Signal Processing 143 (2018) 42–47

Contents lists available at ScienceDirect

Signal Processing

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

Short communication

Speaker localization using direct path dominance test based on sound field directivity☆

Boaz Rafaely*, Koby Alhaiany

Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel

ARTICLE INFO

Article history:
Received 7 March 2017
Revised 8 August 2017
Accepted 12 August 2017
Available online 23 August 2017

Keywords:
Speaker localization
Reverberation
Spherical microphone arrays
Directivity

ABSTRACT

Estimation of the direction-of-arrival (DoA) of a speaker in a room is important in many audio signal processing applications. Environments with reverberation that masks the DoA information are particularly challenging. Recently, a DoA estimation method that is robust to reverberation has been developed. This method identifies time-frequency bins dominated by the contribution from the direct path, which carries the correct DoA information. However, its implementation is computationally demanding as it requires frequency smoothing to overcome the effect of coherent early reflections and matrix decomposition to apply the direct-path dominance (DPD) test. In this work, a novel computationally-efficient alternative to the DPD test is proposed, based on the directivity measure for sensor arrays, which requires neither frequency smoothing nor matrix decomposition, and which has been reformulated for sound field directivity with spherical microphone arrays. The paper presents the proposed method and a comparison to previous methods under a range of reverberation and noise conditions. Result demonstrate that the proposed method shows comparable performance to the original method in terms of robustness to reverberation and noise, and is about four times more computationally efficient for the given experiment.

1. Introduction

The estimation of the direction-of-arrival (DoA) of speakers in reverberant environments using microphone arrays is important in a wide range of applications, including speech enhancement, source separation, robot audition and video conferencing. Methods for DoA estimation have been previously studied extensively. These include beamforming [1], subspace methods such as multiple signal classification (MUSIC) [2], and time-delay of arrival estimation methods [3]. DoA estimation methods specifically developed for speech signals exploit the non-stationarity and sparsity of speech in the short-time Fourier transform (STFT) domain and enable DoA estimation even for under-determined systems with more sources than microphones [4–6]. However, in reverberant environments, room reflections mask the direct sound that carries DoA information, thus degrading the DoA estimation performance.

Recently, a method for DoA estimation of multiple speakers that is robust to reverberation has been developed [7]. This method processes the signals in the time-frequency domain, and employs the direct-path dominance (DPD) test to identify time-frequency bins dominated by the direct path. Unlike the earlier coherence test [5], the DPD test applies frequency smoothing to significantly reduce the effect of coherent early room reflections. Furthermore, the method is designed for spherical microphone arrays [8], such that DoAs of sources in all directions can be estimated, and frequency smoothing is applied without focusing [7,9]. Experimental investigations validated the robustness of the method to reverberation and its advantages over previous methods [7]. The method has been further developed recently, with Gaussian mixture modeling (GMM) applied to the DoA samples to reduce the estimation bias due to the non-normal distribution of the data [10]. Although the proposed reverberation-robust method is useful, it is computationally costly, as it requires frequency-smoothing to compute the spatial spectrum matrix for each time-frequency bin, eigen-value decomposition of each of these matrices, and, finally, GMM for data clustering. However, some applications of audio signal processing that employ DoA estimation may be computationally restricted, e.g. robot audition, due to the limited computational resources and the strict real-time requirements.

This paper proposes a novel alternative to the DPD test which is more computationally efficient than current methods. The new test is based on the directivity measure for sensor arrays [11], reformulated for the sound field directivity as measured by the array. This measure has a maximum value for a single-source sound field, and is computed independently for each time-frequency bin.

☆ This work was supported by the Israel Science Foundation (ISF) under Grant 146/13.

* Corresponding author.
E-mail address: br@bgu.ac.il (B. Rafaely).

http://dx.doi.org/10.1016/j.sigpro.2017.08.010
0165-1684/

It is shown that the proposed method offers DoA estimation performance that is comparable to that of the recent version of the algorithm [10], while considerably reducing the computation requirements. The paper is structured as follows. An overview of previous methods is outlined in Section 2, after which the development of the new DPD test is presented in Sections. 3 and 4. A simulation study is detailed in Section 5, followed by the conclusions in Section 6.

2. DoA estimation based on DPD test and GMM clustering

The current method for speaker localization, recently presented in [7,10] and developed for spherical microphone arrays [8], is outlined in this section. The method is based on the DPD test to identify time-frequency bins that are dominated by the direct sound [7], and GMM for data clustering to reduce estimation bias [10]. Consider a spherical microphone array with Q microphones arranged on the surface of a sphere of radius r. The sound pressure at the microphones is denoted by $p(k, r, \theta_q, \phi_q)$, with k denoting the wave number and (θ_q, ϕ_q) the spherical coordinates of the angular position of microphone q. $\Omega_q \equiv (\theta_q, \phi_q)$ is employed to simplify notation. The sound pressure at the microphones due to L sound sources that surround the array is written in a matrix from as [11]

$$\mathbf{p}(k) = \mathbf{V}(k)\mathbf{s}(k) + \mathbf{n}(k), \tag{1}$$

with the $Q \times L$ steering matrix $\mathbf{V}(k)$ representing the frequency response from each source to each microphone, the $Q \times 1$ vector $\mathbf{p}(k)$ is given by $\mathbf{p}(k) = [p(k, r, \Omega_1), \ldots, p(k, r, \Omega_Q)]^T$, and the $Q \times 1$ vector $\mathbf{n}(k) = [n_1(k), \ldots, n_Q(k)]^T$, represents sensor noise at the microphones. The $L \times 1$ vector $\mathbf{s}(k)$ is given by $\mathbf{s}(k) = [s_1(k), \ldots, s_L(k)]^T$, and represents the signal at the L sources.

For spherical arrays, the steering matrix can be decomposed into frequency-dependent and direction-dependent components, and so Eq. (1) can be written as [7]

$$\mathbf{p}(k) = \mathbf{Y}(\mathbf{\Omega})\mathbf{B}(k)\mathbf{Y}^H(\mathbf{\Psi})\mathbf{s}(k) + \mathbf{n}(k). \tag{2}$$

The $Q \times (N+1)^2$ matrix $\mathbf{Y}(\mathbf{\Omega})$ holds the spherical harmonics functions $Y_n^m(\Omega_q)$ of order n and degree m, for all $0 \le n \le N$ and $-n \le m \le n$. Spherical array design considerations typically lead to an array order N satisfying $(N+1)^2 \le Q$ and a frequency range of operation satisfying $kr < N$ (see [8] for further details). All orders and degrees in $\mathbf{Y}(\mathbf{\Omega})$ are arranged in a single dimension, leading to a running column index of $n^2 + n + m + 1$. The $(N+1)^2 \times (N+1)^2$ diagonal matrix $\mathbf{B}(k)$ holds radial functions that represent the scattering of a plane wave from a rigid sphere, in the case of a rigid-sphere array, or the elements of the phase response in the case of an open-sphere array [8], and $\mathbf{Y}(\mathbf{\Psi})$ is similar to $\mathbf{Y}(\mathbf{\Omega})$, but of size $L \times (N+1)^2$, with $\mathbf{\Psi}$ representing source arrival directions, and $(\cdot)^H$ representing the Hermitian transpose.

Multiplying Eq. (2) from the left by the pseudo-inverse $[\mathbf{Y}(\mathbf{\Omega})]^\dagger$, and the inverse $\mathbf{B}^{-1}(k)$ (regularization may be required at low frequencies or near the nulls of the radial functions [8]), leads to plane wave decomposition of the sound field [8],

$$\mathbf{a}(k) = \mathbf{Y}^H(\mathbf{\Psi})\mathbf{s}(k) + \tilde{\mathbf{n}}(k), \tag{3}$$

where vector $\mathbf{a}(k)$ of size $(N+1)^2 \times 1$ is given by $\mathbf{a}(k) = [a_{00}(k), \ldots, a_{NN}(k)]^T$. Terms $a_{nm}(k)$ represent the frequency-dependent spherical harmonics coefficients of the plane wave density function measured by the array. Vector $\tilde{\mathbf{n}}(k)$ is the modified sensor noise, given by $\tilde{\mathbf{n}}(k) = \mathbf{B}^{-1}(k)[\mathbf{Y}(\mathbf{\Omega})]^\dagger \mathbf{n}(k)$.

In the next step, STFT is applied to the plane-wave decomposition signal, leading to

$$\mathbf{a}(\tau, \nu) = \mathbf{Y}^H(\mathbf{\Psi})\mathbf{s}(\tau, \nu) + \tilde{\mathbf{n}}(\tau, \nu), \tag{4}$$

with τ denoting the time index and ν denoting the frequency index. Now, at each time-frequency bin a spatial spectrum matrix is computed, denoted by $\mathbf{R}(\tau, \nu)$, which averages the spatial information over T time bins and F frequency bins in the neighborhood of the selected bin,

$$\mathbf{R}(\tau, \nu) = \frac{1}{TF} \sum_{\tau'=\tau}^{\tau+T-1} \sum_{\nu'=\nu}^{\nu+F-1} \mathbf{a}(\tau', \nu')\mathbf{a}^H(\tau', \nu'). \tag{5}$$

Time averaging is employed to approximate the expectation operation in the computation of the spatial spectrum matrix [11]. Frequency averaging, also referred to as frequency smoothing, is applied so that the direct sound can be distinguished from coherent reflections. This is important because without frequency smoothing early reflections may bias the DoA estimation, and performance may be significantly degraded under reverberation [7].

The eigen-value decomposition of $\mathbf{R}$ is computed next. If $\mathbf{R}$ is dominated by a single source it is of unit rank, and so the ratio between the first and second singular values is infinite. This motivated the introduction of the DPD test to identify time-frequency bins dominated by a single source, i.e. the direct sound from a speaker [7],

$$\mathcal{D} = \left\{ (\tau, \nu) : \frac{\sigma_1(\mathbf{R}(\tau, \nu))}{\sigma_2(\mathbf{R}(\tau, \nu))} \ge \mathcal{TH} \right\}, \tag{6}$$

where σ_1, σ_2 denote the largest and second largest singular values, respectively. $\mathcal{TH}$ is a threshold value, chosen to be sufficiently larger than one to guarantee that $\mathbf{R}$ is dominated by a single singular vector.

The multiple signal classification (MUSIC) spectrum [2], $P(\Theta, \tau, \nu)$, is computed next for bins that pass the DPD test, i.e. for all $(\tau, \nu) \in \mathcal{D}$, to identify the DoA associated with these bins,

$$P(\Theta, \tau, \nu) = \frac{1}{\|\mathbf{U_n}^H(\tau, \nu)\mathbf{y}^*(\Theta)\|^2}, \ (\tau, \nu) \in \mathcal{D} \tag{7}$$

where $\Theta = (\theta, \phi)$ represents a direction under analysis, and $\mathbf{y}$, of size $(N+1)^2 \times 1$, holds the spherical harmonics functions $Y_n^m(\Theta)$ at element number $n^2 + n + m + 1$. With $\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ representing eigen-value decomposition, matrix $\mathbf{U_n}$ is composed of all columns of $\mathbf{U}$ except the first column, which is related to the largest singular value. It therefore represents the noise subspace of $\mathbf{R}$, assuming a single source, and is of dimensions $(N+1)^2 \times [(N+1)^2 - 1]$. DoA estimation for all bins that passed the DPD test is then computed by

$$\Theta_{\mathcal{D}} = \left\{ \Theta : \arg\max_{\Theta} P(\Theta, \tau, \nu), \ \forall (\tau, \nu) \in \mathcal{D} \right\}. \tag{8}$$

Finally, the desired DoA can be computed directly as the mean of the set $\Theta_{\mathcal{D}}$, although for a more accurate DoA estimation, GMM clustering is applied to $\Theta_{\mathcal{D}}$ and the mean of the dominant cluster is computed as the desired DoA. This was shown to reduce estimation bias [10].

It should be noted that the method presented here has been developed for spherical arrays, but can also be applied to any volumetric array that can extract the spherical harmonics coefficients of the sound field [8]. The application of the method to other standard arrays, such as uniform linear arrays, may require further study. First, to formulate the DPD test, including frequency-smoothing, for these array configurations, and second, to account for the ambiguities in DoA estimation when linear or planar arrays are positioned in a 3D sound field.

3. Sound field directivity

In this section a measure for the directivity of a sound field measured by a spherical array is formulated. The new DPD test developed in the following section is based on this measure. The directivity is widely used to characterize the spatial selectivity of