


Fig. 4. A three-level and two-channel iterative filter band for DWT.

The Continues Wavelet Transform (CWT) for signal $x(t)$ is described by [11]:

$$W_\Psi(s,\tau) = \int_{-\infty}^{+\infty} x(t)\psi^*_{s,\tau}(t)dt \qquad (7)$$

where,

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-\tau}{s}) \qquad (8)$$

s and τ are scale and translation parameters, respectively. $W_\Psi(s,\tau)$ denotes the wavelet transform coefficients and ψ is the fundamental mother wavelet. DWT is a sampled version of CWT. The only difference is for scale and position value where are based on power of two. The value of s and τ are: $s = 2^j$, $\tau = k * 2^j$ and $(j,k) \in Z^2$. It means:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{2^j}}\psi(\frac{t-k*2^j}{2^j}) \qquad (9)$$

The key point in DWT is signal decomposition. The main idea for signal decomposition is to use low-ass filter (LPF) with down sampling. As shown in Fig. 4, recursive filter banks are used for implementation where $h_0[n]$ and $h_1[n]$ are analysis LPF and HPF, respectively and $\downarrow 2$ represents the down sampling operator by a factor of 2. Based on Fig. 4, microphone signals are divided to 8 subbands for frequency range B1 and B2 where it needs a three-level forward DWT based on two-channel recursive filter band $(f_1[n]-f_8[n])$. A two-level forward DWT is used for bands B3 and B4 $(d_1[n]-d_4[n])$ because we have less speech information in these bands. The relation between filters $h_0[n]$ and $h_1[n]$ is:

$$h_1[n] = (-1)^n h_0[\mathrm{L}+1-n] \qquad (10)$$

where $h_0[n]$ is coefficients for a LPF and $h_1[n]$ is a designed LPF based on $h_1[n]$. There is no certain way to choose an specific wavelet. The type of wavelet depends upon the signal to be analyzed and application. Daubechies (Db4) wavelet has been used because it prepares more accurate information of speech signal. Also, this wavelet detect very well the energy spectrum in low frequencies.

In following, the subband signals by wavelet transform are entered to GCC-PHAT function. In this step, the GCC-PHAT function is calculated for all subbands and also, all microphone pairs related to this subband in X direction of microphone array. Then the *N*-first peak positions of GCC-PHAT function are extracted.

$$\begin{aligned}
\hat{\theta}_{1,(\mathrm{m}_l,\mathrm{m}_q),k}(X) &= \underset{\theta}{\arg\max}\ \tilde{P}^{PHAT}_{x_{h_{\mathrm{a,b}}(m_l,m_q,k)}(X)}(\theta) \\
\hat{\theta}_{2,(\mathrm{m}_l,\mathrm{m}_q),k}(X) &= \underset{\substack{\theta \\ \theta\neq\hat{\theta}_1}}{\arg\max}\ \tilde{P}^{PHAT}_{x_{h_{\mathrm{a,b}}(m_l,m_q,k)}(X)}(\theta) \\
&. \\
\hat{\theta}_{\mathrm{N},(\mathrm{m}_l,\mathrm{m}_q),k}(X) &= \underset{\substack{\theta \\ \theta\neq\hat{\theta}_1,\ldots,\hat{\theta}_{N-1}}}{\arg\max}\ \tilde{P}^{PHAT}_{x_{h_{\mathrm{a,b}}(m_l,m_q,k)}(X)}(\theta)
\end{aligned} \qquad (11)$$

The peak position histogram for GCC-PHAT function is calculated for each subband and in direction X.

$$\begin{aligned}
\mathrm{A}_k = \{&\hat{\theta}_{1,(\mathrm{m}_l,\mathrm{m}_q),k}(\mathrm{X}), \hat{\theta}_{2,(\mathrm{m}_l,\mathrm{m}_q),k}(\mathrm{X}), \\
&\ldots, \hat{\theta}_{N,(\mathrm{m}_l,\mathrm{m}_q),k}(\mathrm{X}); \forall\ (\mathrm{m}_l,\mathrm{m}_q)\}
\end{aligned} \qquad (12)$$

$$Hist_{Ave.}(\theta) = \underset{k=1,2,\ldots,K}{fusion}\ \left(Hist\left(\mathrm{A}_k\right)\right) \qquad (13)$$

Finally, the weighted average method [9] is used to combine the histogram for different subbands.

$$Hist_{W.Ave.}(\theta) = \frac{1}{K}\sum_{k=1}^{K}\frac{S_{1,k}}{\sum_{i=2}^{N} S_{i,k}} Hist\left(\mathrm{A}_k\right) \qquad (14)$$

where $S_{\mathrm{i},k}$ is the magnitude of the *i*-th peak in histogram. Finally, *N*-first peak of the final histogram are extracted to be the DOAs for N multiple speakers in direction X of microphone array.

$$\begin{aligned}
&\hat{\theta}_1(X) = \underset{\theta\in D}{\arg\max}\, Hist_{W.Ave.}(\theta) \text{ first speaker} \\
&\hat{\theta}_2(X) = \underset{\substack{\theta\in D \\ \theta\neq\hat{\theta}_1}}{\arg\max}\, Hist_{W.Ave.}(\theta) \text{ second speaker} \\
&. \\
&\hat{\theta}_N(X) = \underset{\substack{\theta\in D \\ \theta\neq\hat{\theta}_1,\ldots,\hat{\theta}_{N-1}}}{\arg\max}\ Hist_{W.Ave.}(\theta)\ N\text{-th speaker}
\end{aligned} \qquad (15)$$

where *D* is the area for all possible direction of sound sources. This process is repeated for microphone signals in direction Y and Z of 3D nested microphone array. One DOA is extracted for each direction of microphone array. Each DOA shows a plane in 3D space for all possible location of speaker, where the intersection of planes for 3 directions of microphone array prepares an area for location of speaker. Then, we estimate a point in this area that have minimum distance to each of 3 planes for direction X, Y and Z.

IV. Simulations and Results

In this paper, we will show the evaluations on both simulated (from TIMIT database) and real (recorded data on sound processing laboratory in FBK at Trento, Italy) data. It should be noted that almost 90% of the time of overlapping speech is related to just two simultaneous speakers. Also, about 10% of overlapping part is related to three simultaneous speakers. Four or more simultaneous speakers rarely happens in real conditions. For this reason, in the simulations of this research, we consider the cases of one, two, and three simultaneous speakers. In this research, 40s speech signal is used for each speaker. 15s of overlapping speech is related to the case of two simultaneous speakers, and 2.3s is related to the case of three simultaneous speakers. Also, we use Image algorithm [12] to model sound propagation environment. We assume an 13-microphone 3D nested array (Fig. 2) with 2.8*cm* inter-microphone distance to eliminate the spatial aliasing. In simulation, room dimension is considered (472,592,420)*cm*. First, second and third speaker are located at (404,132,110)*cm*, (309,128,120)*cm* and (174,140,140)*cm*; so, the correct DOA for three speakers with respect to the array reference axis (X direction of nested array) are 40°, 85° and 145°, respectively. Then, first, second and third speakers have 198*cm*, 141*cm* and 176*cm* distance to the center of microphone array (microphone m7). Three scenarios are considered for simulations: reverberant scenario ($RT_{60} = 650ms,\ SNR = 20dB$), noisy scenario ($RT_{60} = 200ms,\ SNR = 5dB$) and noisy-reverberant scenario ($RT_{60} = 650ms,\ SNR = 5dB$). The proposed method in this paper is compared with Fullband (sources are localized by considering entire signal spectrum) and subband (with uniform array and uniform filters) methods. Simulations are implemented on 50 continues frames of speech signal. Also Fig. 5 shows a view of simulated room.

The simulations have been done for single, two and three simultaneous speakers. Fig. 6 shows the Mean Square Error (MSE) of obtained 3D estimation of location. We consider error in estimation for the condition with more than 5*cm* distance to correct position. Fig. 6(a)-(c) shows the results of single speaker, two and three simultaneous speakers respectively, for real and simulated data and also for different scenarios. The MSE of our proposed method for source location in less than two other methods. It means our proposed method localizes the location of sound sources better than Fullband and subband method with uniform filters.

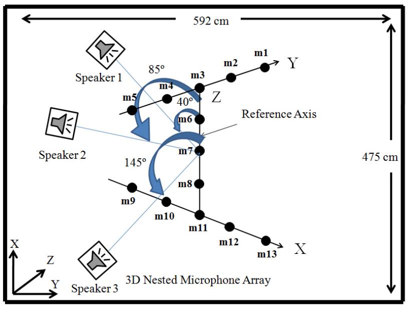
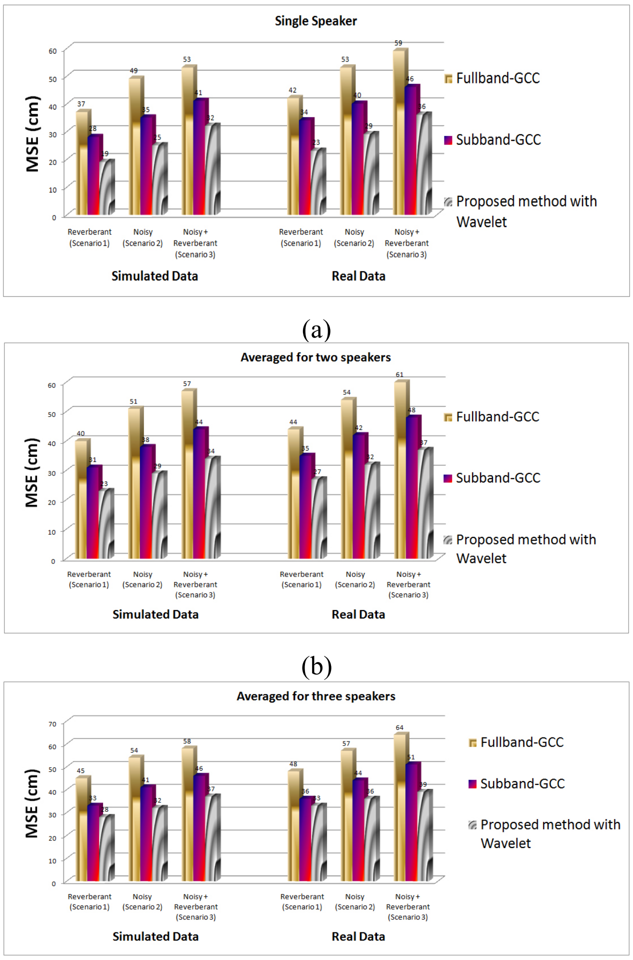


Fig. 5. A view of room for simulations and similar to real conditions



(a)

(b)

(c)

Fig. 6. The obtained MSE for real and simulated data in 3 different scenarios a) Single speaker, b) Two simultaneous speakers, and c) Three simultaneous speakers.

Table II shows the computational complexity comparison for two simultaneous speakers and different scenarios for our proposed method, Fullband and subband with uniform filters. As we mentioned, our proposed method has better results in compare with other methods but this improvement in accuracy has the cost of slightly increasing the computational complexity because of using wavelet transform.

V. Conclusion

Sound source localization is one of the most important topic in speech processing specially in the condition with multiple simultaneous speakers. GCC function is a common and traditional method for sound source localization but it is not accurate in noisy-reverberant environments and also, with multiple simultaneous speakers conditions. In this paper, firstly, we proposed the 3D nested microphone array. This array localize 3D position of speakers. Also, it eliminates the spatial aliasing based on its nested format. Then, we proposed a method for subband processing based on DWT to have better attention on low frequency component of speech signal. The advantage of wavelet transform is the focus on different frequency components of speech signal. Wavelet filters were designed in a way to have better resolution on low frequency band of speech signal to extract more information. Finally, the proposed method was compared with Fullband and subband method with uniform filters and microphone array. Although our proposed method has a few more computational complexity in comparison with other methods, but it localized the position of simultaneous speakers with lower MSE.

TABLE II. Computational Complexity for Two Simultaneous Speakers and Different Scenarios for Our Proposed Method, Fullband and Subband.

	Fullband-GCC	**Subband-GCC**	**Proposed method-Wavelet**
Scenario 1	351	412	485
Scenario 2	369	425	507
Scenario 3	390	401	498

Acknowledgment

The authors acknowledge financial support from: CONICYT - PIA - Anillo ACT1416, FONDEF project code IT17M10012, USACH and Internal grant DIDA L1-03 comunicaciones inalambricas con VoIP en tiempo real sobre canal VLC utilizando SIP y RFC, UTEM.

References

[1] H. Nakashima, and T. Mukai, “3D Sound Source Localization System Based on Learning of Binaural Hearing,” in *Proc. IEEE SMC* 2005, pp. 3534-3539, 2005.

[2] Y. Sasaki, et al, “2D Sound Localization from Microphone Array Using a Directional Pattern,” in *The 25th Annual Conference of The Robotics Society of Japan*, 2007.

[3] M. Brandstein, and D Ward, *Microphone Arrays*, Springer Verlag, 2001.

[4] X. Sheng, and Y.-H. Hu, “Maximum Likelihood Multiple-Source Localization Using Acoustic Energy Measurements with Wireless Sensor Networks,” *IEEE Trans. Signal Process.*, vol. 53, pp. 44 - 53, Jan 2005.

[5] J. H. DiBiase, “*A High-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*,” PhD thesis, Brown University, Providence, RI, May 2000.

[6] E. Di Claudio, R. Parisi, and G. Orlandi, “Multi-Source Localization In Reverberant Environments,” in *Proc. ICASSP*, Istanbul City, Turkey, 2000.

[7] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, “Localization of multiple sound sources based on a CSP analysis with a microphone array,” in *Proc. ICASSP*, vol. 1, pp. 1053-1056, Istanbul, Turkey, 2000.

[8] R. Talmon, I. Cohen, and S. Gannot, “Supervised source localization using diffusion kernels,” in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2011, pp. 245–248.

[9] A. Dehghan Firoozabadi, H. R. Abutalebi, “Subband processing-based approach for the localisation of two simultaneous speakers,” *IET Signal Process*. vol. 8, issue 9, 2014, pp. 996–1008.

[10] C. H. Knapp, and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, 1976, pp. 320-327.

[11] I. Mamatha, S. Tripathi, T. S. B. Sudarshan, “Convolution based efficient architecture for 1-D DWT,” in *International Conference on Computing Communication and Automation 2017*, pp. 1436-1440.

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