Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

# Position estimation of binaural sound source in reverberant environments

Lama Ghamdan *, Mahmoud A. Ismail Shoman, Reda Abd Elwahab, Nivin Abo El-Hadid Ghamry

*Department of Information Technology, Faculty of Computers and Information, Cairo University, Egypt*

ARTICLE INFO

*Article history:*
Received 17 March 2016
Revised 27 April 2016
Accepted 27 May 2016
Available online 7 January 2017

*Keywords:*
Position estimation
Binaural cues
Distance and azimuth
Reverberant rooms

ABSTRACT

Most binaural sound source systems perform localization in either direction or distance perception. However, in real scenarios both perceptions are important to estimate source position in various environment conditions especially with the rapid technological growth in smart machines and their involvement in human daily life. This paper introduces an approach for azimuth and distance of binaural sound source localization in different reverberating environments using only two microphones. The algorithm is based on statistical features of the binaural cues and the difference of the binaural magnitude spectra of the binaural signal. Gaussian Mixture Models (GMMs) are used to jointly learn both distances and azimuths in different reverberant rooms. The proposed system does not require any prior knowledge of head related transfer function (HRTF), acoustical environment or room parameters. The performance has been evaluated at different aspects and conditions and reported effective and robust results, especially in the case of training set mismatch.

## 1. Introduction

Robots and smart machines have involved effectively and widely in human life in the last few years, which raises the demand for more natural communication inspired by the biological human vision and audition. Evident development has been accomplished in the field of vision perception, but audition perception using only two microphones placed in the artificial head is still a significant challenge and considered in its early stages [1]. Most of studies in the last decades used microphones array techniques like beamforming, which lead to high performance as the number of microphones increases, but it is computationally expensive. Binaural sound source localization has gained more focus in its different aspects (e.g. 2D and 3D localization, moving sources, and head movement) to add more realization to the localization task.

The technological growth employs binaural localization in other different and wide applications such as video conferences, smart rooms, virtual reality applications, auditory scene analyzers, hands-free communication, surveillance, and intelligent hearing aid devices; however, the performance of localization degrades in the real environmental conditions, in which human auditory system can robustly avoid. More researches were conducted to treat the conditions such as reverberant rooms, interfering noise, and interfering sources [2,3], rather than ideal conditions.

The human auditory system is capable of extracting the spatial location of objects in the spherical coordinates in terms of direction (azimuth, elevation) and distance. Researches focus on directional perception, mainly, azimuth estimation in various scenarios. Recently, elevation has gained more attention [4], but in distance perception the researches mostly addressed it with microphones arrays, while binaural audition was less considered. Since azimuth and distance are the most effective relevance to human listeners in position estimation [5], several studies were conducted investigating the relation and the influence of the cues of direction and distance on each other. They reported that the combination of azimuth and distance estimation maximizes localization accuracy [6,5,7]. However, the majority of the studies provide either of them as given information that improves the accuracy as test cases or to

* Corresponding author at: Department of Information Technology, Faculty of Computers and Information, Cairo University, Ahmed Zewail, Ad Doqi, Giza Governorate, Egypt.

*E-mail addresses:* l.shujaa@grad.fci-cu.edu.eg (L. Ghamdan), m.essmael@fci-cu.edu.eg (M.A. Ismail Shoman), r.abdelwahab@fci-cu.edu.eg (R.A. Elwahab), nivin@fci-cu.edu.eg (N.A. El-Hadid Ghamry).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2016.05.002
1110-8665/© 2017 Production and hosting by Elsevier B.V. on behalf of Faculty of Computers and Information, Cairo University.

study their influence. Despite the distance represents depth information and destination for mobile robots, most of 3D systems ignore distance. Few researches correlate azimuth and distance for position estimation based on microphone arrays for tracking mobile objects [8,9]. Hence, this work proposed in the direction of jointly estimation of azimuth and distance for position estimation based on a combination of statistical features of binaural signal.

To some extent, closed spaces represent the environment of most human activities and interactions; nevertheless, they suffer from reverberation caused by wave reflections of room surfaces which degrades the localization performance. Positional judgment should perform well regardless the acoustical environment conditions as real scenarios, whereas it is hard to cover all possible room conditions in the training stage. For distance perception, primary cues suggested for estimation including intensity, spectral cues, binaural cues, and Direct to Reverberant Ratio (DRR) which is the ratio of the energy of the direct and reverberant signal that is related to absolute distance estimation. Studies [6–9] considered Direct to Reverberant Ratio (DRR) in the reverberant environments is a significant cue of the received signal that perform better in reverberant environment and extracted from the binaural impulse response of the rooms which needs heavy computations, in addition, its estimation is difficult and inexact in practice. Thus several studies represented algorithms to blindly extract the DRR from the reverberant signal [5,10,9,11]. Cooke in [5] presented an equalization-cancellation technique, in this method the azimuth is estimated and utilized to extract energy arriving from reverberant signal, then the distance information is updated, but it needs the room reverberation time (T60) and works for distances above 2 m. In [10] an analytically relationship was derived between the DRR and the binaural magnitude squared coherence. A most recent method [11] estimated the DRR using a null-steering beamformer for two elements microphone array.

In [6] a position learning of the sound source method is introduced based on the magnitude and the phase difference of cross-spectra, however, it is stated that this method has difficulties in estimating sources sharing the same azimuth with different distances. Vesa improved this drawback in [12] using magnitude squared coherence as a feature for distance estimation, but his algorithm was limited in orientation angles grid and depends on receiver head rotation which does not have exactly the same effect as the source azimuth changes. Recently, Georganti [13] developed a novel feature for distance estimation depends on the standard deviation of the difference of the magnitude spectra of the binaural signal (BSMD STD) which does not need any prior knowledge of room acoustic properties such room impulse response, reverberation time and room volume. This novel feature showed high dependency on direction in the horizontal plane, especially in high reverberation rooms. Georganti also incorporated statistical properties of binaural cues for more robustness.

In azimuth estimation inspired by the human auditory system and based on only two microphones, the primary cues are Interaural Time Difference (ITD) that is the time difference of arrival of the sound signal between left and right ears and Interaural Level Difference (ILD) that is defined as the level of intensity difference between the two ears. These cues have been extensively studied to present localization systems; in recent years researches estimate azimuth based on joint ITD and ILD features as Raspaud in [14]. May [2] developed Gaussian mixture model depending on probabilistic model of ITD and ILD, and Youssef et al. used neural network approach to estimate the azimuth in a humanoid robotic context [15].

In this paper we propose a system that combines two models to predict the position of speech source in terms of direction in the horizontal plane and distance in reverberant rooms based on the statistical properties of the binaural cues and the standard deviation of the spectral magnitude difference of the binaural signal and the rest of the paper is organized as follows: the next section describes the model approach, the details of features extraction and selection process. In Section 3 the classification approach to estimate the source position is explained, and Section 4 demonstrates the simulation and the used database details. The experiment results and evaluation are found in Section 5. Finally the conclusion is in Section 6.

## 2. Model approach

### 2.1. Feature extraction

Toward achieving position estimation of binaural sound source in terms of direction and distance a combination of features has been extracted, which reflect both azimuth and distance information. In this section, the extraction of features is explained, the features selection approach is also demonstrated in detail. The complete process of the proposed system is described in Fig. 1.

#### 2.1.1. Binaural Spectral Magnitude Difference Standard Deviation (BSMD-STD):

The standard deviation of the spectral magnitude difference of the left and the right signal has shown high relation to the Head Related Transfer Function (HRTF) for definite frequency subbands rather than the complete bandwidth, consequently high dependency on distance and azimuth estimation depending on the selected band. The dependency between HRTF and spectral magnitude standard deviation is shown in Fig. 2.

Various tests were conducted, it was found that the range of 200–3000 Hz reflects high distance and azimuth information, and BSMD-STD extracted using hanning window for blocks of 1.2 s. In [13] the BSMD-STD was used for distance detection in reverberation closed rooms. Our approach tends to exploit the azimuth information to jointly estimate distance and azimuth in closed reverberant rooms. Fig. 3 shows BSMD-STD as a function of azimuth. BSMD-STD for specific frequency band is given by

$$\sigma_x^{ij} = \left[\frac{1}{n_j - n_i + 1}\sum_{k=n_i}^{n_j}[\Delta_x^{dB}(k) - \mu_x^{ij}]^2\right] \tag{1}$$

where

$$\mu_x^{ij} = \frac{1}{n_j - n_i + 1}\sum_{k=n_i}^{n_j}\Delta_x^{dB}(k) \tag{2}$$

where $n_i$ and $n_j$ are the bounds of the frequency range, $k$ is the frequency bin and $\mu_x^{ij}$ is the mean of the spectral magnitude.

#### 2.1.2. Binaural cues

The primary cues for binaural perception of human auditory system to localize sound source are ITD (Interaural Time Difference) and ILD (Interaural Level Difference). Most of systems exploited these cues to identify the direction of the sound source, but for distance estimation they were not widely used although their significant performance and distance dependency, especially ILD [3]. The ITD and ILD were extracted for different frequency channels, then statistical measurements were computed for every frequency channel. The following paragraph. will explain the estimation techniques.

- The Auditory Model:
  The input binaural signal is decomposed into S = 32 frequency channels for each left and right ears using phase-compensated

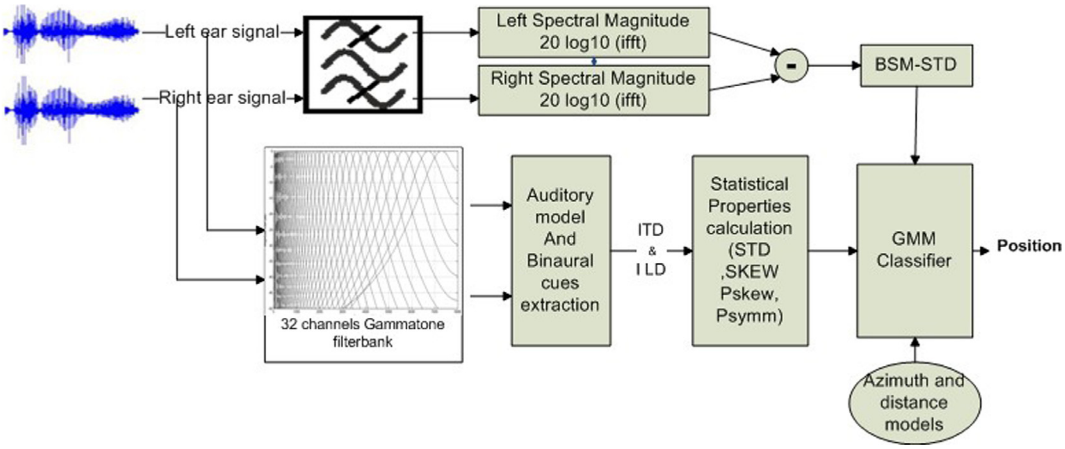

**Fig. 1.** Block diagram of the system.

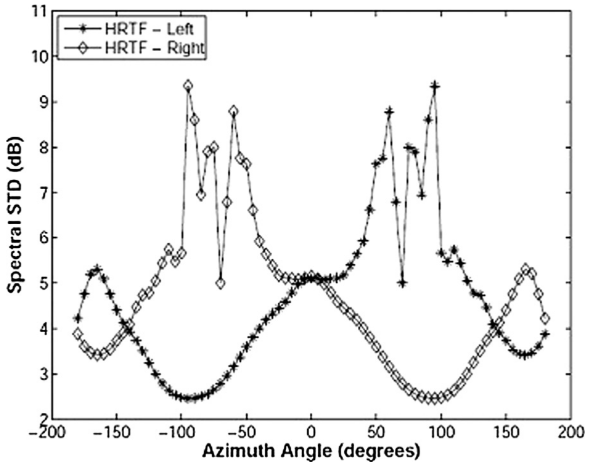

**Fig. 2.** The interdependency between azimuth and spectral magnitude, where the left and right anechoic Head Related Transfer Function estimated as a function of azimuth and spectral magnitude standard deviation, taken from [13].

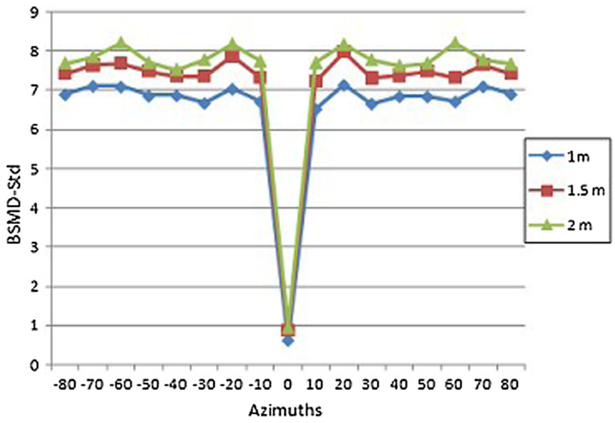

**Fig. 3.** BSMD-STD as a function of azimuth and distance in reverberant room = 0.2 s.

fourth-order gammatone filterbank to simulate the frequency selectivity of human cochlea, and to synchronize the binaural cues across frequency channels at a common time instance [16], with center frequencies from 80 Hz to 5000 Hz spaced on equivalent rectangular bandwidth (ERB) scale [17], this is followed by inner hair neural transduction approximated by half wave rectification, then square root compression. Each frequency channel signal divided into 20 ms frames with an overlap of 10 ms for each successive frame, these binaural cues where calculated for blocks of 1.2 s.

- Interaural Time Difference (ITD):

  ITD is defined as the difference in the time of arrival between the left and right signals that reaches the ears, it is calculated using normalized cross-correlation between left and right signals for every frequency channel $i$ as

$$c_i(k,\tau) = \frac{\sum_{n=1}^{N-1}\left(l_i(k\frac{N}{2}-n)-\bar{l}_i\right)\left(r_i(k\frac{N}{2}-n)-\bar{r}_i\right)}{\sqrt{\sum_{n=1}^{N-1}\left(l_i(k\frac{N}{2}-n)-\bar{l}_i\right)^2}\sqrt{\sum_{n=1}^{N-1}\left(r_i(k\frac{N}{2}-n)-\bar{r}_i\right)^2}} \qquad (3)$$

where $k, N$ are the frame number and the frame length respectively. $\bar{l}_i$ and $\bar{r}_i$ are the means value of the left and the right signals. The time lags of the cross-correlation function calculated within the range of [−1, 1] ms, (−44, 44) in samples. Then ITD (in samples) is given by

$$\hat{\tau}_i(k) = \arg\,\mathrm{max}c_i(k,\tau) \qquad (4)$$

where $\hat{\tau}_i$ is the maximum time lag of the frame $k$ in channel $i$. Towards improve the accuracy of ITD, additional fractional part estimated through exponential interpolation around the estimated $\hat{\tau}_i$ as used in [18]. This step is done due to the probability that the real cross-correlation function maximum may lay between two successive samples since the sampling interval restricts the ITD resolution. The final ITD in seconds for a time frame is estimated as

$$\widehat{itd}_i(k) = \frac{1}{f_s}(\hat{\tau}_i(k) + \hat{\delta}_i(k)) \qquad (5)$$

where

$$\hat{\delta}_i(k) = \frac{log_{10}c_i(k,\hat{\tau}_i(k)+1) - log_{10}c_i(k,\hat{\tau}_i(k)-1)}{4log_{10}c_i(k,\hat{\tau}_i(k)) - 2log_{10}c_i(k,\hat{\tau}_i(k)-1) - 2log_{10}c_i(k,\hat{\tau}_i(k)+1)} \qquad (6)$$

- Interaural Level Difference (ILD):

  Interaural level difference estimated by taking the energy ratio of each frequency band of the signal arrives the left and the right ear:

$$\widehat{ild}_i(k) = 20log_{10}\left(\frac{\sum_{n=0}^{N-1} r_i(k\frac{N}{2}-n)^2}{\sum_{n=0}^{N-1} l_i(k\frac{N}{2}-n)^2}\right) \qquad (7)$$

### 2.1.3. Statistical properties of binaural cues

ITD and ILD have different and complex patterns across the different frequency bands. According to Duplex theory of Lord Rayleigh [19] ITD presents more accurate information at low frequencies while ILD has better localization information at high frequencies. This implies that ITD and ILD distributions contain complementary information about source position; however, factors such reverberation cause deviation and variation in their distribution, whereas reverberation causes temporal fluctuations in ITD and decreases the magnitude of ILD. After [20] measurements of the binaural cues histograms, the interdependency of azimuths and distances is obvious in the binaural cues distributions in different frequencies, wherefore, it is possible to measure the statistical properties of the binaural cues and quantify the variations across frequencies to capture the overall concurrent effect of distance and azimuth on ITD and ILD distributions. Figs. 4 and 5 show the standard deviation and percentile skewness as a function of azimuth and distance at $RT_{60} = 0.2$ s and center frequency channel 2071 Hz. A joint feature space improves position discrimination rather than limiting features to ITD or ILD. In this context, we employ statistical properties of the binaural cues distribution to capture the variations across frequency bands at different reverberation conditions and different angles and distances in the horizontal plane.

- Standard deviation
The standard deviation of the binaural cues for specific frequency channel is defined as

$$\sigma_i(ITD, ILD) = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(B_{(i,k)}(ITD, ILD) - \overline{B}_{(i,k)}(ITD, ILD))^2} \quad (8)$$

where $i, k, m$ are channel index, frame number and total number of frames respectively. $\overline{B}_{(i,k)}$ is the mean of $(B_{(i,k)}$ in gammatone channel over $m$frames.

- Skewness
The skewness is given by

$$\gamma_i(ITD, ILD) = \frac{\frac{1}{m}\sum_{k=1}^{m}((B_{(i,k)}(ITD, ILD) - (\overline{B}_{(i,k)}(ITD, ILD))^3}{\sqrt{\frac{1}{m}\sum_{k=1}^{m}((B_{(i,k)}(ITD, ILD) - (\overline{B}_{(i,k)}(ITD, ILD))^2}^3} \quad (9)$$

Percentile statistical properties, proposed by Ludvigsen [21], are introduced and calculated for the histograms of the estimated binaural cues for every gammatone channel. Instead of comparing the ITD and ILD histograms, our approach is to capture their statistical properties which describe their behavior.

- Percentile Skewness $P_{skew}$
The difference is between the median and the $50_{th}$ percentile. In case of symmetrical distribution, the $P_{skew}$ is zero, and for asymmetrical distribution the difference would be high.

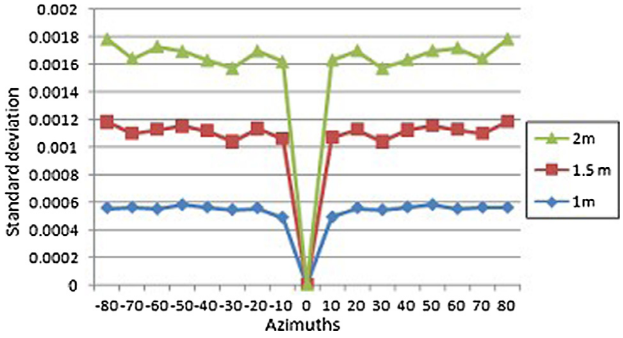

**Fig. 4.** Standard deviation of speech signal as a function of azimuth and distance in reverberant room = 0.2 s at center frequency channel = 2071 Hz.

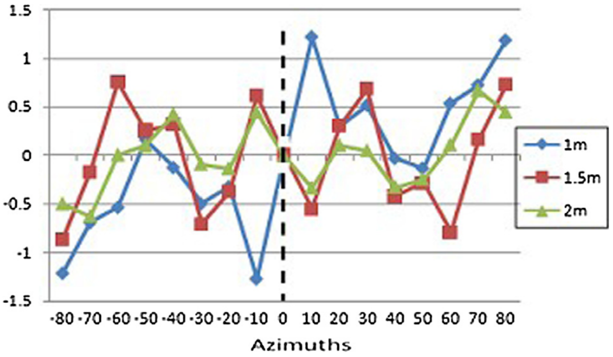

**Fig. 5.** Percentile of symmetry as a function of azimuth and distance in reverberant room = 0.2 s and center frequency channel = 2071 Hz, the difference values of positive and negative azimuths is obvious, distances confusion is decreased with other features.

$$P_{\mathrm{skew}i} = P_i^{50} - \frac{P_i^{90} + P_i^{10}}{2} \quad (10)$$

- Percentile Symmetry $P_{sym}$
To describe the aligning of the distribution, the percentile differences at both sided are calculated, where, positive value indicates left – sided distribution and negative value indicates right sided distribution. $P_{sym}$ is calculated as

$$P_{\mathrm{sym}i} = (P_i^{90} - P_i^{50}) - (P_i^{50} - P_i^{10}) \quad (11)$$

### 2.2. Features selection

The features vector combination determined based on several experiments that investigated the location information they contain. BSMD-STD has been used as distance indicator, it also can discriminate azimuth. The idea is to find out how much azimuth information it may contain, frame size was found that mostly effect azimuth discrimination, frame size of 1.2 s is used which describe strongly both distance and azimuth, however, 1.8 s found the best, but it would consider a long frame size.

The binaural cues have always been used to estimate azimuth in horizontal plane, also they significantly improve distance estimation especially ILD [3,7], therefore, the statistical properties of the binaural cues have been used and investigated. Many properties were studied like average deviation, Kurtosis, percentile kurtosis, percentile width, lower half percentile, to define their dependency and how much these properties can describe joint information about distance and azimuth. Feature selection algorithm, the minimal-redundancy maximal-Relevance (mRMR),was used to find the most dependent and relevant features. The results consisted with the finds that standard deviation, skewness, percentile skewness and percentile symmetry were stronger to discriminate combined classes of distances and directions. Due to the nature of ITD and ILD and their distribution across frequency channels we found that standard deviation and skewness of ITD are more effective while percentile skewness and percentile symmetry of ILD are more descriptive, therefore, the final feature space vector presented to the GMM would be:

$$\upsilon_i(\theta, d) = (BSMD - STD, \sigma_i(\mathrm{ITD}), \gamma_i(\mathrm{ITD}), P_{\mathrm{skew}_i}(\mathrm{ILD}), P_{\mathrm{skew}_i}(\mathrm{ILD})) \quad (12)$$

We focused on exploiting the nature of these features and select the most powerful properties, then combining them in one feature space that is able to define distance and azimuth. These features were used in one classification context to improve classification performance and time without decreasing the localization accuracy and robustness in both distance and direction perception.

## 3. Classification approach

Gaussian Mixture Models are used to estimate the azimuths and distances of sound sources; it is statistical approach that depends on probability density modeling which fits the nature of the binaural features space extracted in the previous section. GMMs are used to train azimuth and distant dependent patterns and expected to be less sensitive in case of untrained room, source and receiver conditions. K-means algorithm is used to initialize the GMM parameters for a specific sound source direction and distance within gammatone channel $i$:

$$\lambda = (\omega_i, \vec{\mu}_i, \Sigma_i) \text{ where } i = 1, \ldots, S. \tag{13}$$

$\omega_i$ represents the Gaussian component weight, $\vec{\mu}_i$ is the mean vector and $\Sigma_i$ is the covariance matrix. Diagonal matrix is used to describe the relation and the dependency between the features instead of full covariance matrix which is computationally more complex and expensive. The expectation maximization algorithm is used with maximum number of iterations 300 iterations to estimate the parameters. Since the features vary in scales, variance normalization is applied prior to classification process.

The true number of Gaussian components is hard to determine, and large number decreases the GMM ability of generalization to untrained data while selecting small number of Gaussian components leads to inappropriate learning of features characteristics; therefore, different algorithms appeared to allow automatic selection of the optimum number of components rather than manual or visual selection, and both approaches have been examined. Gaussian models of 5 components were found to be sufficient to localization performance.

## 4. Simulation and database

The binaural reverberant signals were simulated by generating binaural room impulse response (BRIR) using Roomsim package [22], which uses image method [23] to simulate room acoustics and integrate the Head Related Transfer Function (HRTFs) [24] measurements of KEMAR manikin dummy head in anechoic condition. The monoral source speech signals were selected from TIMIT corpus database [25], the signals upsampled from 16 kHz to 44.1 kHz and convolved with the generated reverberant BRIRs. Simulated room dimensions were 6 m × 4 m × 3 m, and the source was placed at azimuths with the range [−90, 90] in the horizontal plane with step of 10° and radial distances 1 m, 1.5 m and 2 m. The receiver was placed at 1.5 m × 2 m and 1.5 m above the ground as shown in Fig. 6. Reverberation time ($RT_{60}$) is defined as the time the sound signal needs to decay by 60 dB at a particular frequency after the sound signal stopped emitting; it is frequency dependent and directly affects the speech intelligibility, localization and quality in closed spaces. The longer the reverberation time, the harder the conditions of room acoustics, thus the localization performance. Different reverberation time means were simulated $RT_{60}$ = 0.2 s, 0.3 s, 0.4 s, 0.5 s, 0.6 s, 0.7 s, 0.8 s, 0.9 s for different training and testing conditions. Other acoustic parameters of the simulated rooms such surface absorption coefficients, humidity, temperature and distance attenuation were taken in consideration and adjusted to simulate real rooms.

In experiments different room, position, distances and azimuths were chosen to evaluate the performance and the accuracy of the proposed system.

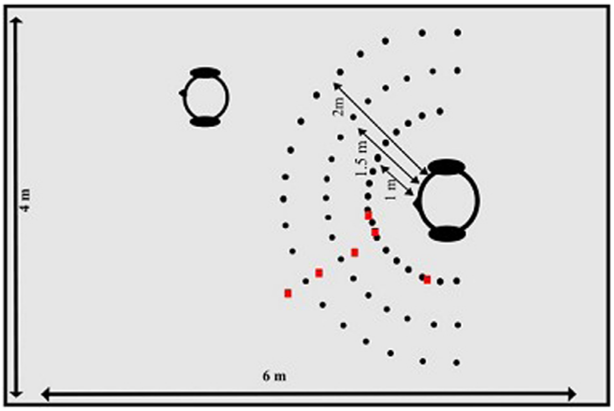

**Fig. 6.** Simulated room with all sources and receiver positions.

## 5. Experiments and results

In this section, we discuss the performance of the statistical properties of the binaural cues and the BSMD-STD features of the signals, and we evaluate the system ability and performance to estimate the position using the database described in the previous section in different reverberant conditions, positions, azimuths and distances. The performance measure is the mean classification performance which is calculated by taking the mean of the diagonal of the confusion matrix.

### 5.1. Experiment 1: reverberation time ($RT_{60}$)

#### 5.1.1. Identical conditions

In this experiment, the system performance of testing data set was evaluated in known parameters and conditions of room size, reverberation time, receiver position and source positions, which provided in training stage as in Fig. 6. The system trained for $RT_{60} = 0.2$ s, 0.5 s, 0.8 s individually, and for all reverberation times while the testing conducted separately. The performance rate is shown in Fig. 7. It is seen that the correct estimated position performance degrades when $RT_{60}$ differs between the training and testing set, especially for low reverberation time ($RT_{60}$ = 0.2 s); however, when we applied generalization with multi $RT_{60}$'s in training, the result improved significantly. The performance rate improved compared to results in [13] with the same number of features and similar room conditions with GMM classifier in known azimuth (training and testing performed separately on specific azimuth), where the performance was less than 90%, and it increased as the number of features employed in the classification increased. Also, when the training and testing is applied to a mixture of azimuths GMM performance degrades to less than 85%, but it always above 75%.

#### 5.1.2. Different conditions

Here, the system generalization ability verified in different $RT_{60}$'s that have not been trained for at all. The training was performed with extracted data at $RT_{60}$'s of 0.2 s, 0.5 s and 0.8 s, while the testing was performed for speech signals at $RT_{60}$'s of 0.3 s, 0.6 s, 0.9 s. The result is shown in Fig. 8. It is seen that small $RT_{60}$ = 0.3 is harder to generalize as previous in case of mismatch of known trained $RT_{60}$, and higher $RT_{60} = 0.6$ s has higher performance rate while it decreases at $RT_{60} = 0.9$ s. This result is accepted since it is unknown acoustic conditions where the classifier has not trained for, especially with GMM classifier which is supervised learning algorithm.

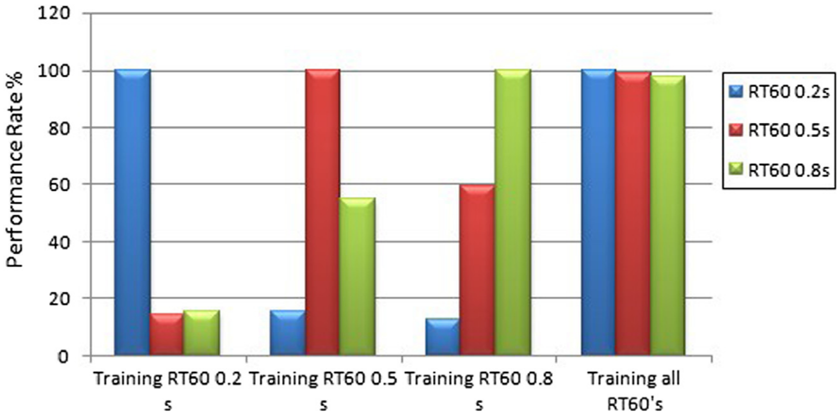

**Fig. 7.** Performance rate of different azimuths and distances, The training is on one $RT_{60}$ and all $RT_{60}$'s, testing on each separated $RT_{60}$.

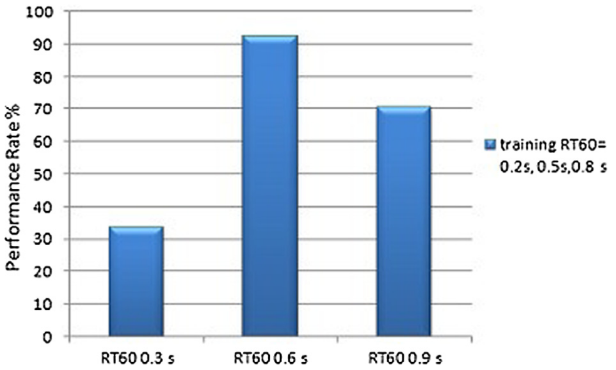

**Fig. 8.** Performance rate of different azimuths and distances, The training is on all $RT_{60}$'s, testing on different $RT_{60}$'s which are not provided in training.

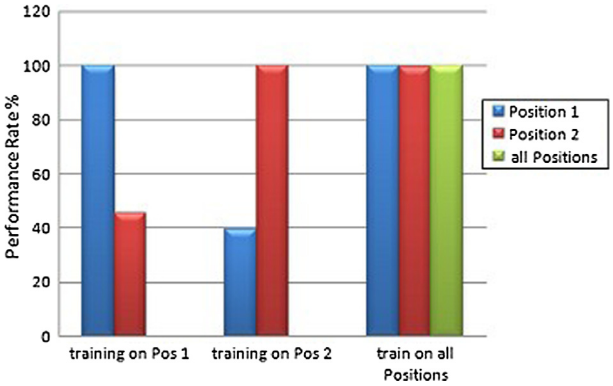

**Fig. 9.** Performance rate of different azimuths and distances at different receiver positions, The training is on all positions and each position separately, testing on each different position.

### 5.2. Experiment 2: receiver positions

The receiver position changing inside the room is important issue to take into consideration and test how the system will perform in such case. Thus, in this experiment the receiver placed in two different positions, position 1 (original receiver position) and position 2, where the receiver placed at 4 m × 1 m × 1.5 m. Position 2 was chosen to be closer to the wall to examine close wall reflections effects. The distance at position 2 is limited to 1 m for some source azimuths because of the room dimensions. The training was performed for each position separately, then on all positions with a reverberation time of 0.2 s. The results can be seen in Fig. 9, in case of positions mismatching in training and testing sets, the performance decreased almost to the half, this can be due to acoustic constrains of reflective wall closeness in Position 2. It is notable that the correct estimation rate increased in case of generalization training sets (mixed sources positions training) which results high and stable estimation.

### 5.3. Experiment 3: source positions

In this section, the system is tested and evaluated in case of unknown azimuths and distances which have not been trained. This is important to examine since it generalizes the system ability especially if the sound source is moving. The triangles in Fig. 6 exhibited azimuths and distances used with reverberation time of 0.2 s. the mean error of estimated azimuth with distance of 1 m is almost 3.5° compared to results in [15] that reported approximately 2° in anechoic room and 5° for $RT_{60}$ = 0.7 s but with identical training and testing set conditions. Our result is acceptable due to the step of azimuth which is 10°, and also the result is expected to decrease in case of decreasing the step in trained azimuths. The distance mean error is approximately 0.2 m.

## 6. Conclusion

This paper presented a system that robustly estimates the position of a sound source in both distance and direction perception in reverberant environments based on a BSMD-STD and set of statistical properties of binaural cues. A combined feature vector is provided as input to Gaussian Mixture Models (GMMs) for classification, then the corresponding position of the sound source is estimated. Various reverberation conditions and positions were tested and evaluated. The system provided robust and high accuracy performance results in different scenarios and the ability to adjust untrained positions.

## References

[1] Keyrouz F. Advanced binaural sound localization in 3-d for humanoid robots. IEEE Trans Instrum Meas 2014;63(9):2098–107. doi: http://dx.doi.org/10.1109/TIM.2014.2308051.

[2] May T, van de Par S, Kohlrausch A. A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Trans Audio Speech Lang Process 2011;19(1):1–13. doi: http://dx.doi.org/10.1109/TASL.2010.2042128.

[3] Woodruff J, Wang D. Binaural localization of multiple sources in reverberant and noisy environments. IEEE Trans Audio Speech Lang Process 2012;20(5):1503–12. doi: http://dx.doi.org/10.1109/TASL.2012.2183869.

[4] Liu H, Zhang J, Fu Z. A new hierarchical binaural sound source localization method based on interaural matching filter. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). p. 1598–605. doi: http://dx.doi.org/10.1109/ICRA.2014.6907065.

[5] Lu Y-C, Cooke M. Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources. IEEE Trans Audio Speech Lang Process 2010;18(7):1793–805. doi: http://dx.doi.org/10.1109/TASL.2010.2050687.

[6] Smaragdis P, Boufounos P. Position and trajectory learning for microphone arrays. IEEE Trans Audio Speech Lang Process 2007;15(1):358–68. doi: http://dx.doi.org/10.1109/TASL.2006.876758.

[7] Rodemann T. A study on distance estimation in binaural sound localization. In: IEEE/RSJ international conference on Intelligent Robots and Systems (IROS). p. 425–30. doi: http://dx.doi.org/10.1109/IROS.2010.5651455.

[8] Gontmacher J, Yarhi A, Havkin P, Michri D, Fisher E. Dsp-based audio processing for controlling a mobile robot using a spherical microphone array. In: 2012 IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI). p. 1–5. doi: http://dx.doi.org/10.1109/EEEI.2012.6377070.

[9] Olivier C, Fouillade J, Felon A, Cole J, Clinton N, Sanchez R, et al. The tracking of a moving object by a mobile robot following the objects sound. J Intell Robot Syst 2013;71(1):31–42.

[10] Kuster M. Estimating the direct-to-reverberant energy ratio from the coherence between coincident pressure and particle velocity. J Acoust Soc Am 2011;130(6):3781–7.

[11] Eaton J, Moore A, Naylor P, Skoglund J. Direct-to-reverberant ratio estimation using a null-steered beamformer. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 46–50. doi: http://dx.doi.org/10.1109/ICASSP.2015.7177929.

[12] Vesa S. Binaural sound source distance learning in rooms. Trans Audio Speech Lang Proc 2009;17(8):1498–507. doi: http://dx.doi.org/10.1109/TASL.2009.2022001.

[13] Georganti E, May T, van de Par S, Mourjopoulos J. Extracting sound-source-distance information from binaural signals. In: Blauert J, editor. The technology of binaural listening, modern acoustics and signal processing. Berlin, Heidelberg: Springer; 2013. p. 171–99.

[14] Raspaud M, Viste H, Evangelista G. Binaural source localization by joint estimation of ild and itd. IEEE Trans Audio Speech Lang Process 2010;18(1):68–77. doi: http://dx.doi.org/10.1109/TASL.2009.2023644.

[15] Youssef K, Argentieri S, Zarader J-L. A learning-based approach to robust binaural sound localization. In: 2013 IEEE/RSJ international conference on Intelligent Robots and Systems (IROS). p. 2927–32. doi: http://dx.doi.org/10.1109/IROS.2013.6696771.

[16] Brown GJ, Cooke M. Computational auditory scene analysis. Comput Speech Lang 1994;8(4):297–336. doi: http://dx.doi.org/10.1006/csla.1994.1016.

[17] Roman N, Wang D, Brown GJ. Speech segregation based on sound localization. International Joint Conference on Neural Networks, 2001. Proceedings. IJCNN '01., vol. 4. p. 2861–6.

[18] Zhang D, Wu X. On cross correlation based-discrete time delay estimation. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05), vol. 4. springer; 2005. p. iv/981–4.

[19] Rayleigh L. On our perception of sound direction. Phil Mag 1907;13(74):214–32.

[20] Qu T, Xiao Z, Gong M, Huang Y, Li X, Wu X. Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap. IEEE Trans Audio Speech Lang Process 2009;17(6):1124–32. doi: http://dx.doi.org/10.1109/TASL.2009.2020532.

[21] Ludvigsen C. Schaltungsanordnung für die automatische regelung von hörhilfsgeräten [an algorithm for an automatic program selection mode].

[22] D Campbell KP, Brown G. A matlab simulation of shoebox room acoustics for use in research and teaching. Comput Inform Syst 2005;9(3):48–51.

[23] Allen JB, Berkley DA. Image method for efficiently simulating small-room acoustics. J Acoust Soc Am 1979;65(4):943–50.

[24] Gardner WG, Martin KD. Hrtf measurements of a kemar. J Acoust Soc Am 1995;97(6):3907–8.

[25] Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL, etal. DARPA TIMIT acoustic phonetic continuous speech corpus; 1993.