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

Hanan Beit-On1, Vladimir Tourbabin2, Boaz Rafaely3,

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

\[
p(\tau, \omega) = p_d(\tau, \omega) + p_r(\tau, \omega) = h(\omega, \psi_0)s_0(\tau, \omega) + \sum_{q=1}^{Q-1} h(\omega, \psi_q)s_q(\tau, \omega),
\]

where \(\tau\) and \(\omega\) denote the time and the frequency indices, respectively. \(p(\tau, \omega) = [p^1(\tau, \omega), p^2(\tau, \omega)]^T\), where \(p^1(\tau, \omega)\) and \(p^2(\tau, \omega)\) are the STFT of the left and right microphone signals, respectively. \(s_q(\tau, \omega)\) denotes the STFT of the \(q\)-th source signal that originates from direction \(\psi_q\), and \(h(\omega, \psi_q)\) is the head-related transfer function (HRTF) from \(\psi_q\). The formulation in (1) assumes that the multiplicative transfer function approximation (MTF) holds, i.e., the length of the HRFT filters \(h(\omega, \psi)\) (in time), is significantly shorter than the length of the STFT window. The representation in (1)
can be decomposed into \( p_\tau(\tau, \omega) = h(\omega, \psi_0)s_0(\tau, \omega) \) and \( p_\tau(\tau, \omega) = \sum_{q=1}^{Q-1} h(\omega, \psi_q)s_0(\omega, \tau) \), 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 \([15, 19, 20, 21]\).

### 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. This assumption also leads to a high sigma-ratio value. For bins with a low LSDD value, the set of bins selected by the sigma-ratio test is

\[
A_{\text{SIGMA-RATIO}} = \left\{ (\tau_0, \omega_0) : \frac{\sigma_1(\tau_0, \omega_0)}{\sigma_2(\tau_0, \omega_0)} > T_{\text{SIGMA-RATIO}} \right\},
\]

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 \(S_p(\tau_0, \omega_0)\) and \(T_{\text{SIGMA-RATIO}}\) 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 (LSD) DPD test \([22, 23]\). The LSD-DPD test \([23]\) is presented next and is used hereafter for studying tests that do not employ averaging.

The LSD-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 \(p(\tau, \omega)\) and the HRTF \(h(\omega, \psi)\) for an individual bin is used to quantify this similarity. The LSD measure at \((\tau, \omega)\) is defined as \([23]\)

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

where \(\|\cdot\|\) is the 2-norm. The LSD measure ranges between 0 and \(\frac{\pi}{2}\), where low LSD values indicate high similarity to a single source. TF bins with a low LSD 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 LSD-DPD test is

\[
A_{\text{LSD}} = \{ (\tau, \omega) : \text{LSDD}(\tau, \omega) < T_{\text{LSD}} \},
\]

where \(T_{\text{LSD}}\) 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 LSD-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 \(p(\tau, \omega) \approx h(\omega_0, \psi_0)s_0(\tau, \omega)\). Therefore, the unit rank matrices \(p(\tau, \omega)p^H(\tau, \omega)\), \((\tau, \omega) \in \Omega_{(\tau_0, \omega_0)}\) are also similar, leading to a spatial covariance matrix \(S_p(\tau_0, \omega_0)\) of unit numerical rank and to a high sigma-ratio value. For
The results in the previous section suggest that bins with significant reverberation may be selected by tests that do not incorporate rate 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 × 5 × 3 m³ was simulated using the image method [26] with a wall reflection coefficient of \( R = 0.92 \) that leads to an approximate reverberation time of \( T_\text{60} = 0.8 \) 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 height with azimuths varying from \( -70^\circ \) to \( 70^\circ \), spaced by \( 5^\circ \). 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 order of one. The Neumann KU-100 HRTF data set, which includes HRTF samples from 2702 directions, was used for computing the LSDD measure [22]. The argument \( \psi \) in (3) 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 \( \psi \) 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 \((\tau, \omega)\) was computed as \( |p^\ell(\tau, \omega)| + |p^\ell(\tau, \omega)|^2 \). 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.
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 two hypotheses, denoted by $H_0$ and $H_1$, where $H_0$ corresponds to the null hypothesis that no sound is present, and $H_1$ corresponds to the alternative hypothesis that a sound is present. The DRR of the plane-wave density at the origin due to the direct and reverberant parts is defined. A two hypotheses detection problem is defined with two hypotheses, denoted by $H_0$ and $H_1$, where $H_0$ corresponds to the null hypothesis that no sound is present, and $H_1$ corresponds to the alternative hypothesis that a sound is present. The ROCs of the sigma-ratio and modified LSDD-DPD tests are plotted in Fig. 1 as a function of the 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 2 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


