Egyptian Informatics Journal xxx (2017) xxx–xxx

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 xxxx

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

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

Production and hosting by Elsevier

* 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).

http://dx.doi.org/10.1016/j.eij.2016.05.002
1110-8665/

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

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

where

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

where n_i and n_j are the bounds of the frequency range, k is the frequency bin and μ_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

Please cite this article in press as: Ghamdan L et al. Position estimation of binaural sound source in reverberant environments. Egyptian Informatics J (2017), http://dx.doi.org/10.1016/j.eij.2016.05.002