PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Binaural Direction-of-Arrival Estimation in Reverberant Environments Using the Direct-path Dominance test

Hanan BEIT-ON[(1)], Boaz RAFAELY[(2)]

[(1)]Ben-Gurion University of the Negev, Israel, hananb@post.bgu.ac.il
[(2)]Ben-Gurion University of the Negev, Israel, br@bgu.ac.il

Abstract

Abstract— Speaker localization in reverberant environments is currently an important challenge in audio signal processing, and several methods have recently been proposed to address it. One such family of methods is based on the direct-path dominance (DPD) test that was developed specifically for spherical arrays and recently extended to arbitrary arrays. In this paper, an extension of the DPD test for a binaural array is examined. This extension uses a focusing technique to remove the frequency dependence of the steering matrix within frequency bands. This facilitates frequency smoothing, which is necessary for decorrelating coherent reflections. Unlike other focusing approaches, the proposed focusing approach does not rely on the initial estimation of source directions, facilitating its application in reverberant environments. However, in the binaural case, there is no guarantee of focusing performance. This paper presents an analysis of this focusing process for the binaural case and shows that good focusing performance can be obtained in the binaural case. A simulation study examines the effect of the focusing and smoothing and demonstrates the proposed method's capabilities in overcoming adverse reverberation conditions.

1 Introduction

Binaural direction-of-arrival (DOA) estimation of sound from speakers in a room is important for speech enhancement and source separation in applications such as hearing aids and robot audition. DOA estimation methods for sources, in general, include maximum likelihood estimation [16], beamforming techniques [16] and subspace methods, such as MUSIC [16]. Because speakers in a room generate coherent signals due to room reflection, DOA estimation methods that are designed for coherent signals are more suitable in this case. One such method is the coherent signal subspace method (CSSM) [17], which employs focusing and frequency smoothing to decorrelate the sources, enabling the DOA estimation of coherent signals using subspace methods. Coherent processing of the CSSM has recently been applied in a DOA estimation method for speech signals, called the direct-path dominance (DPD) test [10]. This method tackles the issue of reverberation by using only the direct-path bins in the short-time Fourier transform (STFT) domain for the estimation task. The need of focusing matrices has been circumvented by employing spherical arrays, for which the frequency dependence of the steering matrix is removed by the plane-wave decomposition operation [11]. Although the DPD test method has shown robustness to diverse reverberation and noise conditions [12], its application is limited to spherical arrays.

An extension of the DPD test for arbitrary arrays was recently proposed [3]. This extension incorporates a focusing process that does not require an initial DOA estimation facilitating its application in reverberant environments. Unlike the bin selection method proposed in [15], which is also available for arbitrary arrays, the DPD test employs coherent processing. The coherent processing increases the accuracy of the spatial spectrum [6], and this, in turn, leads to an accurate estimation when using a subspace localization method such as MUSIC. Unlike the binaural localization method proposed in [9], the DPD test does not require a training phase. Furthermore, in contrast to the speaker localization method proposed in [8], the DPD test does not require speech-free signal segments for estimating the noise statistics and can operate on short time segments.

Although the proposed extension of the DPD test is useful, it has limitations. In particular, poor focusing performance may result in DOA estimation errors. For binaural arrays, the focusing matrices are obtained as the least squares solution to an over-determined system, thus incorporating a fitting error. Moreover, the development of the focusing matrices assumes knowledge of the spherical harmonics coefficients matrix. For binaural arrays, this matrix is obtained via the discrete spherical Fourier transform (DSFT) of the head related transfer function (HRTF) and its accuracy depends on the sampling of the HRTF. In this paper, the performance of the focusing for binaural arrays is investigated, and guidelines for sampling the HRTF are suggested. A simulation study examines the effect of focusing and smoothing and demonstrates binaural DOA estimation with the DPD test under adverse reverberation conditions.

2 System model

Consider an acoustic scene of a single speaker in a reverberant environment. Also, consider a sound field composed of L far field sources, where a source can represent a direct-sound or, for example, a reflection due to room boundaries. Employing the multiplicative transfer function (MTF) approximation [2], according to which if the length of the time windows are sufficiently large with respect to the length of the HRTF (in time), the binaural signal can be expressed in the STFT domain as

$$p(\tau,\nu) = H(\nu,\psi)\, s(\tau,\nu) + n(\tau,\nu), \tag{1}$$

where τ and ν are the time frame and the frequency indices, respectively. $p(\tau,\nu) = [p_l(\tau,\nu), p_r(\tau,\nu)]^T$, with $p_l(\tau,\nu)$ and $p_r(\tau,\nu)$ denoting the signal at the left and the right binaural microphones, respectively. $s(\tau,\nu) = [s_1(\tau,\nu),...,s_L(\tau,\nu)]^T$ is the source signal vector, where $s_l(\tau,\nu)$ is the complex amplitude of the l'th source at the origin. $H(\nu,\psi) = [h(\nu,\psi_1),...,h(\nu,\psi_L)]^T$ is a $2\times L$ HRTF matrix, where $h(\nu,\psi_l)$ is the HRTF that corresponds to the l'th source with DOA $\psi_l = (\theta_l,\phi_l)$, where θ_l and ϕ_l are the elevation and the azimuth angles, respectively. $n(\tau,\nu) = [n_l(\tau,\nu), n_r(\tau,\nu)]^T$ is the additive sensor noise, which is assumed to be spatially white and uncorrelated with $s(\tau,\nu)$.

3 The DPD test for binaural arrays

The first stage of the DPD test method is to construct the smoothed spectrum at each time frequency (TF) bin. To construct the smoothed spectrum in bin (τ_0,ν_0), focusing is applied to a rectangular window of J_τ time frames and J_ν frequencies around (τ_0,ν_0), followed by an averaging of the spectrum over the window. The averaging over time frames approximates the expectation operation, and the averaging over frequencies implements frequency smoothing. Focusing is applied to remove the frequency dependence of the HRTF matrices within the averaging window to preserve the spatial information in the smoothed HRTF matrix. Focusing is performed by aligning the HRTF matrices within the averaging window to the HRTF matrix from the central frequency using a focusing matrices $T(\nu,\nu_0)$ that satisfies

$$T(\nu,\nu_0)\, H(\nu,\psi) = H(\nu_0,\psi). \tag{2}$$

The focusing is applied by multiplying the binaural signal $p(\tau,\nu)$ with the corresponding focusing matrix $T(\nu,\nu_0)$, such that the transformed binaural signal $\tilde{p}(\tau,\nu)$ can be expressed as

$$\tilde{p}(\tau,\nu) = T(\nu,\nu_0)\, p(\tau,\nu) = H(\nu_0,\psi)\cdot s(\tau,\nu) + \tilde{n}(\tau,\nu), \tag{3}$$

where $\tilde{n}(\tau,\nu) = T(\nu,\nu_0)\, n(\tau,\nu)$. The smoothed spectrum is then constructed by averaging the spectrum within the averaging window. Assuming $s(\tau,\nu)$ and $n(\tau,\nu)$ are uncorrelated, the smoothed spectrum can be expressed as

$$\overline{S_{\tilde{p}}}(\tau_0,\nu_0) = H(\nu_0,\psi)\,\overline{S_s}(\tau_0,\nu_0)\, H^H(\nu_0,\psi) + \overline{S_{\tilde{n}}}(\tau_0,\nu_0), \tag{4}$$

where $\overline{S_s}(\tau_0,\nu_0)$ and $\overline{S_{\tilde{n}}}(\tau_0,\nu_0)$ are the smoothed covariance matrices of the source and noise signals, respectively. The frequency smoothing is employed to decorrelate coherent sources, such that after smoothing the matrix $\overline{S_s}(\tau_0,\nu_0)$ is assumed to be of full-rank with a low condition number. This facilitates the application of subspace localization methods such as MUSIC.

The DPD test selects TF bins with one dominant source and assumes that these bins contain a significant contribution from the direct-sound and insignificant contributions from room reflections. Thus, the DOA of the actual source can be correctly estimated using these bins. The group of bins selected by the DPD test is

$$\mathscr{A}_{\mathrm{DPD}} = \left\{ (\tau_0,\nu_0) : \frac{\sigma_1(\tau_0,\nu_0)}{\sigma_2(\tau_0,\nu_0)} > \mathrm{TH}_{\mathrm{DPD}} \right\},$$

where $\sigma_1(\tau_0,\nu_0)$ and $\sigma_2(\tau_0,\nu_0)$ are the singular values of $\overline{S_{\tilde{p}}}(\tau_0,\nu_0)$ and $\mathrm{TH}_{\mathrm{DPD}}$ is a chosen threshold. Several approaches for estimating the speaker DOA from the selected bins have been proposed, including MUSIC with coherent and incoherent integration of the signal sub-spaces from the different bins [10], and bin-wise DOA estimation followed by a statistical analysis to fuse the estimates [14].

The formulation of the focusing matrices proposed in [3] assumes an order-limited HRTF, i.e. the HRTF can be fully represented using spherical harmonics up to order N. As an order-limited function, the decomposition of the HRTF matrix using spherical harmonics is given by

$$H(\nu,\psi) = H_{nm}(\nu)\,Y(\psi), \tag{5}$$

where

$$Y(\psi) = \begin{bmatrix} Y_0^0(\psi_1) & Y_0^0(\psi_2) & \cdots & Y_0^0(\psi_L) \\ Y_1^{-1}(\psi_1) & Y_1^{-1}(\psi_2) & \cdots & Y_1^{-1}(\psi_L) \\ \vdots & \vdots & \ddots & \vdots \\ Y_N^N(\psi_1) & Y_N^N(\psi_2) & \cdots & Y_N^N(\psi_L) \end{bmatrix}$$

is an $(N+1)^2 \times L$ matrix of the spherical harmonics functions $Y_n^m(\psi)$ of order n and degree m [11] and

$$H_{nm}(\nu) = \begin{bmatrix} h^l_{0,0}(\nu) & h^l_{1,-1}(\nu) & \cdots & h^l_{N,N}(\nu) \\ h^r_{0,0}(\nu) & h^r_{1,-1}(\nu) & \cdots & h^r_{N,N}(\nu) \end{bmatrix}$$

is a $2\times(N+1)^2$ matrix of the spherical harmonics coefficients $h^l_{n,m}(\nu)$, $h^r_{n,m}(\nu)$ of order n and degree m of the HRTF of the left (l) and right (r) binaural microphones, respectively. Substituting (5) into (2) it follows that a focusing matrix that aligns $H(\nu,\psi)$ to $H(\nu_0,\psi)$ for any set of sources should satisfy

$$T(\nu,\nu_0)\,H_{nm}(\nu) = H_{nm}(\nu_0). \tag{6}$$

The general solution to this system is given by

$$T(\nu,\nu_0) = H_{nm}(\nu_0)\,{H_{nm}}^{\dagger}(\nu), \tag{7}$$

where $(\cdot)^{\dagger}$ denotes the pseudo-inverse operation.

4 Focusing formulation for binaural arrays

The development of the focusing matrices presented in the previous section assumes an order-limited HRTF and knowledge of the spherical harmonics matrix $H_{nm}(\nu)$. Conditions for which these assumptions hold are formulated in this section.

The resemblance of the human head to a rigid sphere suggests that, like a rigid sphere, the HRTF is nearly order-limited with order $N = kr$, where $k = \frac{2\pi f}{c}$ is the wave number with c denoting the speed of sound, and r is the sphere radius. Thus, the assumption according to which the HRTF can be represented with spherical harmonics coefficients up to order N is approximately maintained for $kr < N$.

While for some arrays, such as spherical arrays with a rigid-sphere, the spherical harmonics coefficients matrix follows an analytical expression, in the binaural case, this matrix is computed using the discrete spherical Fourier transform (DSFT) of the HRTF matrix. The DSFT is given by [11]

$$H_{nm}(\nu) = H(\nu, \psi) Y^{\dagger}(\psi), \tag{8}$$

where $\psi = [\psi_1, ..., \psi_L]^T$ is a set of sampling points of the HRTF over a sphere of directions. Such samples can be obtained with a numerical simulation or using measurements. Accurate computation of $H_{nm}(\nu)$, requires more samples than coefficients, i.e. $L \geq (N+1)^2$, and a matrix $Y(\psi)$ of full-rank [11]. Table 1 details the necessary number of required samples L for several sampling schemes in order that these conditions are satisfied.

Table 1. Number of samples for accurate computation of $H_{nm}(\nu)$ for several sampling schemes

Sampling scheme	equal angle	Gaussian	uniform
Number of samples	$4(N+1)^2$	$2(N+1)^2$	$(N+1)^2$

Even if the assumptions discussed above approximately hold, there is no guarantee for ideal focusing in the binaural case. This is because the system in (6) is over-determined for all N (excluding $N = 0$ for which the HRTF contains no directional information); hence, the solution in (7) is obtained in the least squares sense, thus incorporating a fitting error. The next section examines the focusing performance for a binaural array.

5 Focusing performance analysis

To evaluate the performance of focusing for simple binaural arrays, an array of two microphones mounted on a rigid sphere was employed. The array steering function from a set of $L = 338$ points in a Gaussian sampling scheme was obtained according to [11] up to order $N = 12$, with a sampling frequency of $Fs = 16\,\text{kHz}$ and with a frequency resolution of 512 points. The mean (over directions) of the normalized focusing error is defined as [7]

$$\varepsilon_f(\nu_0) = \frac{1}{L} \sum_{l=1}^{L} \frac{\overline{\|T(\nu, \nu_0) h(\nu, \psi_l) - h(\nu_0, \psi_l)\|^2}}{\|h(\nu_0, \psi_l)\|^2}, \tag{9}$$

where $\overline{(\cdot)}$ denotes averaging over the frequency within the focusing window.

The focusing matrices were computed according to (7) with several spherical harmonics orders N for a focusing windows of 15 frequencies. Figure 1 shows $\varepsilon_f(\nu_0)$ as a function of the central frequency ν_0. To examine the effect of the order-limited assumption, the values of kr are depicted at the top of figure 1. The dashed line denotes the values of $\varepsilon_f(\nu_0)$ when the focusing matrices were chosen to be an identity matrix, i.e. no focusing was employed.

Figure 1 shows that, although the focusing computation incorporates the solution of an over-determined system, good performance can still be achieved, and a significant improvement is evident with respect to the case of no focusing. In addition, Figure 1 shows that the performance of focusing matrices of order N deteriorates for $kr > N$. This is because in this range a truncated representation of the HRTF is used in (7), resulting in an increased focusing error. Nevertheless, good focusing performance is obtained at frequencies in the range $kr < N$. This implies that in the case for which the DPD test does not exhibit high sensitivity to focusing errors,

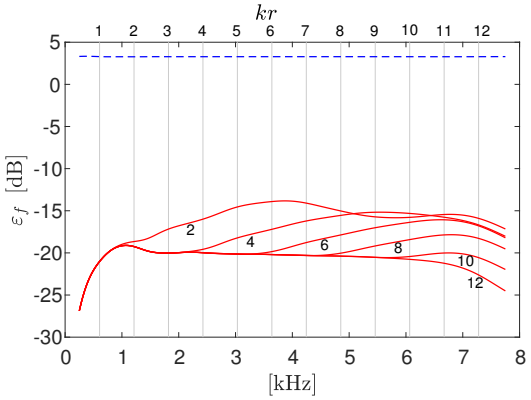


Figure 1. Focusing error for several spherical harmonics orders $N = 2,4,6,8,10,12$ and for focusing window of 15 frequencies. The dashed line depicts the error when no focusing is employed

sparse sampling of the HRTF may be sufficient for good performance. The next section examines the effect of focusing performance on the DOA estimation.

6 Simulation study

In this section the application of the DPD test for the binaural case is examined through simulation. A simple acoustic condition of single a speaker and a single room reflection is initially assumed in Section 6.1 in order to examine the effect of smoothing and focusing. Then, a more complex acoustic condition of a highly reverberant room is assumed in Section 6.2 in order to examine the effect of smoothing and focusing in a more realistic condition and to demonstrate the DOA estimation with the DPD test under adverse reverberation conditions. Azimuth estimation only was considered, which is more natural for the binaural array.

The Numman KU-100 [4] was used as a binaural array. The speaker, simulated as a point source, was positioned 1.5 m away from the array at the array frontal horizontal plane. A four-seconds segment of a speech signal from the TIMIT database [5] was used as speech material. The spherical harmonics coefficients up to order $N = 12$ of the plane waves' density at the array position were synthesized using the image method [1] at a sampling frequency of 16 kHz. The binaural signal was simulated according to [13].

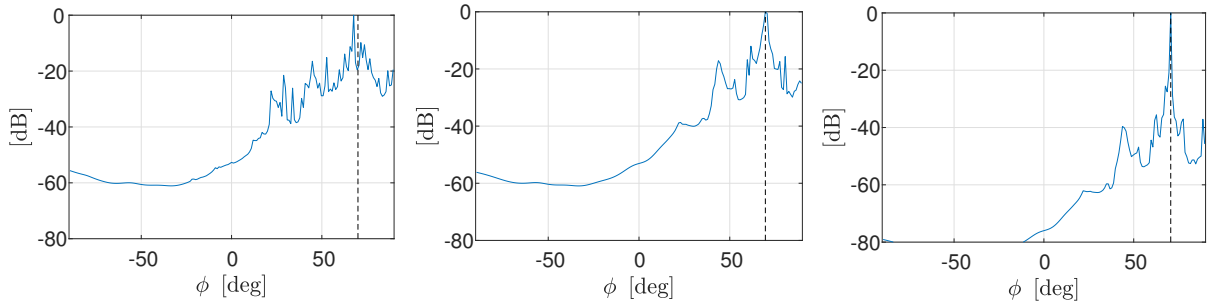
The DPD test was applied to the simulated signal with the following parameters: STFT with Hann window of 512 samples length (32 ms) and 50% overlap between adjacent frames, $J_\tau = 3$ time frames and $J_\nu = 15$ frequencies were used for averaging, and the focusing matrices were computed with spherical harmonics order of $N = 12$. The threshold was selected separately for each frequency such that 5% of the TF bins from each frequency would pass the test. TF bins from frequencies below 1 kHz and beyond 6.5 kHz were automatically rejected by the test.

6.1 A single reflection

A simple acoustic condition of a single speaker and a single room reflection is studied to examine the effect of smoothing and focusing on DOA estimation with the DPD test. The speaker and the reflection DOAs were $\psi_1 = (90°, 70°)$ and $\psi_2 = (90°, 20°)$, respectively. The amplitude ratio at the array between the reflection and the direct-path (speech signal) was 0.6, and the reflection arrives at the array with a delay of 5 ms, which represents the delay of a typical early room reflection. A noise free condition was assumed.

To examine the effect of smoothing and focusing, the DPD test was applied with and without focusing and smoothing. Under all conditions, the spectrum of MUSIC was computed for the frontal azimuth direction at each of the selected TF bins. An incoherent integration of the spectrum is then applied by averaging the

spectra from the different bins. Figure 2 shows the mean MUSIC spectrum under the different conditions. The true azimuth is marked by a dashed black line.



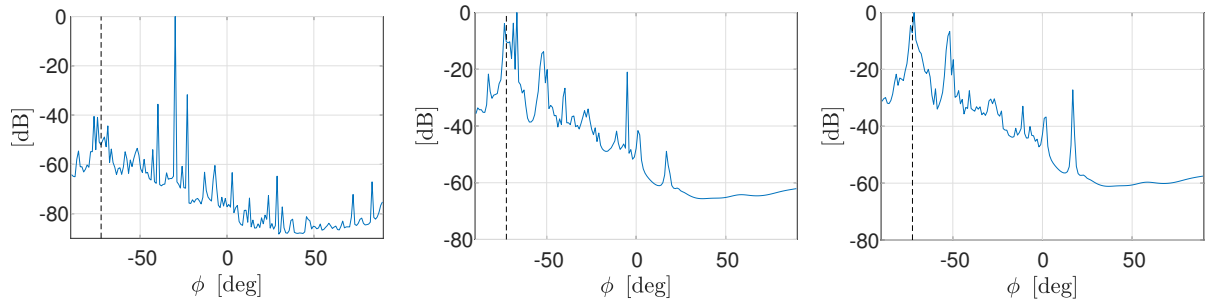
(a) (b) (c)

Figure 2. Mean MUSIC spectrum from the selected TF bins in each condition: (a) without smoothing, (b) with smoothing without focusing, (c) with smoothing and focusing. The dashed line represent the direction of the speaker.

Figure 2a shows that without smoothing the peak in the MUSIC spectrum is away from the speaker azimuth. This is because, with no smoothing the spatial spectrum in all TF bins is of unit rank; consequently, a significant number of selected TF bins contain DOA information that is distorted by the reflection. Figures 2b, 2c show that the MUSIC spectra for both of the conditions that apply smoothing have a peak very close to the speaker azimuth. Nevertheless, the peak in the condition with focusing is sharper and more prominent. Incorporating focusing better preserves the spatial information in the smoothed HRTF matrix and, consequently, improves the DOA estimation from the selected TF bins.

6.2 Reverberant room

The binaural signal due to a speaker in a reverberant room is simulated. A room of dimensions $8 \times 5 \times 3\,\mathrm{m}$ and with a reverberation time of $T_{60} = 1\,\mathrm{sec}$ was considered. The array position was selected to be $(x, y, z) = (2, 1.5, 1.7)\,\mathrm{m}$. A sensor noise of $40\,\mathrm{dB}$ SNR was simulated. Figure 3 shows the mean MUSIC spectra under the different conditions.



(a) (b) (c)

Figure 3. Mean MUSIC spectrum from the selected TF bins in each condition: (a) without smoothing, (b) with smoothing without focusing, (c) with smoothing and focusing. The dashed line represent the direction of the speaker.

The results suggest that without smoothing, TF bins that contain a significant contribution from reflections are selected. Figures 2b, 2c shows that the spectrum for both conditions that apply smoothing are quite similar,

excluding the area around the speaker azimuth. In this area the condition with focusing has a peak very close to the speaker azimuth, and the condition without focusing exhibits a couple of peaks with deviation from the speaker azimuth. This result suggests that focusing have a positive effect the accuracy of the DOA estimation from the selected TF bins.

Figure 4 shows the spectrogram of the signal obtained in the left microphone and the TF bins that were selected by the DPD test in black.

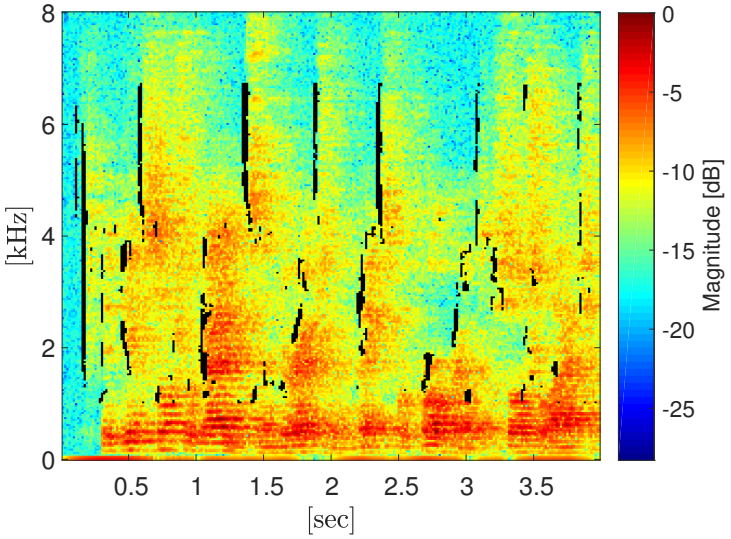


Figure 4. Spectrogram of reverberant speech with selected bins in black

Figure 4 shows that the majority of the selected TF bins are located at the onset of the speech signal. This is because the direct-sound arrives earlier at the array than reflections that have a longer traveling time. This result supports the validity of the used measure in selecting TF bins in which the direct-path is dominant.

Figure 5a shows the histogram of the azimuth estimates from the selected TF bins and figure 5b shows the histogram of the azimuth estimates from all TF bins. The histograms were normalized such that the area below the graph will equal 1.

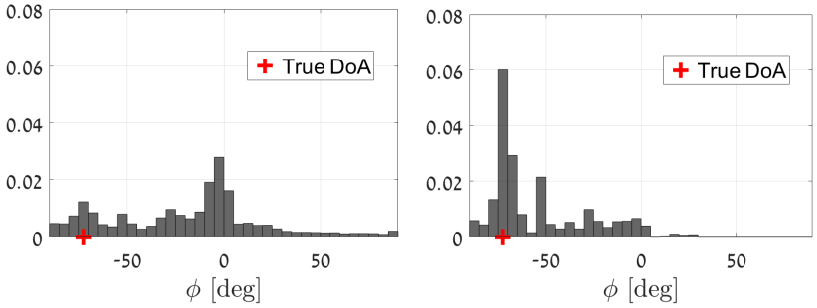


Figure 5. Histogram of azimuth estimates from (a) all bins, (b) selected bins

Comparing the histograms, it is clear that using the TF bins selection with the DPD test improves the accuracy

of the DOA estimation.

7 Conclusions

In this paper, the DPD test method for speaker localization is examined for the binaural case. Central to the DPD test is a focusing process. The assumptions at the basis of this focusing process are examined for the binaural case, and conditions for which these assumptions hold are formulated. It has been shown that good focusing performance can be obtained in the binaural case. An experiment with simulated data verifies the necessity of focusing and smoothing and demonstrates the capabilities of the proposed method of overcoming adverse reverberation conditions with the DPD test.

REFERENCES

[1] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[2] Y. Avargel and I. Cohen. On multiplicative transfer function approximation in the short-time fourier transform domain. *IEEE Signal Processing Letters*, 14(5):337–340, 2007.

[3] H. Beit-On and B. Rafaely. Speaker localization using the direct-path dominance test for arbitrary arrays. In *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, pages 1–4. IEEE, 2018.

[4] B. Bernschütz. A spherical far field hrir/hrtf compilation of the neumann ku 100. In *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*, page 29. AIA/DAGA, 2013.

[5] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

[6] H. Hung and M. Kaveh. Focussing matrices for coherent signal-subspace processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(8):1272–1281, 1988.

[7] T.-S. Lee. Efficient wideband source localization using beamforming invariance technique. *IEEE Transactions on Signal Processing*, 42(6):1376–1387, 1994.

[8] X. Li, L. Girin, R. Horaud, and S. Gannot. Estimation of the direct-path relative transfer function for supervised sound-source localization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2171–2186, 2016.

[9] N. Ma, T. May, and G. J. Brown. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(12):2444–2453, 2017.

[10] O. Nadiri and B. Rafaely. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1494–1505, 2014.

[11] B. Rafaely. *Fundamentals of spherical array processing*, volume 8. Springer, 2015.

[12] B. Rafaely and K. Alhaiany. Speaker localization using direct path dominance test based on sound field directivity. *Signal Processing*, 143:42–47, 2018.

[13] B. Rafaely and A. Avni. Interaural cross correlation in a sound field represented by spherical harmonics. *The Journal of the Acoustical Society of America*, 127(2):823–828, 2010.

[14] B. Rafaely, C. Schymura, and D. Kolossa. Speaker localization in a reverberant environment using spherical statistical modeling. *The Journal of the Acoustical Society of America*, 141(5):3523–3523, 2017.

[15] V. Tourbabin, D. L. Alon, and R. Mehra. Space domain-based selection of direct-sound bins in the context of improved robustness to reverberation in direction of arrival estimation. 2018.

[16] H. L. Van Trees. *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[17] H. Wang and M. Kaveh. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(4):823–831, 1985.