INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

The importance of time-frequency averaging for binaural speaker localization in reverberant environments

Hanan Beit-On[1], *Vladimir Tourbabin*[2], *Boaz Rafaely*[1]

[1]School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[2]Facebook Reality Labs

hananb@post.bgu.ac.il, vtourbabin@fb.com, br@bgu.ac.il

Abstract

A common approach to overcoming the effect of reverberation in speaker localization is to identify the time-frequency (TF) bins in which the direct path is dominant, and then to use only these bins for estimation. Various direct-path dominance (DPD) tests have been proposed for identifying the direct-path bins. However, for a two-microphone binaural array, tests that do not employ averaging over TF bins seem to fail. In this paper, this anomaly is studied by comparing two DPD tests, in which only one has been designed to employ averaging over TF bins. An analysis of these tests shows that, in the binaural case, a TF bin that is dominated by multiple reflections may be similar to a bin with a single source. This insight can explain the high false alarm rate encountered with tests that do not employ averaging. Also, it is shown that incorporating averaging over TF bins can reduce the false alarm rate. A simulation study is presented that verifies the importance of TF averaging for a reliable selection of direct-path bins in the binaural case.

1. Introduction

Binaural speaker localization or direction-of-arrival (DOA) estimation is an important component in speech enhancement in various head-mounted communication devices, e.g. hearing aids and robot audition. These devices often operate in reverberant environments such as offices and living rooms. Under reverberant conditions, DOA estimation becomes a challenge due to reflections from room boundaries that mask the true DOA.

Traditional methods for binaural localization use the interaural level difference (ILD) and interaural time (or phase) difference (ITD or IPD) for estimating source direction, e.g. [1, 2, 3]. However, these features are not robust to reverberation. Several methods have been proposed to overcome the effect of reverberation. The learning-based method in [4] gains robustness to reverberation by using speech signals that are corrupted by diffuse noise to train a deep neural network classifier. However, this method requires training and returns DOA estimates even for time segments that do not contain information about the speaker's direction. Another recently proposed method overcomes reverberation issues by using the direct-path component of the relative transfer function (RTF) [5, 6, 7] for estimating source direction. However, this method requires a speech-free segment for estimating noise statistics. Therefore, its performance strongly depends on the accuracy of the noise and the RTF estimates.

One effective approach to overcome the effects of reverberation, that does not require training or transfer function estimation, is based on the direct-path dominance (DPD) test [8]. With this approach, DOA estimation is performed in the time-frequency (TF) domain by selecting TF bins in which the direct-path signal is dominant, and then only these bins are used for estimating the speaker's DOA. Various DPD tests have been proposed for identifying direct-path bins [8, 9, 10, 11, 12, 13, 14, 15], one of which was tested for a two-microphone binaural array [11, 16]. These tests can be classified into two classes, where one class employs averaging over TF bins, while the other does not incorporate averaging, i.e. tests in which a decision on an individual TF bin is based on the bin itself. Although all tests have demonstrated improvement in the robustness to reverberation, tests that do not incorporate averaging over TF bins seem to fail when applied to binaural arrays.

In this paper, two tests, one from each class, are investigated under the binaural setup. An analysis of these tests is presented, showing that in the binaural case, a TF bin that is dominated by multiple reflections may be similar to a bin with a single dominant source. This insight can explain the high false alarm rate with tests that do not employ averaging, and which select TF bins based on their similarity to a single source. Also, it is shown that incorporating averaging over TF bins can reduce the false alarm rate. A simulation study is presented that compares the performance of the studied tests and verifies the importance of TF averaging for a reliable selection of direct-path bins in the binaural case.

2. System model

Consider a static scenario in which a speech source and a binaural array are located in a reverberant environment. The soundfield at the array's position can be modeled as a composition of Q plane-waves emitted by Q far-field sources, where a source can represent a direct sound, or, for example, a reflection due to room boundaries. The binaural signal in the short-time-Fourier-transform (STFT) domain can be expressed as

$$\begin{aligned}\mathbf{p}\left(\tau,\omega\right) &= \mathbf{p}_d\left(\tau,\omega\right) + \mathbf{p}_r\left(\tau,\omega\right) \\ &= \mathbf{h}(\omega,\psi_0)s_0\left(\tau,\omega\right) + \sum_{q=1}^{Q-1}\mathbf{h}(\omega,\psi_q)s_q\left(\omega,\tau\right), \quad (1)\end{aligned}$$

where τ and ω denote the time and the frequency indices, respectively. $\mathbf{p}\left(\tau,\omega\right) = \left[p^l\left(\tau,\omega\right), p^r\left(\tau,\omega\right)\right]^T$, where $p^l\left(\tau,\omega\right)$ and $p^r\left(\tau,\omega\right)$ are the STFT of the left and right microphone signals, respectively. $s_q\left(\tau,\omega\right)$ denotes the STFT of the q-th source signal that originates from direction ψ_q, and $\mathbf{h}\left(\omega,\psi_q\right)$ is the head-related transfer function (HRTF) from ψ_q. The formulation in (1) assumes that the multiplicative transfer function (MTF) approximation [17] holds, i.e. the length of the HRTF filters $\mathbf{h}\left(\omega,\psi\right)$ (in time), is significantly shorter than the length of the STFT window. The representation in (1)

http://dx.doi.org/10.21437/Interspeech.2020-2256

can be decomposed into $\mathbf{p}_d(\tau,\omega) = \mathbf{h}(\omega,\psi_0)s_0(\tau,\omega)$ and $\mathbf{p}_r(\tau,\omega) = \sum_{q=1}^{Q-1}\mathbf{h}(\omega,\psi_q)s_q(\omega,\tau)$, which denotes the direct and the reverberant parts, respectively. The direct part bears the DOA information of the speaker. Therefore, DOA estimates for bins with a dominant direct part are expected to be more accurate than DOA estimates for bins with a dominant reverberant part, in which the desired DOA information is distorted by reflections. To improve localization accuracy under reverberation, various direct-path dominance (DPD) tests have been proposed that aim to identify the direct-path bins. Given a set of direct-path bins selected by a DPD test, bin-wise DOA estimation is typically performed, followed by statistical analysis to fuse the estimates [18, 19, 20, 21].

3. Overview of direct-path dominance tests

In this section, an overview of current DPD tests is presented. The various tests are classified into two classes, where one class employs averaging over TF bins, while the other class does not incorporate averaging. Two tests, one from each class, are presented in more detail for the binaural case.

3.1. DPD tests incorporating averaging

Most of the current DPD tests incorporate averaging. The speech onset detection based test [13] uses subtraction of the signal power from consecutive time frames to detect drastic increments in the signal envelope. The consistency based tests in [14, 15] use averaging of DOA estimates (one per bin) in order to assess the estimates' spread, which is used to grade the bins. The direct-to-reverberant ratio (DRR) based test in [12] uses averaging over time frames to estimate the spatial covariance matrix, while in the tests in [8, 11, 10], covariance matrices are further smoothed over frequencies to decorrelate coherent reflections. The DPD test presented in [11], referred to here as the sigma-ratio test, is presented next and used in the remainder as an example for studying DPD tests that employ averaging.

The sigma-ratio test selects bins with a spatial covariance matrix of unit rank, suggesting the existence of a single dominant source, which is assumed to be the direct sound from the speaker. The sigma-ratio test incorporates frequency smoothing and a focusing process. The frequency smoothing is applied prior to the rank test in order to decorrelate coherent sources preventing the selection of bins with multiple coherent reflections. The focusing process aims to remove the frequency dependence of the HRTF $\mathbf{h}(\omega,\psi)$ in order to preserve the spatial information during the successive frequency smoothing operation. To construct the spatial covariance matrix at (τ_0,ω_0), focusing is first applied to a rectangular window $\Omega_{(\tau_0,\omega_0)}$ centered about (τ_0,ω_0). The focusing is performed by multiplying $\mathbf{p}(\tau,\omega)$ for each bin by a corresponding focusing transformation $\mathbf{T}(\omega,\omega_0)$ that satisfies $\mathbf{T}(\omega,\omega_0)\mathbf{h}(\omega,\psi) = \mathbf{h}(\omega_0,\psi)$, $\forall\psi$. With ideal focusing, the transformed array signal at $(\tau,\omega)\in\Omega_{(\tau_0,\omega_0)}$ is given by

$$\tilde{\mathbf{p}}(\tau,\omega) = \mathbf{T}(\omega,\omega_0)\mathbf{p}(\tau,\omega) = \mathbf{H}(\omega_0)\mathbf{s}(\tau,\omega), \quad (2)$$

where $\mathbf{H}(\omega_0) = [\mathbf{h}(\omega_0,\psi_0),...,\mathbf{h}(\omega_0,\psi_Q)]$ is a $2\times Q$ HRTF matrix. The smoothed spatial covariance matrix of the transformed array signal at (τ_0,ω_0) can be expressed as [11]

$$\begin{aligned}\mathbf{S}_{\mathbf{p}}(\tau_0,\omega_0) &= \sum_{(\tau,\omega)\in\Omega}\tilde{\mathbf{p}}(\tau,\omega)\tilde{\mathbf{p}}^H(\tau,\omega)\\ &= \mathbf{H}(\omega_0)\mathbf{S}_{\mathbf{s}}(\tau_0,\omega_0)\mathbf{H}^H(\omega_0), \quad (3)\end{aligned}$$

where H denotes the conjugate transpose, and $\mathbf{S}_{\mathbf{s}}(\tau_0,\omega_0) = \sum_{(\tau,\omega)\in\Omega}\mathbf{s}(\tau,\omega)\mathbf{s}^H(\tau,\omega)$ is the $Q\times Q$ cross-correlation matrix of the source signals, where $\mathbf{s}(\tau,\omega) = [s_0(\tau,\omega),...,s_{Q-1}(\tau,\omega)]^T$ is the source signals vector. After smoothing, $\mathbf{S}_{\mathbf{s}}(\tau_0,\omega_0)$ is expected to be of full rank with a low condition number, such that TF bins with a single dominant source can be identified by examining the rank of $\mathbf{S}_{\mathbf{p}}(\tau_0,\omega_0)$. The set of bins selected by the sigma-ratio test is

$$\mathcal{A}_{\text{SIGMA-RATIO}} = \left\{(\tau_0,\omega_0) : \frac{\sigma_1(\tau_0,\omega_0)}{\sigma_2(\tau_0,\omega_0)} > \mathcal{TH}_{\text{SIGMA-RATIO}}\right\}, \quad (4)$$

where $\sigma_1(\tau_0,\omega_0)$ and $\sigma_2(\tau_0,\omega_0)$ are the largest and second largest (which is also the smallest in the binaural case) singular values of $\mathbf{S}_{\mathbf{p}}(\tau_0,\omega_0)$ and $\mathcal{TH}_{\text{SIGMA-RATIO}}$ is a chosen threshold.

3.2. DPD tests not incorporating averaging

Tests that do not incorporate averaging include the directivity based DPD test [9] and the local space-domain distance (LSDD)-DPD test [22, 23]. The LSDD-DPD test [23] is presented next and is used hereafter for studying tests that do not employ averaging.

The LSDD-DPD test selects bins in which the microphone signal is similar to an HRTF, suggesting the existence of a single source. The Hermitian angle [24] between the microphone signal $\mathbf{p}(\tau,\omega)$ and an HRTF $\mathbf{h}(\omega,\psi)$ for an individual bin is used to quantify this similarity. The LSDD measure at (τ,ω) is defined as [23]

$$\text{LSDD}(\tau,\omega) = \min_{\psi}\left\{\cos^{-1}\left(\frac{\left|\mathbf{h}^H(\omega,\psi)\mathbf{p}(\tau,\omega)\right|}{\|\mathbf{h}(\omega,\psi)\|\,\|\mathbf{p}(\tau,\omega)\|}\right)\right\}, \quad (5)$$

where $\|\cdot\|$ is the 2-norm. The LSDD measure ranges between 0 and $\frac{\pi}{2}$, where low LSDD values indicate high similarity to a single source. TF bins with a low LSDD value are therefore assumed to be dominated by a single source. This assumption is examined in the next section. The set of bins selected by the LSDD-DPD test is

$$\mathcal{A}_{\text{LSDD}} = \{(\tau,\omega) : \text{LSDD}(\tau,\omega) < \mathcal{TH}_{\text{LSDD}}\}, \quad (6)$$

where $\mathcal{TH}_{\text{LSDD}}$ denotes a chosen threshold.

4. Analysis of DPD tests with and without TF averaging

The various DPD tests have been shown to perform well in the original papers with the arrays to which they were applied. However, as shown here, tests that do not incorporate averaging over TF bins seem to fail when applied to binaural arrays. This section presents an analysis of sigma-ratio and of the LSDD-DPD measures for two extreme cases, in which the microphone signals are dominated by the direct sound or by reverberation, in order to provide an insight into the effect of TF averaging in the binaural case.

For TF region $\Omega_{(\tau_0,\omega_0)}$ with a dominant direct part, the microphone signals at $(\tau,\omega)\in\Omega_{(\tau_0,\omega_0)}$ are similar and approximately equal to $\tilde{\mathbf{p}}(\tau,\omega)\approx\mathbf{h}(\omega_0,\psi_0)s_0(\tau,\omega)$. Therefore, the unit rank matrices $\tilde{\mathbf{p}}(\tau,\omega)\tilde{\mathbf{p}}^H(\tau,\omega)$, $(\tau,\omega)\in\Omega_{(\tau_0,\omega_0)}$ are also similar, leading to a spatial covariance matrix $\mathbf{S}_{\mathbf{p}}(\tau_0,\omega_0)$ of unit numerical rank and to a high sigma-ratio value. For

TF region $\Omega_{(\tau_0,\omega_0)}$ with a significant reverberant part, the microphone signals at $(\tau,\omega) \in \Omega_{(\tau_0,\omega_0)}$ approximately equal $\tilde{\mathbf{p}}(\tau,\omega) \approx \sum_{q=1}^{Q-1} \mathbf{h}(\omega_0,\psi_q) s_q(\tau,\omega)$. The reflections' amplitudes $\{s_q(\tau,\omega)\}_{q=1}^{Q}$ are affected by factors such as distance dependent attenuation and phase, which has a linear dependence on ω, and thus varying with ω. These variations imply that, for a TF region $\Omega_{(\tau_0,\omega_0)}$ with a sufficiently wide frequency range and with two or more dominant reflections, the vectors $\tilde{\mathbf{p}}(\tau,\omega)$ in $\Omega_{(\tau_0,\omega_0)}$ are likely to be diverse, which is likely to lead to a matrix $\mathbf{S}_{\mathbf{p}}(\tau,\omega)$ of full rank and to a small sigma-ratio value.

The LSDD measure, which does not employ averaging, is now analyzed. Direct-path bins will yield low LSDD values due to the small angle obtained for speaker direction. For bins with significant reverberation, the LSDD (or other measure that does not employ averaging and that is based on the similarity to a single source) will exhibit small values if there exists some direction ψ for which

$$\mathbf{p}(\tau,\omega) \approx z(\tau,\omega)\,\mathbf{h}(\omega,\psi), \tag{7}$$

where $z(\tau,\omega)$ is an arbitrary complex scalar. Let us assume, without loss of generality, that $\mathbf{h}(\omega,\psi)$ and $\mathbf{p}(\tau,\omega)$ are normalized such that their first entry equals 1, namely $\mathbf{h}(\omega,\psi) = \left[1, \Delta L_{\mathbf{h}}(\omega,\psi)\, e^{j\Delta\phi_{\mathbf{h}}(\omega,\psi)}\right]^T$ and $\mathbf{p}(\tau,\omega) = [1, \Delta L_{\mathbf{p}}(\tau,\omega)\, e^{j\Delta\phi_{\mathbf{p}}(\tau,\omega)}]^T$, where $\Delta L_{\mathbf{h}}(\omega,\psi) = \left|\frac{h^l(\omega,\psi)}{h^r(\omega,\psi)}\right|$ and $\Delta L_{\mathbf{p}}(\tau,\omega) = \left|\frac{p^l(\tau,\omega)}{p^r(\tau,\omega)}\right|$ and $\Delta\phi_{\mathbf{h}}(\omega,\psi) = \angle\frac{h^l(\omega,\psi)}{h^r(\omega,\psi)}$ and $\Delta\phi_{\mathbf{p}}(\tau,\omega) = \angle\frac{p^l(\tau,\omega)}{p^r(\tau,\omega)}$ are the ILDs and IPDs of $\mathbf{p}(\tau,\omega)$ and $\mathbf{h}(\omega,\psi)$, respectively. With this normalized representation it follows that (7) is maintained iff there exist some ψ for which

$$\Delta\phi_{\mathbf{p}}(\tau,\omega) \approx \Delta\phi_{\mathbf{h}}(\omega,\psi) \tag{8}$$

and

$$\Delta L_{\mathbf{p}}(\tau,\omega) \approx \Delta L_{\mathbf{h}}(\omega,\psi). \tag{9}$$

However, at frequencies for which the interaural distance d is larger than half a wavelength λ, typically above 1 kHz for a human head, $\Delta\phi_{\mathbf{h}}(\omega,\psi)$ can have any value between $-\pi$ and π, so that (8) is certainly met for at least one direction ψ^*. As frequency increases, the IPD is ambiguous and (8) is satisfied for an increasing number of directions $\psi_1^*, ..., \psi_N^*$. Therefore, at these frequencies, low LSDD values are obtained if (9) is satisfied for some $\psi^* \in \{\psi_1^*, ..., \psi_N^*\}$. This is a mitigating condition that may be satisfied even for bins with significant reverberation, leading to their selection, and consequently degrading the performance of tests that do not incorporate averaging. At frequencies for which $d < \frac{\lambda}{2}$, $\Delta\phi_{\mathbf{h}}(\omega,\psi)$ can take any value between $-\frac{2\pi d}{\lambda}$ and $\frac{2\pi d}{\lambda}$. However, $\Delta\phi_{\mathbf{p}}(\tau,\omega)$ also tends to be closer to zero due to the correlation between the reverberant signals $p^l(\tau,\omega)$ and $p^r(\tau,\omega)$, which increases as frequency decreases [25]. Therefore, (8) is likely to be maintained also for $d < \frac{\lambda}{2}$ and so the above analysis may be valid for the entire frequency range.

Incorporating averaging can lead to a stricter condition than (9), and, consequently, to a reduction in the false alarm rate. The modified LSDD measure is defined as

$$\text{modified LSDD}(\tau_0,\omega_0) = \min_{\psi}\left\{\frac{1}{J}\sum_{(\tau,\omega)\in\Omega_{(\tau_0,\omega_0)}} \cos^{-1}\left(\frac{\left|\mathbf{h}^H(\omega,\psi)\,\mathbf{p}(\tau,\omega)\right|}{\|\mathbf{h}(\omega,\psi)\|\,\|\mathbf{p}(\tau,\omega)\|}\right)\right\}, \tag{10}$$

where J denotes the cardinality of $\Omega_{(\tau_0,\omega_0)}$. From (10) it follows that in order for the modified LSDD measure to be small the angle between $\mathbf{p}(\tau,\omega)$ and $\mathbf{h}(\omega,\psi)$ should be small for the same ψ for each of the bins in $\Omega_{(\tau_0,\omega_0)}$. This condition is satisfied for $\Omega_{(\tau_0,\omega_0)}$ with a dominant direct part and it is unlikely to exist for $\Omega_{(\tau_0,\omega_0)}$ with a significant reverberant part for which the vectors $\mathbf{p}(\tau,\omega)$ in $\Omega_{(\tau_0,\omega_0)}$ are diverse.

5. Simulation study

The results in the previous section suggest that bins with significant reverberation may be selected by tests that do not incorporate averaging. The current section summarizes the simulations that have been conducted to examine the effect of averaging on a test performance. A scenario of a single speaker in a typical reverberant room was considered and the performance of the studied DPD tests was examined. For binaural arrays, estimating the DOA in 3D may be challenging due to the small number of microphones. Therefore, to prevent errors due to the fundamental limits of the array, a speaker was placed at the array frontal horizontal plane and only speaker azimuth was estimated.

5.1. Setup

Reverberant recordings due to a single speaker in a room were simulated. A rectangular room of dimensions $8 \times 5 \times 3\,\text{m}^3$ was simulated using the image method [26] with a wall reflection coefficient of $R = 0.92$ that leads to an approximate reverberation time of $T_{60} = 0.8\,\text{s}$. The Neumann KU-100 binaural array [27] was located at $[x, y, z] = [2, 1.5, 1.7]$ and the speaker, simulated as a point source, was placed 1.5 m away from the array at the same hight with azimuths varying from –70° to 70°, spaced by 5°. For each speaker location, a speech signal with a length of approximately 4 s, and a sampling frequency of 16 kHz, was randomly selected from a set of fifteen speech signals, that were taken from the TIMIT database [28]. Finally, white Gaussian sensor noise with an SNR of 30 dB was added to the binaural signal.

The binaural signal was transformed to the STFT domain using a 512 samples (32 ms) Hann window with an overlap of 16 ms. The operating frequency range for all tests was $0.5 - 8$ kHz. The sigma-ratio based DPD test was implemented with the focusing transformations proposed in [11]. The spatial covariance matrices were computed using (3) with a window $\Omega_{(\tau_0,\omega_0)}$ of 3 time frames and 15 frequencies. For the sigma-ratio test, speaker azimuth was estimated at the selected bins using the MUSIC algorithm [29] with a source subspace of a single dimension. The Neumann KU-100 HRTF data set, which includes HRTF samples from 2702 directions, was used for computing the LSDD measures [27]. The argument ψ in (5) and (10) was restricted to take only directions along the array frontal horizontal plane. The modified LSDD-DPD test measure was computed using (10) with a window $\Omega_{(\tau_0,\omega_0)}$ of 1 time frame and 15 frequencies. In both the LSDD-DPD test and its modification, speaker azimuth at the selected bins was estimated by the argument ψ that yields the minimum angle. For all tests, an energy threshold was employed, automatically rejecting 10% of bins with the lowest power, where the power at (τ,ω) was computed as $\left|p^l(\tau,\omega)\right| + \left|p^r(\tau,\omega)\right|$. The threshold of the various tests was set such that only a percentile of top-rated bins will pass the test. This approach to threshold selection was chosen so that the different tests, which use a diverse set of measures, could be evaluated on a common basis.

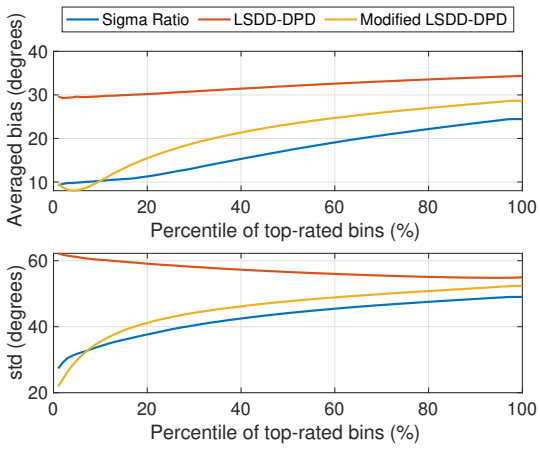

Figure 1: *Averaged (over speaker directions) bias and standard deviation of azimuth estimates from selected bins as a function of the percentile of selected bins.*

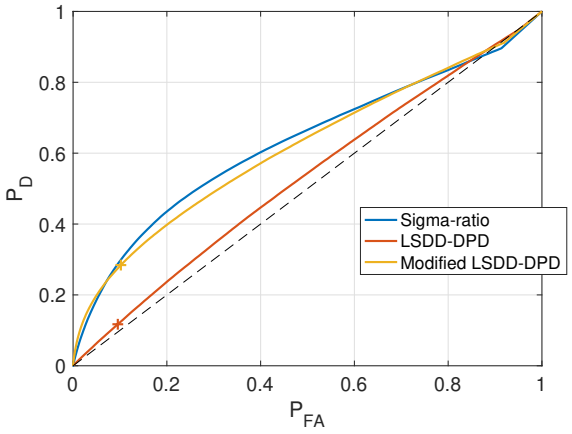

Figure 2: *ROC for the studied tests.*

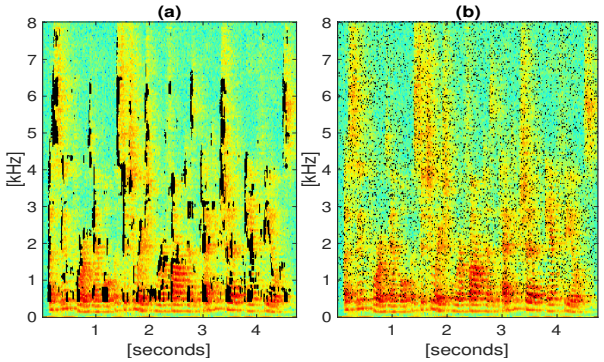

Figure 3: *Spectrogram of the signal at the left ear. Top 10%: (a) modified LSDD, (b) LSDD; selected bins are marked in black.*

5.2. Results

The averaged (over speaker directions) bias and standard deviation of azimuth estimates from selected bins with the studied tests is plotted in Fig. 1 as a function of the percentile of top-rated bins. Figure 1 shows that the performance of the LSDD-DPD test is inferior to that of the sigma-ratio and the modified LSDD-DPD tests. Figure 1 also shows that the standard deviation in the LSDD test increases as the percentage of selected bins decreases. These results suggest that the LSDD measure is not reliable for identifying the direct-path bins in the binaural case, validating the arguments presented in Section 4.

To specifically investigate the false alarm rate, the receiver operating characteristic (ROC) of the studied tests is examined. A two hypotheses detection problem is defined with $\mathcal{H}_0$: DRR $>$ 3 dB and $\mathcal{H}_1$: DRR $\leq$ 3 dB. The DRR of the sound pressure at the origin (the center of the head) in free-field is computed at (τ, ω) as $10 \log_{10} \frac{|a^d_{00}(\tau,\omega)|^2}{|a^r_{00}(\tau,\omega)|^2}$, where $a^d_{00}(\tau, \omega)$ and $a^r_{00}(\tau, \omega)$ denote the zeroth order spherical harmonics coefficient of the plane-wave density at the origin due to the direct and the reverberant part, respectively. Letting $\mathcal{A}(\mathcal{TH})$ denote the set of selected bins and $\mathcal{A}_{\text{DRR}>3}$ denote the set of bins with DRR $>$ 3, the probability of detection and the probability of false alarm can be assessed by defining

$$P_D(\mathcal{TH}) = \frac{|\mathcal{A}(\mathcal{TH}) \cap \mathcal{A}_{\text{DRR}>3}|}{|\mathcal{A}_{\text{DRR}>3}|} \tag{11}$$

$$P_{FA}(\mathcal{TH}) = \frac{|\mathcal{A}(\mathcal{TH}) \cap \overline{\mathcal{A}_{\text{DRR}>3}}|}{|\overline{\mathcal{A}_{\text{DRR}>3}}|}, \tag{12}$$

where $|\cdot|$ and $\overline{(\cdot)}$ denote the cardinality and the complement of a set, respectively. Figure 2 depicts the ROC of the studied tests. The dashed line denotes the 45° ROC, which can be attained by a detector that randomly selects x percent of the bins, ignoring all data, leading to $P_{FA} = P_D = \frac{x}{100}$. Figure 2 shows that the ROC of the LSDD-DPD test is close to the 45° ROC, and that the ROCs of the sigma-ratio and modified LSDD-DPD tests are higher, implying that for a given detection rate their false alarm rate will be lower than that of the LSDD-DPD test. The plus marks on the ROCs of the LSDD-DPD test and its modification correspond to thresholds that lead to a selection of a 10% percentile of top-rated bins. These points correspond to a false alarm of about 0.1 and to detection of about 0.1 for the LSDD-DPD test and about 0.25 for its modification.

Figure 3 depicts the bins selected by the LSDD-DPD test and its modification with these thresholds (10% false-alarm rate) for one of the simulated scenarios in which the speaker was located at an azimuth of 70°. Figure 3 shows that the bins selected by the modified LSDD-DPD test are concentrated around speech onsets, where the direct path tends to be dominant, while the bins selected by the LSDD-DPD test are spread over different TF regions, suggesting that bins with significant reverberation are selected.

6. Conclusions

DPD tests that do not incorporate averaging have been shown to perform well with the original arrays on which they were applied. The current work has highlighted the weaknesses of these tests for binaural arrays. The cause of this failure was shown to be the similarity of bins with significant reverberation to bins with single source signals, leading to their selection by the test. This similarity is shown to occur occasionally, leading to a high false alarm rate. It is further demonstrated that incorporating averaging over TF bins can improve the performance.

7. Acknowledgments

This work was supported by Facebook Reality Labs.

8. References

[1] Ozgur Yilmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[2] Martin Raspaud, Harald Viste, and Gianpaolo Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *Ieee transactions on audio, speech, and language processing*, vol. 18, no. 1, pp. 68–77, 2009.

[3] Stanley T Birchfield and Rajitha Gangishetty, "Acoustic localization by interaural level difference," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* IEEE, 2005, vol. 4, pp. iv–1109.

[4] Ning Ma, Tobias May, and Guy 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)*, vol. 25, no. 12, pp. 2444–2453, 2017.

[5] Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.

[6] Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1997–2012, 2017.

[7] Xiaofei Li, Bastien Mourgue, Laurent Girin, Sharon Gannot, and Radu Horaud, "Online localization of multiple moving speakers in reverberant environments," in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. IEEE, 2018, pp. 405–409.

[8] Or Nadiri and Boaz 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*, vol. 22, no. 10, pp. 1494–1505, 2014.

[9] Boaz Rafaely and Koby Alhaiany, "Speaker localization using direct path dominance test based on sound field directivity," *Signal Processing*, vol. 143, pp. 42–47, 2018.

[10] Lior Madmoni and Boaz Rafaely, "Direction of arrival estimation for reverberant speech based on enhanced decomposition of the direct sound," *IEEE Journal of Selected Topics in Signal Processing*, 2018.

[11] Hanan Beit-On and Boaz 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)*. IEEE, 2018, pp. 1–4.

[12] Andreas Brendel, Chengyu Huang, and Walter Kellermann, "Stft bin selection for localization algorithms based on the sparsity of speech signal spectra," *ratio*, vol. 2, pp. 6, 2018.

[13] Hao Wang and Jing Lu, "A robust doa estimation method for a linear microphone array under reverberant and noisy environments," *arXiv preprint arXiv:1904.06648*, 2019.

[14] Shaowei Ding and Huawei Chen, "Doa estimation of multiple speech sources by selecting reliable local sound intensity estimates," *Applied Acoustics*, vol. 127, pp. 336–345, 2017.

[15] Sina Hafezi, Alastair H Moore, and Patrick A Naylor, "Spatial consistency for multiple source direction-of-arrival estimation and source counting," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4592–4603, 2019.

[16] Hanan Beit-On and Boaz Rafaely, "Binaural direction-of-arrival estimation in reverberant environments using the direct-path dominance test," in *23nd International Congress on Acoustics (ICA)*, Aachen, Germany, Sept. 2019.

[17] Yekutiel Avargel and Israel Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[18] Boaz Rafaely, Christopher Schymura, and Dorothea Kolossa, "Speaker localization in a reverberant environment using spherical statistical modeling," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3523–3523, 2017.

[19] Boaz Rafaely and Dorothea Kolossa, "Speaker localization in reverberant rooms based on direct path dominance test statistics," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 6120–6124.

[20] Symeon Delikaris-Manias, Despoina Pavlidi, Ville Pulkki, and Athanasios Mouchtaris, "3d localization of multiple audio sources utilizing 2d doa histograms," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1473–1477.

[21] Boaz Rafaely, Dorothea Kolossa, and Yanir Maymon, "Towards acoustically robust localization of speakers in a reverberant environment," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 96–100.

[22] Vladimir Tourbabin, David Lou Alon, and Ravish Mehra, "Space domain-based selection of direct-sound bins in the context of improved robustness to reverberation in direction of arrival estimation," in *Proc. 11th European Congress and Exposition on Noise Control Engineering (EURONOISE18)*, 2018, pp. 2589–2596.

[23] Vladimir Tourbabin, Jacob Donley, Boaz Rafaely, and Ravish Mehra, "Direction of arrival estimation in highly reverberant environments using soft time-frequency mask," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 383–387.

[24] K. Scharnhorst, "Angles in complex vector spaces," *Acta Applicandae Mathematica*, vol. 69, no. 1, pp. 95–103, Oct. 2001.

[25] Boaz Rafaely, "Spatial-temporal correlation of a diffuse sound field," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3254–3258, 2000.

[26] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[27] Benjamin 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*. AIA/DAGA, 2013, p. 29.

[28] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[29] Ralph Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.