

Multiple sound source localization using gammatone auditory filtering and direct sound component detection

Huaiyu Chen and Li Cao

Department of Automation, Tsinghua University, Beijing 100084, China

Abstract. In order to research multiple sound source localization with room reverberation and background noise, we analyze the shortcomings of traditional broadband MUSIC and ordinary auditory filtering based broadband MUSIC method, then a new broadband MUSIC algorithm with gammatone auditory filtering of frequency component selection control and detection of ascending segment of direct sound component is proposed. The proposed algorithm controls frequency component within the interested frequency band in multichannel bandpass filter stage. Detecting the direct sound component of the sound source for suppressing room reverberation interference is also proposed, whose merits are fast calculation and avoiding using more complex de-reverberation processing algorithm. Besides, the pseudo-spectrum of different frequency channels is weighted by their maximum amplitude for every speech frame. Through the simulation and real room reverberation environment experiments, the proposed method has good performance. Dynamic multiple sound source localization experimental results indicate that the average absolute error of azimuth estimated by the proposed algorithm is less and the histogram result has higher angle resolution.

1 Introduction

Sound source localization (SSL) is a key research point in the field of speech signal processing. It usually uses an array of microphones to receive the acoustic signal and applies a series of signal processing techniques to estimate the direction of arrival (DOA) of active sound sources. It plays a significant role in many other application scenarios such as sound signal separation, speech enhancement and recognition, speech de-noising and echo cancellation, robot and human-computer interaction, remote video conferencing system, smart home monitoring system and so on.

According to the relevant research results, most state-of-the-art sound source localization methods can be divided into three categories. The first category is the localization method based on the time difference of arrival (TDOA) between different microphone pairs. It is easy to implement and has low complexity. TDOA-based methods can meet the real-time requirements of a single sound source azimuth estimation, but fail when multiple active sound sources exist ^[1]. The second category is beamforming method, which is one of frequency domain computing technique. While this kind of localization method has bigger angle ambiguity and high complexity in terms of the existence of multiple active sound sources. The last category is subspace decomposition based technique, which is capable of processing the situation of multiple sources. The typical approach is the well-known multiple signal classification (MUSIC) algorithm. Broadband MUSIC, deformed MUSIC algorithm, can give the estimated value of multiple sound source azimuths simultaneously, which depends on the eigenvalue decomposition of spectrum covariance matrix. This method has high angle resolution.

Traditional MUSIC approach is usually applied to the radar antenna array signal processing, which requires stationary far-field narrowband signal. The majority of broadband MUSIC deformation



algorithms are gradually put forward for different wideband application situations [2,3]. Although broadband MUSIC was improved to obtain estimation value of DOA, the position accuracy is not high. Because broadband MUSIC method divide the whole frequency band into several equal length sub bands or frequency bins, but some frequency bins not always satisfy the narrowband characteristic. To be specific, parts of frequency bins within low frequency components cannot be treated as narrow band signal componence. Because the corresponding center frequency (denoted with f_0) is comparatively small, while bandwidth (denoted with B) is fixed, which can't meet the narrowband approximating condition: $B/f_0 < 0.1$. This broadband MUSIC approach exists the estimated error essentially. Auditory filtering based broadband MUSIC (Gammatone-BroadbandMUSIC) was presented to localize two sound source in simulation environment without reverberation [4]. Gammatone auditory filtering is a multi-channel bandpass filtering technique with different variable bandwidth, which has smaller bandwidth when center frequency is low. Moreover, pseudo-spectrums of all the frequency channels need to be calculated, causing huge calculation quantity. What's more, this algorithm is only suitable for no reverberation environment.

Considering the above existed shortcomings, this paper comes up with a new broadband MUSIC algorithm with gammatone auditory filtering of frequency component selection control and detection of ascending segment of direct sound componence, aiming at estimating DOA of multiple active sound sources in the room reverberation environment. This paper assumes the far field model. The proposed algorithm controls frequency component within the interested frequency band in multi-channel bandpass filter stage. Because the frequencies of speech voices generally change from 200Hz to 3000Hz. On the other hand, detecting direct componence of sound source supresses the room reverberation, which will enhance the accuracy and reduce the computing time owing to less sound frame to be processed.

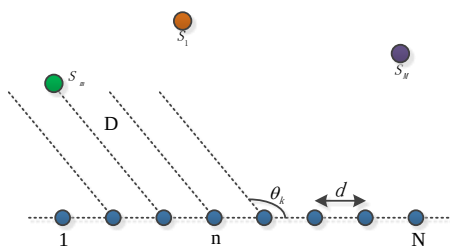


Figure 1: Far field sound source localization model

2 Proposed Method

2.1 Gammatone filtering based MUSIC algorithm

Gammatone filter is a bank of bandpass filters, having different central frequencies and bandwidths. As to gammatone filter, lower the center frequency is, narrower the corresponding bandwidth is. The impulse response of the k th filter is as follows [5,6,10]:

$$h_k(t) = \begin{cases} c_k t^{n-1} e^{-2\pi b_k t} \cos(2\pi f_k t + \phi_k), & t > 0 \\ 0, & t \leq 0 \end{cases} \quad (1)$$

where $k = 1, \dots, C$, C is the total channels, c_k is the gain coefficient, n is the the filter order, b_k is the decay coefficient, f_k is center frequency of the filter and ϕ_k is the phase. The frequency response of four-order and 64 channels gammatone filter is shown in Figure 2, which is used during all simulations and experiments in this paper. The decay coefficient b_k can be calculated by the following equation [6]:

$$b_k = 1.019 \times \text{ERB}(f_k) = 1.019 \times (24.7 + 0.108 f_k) \quad (2)$$

which ERB is short for the equivalent rectangular bandwidth and is a psychoacoustic measure parameter of auditory filter.

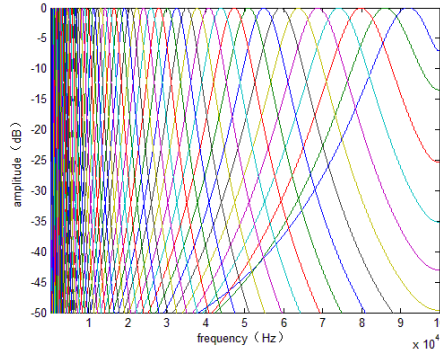


Figure 2: Frequency response of the gammatone filter bank

This paper assumes the far field model. Suppose that the number of sound sources is M and uniform linear microphone array has N elements. The schematic diagram is in Figure 1. After the q th speech frame received by the n th microphone $\mathbf{x}_n(q)$ is filtered by gammatone filter with C channels, the filtered signal matrix can be denoted as:

$$\mathbf{y}_n(q, f_k) = [\mathbf{g}_n(q, f_1) \quad \mathbf{g}_n(q, f_2) \quad \cdots \quad \mathbf{g}_n(q, f_C)]^T \quad (3)$$

$$\mathbf{g}_n(q, f_k) = [g_n^k(q, 1) \quad g_n^k(q, 2) \quad \cdots \quad g_n^k(q, L)] \quad (4)$$

where q denotes number of the speech frame, L is the length of each frame. $\mathbf{g}_n(q, f_k)$ is the filtering result of k th frequency channel of $\mathbf{x}_n(q)$, $k = 1, \dots, C$, $n = 1, \dots, N$. Then the signal matrix expression of k th frequency channel of the q th frame signal received by microphone array is as follows:

$$G(q, f_k) = [\mathbf{g}_1(q, f_k) \quad \mathbf{g}_2(q, f_k) \quad \cdots \quad \mathbf{g}_N(q, f_k)]^T \quad (5)$$

Broadband MUSIC algorithm requires a complex signal, whereas $G(q, f_k)$ is the real signal. So the Hilbert transform is performed on $G(q, f_k)$ to convert the real signal into a complex signal. The expression of transformed signal is:

$$X_{GH}(q, f_k) = \begin{bmatrix} z_1^k(q, 1) & z_1^k(q, 2) & \cdots & z_1^k(q, L) \\ z_2^k(q, 1) & z_2^k(q, 2) & \cdots & z_2^k(q, L) \\ \vdots & \vdots & z_n^k(q, l) & \vdots \\ z_N^k(q, 1) & z_N^k(q, 2) & \cdots & z_N^k(q, L) \end{bmatrix} \quad (6)$$

where $z_n^k(q, l) = \text{Hilbert}(g_n^k(q, l))$, l is the intra-frame offset, $l = 1, \dots, L$. The subscript GH denotes the result after Gammatone filtering and Hilbert transform processing.

Then the expression of k th frequency channel of the q th frame signal received by microphone array is changed into:

$$X_{GH}(q, f_k) = A(\theta, f_k) S_{GH}(q, f_k) + N_{GH}(q, f_k) \quad (7)$$

$$A(\theta, f_k) = [\mathbf{a}(\theta_1, f_k) \quad \mathbf{a}(\theta_2, f_k) \quad \cdots \quad \mathbf{a}(\theta_M, f_k)] \quad (8)$$

$$\mathbf{a}(\theta_m, f_k) = \begin{bmatrix} 1 & e^{-j2\pi f_k \frac{d \sin(\theta_m)}{c}} & \cdots & e^{-j2\pi f_k \frac{(N-1)d \sin(\theta_m)}{c}} \end{bmatrix}^T \quad (9)$$

where $A(\theta, f_k)$ is the direction controlling matrix of whole sound source system in f_k frequency component^[3,4]. $\mathbf{a}(\theta_m, f_k)$ is the direction controlling vector in frequency f_k and direction θ_m from the m th sound source and c is the acoustic velocity.

By computing the covariance matrix of $X_{GH}(q, f_k)$ and carrying out eigenvalue decomposition, we can obtain the orthogonal signal subspace U_s and noise subspace U_N . Then the pseudo-spectrum of k th frequency channel of the q th frame signal can be given by:

$$P_{music_k}(\theta) = \frac{a^H(\theta, f_k)a(\theta, f_k)}{a^H(\theta, f_k)U_N(f_k)U_N^H(f_k)a(\theta, f_k)} \quad (10)$$

where the superscript H denotes conjugate transpose.

2.2 Frequency selection control and detection of ascending segment of direct sound componence

As mentioned above, pseudo-spectrums of all the frequency channels need to be calculated when using Gammatone-BroadbandMUSIC algorithm. Besides its high complexity, parts of frequency channels are noise interference probably. We advise that filtering and calculation are only in the interested frequency band. In this paper, the band is limited to 200 ~ 3000Hz. This will reduce the error interference in essence, and also reduce the processing time.

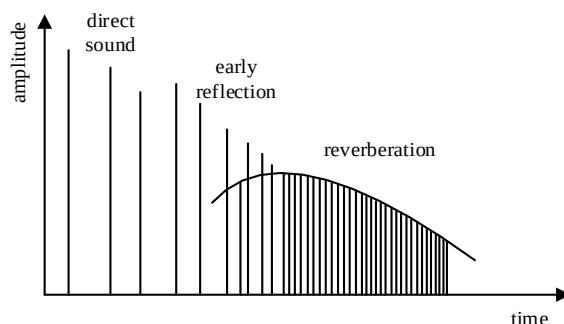


Figure 3: Schematic diagram of sound source componence

The article [4] mentioned that Gammatone-BroadbandMUSIC algorithm was only suitable for no reverberation environment. However, the reverberation in the room is unavoidable, usually very strong, which will cause poor results. In fact, the voice signal transmission will encounter tables and chairs, wall absorption and multiple reflections, so the signal received by microphone is the mixed signal along multiple propagation paths. It mainly consists of three parts: direct sound componence, early reflection and reverberation componence.

This paper proposes a bran-new method based on local maximal judgment to detect the ascending segment of direct sound componence. The termination condition of the searching is that if there is a continuous number of K maximal points (K was set to 10 in this paper.), whose amplitudes are less than the certain scanned maximal point. This point will be regarded as the end point of speech ascending segment. Although the maximal point detected by this method is not the actual position, the effect of the late reflection and reverberation is greatly reduced. Figure 4 shows the detection result of ascending segment of direct sound. The simulation results in the next section also show that this detection method is valid and improves the SSL accuracy.

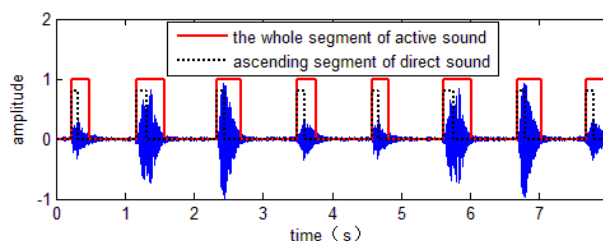


Figure 4: Detection result of ascending segment of direct sound

Especially to deserve to be mentioned, we found that the estimation results of different frequency channels were discrepant. The pseudo-spectrum of the lower frequency region had a larger peak and the number of peaks was more reasonable. However, many pseudo peaks appeared in the higher frequency region, and the amplitude is relatively small. So the weighted technique is proposed according to the maximum value of each pseudo-spectrum.

$$\omega_k = \frac{\max(P_{music_k})}{\max\{\max(P_{music_1}), \dots, \max(P_{music_C})\}} \quad (11)$$

Then the final pseudo-spectrum of each speech frame can be obtained by:

$$P_{GammatoneMUSIC}(\theta) = \sum_{k=k_1}^{k_2} \omega_k P_{music_k} \quad (12)$$

where $1 \leq k_1 < k_2 \leq C$, the number of frequency channel belonging to the interested frequency band is $k_2 - k_1 + 1$.

3 Simulation and experiment

In this section, the performance of proposed method from the perspective of both computer simulations and real position experiments is shown. We compared three performance indicators of the proposed method with broadband MUSIC and Gammatone-BroadbandMUSIC method.

3.1 Performance indicators

The first performance indicator is the Frame Accuracy Rate (FAR), used to evaluate the frame estimation accuracy rate after frame framing of speech from the given position. The specific expression is given by:

$$FAR = \frac{1}{N_{frame}} \sum_{q=1}^{N_{frame}} T(\hat{\theta}_q - \theta_0), T(\theta) = \begin{cases} 1, & |\theta| \leq 6^\circ \\ 0, & |\theta| > 6^\circ \end{cases} \quad (13)$$

The second indicator is the Mean Absolute Estimated Error (MAEE), which is used to evaluate the absolute deviation of the estimated results from the reference angle for each speech frame [7]. The specific expression is given by:

$$MAEE = \frac{1}{N_{frame}} \sum_{q=1}^{N_{frame}} |\hat{\theta}_q - \theta_0| \quad (14)$$

The third indicator is the Histogram Absolute Estimated Error (HAEE), which is the histogram of total estimated results of all the frames.

$$HAEE = |\tilde{\theta} - \theta_0| \quad (15)$$

where θ_0 is the reference value of the azimuth, $\hat{\theta}_q$ is the estimated value of the qth speech frame, $\tilde{\theta}$ is the maximum position after spline interpolation on histogram and N_{frame} is the total number of frames for the calculation process. HAEE is a statistical average of all results, smaller relative to MAEE value.

3.2 Simulation results

The simulation is based on Lehmann's improved Image Source Method [8,9]. The simulated microphone array is a uniform linear array with 8 elements ($N = 8$), the array interval $d = 0.06$ m. The room size is $6.50 \text{ m} \times 7.80 \text{ m} \times 2.90 \text{ m}$. The array is located at (3.0 m, 1.5 m, 1.5 m). The height of microphone array and the sound source is 1.5m in the same horizontal plane, which will help reduce estimation error. The sampling frequency is 200 kHz. The reflection coefficient of four walls is 0.7, while the floor and the ceiling possess 0.8 and 0.6. The acoustic velocity is 343 m/s. The three sound sources used in the simulation are the Chinese pronunciation recorded signal of "Tsinghua University" (referred to as sample 1), TV snow noise (referred to as sample 2) and English pronunciation signal of digital 1 to 10 (referred to as sample 3). Three typical azimuths are -25 degree, +5 degree and +40 degree.

Table.1: Frame accuracy under different reverberation time

FAR	200ms	300ms	400ms	500ms	600ms
not detect	0.8155	0.7143	0.6487	0.5422	0.4893
do detect	0.8488	0.8000	0.7857	0.7555	0.7292

The first simulation was carried out to illustrate the effectiveness of the method of direct sound detection. Sound source sample 1 was placed +40 degree with 2m distance. The estimated results before and after the direct sound detection were compared when changing the reverberation time parameter T_{60} from 200 to 600 ms. The specific result is shown in Figure 5 and Table 1.

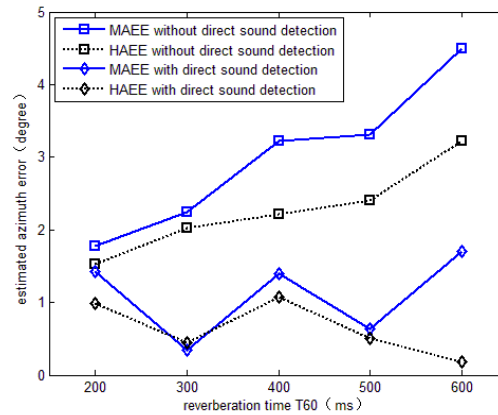


Figure 5: Influence of direct sound component detection on azimuth estimation

Both Figure 5 and Table 1 show that the improved Gammatone-BroadbandMUSIC algorithm with detecting ascending segment has good performance, because three performance indicators all are better even in the strong reverberation condition. This indicate that the bran-new detecting method of direct sound is effective.

In order to compare the proposed method with broadband MUSIC (also called as algorithm 1) and ordinary Gammatone-BroadbandMUSIC (also called as algorithm 2), we performed the second simulation: the position of microphone array was fixed, reverberation time $T_{60}=200\text{ms}$, the distance between the array and sound source changed from 1m to 5m and distance interval was 1m, each location were simulated for three different directions with using three different sound source signal. So there are 45 groups of simulation data totally. Three algorithm were performed for every group simulation data. Their average performance indicators are presented in Table 2. It's clear that the proposed algorithm has higher frame accuracy rate and less estimated absolute error.

Table.2: Comparison of single sound source simulation estimation results of different algorithms

method	Average FAR	Average MAEE	Average HAAE
algorithm 1	0.8333	1.94°	1.57°
algorithm 2	0.8451	2.14°	1.05°
proposed algorithm	0.8991	1.52°	0.56°

3.3 Experimental results

This section presents two parts of experiment, including single sound source localization and dynamic multiple sound source localization. All experiments were implemented in the real meeting room with the about size of 6.50 m*7.80 m*2.90m. The room environment was complicated because there were many chairs, several sofas and a big meeting desk, which would arise strong reflection and reverberation and worsen the experimental localization results.

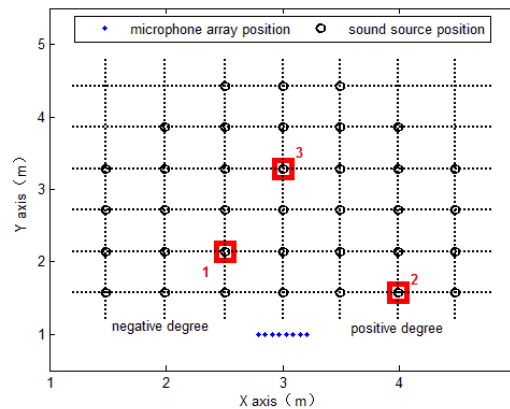


Figure 6: Different positions for sound source acquisition

Figure 6 shows the 36 different positions on the same side of the microphone. Their azimuth angle is in the range of -70 degree to +70 degree in same horizontal plane. We collected three sound source samples at each position, a total of 108 groups of sound data. For each group of data, three different sound source localization algorithms were used for azimuth estimation, the results were shown in Table 3 below. As can be seen from Table 3, the proposed algorithm has better performance than the other two algorithms, and the frame estimation accuracy of the three speech samples is close to 70%.

Table.3: Comparison of FAR of single sound source experiment estimation results of three different algorithms

FAR	sound sample 1	sound sample 2	sound sample 3
algorithm 1	0.5405	0.5840	0.5775
algorithm 2	0.5929	0.6236	0.5868
proposed algorithm	0.6793	0.6842	0.6904

On the other hand, in order to illustrate the effectiveness and expansibility of this algorithm, multiple sound source localization experiments were continued in the same indoor environment. Here the specific description of the dynamic sound source experiment is as follows: three specific sound source positions corresponding to the red rectangle(numbered as 1,2,3) in Figure 9 were selected, whose azimuths are approximately -21 degrees, + 58 degrees and 0 degrees, respectively. Three sound source phonated in turn, then followed by two mixed voices and three sound sources at the same time. So the number of active sound source is variable.

Table.4 Comparison of multiple sound source experiment estimation results of three different algorithms

method	average HAEE		
	-21°	0°	+58°
algorithm 1	1.02°	0.61°	0.02°
algorithm 2	0.23°	0.48°	2.27°
proposed algorithm	0.93°	0.40°	0.75°

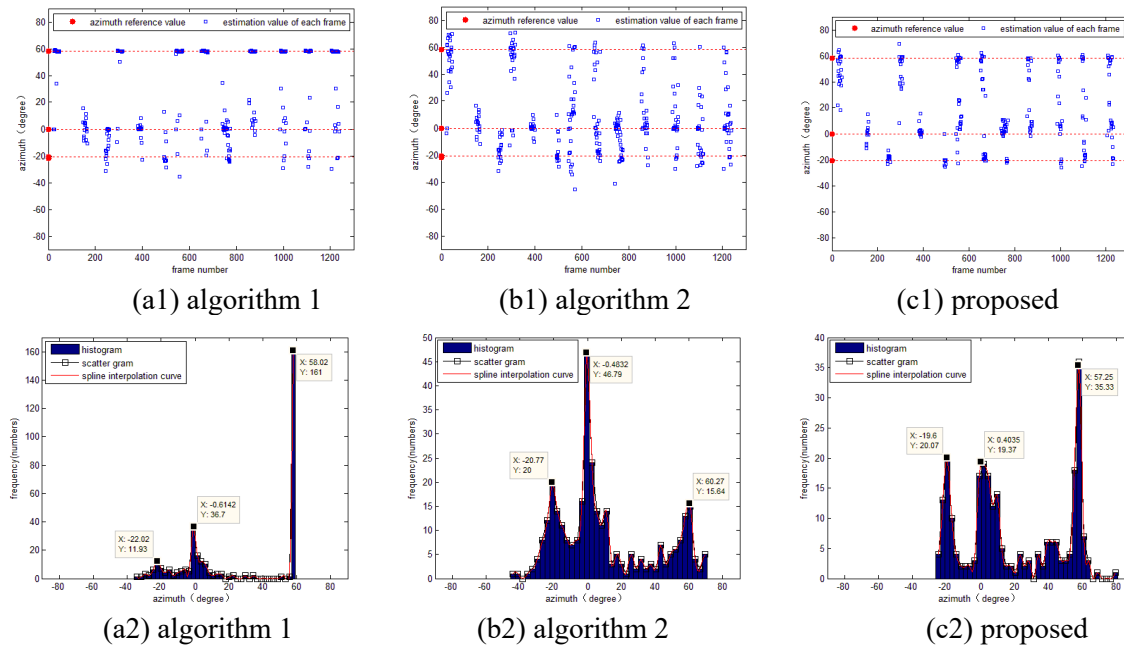


Figure 7 Comparison of multiple sound sources localization experimental results of three different algorithms

The experimental results of the dynamic estimation and histogram of three different algorithms are shown in Figure 7.

Table 4 shows three average HAEE indicators contrast with different algorithms, the proposed algorithm has higher estimation accuracy. The dynamic estimation results of different algorithms can be seen in Figure 7(a1)-(c1). The proposed algorithm not only has the advantage of the two other methods in the dynamic estimation of each frame, but also has the better angle resolution of the statistical histogram.

4 Conclusion

In this paper, a new broadband MUSIC algorithm with gammatone auditory filtering of frequency component selection control and detection of ascending segment of direct sound component is proposed. Based on ordinary auditory Gammatone filtering broadband MUSIC algorithm, three improvements are made. First of all, controlling frequency component selection when gammatone filtering, which is limited into interested band. Secondly, detecting the ascending segment of direct sound component is extremely necessary. It can suppress the reflection and reverberation validly. Thirdly, the pseudo-spectrums of different frequency channels of each frame are weighted according to the maximum value of each pseudo-spectrum. Under three performance indicators given in this paper, the simulation and experimental results show that the proposed algorithm has better performance. This paper only gives the azimuth estimation of multiple sound sources. In the follow-up study, we will focus on the estimation of the corresponding distance.

Acknowledgements

The authors are truly grateful to Shenzhen Horn Audio Co. Ltd. for their support.

References

- [1] Knapp C, Carter G. "The generalized correlation method for estimation of time delay", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **24**, pp. 320-327, (1976).
- [2] Asono F, Asoh H, Matsui T. "Sound source localization and signal separation for office robot 'Jijo-2'", *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 243-248, (1999).

- [3] Potamitis I, Kokkinakis G. "Speech separation of multiple moving speakers using multisensor multitarget techniques", *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, **37**, pp. 72-81, (2007).
- [4] Liao Fengchai, Li Peng, Liu Wenju. "Auditory filter based broadband MUSIC algorithm for sound source localization", *Acta Acustica*, **6**, pp. 642-650, (2012).
- [5] Patterson R D, Nimmo-Smith I, Holdsworth J, et al. "An efficient auditory filterbank based on the gammatone function", *IOC Speech Group on Auditory Modelling at RSRE*, **2**, (1987).
- [6] Roman N, Wang D L, Brown G J. "Speech segregation based on sound localization", *The Journal of the Acoustical Society of America*, **114**, pp. 2236-2252, (2003).
- [7] Pavlidi D, Griffin A, Puigt M, et al. "Real-time multiple sound source localization and counting using a circular microphone array", *IEEE Transactions on Audio, Speech, and Language Processing*, **21**, pp. 2193-2206, (2013).
- [8] Allen J B, Berkley D A. "Image method for efficiently simulating small-room acoustics", *The Journal of the Acoustical Society of America*, **65**, pp. 943-950, (1979).
- [9] Lehmann E A, Johansson A M. "Diffuse reverberation model for efficient image-source simulation of room impulse responses", *IEEE Transactions on Audio, Speech, and Language Processing*, **18**, pp. 1429-1439, (2010).
- [10] Ellis D Gammatone-like spectrograms, web resource, (2009).
<http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram/>