Applied Acoustics 138 (2018) 28–38

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Multiple-to-single sound source localization by applying single-source bins detection

Maoshen Jia[a,*], Jundai Sun[a], Changchun Bao[a], Christian Ritz[b]

[a] Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[b] ICT Research Institute and School of Electrical Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2500, Australia

ARTICLE INFO

Keywords:
Source localization
Sparsity
B-format microphone

ABSTRACT

This paper proposes a novel localization scheme for multiple sound sources that imposes the relaxed sparsity constrains (not all time-frequency coefficients are overlapped) on the source signals. First, a "DOA convergence" assumption is proposed, which means that if most of the time-frequency (T-F) bins in a T-F zone are derived from only one source – defined as single source bins (SSBs), the corresponding direction of arrival (DOA) estimates are relatively concentrated with a heavy density. This assumption is validated through statistical analysis by applying a quantitative measure of convergence. Accordingly, by applying the "DOA convergence" assumption, the detection of SSBs is converted to a clustering problem, K-means clustering and density-based spatial clustering of applications with noise (DBSCAN) algorithms are utilized to complete the task in this paper. The cross distortions (localization error due to the cocktail party phenomenon) in localization caused by multiple simultaneously occurring sources is significantly weakened by conducting DOA estimation among these SSBs, i.e., the multiple source localization is rewritten to a single source one among these SSBs. Moreover, the proposed SSBs detection is applicable to other localization methods and not limited to specific microphone topology. Experimental results demonstrate the localization accuracy of the proposed method outperforms the state-of-the- art localization approaches which are based on single source zone detection. However the proposed method is capable of real-time processing, the accuracy is insufficient in the current system. If non-real time processing is allowed, our method can be realized with higher accuracy than the conventional ones.

1. Introduction

Spatial audio gives the listener a sensation that sound comes from a particular direction in 3D space and helps to create immersive virtual soundscapes. Recently, virtual/augmented reality (V/AR) has attracted renewed interest, it aims to improve the reality of sound scene as well.

For non-real time applications, each source objects in the sound scene can be recorded through an exclusive microphone. The immersive listening of original sound scene can be reproduced by a convolution processing between all the mono signals of each source objects and their head-related transfer functions (HRTF) filters, respectively. However, it does not work in real-time applications, where there are only recording mixtures can be provided for processing. Considering this situation, a more accurate localization approach is needed to obtain the spatial parameter of sound scene at low-delay procedure. Specifically, source localization aims to estimate the direction of arrival (DOA) s of unknown source signals from the recording mixtures. It plays an essential role in various audio applications, such as: automated camera steering and teleconferencing systems [1], speaker separation [2] and robot audition [3]. Moreover, the information obtained by sound source localization could be widely used for scalable audio coding and reconstruction of the sound scene.

Broadly, the existing localization methods can be divided into four main categories. The first one is time difference of arrival (TDOA) estimation based methods and its extensions [4–6]. However, most of the localization methods based on TDOA require excessive microphones to improve the reliability [7]. The second one is high-resolution spectral estimation techniques, such as multiple signal classification (MUSIC) [8–10] and estimation of signal parameters via rotational invariance (ESPRIT) [11] algorithms, that are based on the spectral analysis of the correlation matrix of the measured signals. Nevertheless, there are two issues regarding the MUSIC algorithm, one is the computational cost, while the other is the requirement of previous knowledge about the number of actual sources [10]. The third category of localization methods are based on maximization of the steered response power (SRP) of a beamformer output, the maximum likelihood (ML) criterion is applied, which in the case of a single source, culminates in inspecting the output power of a beamformer steered to different locations and in

* Corresponding author.
E-mail address: jiamaoshen@bjut.edu.cn (M. Jia).

https://doi.org/10.1016/j.apacoust.2018.03.015
Received 6 November 2017; Received in revised form 5 February 2018; Accepted 17 March 2018
Available online 30 March 2018
0003-682X/

searching the points where it receives its maximum value [12]. The last one, derive as a solution to the blind source separation (BSS) problem, Independent Components Analysis (ICA) [13] and Sparse Components Analysis (SCA) [14,15] are applied to localization for multiple sound sources, but the w-disjoint orthogonally (W-DO) property they are relying on is less accuracy when there are more sound sources or with reverberation/noise.

Recently, a "single source" zone (SSZ) detecting based DOA estimation method is proposed which works by conducting DOA estimation in single-source zones where there is only one source absolutely dominant than others. It has been proved to have good performance by using circular microphone, B-format microphone and so on [14,16]. However, the T-F bin belongs to the detected SSZ is not all strictly come from one source, there are a few overlapped T-F bins which will badly affect the accuracy of localization especially for the underdetermined case.

In this paper, a novel localization scheme for multiple sources is presented by converting the multi-source problem to a single-source one. Specifically, based on the relaxed sparsity constrains on the source signals [16], a "DOA convergence" assumption is definitely proposed. By applying this assumption, the scheme can detect all the SSBs by using a clustering algorithm and then the multiple source localization can be achieved by only retaining the DOA estimates among SSBs. In detail, the main contribution of this work includes the following respects:

1. Based on the statistical results of the existence of SSBs among different number of sources, a conclusion is drawn that there are still SSBs even when the number of sources exceeds the number of microphones.
2. A "DOA convergence" assumption is proposed that the DOA estimates among the SSBs in a T-F zone will converge to a central direction which corresponds to a true source with a high probability. Conversely, the DOA estimates among the non-SSBs would be badly influenced due to the overlapped T-F coefficients of multiple sources and uniformly distributed over all directions. Besides, this assumption is definitely validated via a statistical experiment.
3. In this paper, k-means clustering and DBSCAN algorithms are adopted as two alternative clustering algorithms for comparison. The implementation principles of these two methods (k-means clustering and DBSCAN algorithm based SSBs detection) are also illustrated in this paper. Moreover, a series of aspects including localization and source counting accuracy, low-delay are considered in the evaluation. Experimental results demonstrate the localization accuracy of the proposed method outperforms existing SSZ based localization approaches.

The remainder of the paper is organized as follows: Section 2 investigates the proportion of SSBs among multiple sources and the definition and verification of "DOA convergence" assumption. Section 3 presents the proposed method. Experimental results are presented in Section 4, while conclusions are drawn in Section 5.

2. Basis of the proposed method

2.1. Exploring the proportion of SSBs among multiple sources

From the investigation of [16], it can be concluded that the sparsity/W-DO assumption was found to be satisfied more often for multiple source signals recorded by pairs of directional microphones in a B-format array compared with analyzing the raw signals. That is, multiple source signals satisfy the relaxed sparsity that there are always some SSZs over the recording mixture signals where one source is absolutely dominant over the others. Similar to the principle of calculus, when the width of the T-F zone gradually reduces to one, the zone is automatically converted to a T-F bin. If the T-F bins in the mixture signals only derives from one source, it is defined as SSBs in this paper. The other T-F bins that derive from more than one sources are defined as non-SSBs. Aiming to exploring the proportion of SSBs versus different source numbers, the difference of distribution characteristics on DOA estimation between SSBs and non-SSBs is adopted to define the measurement in this subsection.

Specifically, suppose there are Q sources simultaneously occurring, mixture signals are recorded via a microphone array (linear, circular or others). The corresponding DOA estimates at each T-F bin can be calculated by different localization methods which are based on the choice of microphone array [14,16]. In order to investigate the proportion of SSBs among different source number, the difference of distribution characteristics on DOA estimation between SSBs and non-SSBs is considered, i.e., the DOA estimates at SSBs will correspond to a true source and locate in a range $[\mu_i-\Delta\mu,\mu_i+\Delta\mu]$ with a very high probability, where μ_i denotes the DOA of the source i and $\Delta\mu$ is a range threshold to get a better tolerance; while for non-SSBs, the DOA estimates will hardly correspond to a true source and located outside the above range [17]. Therefore, for any one of the mixture signal $X(n,k)$ (the TF representation of speech signal x, where n and k are the time and frequency index, respectively), a sparse signal of it denoted by $X_i'(n,k)$ for source i, can be obtained by extracting the TF components whose DOA estimates locate in $[\mu_i-\Delta\mu,\mu_i+\Delta\mu]$, i.e.,

$$X_i'(n,k)=\begin{cases}X(n,k), & \text{if } \hat{\mu}(n,k)\in[\mu_i-\Delta\mu,\mu_i+\Delta\mu]\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

where $i=1,2,\ldots Q, k=1,2,\ldots L, L$ is the number of Short Time Fourier Transform (STFT) points, and $\hat{\mu}(n,k)$ denotes the DOA estimate at TF instant (n,k). All the preserved TF points of $X(n,k)$ by using (1) are SSBs. So the Frame SSB Proportion (FSP) is defined as the occupied ratio of the SSBs among the total TF bins, i.e.,

$$FSP=\frac{\|\sum_{i=1}^{Q}\boldsymbol{X}_i'\|_0}{\|\boldsymbol{X}\|_0} \tag{2}$$

where the bold-type letter $\boldsymbol{X}=[X(n,1),X(n,2),\ldots,X(n,L)]$,$\boldsymbol{X}_i'=[X_i'(n,1),X_i'(n,2),\ldots,X_i'(n,L)]$, and $\|\cdot\|_0$ counts the number of non-zero components in its argument. It can be concluded from the definition of FSP that a higher value will be obtained when there are more SSBs in the mixture signal.

It should be noted that SSBs detected by this methods is not the best one, if two sources are active in the same time frequency bin, but the estimated DOA is within the threshold range, the overlapped bins will be counted as belonging to only 1 source, when actually it is not a single source bin. The other way to measure the truth number of non-SSBs may like the method of SSZ detection [16]. For the original sources, that might mean just checking for each bin and each source if the amplitude is above a given threshold indicating a source is active. This gives a binary mask time-frequency mask for each source (of 1 or 0 for each bin). Adding the masks together for each source will then tell us if we have 1, 2, 3, …, M sources active in each bin. The latter one may be more accurate, but it does not correspond to the use in practical localization procedure, because we cannot know the source signals priorly. Thus, though the former measurement may include a few non-SSBs, it can directly indicate the proportion of SSBs when using the mixture signals for detection.

By using the quantitative measure defined in (1) and (2), a statistical analysis is taken to investigate the ratio of SSBs. A total of 40 sentences (the sampling frequency is 16 kHz, both male and female speakers are included) from the NTT speech database were used for testing. The B-format microphone is used for recording due to its simplicity of DOA estimation ([16,18]) and the testing anechoic environment is simulated by Roomsim toolbox [19]. The source number ranges from 1 to 6, the angle between adjacent sources keeps 60°, four values, i.e., 6°,8°,10°,12°, of $\Delta\mu$ are chosen for statistic.

The results of average FSP are shown in Fig. 1, it can be seen that