SPEAKER LOCALIZATION IN REVERBERANT ROOMS BASED ON DIRECT PATH DOMINANCE TEST STATISTICS

Boaz Rafaely[a] *and Dorothea Kolossa*[b]

[a] Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel, br@bgu.ac.il
[b] Department of Electrical Engineering and Information Technology, Ruhr-Universität Bochum, 44801 Bochum, Germany, dorothea.kolossa@rub.de

ABSTRACT

Speaker localization using microphone arrays is typically based on the expected phase and amplitude differences between microphones as a function of the wave arrival direction. However, in rooms with significant reverberation, the direct sound is contaminated by reflections and localization often fails. Recently, a reverberation-robust localization method was proposed, which uses only the direct-path bins in the short-time Fourier transform (STFT) of the speech signals. The method is based on thresholding according to the ratio between the first two singular values of the spatial spectrum matrix. In this work, a confidence measure is developed based on this ratio, which is then used for speaker localization in a statistical estimation framework, based on a Gaussian mixture model. The paper presents the theory of the proposed method and simulation examples validating the advantages of the new approach.

Index Terms— Speaker localization, reverberation, spherical microphone arrays, multiple signal classification, Gaussian mixture model

1. INTRODUCTION

The estimation of the direction of arrival (DoA) of speech signals in a room, using an array of microphones, is important in applications such as speech enhancement, source separation, robot audition and video conferencing. A wide range of methods has been developed for DoA estimation of sources in general, based on beamforming [1], and on subspace methods such as multiple signal classification (MUSIC) [2] and estimation of signal parameters via rotational invariance techniques (ESPRIT) [3]. Speakers in a room typically generate coherent signals due to room reflections, and so DoA estimation methods designed for coherent sources are more suitable for speaker localization. The coherent signal subspace method (CSSM) [4] employs frequency smoothing to restore the reduction in rank of the spatial spectrum matrices, thereby overcoming the degradation in performance due to source coherence. Recent DoA estimation methods for speech signals exploit their non-stationarity and sparsity in the short-time Fourier transform domain to improve DoA estimation performance even for under-determined systems with more sources than microphones [5, 6]. Although well-suited for speech, the performance of these methods degrades significantly under reverberation, as the sparseness property is easily affected by reverberation [7]. Mohan et al. [8] developed a coherence test to identify time-frequency bins that are dominated by a contribution from a single source. Although this method facilitates the localization of more sources than microphones, it fails at moderate levels of reverberation due to the failure of the coherence test.

Recently, a method for DoA estimation of multiple speakers in a room has been developed [9]. Similarly to the coherence test [8], the method processes the microphone signal in the time-frequency domain, and employs a test, referred to as the direct-path dominance (DPD) test, to identify time-frequency bins dominated by a single source. However, unlike the coherence-test-based method, the DPD-based method employs frequency smoothing such that direct sound and room reflections become incoherent, therefore overcoming the effect of room reverberation. Furthermore, the method is designed for spherical microphone arrays [10], so that DoA estimation can be applied for sources in all directions, and the frequency smoothing operation can be employed without the need for focusing matrices [9]. Experimental investigations have validated the robustness of the method to reverberation [9].

Although the DPD-test-based method is useful, it has limitations. In particular, under real-life conditions of high reverberation, only a few time-frequency bins pass the DPD test and DoA estimation is based on only a few samples, leading to increased errors due to the limited statistical information. This paper proposes an approach to overcome this limitation. The threshold of the DPD test, based on the ratio of

This work was supported by the European Union's Seventh Framework Programme (FP7/2007-2013) under grants agreement no. 609465, and no. 618075.

the first two eigenvalues of the spatial spectrum matrix, is relaxed, allowing for more time-frequency bins to pass the DPD test. Then, statistical analysis is employed to estimate the DoA. Since this relaxation, may, however, lead to inclusion of biased samples in the computation, simple linear estimators, such as straightforward maximum-likelihood estimation, will fail. Hence, a Gaussian mixture model is applied to the DoA samples, leading to an even more robust and accurate DoA estimation. The paper presents the development of the new algorithm, and simulation results for DoA estimation of a speaker in a room, comparing the proposed method to the original DPD-test-based method.

2. SPHERICAL ARRAY PROCESSING

This section presents the system model and the array processing employed before DoA estimation is performed. Consider a spherical microphone array with Q microphones arranged on the surface of a rigid sphere. Other array configurations can also be considered using a similar formulation [10]. The sound pressure at the microphones is denoted by $p(k, r, \theta_q, \phi_q)$, with k the wave number, r the sphere radius, and (θ_q, ϕ_q) denoting the spherical coordinates of microphone q. Denoting $\Omega_q \equiv (\theta_q, \phi_q)$, the sound pressure at the microphones can be written as [9]

$$\mathbf{p} = \mathbf{V}\mathbf{s} + \mathbf{n} \tag{1}$$

where $\mathbf{p} = [p(k, r, \Omega_1), ..., p(k, r, \Omega_Q)]^T$, denotes the sound pressure at the Q microphones, $\mathbf{s} = [s_1(k), ..., s_L(k)]^T$ denotes the L sources composing the sound field, $\mathbf{V}$ is the $Q \times L$ steering matrix denoting the transfer function from each source to each microphone, and $\mathbf{n} = [n_1(k), ..., n_Q(k)]^T$ represents the sensor noise. Equation (1) can be rewritten by representing the steering matrix in the spherical harmonics domain [9]

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

where $\mathbf{Y}(\mathbf{\Omega})$ is the $Q \times (N+1)^2$ spherical harmonics matrix with the complex spherical harmonics functions $Y_n^m(\Omega_q)$ of order n and degree m as its elements at row number q and column number $n^2 + n + m + 1$. The $(N+1)^2 \times (N+1)^2$ diagonal matrix $\mathbf{B}$ holds the radial functions that represent the scattering of a plane wave from a rigid sphere [10], and the $L \times (N+1)^2$ matrix $\mathbf{Y}(\mathbf{\Psi})$ has a similar structure to matrix $\mathbf{Y}(\mathbf{\Omega})$, with $\mathbf{\Psi}$ representing source arrival directions.

Equation (2) is now reformulated in a process referred to as plane wave decomposition [10]. It is first multiplied from the left by the pseudo-inverse $[\mathbf{Y}(\mathbf{\Omega})]^\dagger$, and then by the inverse $\mathbf{B}^{-1}$, leading to

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

with the $(N+1)^2 \times 1$ vector $\mathbf{a} = [a_{00}(k), ..., a_{NN}(k)]^T$ representing the plane wave density function in the spherical harmonics domain, and $\tilde{\mathbf{n}} = \mathbf{B}^{-1}[\mathbf{Y}(\mathbf{\Omega})]^\dagger\mathbf{n}$. In the final stage, Eq. (3) is written in the short-time Fourier transform domain,

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

with τ denoting the time index and ν denoting the frequency index.

3. DOA ESTIMATION USING THE DPD TEST

Equation (4) forms the basis of DoA estimation using the DPD test. Details of this algorithm can be found in [9]. In this section the algorithm is presented briefly. The algorithm employs MUSIC for DoA estimation, and so a spatial spectrum matrix is first composed as

$$\mathbf{R}(\tau, \nu) = \overline{[\mathbf{a}(\tau, \nu)\mathbf{a}^H(\tau, \nu)]}. \tag{5}$$

$\overline{[\cdot]}$ denotes averaging, which in practice is performed over a predefined range over time and frequency. Frequency averaging, or smoothing, is required so that the direct sound can be distinguished from coherent room reflections [9]. Following the procedure of MUSIC, the singular-value decomposition of $\mathbf{R}$ is computed, and the time-frequency bins, (τ, ν), that pass the DPD test are identified,

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

where σ_1 and σ_2 denote the largest and the second largest singular values, respectively, and $\mathcal{TH}$ is a threshold value, sufficiently larger than one, therefore guaranteeing that $\mathbf{R}$ is dominated by a single singular vector. This is an important stage in the algorithm, because this condition guarantees that the signals measured by the microphones are dominated by a single plane wave sound field. For speakers in a room, this is typically produced by the direct sound from the source, because room reflections tend to arrive with a delay relative to the direct sound.

In the following stage, the MUSIC spectrum, $P(\Theta)$, is calculated for all bins that pass the DPD test, i.e. for all $(\tau, \nu) \in \mathcal{D}$,

$$P(\Theta) = \frac{1}{||\mathbf{U_n}^H \mathbf{y}^*(\Theta)||^2} \tag{7}$$

with $\Theta = (\theta, \phi)$ representing the search grid for source directions, and $(N+1)^2 \times 1$ vector $\mathbf{y}$ having elements of the spherical harmonics functions $Y_n^m(\Theta)$ at column number $n^2 + n + m + 1$. Matrix $\mathbf{U_n}$ of size $(N+1)^2 \times [(N+1)^2 - 1]$ represents the noise subspace assuming a single source, and so its columns hold the singular vectors of matrix $\mathbf{R}$ corresponding to the smallest singular values.

In the final stage of the algorithm, source DoAs are estimated using two different approaches. In the first approach, the MUSIC spectrum is averaged for all time-frequency bins that pass the DPD test, and dominant peaks are then identified in this averaged spectrum, representing DoAs of dominant sources. In the second approach, the DoA is found for

each time-frequency bin by clustering singular vectors with a similar direction, constructing an averaged MUSIC spectrum for each cluster, and identifying the DoAs for each cluster separately. Furthermore, whitening is applied whenever the noise variance of signal $\tilde{\mathbf{n}}$ is significant. This approach was shown to be robust to reverberation, and capable of DoA estimation of several speakers in a reverberant room [9].

4. DOA STATISTICS

This section presents an analysis of the statistics of DoA estimation on a segment of speech. Insights from this analysis will motivate the development of an improved statistical method for the DPD-test-based DoA estimation algorithm. Consider a single speaker in a reverberant room. The room size is $8 \times 5 \times 3$ meters, the reverberation time is about 1 s, and the critical distance is 0.6 m. Room impulse responses from a point source in the room to a spherical array with 32 microphones of order $N = 3$, positioned 2 m away from the source are simulated using the image method [11]. Microphone signals are produced by convolving the room impulse responses with speech from the TIMIT database [12]. The effects of spatial aliasing and sensor noise are assumed to be negligible [10].

Figure 1 presents a spectrogram of the clean and reverberant speech signals as measured at the center of the array. The figure clearly illustrate the time-smearing effect of reverberation. The DPD-test-based algorithm as presented in the previous section has been applied to the speech data. Algorithm parameters include: a sampling frequency of 16 kHz, an FFT size of 512 samples, STFT analysis using a Hanning window with 50% overlap, and averaging of the spatial spectrum matrix using a 2×15 window over time and frequency. The output was a DoA estimate for each bin in the STFT map, and an associated σ_1/σ_2 ratio, see Eq. (6). Figure 2 shows the resulting sigma-ratio map for each bin. The figure shows that most values are relatively small, i.e. smaller than 3 (10 dB), while only a small proportion have much higher values. In the second part of the analysis, DoA estimation was performed using 3 different threshold values, i.e. $\mathcal{TH} \in [2, 5, 10]$, see Eq. (6). Then, mean and standard deviation for each group of DoA estimates, for both θ and ϕ, were compared to the true DoA. Figure 3 shows the results of this analysis. Three important observations can be made. As the value of σ_1/σ_2 increases, (i) the mean value becomes more accurate, i.e. bias is reduced, (ii) the variance decreases, and (iii) the population becomes significantly smaller (see figure caption for details).

In the final part of this analysis, the histogram of DoAs for $\mathcal{TH} = 2$ is presented in Fig. 4 as an example. While the figure shows high concentration of DoAs around the true values, the distribution seems to have more than a single peak. This may well be the cause of the bias when simple averaging is used for estimation.

Although the analysis in this section was presented through a single example, a similar behavior was observed for other room dimensions, source and microphone locations, reverberation times, and speech signals. The main conclusion from this analysis is that while the use of the DPD test with high threshold levels may produce good DoA estimates, the population becomes rapidly smaller, which may impose a limit if shorter speech signals are available, for example. In addition, the current method employs simple DoA averaging and does not exploit the full statistics of bin-level estimates. These limitations provide a motivation for improved methods, as presented in the following section.

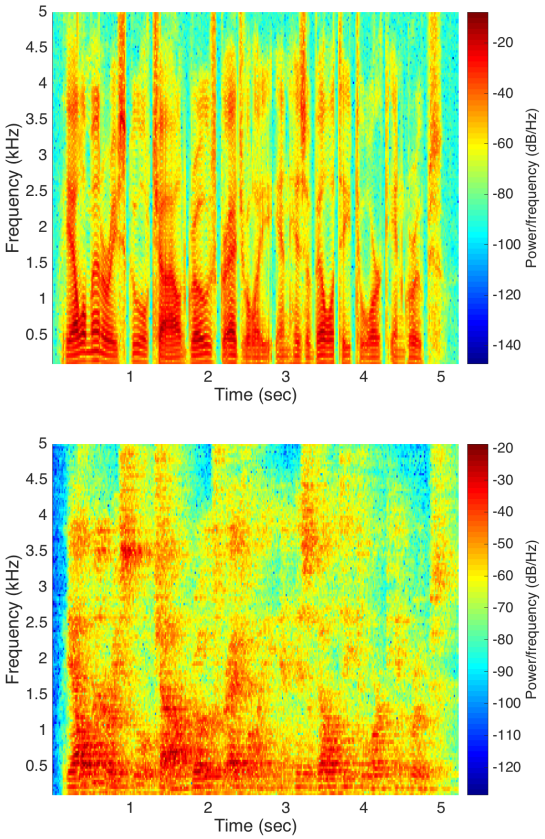

Fig. 1. Spectrogram of clean and reverberant speech.

5. GMM BASED ESTIMATION

In order to allow for the use of more DoA estimation samples, it becomes necessary to deal with their potential bias, ideally through consideration of their complete statistical properties. Hence, following the observation that the probability density functions of the DoA estimates are non-Gaussian and seem to be multi-modal, a Gaussian mixture model (GMM) appears suitable for the distribution. The use of other models designed specifically for circular and spherical distributions is left for a future study. In this study, a GMM with 5 Gaussians was fitted

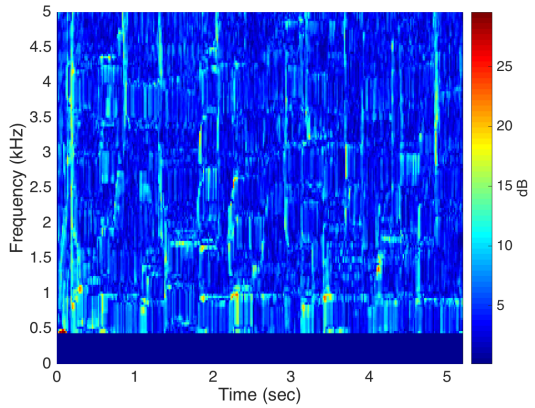

Fig. 2. Time-frequency map of σ_1/σ_2.

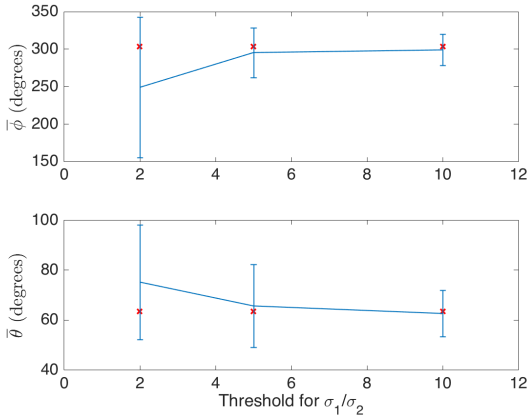

Fig. 3. Mean and standard deviation of the estimated DoAs, $\bar{\theta}$ and $\bar{\phi}$, for threshold values $\mathcal{TH} \in [2, 5, 10]$. Computed using $[18643, 766, 38]$ samples for the 3 threshold values, respectively. The true DoAs are denoted by x-marks.

jointly to (θ, ϕ) using an EM algorithm [13]. The dominant Gaussian, i.e. the one with the largest distribution peak, was selected as the one representing the DoA, while the other 4 Gaussians were excluded from the analysis. Matlab functions `fitgmdist` and `cluster` were employed in this analysis. The estimated DoA was then computed as the mean of the selected Gaussian.

6. SIMULATION STUDY

A simulation study was performed to assess the performance of the two approaches. Simulation parameters are the same as in Sec. 4, but with room dimensions, and source and array locations, perturbed by $10\,\%$, and with a randomly selected speaker from the TIMIT database, for each realization. The results of 20 realizations were averaged and the mean errors have been computed for the two methods studied, (i) DPD test with threshold set to $\mathcal{TH} = 10$, which was selected as it achieved smaller errors than lower threshold values, while providing a sufficient number of time-frequency bins that passed the test, e.g. dozens of bins for several seconds of speech, see Fig. 3; and (ii) GMM-based estimation. Table 1 presents the root-mean squared values of the errors for both methods. The table shows that the reference method of threshold and mean [9] has limited performance. This can be explained by the bias that was observed in the preceding analysis. The new method based on GMM distribution modeling shows a significant improvement, probably due to the robustness to the bias, which is the result of clustering of the data into several Gaussian distributions.

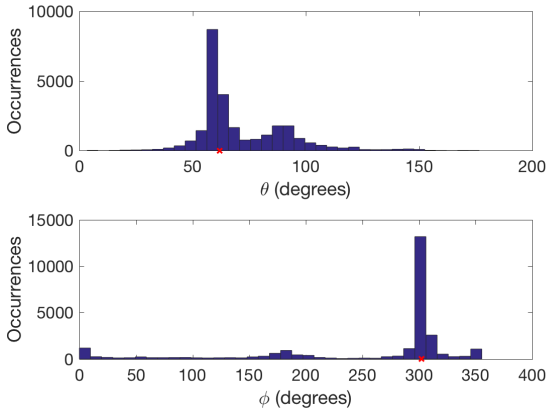

Fig. 4. Histograms of estimated DoAs, θ and ϕ, for threshold values $\mathcal{TH} = 2$. The true DoAs are denoted by x-marks.

Method	$\mathcal{TH}$	Error θ	Error ϕ
Mean [9]	10	$6.5°$	$18.9°$
GMM [new]	2	$2.1°$	$3.1°$

Table 1. DoA estimation errors for both methods, computed as the root-mean squared error between the estimated and true directions for each realization

.

7. CONCLUSIONS AND FUTURE WORK

This paper has presented an analysis of the statistical properties of the DoA estimates in the DPD-test-based method. It was shown that the DoAs have a multi-modal distribution, so that computing the mean over multiple DoA samples potentially leads to a biased estimate. Based on this understanding, a new estimator has been introduced, which uses Gaussian mixture modeling to overcome this issue, and it has been shown to significantly improve the DoA estimation. The study of the effects of room acoustics and speech signal characteristics on the DoA statistics is proposed for future work, with the goal of deriving reliably unbiased estimators in the presence of reverberation.

8. REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[2] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, march 1986.

[3] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, july 1989.

[4] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wideband sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 4, pp. 823–831, August 1985.

[5] S. Rickard and F. Dietrich, "DOA estimation of many W-disjoint orthogonal sources from two mixtures using duet," in *Statistical Signal and Array Processing, 2000. Proceedings of the Tenth IEEE Workshop on*, 2000, pp. 311–314.

[6] A. M. Torres, M. Cobos, B. Pueo, and J. J. Lopez, "Robust acoustic source localization based on modal beamforming and time–frequency processing using circular microphone arrays," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1511–1520, 2012.

[7] S. Arkadi, H.Sawada, R. Mukai, and S. Makino, "Performance evaluation of sparse source separation and DOA estimation with observation vector clustering in reverberant environments," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC 2006)*. Paris, 2006.

[8] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Am.*, vol. 123, no. 4, pp. 2136–2147, 2008.

[9] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1494–1505, October 2014.

[10] B. Rafaely, *Fundamentals of Spherical Array Processing*, Springer-Verlag, Berlin, first edition, 2015.

[11] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. S. Dahlgren, "DAPRA TIMIT acoustic-phonetic continuous speech corpus," CDROM, 1993, US National Institute of Standards and Technology.

[13] G. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, Inc., 2000.