# Space domain-based selection of direct-sound bins in the context of improved robustness to reverberation in direction of arrival estimation

Vladimir Tourbabin, David Lou Alon, Ravish Mehra
Oculus & Facebook

**Summary**
Acoustic reverberation has a detrimental effect on many of the currently used microphone array processing methods. In particular, the performance of most direction of arrival estimation approaches is significantly degraded in the presence of coherent reflections. One general approach to improving the robustness to reverberation is to operate in the short-time Fourier transform domain and select the bins dominated by the direct sound, while rejecting those contaminated by reflections. Following this general idea, a number of effective techniques have been proposed recently, many of which exploit spherical arrays and the spherical harmonics processing framework. Those approaches include the direct-path dominance test [1], the computationally-efficient local directivity-based method [2], and a method based on pseudo-intensity vectors with estimation consistency weighting [3]. While these methods are generally effective, they impose hardware and frequency-bandwidth constraints associated with the spherical harmonics domain processing framework. In the current paper, an alternative method is proposed to extract the direct-path time-frequency bins. The method computes the local space-domain distance spectrum [4] for each bin individually, and selects those which have the most clearly distinguishable peaks. This method operates directly in the space domain, thereby effectively avoiding the limitations associated with the spherical array processing framework. The performance of the proposed local space-domain distance approach is demonstrated here by comparing it to the state-of-the-art direct-path dominance method. The ability of the two approaches to select the direct-path bins is analyzed directly by comparing their outcomes to a ground-truth direct-to-reverberant ratio. It is demonstrated that the proposed approach outperforms the state-of-the-art method, while operating entirely in the space domain, and may provide improvement in the average direct-to-reverberant ratio as high as 30 dB.

**PACS no. 43.60.+d**

## 1. Introduction

Direction of Arrival (DoA) estimation is a fundamental microphone array processing method; it is frequently employed in acoustic scene analysis, signal enhancement, and speech processing [5, 6]. In many applications, DoA estimation is required to operate in reverberant environments, such as homes and offices. The multi-path nature of sound propagation in these environments is challenging to most DoA estimation approaches, due to confusion between the direct and the reflected sound and due to the temporal correlation between the two.

One general approach to increasing robustness to reverberation is by operating only on the signal segments that predominantly contain the direct-path sound, while rejecting those contaminated by reflections [1, 7]. For that purpose the array signal is usually transformed into a time-frequency representation, such as the Short-Time Fourier Transform (STFT) domain, followed by selection of the time-frequency bins dominated by the direct path energy. Using this approach, the robustness to reverberation is largely determined by the ability of the algorithm to correctly identify the direct-path time-frequency bins.

Recently, several effective approaches have been proposed for selecting the direct-path time-frequency bins by exploiting the unique properties of the Spherical Harmonics (SH) domain representation. Examples of these methods include the Direct-Path Dominance (DPD) test with frequency smoothing [1], the SH sound-field directivity test [2], and the estimation consistency test [3]. While being generally effective, the SH domain-based solutions require a relatively large number of microphones and impose constraints

(c) European Acoustics Association

on array geometry and the operating frequency band. An attempt to extend the operating frequency range of the DPD test has been proposed in [4], while still partially relying on the SH domain framework. In addition, a space-domain[1] method for time-frequency bins selection is described in [7]. However, this method is specifically tailored for differential microphone arrays and is limited to one-dimensional DoA estimation.

In the current work, a new effective method is proposed to select the direct-path TF bins. The proposed method operates entirely in the space domain and, therefore, effectively avoids the limitations related to SH domain processing. The proposed method draws its inspiration from the Space-Domain Distance (SDD) algorithm described in [4]; it is based on calculation of the Local Space-Domain Distance (LSDD) spectrum for each time-frequency bin independently, followed by selection of the bins with the most clearly distinguishable peaks.

An additional important contribution of the current paper is a new method for general performance analysis and comparison between different direct-path selection approaches. The method is based on the calculation of the ground-truth Direct-to-Reverberant Ratio (DRR) in each time-frequency bin, which is accomplished through the separation of the Room Impulse Response (RIR) into the direct and the reverberant parts.

The remainder of the paper is organized as follows. General definitions and assumptions are presented in Section 2, followed by a description of the proposed method for the direct-path time-frequency bins selection in Sections 3 and 4. Next, in Section 5 an approach for performance assessment of the direct-path selection method is introduced. Using this approach, the proposed direct-path selection method is compared to the state-of-the-art method described in [1] using real-world recorded RIRs. Suggestions for future work and conclusions follow.

## 2. Signal model and the DRR

The current section introduces the signal model, notations, and basic assumptions that are used throughout the paper. In addition, the narrow-band instantaneous DRR is also explicitly defined in this section and its basic properties are briefly discussed.

Consider an $M$-microphone array of arbitrary geometry. The STFT of the signal received by the array in a reverberant environment within time frame $t$ and in frequency bin $f$ can be modeled as

$$\begin{aligned}\mathbf{x}(t,f) &= \mathbf{x}_d(t,f) + \mathbf{x}_r(t,f) \\ &= s_0(t,f)\mathbf{v}(f,\Omega_0) + \sum_{i=1}^{I} s_i(t,f)\mathbf{v}(f,\Omega_i), \quad (1)\end{aligned}$$

[1] Sometimes refered to as the element-space domain.

where $\mathbf{x}_d(t,f)$ and $\mathbf{x}_r(t,f)$ are $M$-dimensional complex-valued vectors that denote the direct and the reverberant parts of the received signal, respectively. Vector $\mathbf{v}(f,\Omega_i)$ is the array manifold (a.k.a. steering vector) indicating the array response to a unit-amplitude wave in frequency bin $f$ and arriving from direction $\Omega_i$, with $\Omega_0$ and $\{\Omega_i\}_{i=1}^{I}$ denoting the arrival directions of the direct and the reverberant sound components, respectively. It is emphasized that $\Omega_i = (\theta_i, \phi_i)$ stands for both elevation and azimuth, and the one-symbol notation is used here for convenience purposes only. Finally, scalars $\{s_i(t,f)\}_{i=0}^{I}$, denote the amplitude of the components arriving from the different directions; they contain factors such as the source amplitude, effect of reflectors, and distance-dependent phase and attenuation. Note that, in practice, a large number of reflections, $I$, may be required in order to obtain an accurate representation. Finally, the set of all time-frequency bins under consideration is given by $\mathcal{D}_{TF} = \{(t,f)|t = 1...T;\ f = 1...F\}$ for some non-zero $T$ and $F$.

Next, using the notation in (1), the DRR in a given time-frequency bin is defined here as

$$\begin{aligned}\mathrm{DRR}(t,f) &= 20\log_{10}\left(\frac{\|\mathbf{x}_d(t,f)\|}{\|\mathbf{x}_r(t,f)\|}\right)\ [dB] \\ &= 20\log_{10}\left(\frac{\|s_0(t,f)\mathbf{v}(f,\Omega_0)\|}{\left\|\sum_{i=1}^{I} s_i(t,f)\mathbf{v}(f,\Omega_i)\right\|}\right), \quad (2)\end{aligned}$$

where $\|\cdot\|$ denotes the vector 2-norm operator. Note from (2) that the DRR is largely determined by the amplitudes of the different components, $s_i(t,f)$, which depend on time and frequency. One case of particular importance is the source signal onset. Assuming that the time interval between the arrivals of the direct path and the first reflections is greater than the STFT step, the direct sound amplitude, $s_0(t,f)$, during an onset is expected to be significantly greater than the reflection amplitudes, $\{s_i(t,f)\}_{i=1}^{I}$. This, in turn, implies that a relatively high DRR is expected around signal onset in the frequency bins carrying significant energy.

The next section employs the notations and the definitions outlined here to introduce the newly proposed method for selection of the time-frequency bins with high DRR.

## 3. Selection of direct-path bins

The current section introduces the new method for selection of the time-frequency bins dominated by the direct-path component. The selection criterion is based on the SDD spectrum calculated locally for each bin.

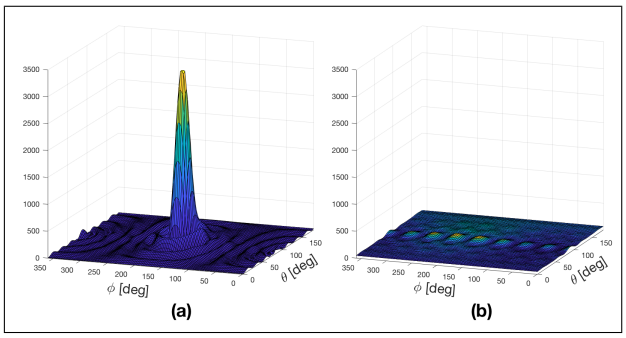


Figure 1. Examples of the LSDD spectra obtained using speech recorded with an 8-microphone array at the frequency 812 Hz. Two different cases are shown: (a) high DRR bin - 28 dB, (b) low DRR bin - 0 dB.

Consider a time-frequency bin $\mathbf{x}(t,f)$; its LSDD spectrum over a grid of arrival directions, $\{\Theta\}_{j=1}^{J}$, is defined as follows [4]:

$$S_{t,f}(\Theta_j) = \frac{1}{d(\mathbf{x}(t,f), \mathbf{v}(f,\Theta_j))},\ j = 1...J, \quad (3)$$

where $d(\cdot,\cdot)$ measures similarity between two vectors and is defined as [4]:

$$d(\mathbf{a},\mathbf{b}) = \min_{\alpha} \left( \frac{\|\mathbf{a} - \alpha\mathbf{b}\|}{\|\mathbf{a}\|} \right). \quad (4)$$

This quantity is symmetric and, in fact, measures the sine of the angle between the two vectors of the same arbitrary dimension. The grid of arrival directions, $\{\Theta\}_{j=1}^{J}$, over which the spatial spectrum is calculated, must be chosen carefully to ensure that it is dense enough and covers all directions of interest. The array manifold over that grid can be obtained in a number of different ways, including theoretical modeling [8], numerical simulation [9], and direct measurements [10].

In order to see how the LSDD spectrum defined in (3) may help selecting the high DRR bins, suppose first that $\mathbf{x}(t,f)$ contains only the direct-sound component, i.e. $\mathbf{x}(t,f) = s_0\mathbf{v}(f,\Omega_0)$. In this case, it is straightforward to see that $d(\mathbf{x}(t,f), \mathbf{v}(f,\Theta_j)) = 0$ for some $j$, provided that the grid $\{\Theta_j\}_{j=1}^{J}$ was chosen to contain the direct-path arrival directions. Hence, in this case, the spectrum, $\{\mathbf{S}_{t,f}(\Theta_j)\}_{j=1}^{J}$, is expected to have an infinite peak in the arrival direction of the direct sound. In practice, the peak is not likely to be infinite because $\mathbf{x}(t,f)$ may contain residual reverberant components or there may be a small mismatch between the direct-path arrival direction and the corresponding grid direction. Nevertheless, the spectrum in this case is expected to contain a strong clearly distinguishable peak. A typical example of such a spectrum using real-world recordings is shown in Fig. 1a. On the other hand, in the case where $\mathbf{x}(t,f)$ contains a significant amount of reverberant energy, the distance $d(\mathbf{x}(t,f),\Theta_j)$ is not guaranteed to be zero for any of the grid directions $\{\Theta_j\}_{j=1}^{J}$. This, in turn, implies that the LSDD spectrum of $\mathbf{x}(t,f)$ in this case is not guaranteed to have a discernible peak in any direction. A typical example of the LSDD spectrum of a bin with 0 dB DRR is shown in Fig. 1b. By observing this figure, the reader may notice several peaks which are hardly distinguishable from the spectrum ripple level, as opposed to the single clearly distinguishable peak in the high DRR case in Fig. 1a.

The apparent difference in the relative peak strengths of the LSDD spectra between the two cases discussed above forms the basis for the direct-path selection technique proposed next. Consider the LSDD spectrum, $\{S_{t,f}(\Theta_j)\}_{j=1}^{J}$, corresponding to the time-frequency bin $\mathbf{x}(t,f)$. It is proposed here to measure the relative strength of the highest peak of the spectrum using the following ratio:

$$R_{t,f} = \frac{S_{t,f}(\Theta_{\bar{j}}) - S_{t,f}(\Theta_{\underline{j}})}{\frac{1}{J-2}\sum\limits_{j\neq\bar{j},\underline{j}} S_{t,f}(\Theta_j) - S_{t,f}(\Theta_{\underline{j}})}, \quad (5)$$

where $\bar{j}$ and $\underline{j}$ are the indices corresponding to the maximum and the minimum values of the spectrum, respectively. The expression in (5) can be thought of as a ratio between the strongest peak height to average spectrum height excluding the peak. Finally, using the relative peak strength, $R_{t,f}$, a criterion for selecting the set of high DRR bins can be formulated as follows:

$$\mathcal{D}^{\star} = \{(t,f) | R_{t,f} > R_{th}\}, \quad (6)$$

with $R_{th}$ indicating a threshold peak strength, to be chosen in accordance with system parameters.

Recall that calculation of the relative peak strength of the LSDD spectrum resembles the directivity-based method proposed in [2]. Similarly to the relative peak strength, the directivity is expected to be high for direct-path bins and low for bins with a significant contribution from multiple reverberant components. However, in the case of a small number of dominant reflections, the directivity may remain high, as opposed to the relative peak strength of the LSDD spectrum, which is expected to resemble the example in Fig. 1b even with a single dominant reflection.

A discussion providing insights into the choice of $R_{th}$ and a performance analysis of the proposed method are presented in Section 5.

## 4. Algorithm summary

The current section summarizes the algorithm proposed above. The algorithm requires two inputs:

- STFT of the audio buffer captured with the array - $\mathbf{x}(t,f)$, $(t,f) \in \mathcal{D}_{TF}$,
- manifold vector of the array over a grid of viable arrival directions - $\mathbf{v}(f,\Theta_j)$, $j = 1...J$; $f = 1...F$.

**Algorithm 1** LSDD algorithm

input $\mathbf{x}(t,f),\ (t,f) \in \mathcal{D}_{TF}$
input $\mathbf{v}(f,\Omega_j),\ f = 1...F,\ j = 1...J$
init $\mathcal{D}^\star \leftarrow \varnothing$
**for** all $(t,f) \in \mathcal{D}_{TF}$ **do**
    compute $S_{t,f}(\Omega_j),\ j = 1...J$ [use Eq. (3)]
    compute $R_{t,f}$ [use Eq. (5)]
    **if** $R_{t,f} > R_{th}$ **then** $\mathcal{D}^\star \leftarrow \mathcal{D}^\star \cup (t,f)$
**end**
output $\mathcal{D}^\star$

Given the two inputs, one can proceed as outlined in Algorithm 1.

The best choice for the value of the parameter $R_{th}$ may depend on the STFT parameters and the DoA estimation algorithm that subsequently operates on $\mathcal{D}^\star$. Experiments performed in the course of the current work lead to the recommendation to choose $R_{th}$ corresponding to a percentile of top-rated bins in the range of 1% to 10%. A deeper exploration of this question is suggested for future work in Section 6.

One possible choice for the DoA estimation method that may operate on the set of the high-DRR bins, $\mathcal{D}^\star$, is the SDD algorithm originally proposed in [4]. According to this algorithm, the DoA estimate is given by:

$$\hat{\Omega} = \underset{\Omega_j}{\operatorname{argmax}} \left\{ \sum_{(t,f)\in\mathcal{D}^\star} S_{t,f}(\Omega_j) \right\}. \tag{7}$$

Note that (7) is not an integral part of the reverberation-robustness improvement scheme proposed in the current work; it is included in this section only for completeness. In fact, many other methods, including maximum likelihood [11], subspace [12], and beamforming-based [13] algorithms can operate on the set of the high-DRR bins, $\mathcal{D}^\star$, and, thereby, benefit from the improved robustness to reverberation.

Finally, it is emphasized that although the examples provided in this paper assume a single source, the LSDD algorithm is strictly applicable to a multi-source scenario. In this case, the resulting set $\mathcal{D}^\star$ is expected to contain multiple subsets, each composed of high-DRR bins corresponding to a different active source.

## 5. Experimental study

The aim of the current section is to study the performance of the proposed LSDD method using real-world recorded data. The experimental setup is described in the first subsection. Then, the ability of the LSDD method to select the high DRR bins is discussed. For this purpose, an approach for obtaining a ground-truth DRR is first described in Subsection 5.2 and, then, used in Subsection 5.3.

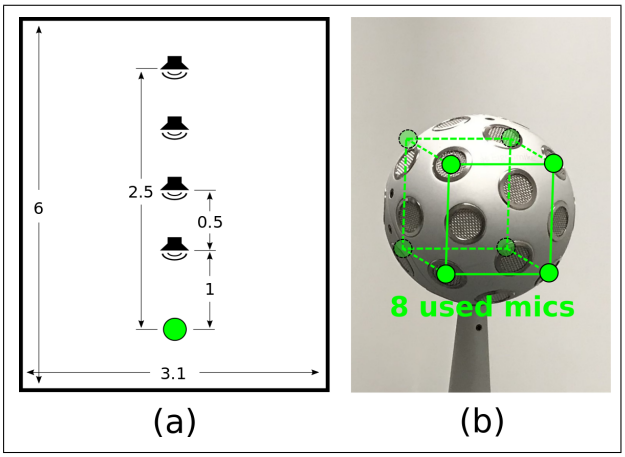


Figure 2. Illustration of the experimental setup: (a) source and receiver (green circle) orientation within the room, (b) 8-element microphone array. Dimensions are in meters.

### 5.1. Experimental setup

The discussion in this section is based on recordings made in an empty rectangular room with the dimensions of $6 \times 3.1 \times 2.9$m and a relatively high wide-band reverberation time of $T_{60} = 1.2$ sec. A Genelec 8340A loudspeaker served as a source. The recordings were made with the source located at 1, 1.5, 2, and 2.5 m from the receiver, as illustrated in Fig. 2a. The recordings were carried out using the Eigenmike® array by mhacoustics. In order to demonstrate the ability of the methods proposed in this work to operate with a limited number of microphones, only 8 elements were used in all of the examples in the current paper. These 8 elements were selected to form a cube with an edge size of approximately 49 mm, as shown in Fig. 2b. In order to obtain the RIRs, the loudspeaker was set to play a logarithmic sweep sine. The impulse responses were estimated from the recorded signals using time-domain deconvolution, as described in [14]. The RIRs are used in Subsections 5.2 and 5.3 for the analysis of the ability of the proposed method to select the high DRR bins.

### 5.2. Ground-truth DRR

Here, it is suggested to obtain the ground-truth DRR by splitting the array RIR into its direct and reverberant parts using time windowing. The approach is illustrated in Fig. 3, which shows a single RIR obtained as described above. The response is shown along with its breakdown into the direct and the reverberant parts. The separation into the two parts can benefit from knowledge of the room geometry and using it to calculate the anticipated arrival times of the early reflections. Next, the two parts can be separately convolved with a dry source signal and, then, transformed into the STFT domain. By doing so for all microphones, one can obtain the direct part, $\mathbf{x}_d(t,f)$, and the reverberant part, $\mathbf{x}_r(t,f)$, of the array signal referred to in

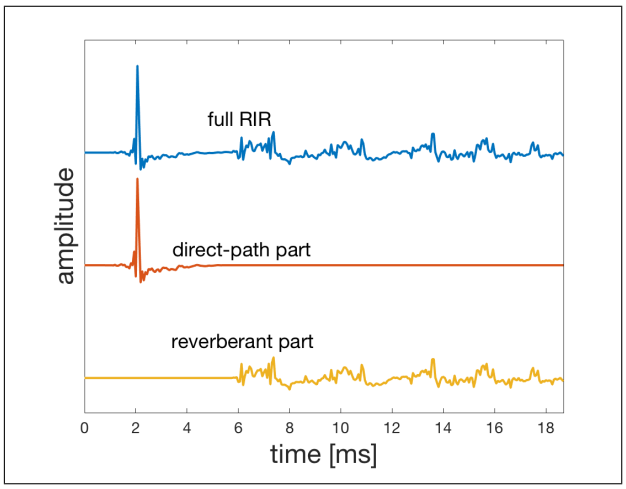


Figure 3. Illustration of an RIR measured with one of the microphones and its breakdown into the direct and the reverberant part obtained through time windowing.

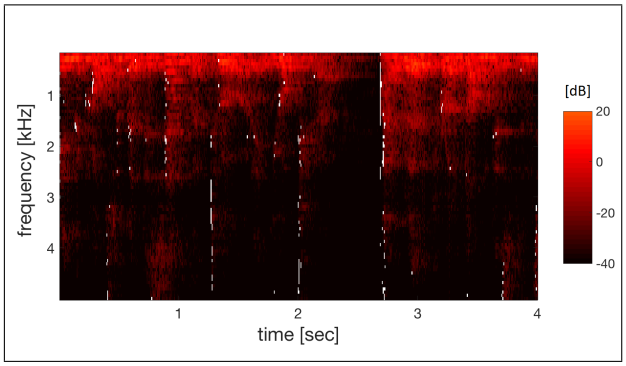


Figure 4. Spectrogram of reverberant speech sampled at 16kHz with frame length of 256 samples and 50% overlap. The bins having DRR $>$ 10dB are indicated in white color.

the signal model in (1). Finally, the DRR can be obtained by computing the ratio of $\mathbf{x}_d(t,f)$ to $\mathbf{x}_r(t,f)$, as suggested by (2).

Using the array RIR and a dry speech recording from [15], the DRR was calculated as described above. To demonstrate the resulting time-frequency distribution of the high DRR bins, the spectrogram of the signal in one of the microphones is plotted in Fig. 4, in which the bins with DRR $>$ 10 dB are overlaid in white color. It demonstrates that the high DRR bins are most common around signal onsets, as expected due to the belated arrival of reflections and rapid sound decay from onset to onset.

Another useful way to analyze the DRR data is by computing the percentage of the time-frequency bins whose corresponding DRR is greater than a certain threshold. A plot of the percentage as a function of the threshold value is shown in Fig. 5. It can be seen that, under the conditions in which the recordings were made, only 10% of the bins pass the 0 dB DRR threshold, and less than 1% pass the 10 dB threshold.

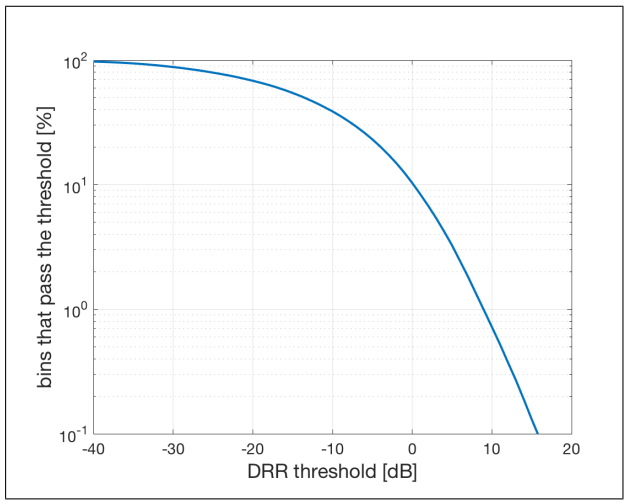


Figure 5. Percentage of bins having a DRR higher than a certain threshold. Based on a spectrogram of reverberant speech of almost 20 seconds, sampled at 16kHz with a frame length of 256 samples and 50% overlap - 300 thousand bins in total. The reverberant speech was produced using impulse responses of a room having a relatively high reverberation time of 1.2 sec with source-receiver separation of 2 m.

### 5.3. DRR improvement

In the current subsection, the ground-truth DRR is exploited in order to analyze the ability of the LSDD method to select the bins dominated by the direct-path component. Furthermore, the LSDD method is compared here to another state-of-the-art method, namely the DPD test [1]. Both methods attempt to assess the DRR of each time-frequency bin by computing a corresponding metric. The metric used by the LSDD is the relative peak strength, as defined in (5), while the metric used by the DPD test is the ratio between the first and the second singular values of the local correlation matrix, as introduced in [1]. Using the above experimental data, the ground-truth DRR and the metrics of both methods were calculated. The DPD test applied here involved the first-order SH decomposition and 8 nearest neighbors for local covariance estimation. A typical example of the resulting ground-truth DRR obtained with a source distance of 1 m and a frequency range of $1500-1750$ Hz is plotted in Fig. 6 versus both of the metrics. The 250 Hz band corresponds to 4 adjacent frequency bins of the spectrogram. From this example, it can be seen that the time-frequency bins can be loosely divided into two general groups. The first group, having DRR $\leq 0$ dB, appear to be largely uncorrelated with the metrics of both methods. Nevertheless, the bins in the second group with DRR $>$ 0 dB display observable correlation. This correlation forms the basis for the ability of both methods to sift the bins with high DRR.

Recall that both methods (LSDD and DPD) suggest to select the high DRR bins by rejecting all the bins whose corresponding metric is lower than a cer-

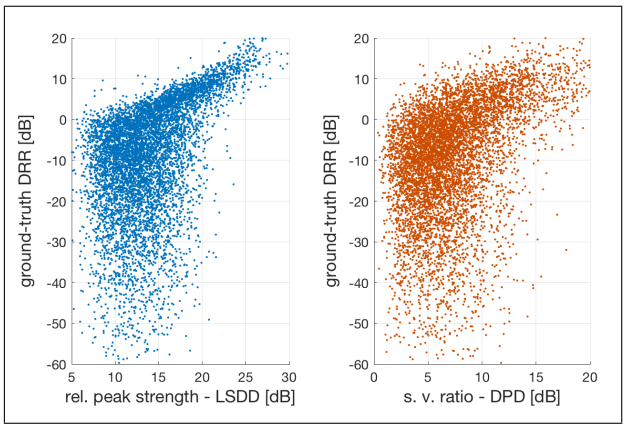


Figure 6. Ground-truth DRR versus the corresponding metrics of (a) the LSDD and (b) the DPD methods.

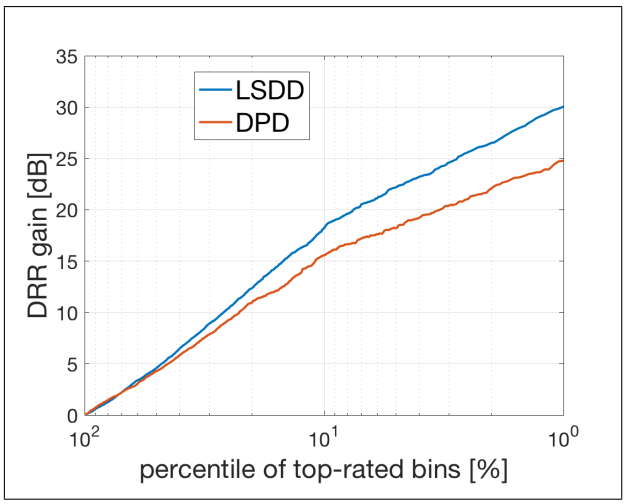


Figure 7. Gain in the average DRR as a function of the selection threshold percentage. Generated from the data shown in Fig. 6

tain threshold value. Note from Fig. 6 that doing so is expected to increase the average DRR of the remaining bins, as compared to the global average. Hence, a natural way to analyze the performance of the methods is by studying the improvement in the average DRR of a certain percentile of the top-rated bins, as plotted in Fig. 7. It can be seen that, as expected, when selecting fewer bins (higher average metric), the gain in the average DRR grows rapidly. Note that, in this example, there is a consistent advantage to the LSDD method reaching 5 dB at 1% of top-rated bins.

The average DRR of the top-rated 1% of the bins has been calculated as a function of the source-receiver distance and for three different frequencies. The results, presented in Fig. 8, were averaged over 6 different speech signals of roughly 20 second duration each. As before, the first-order SH decomposition and 8 nearest neighbors were employed in the DPD test.

First, by observing the results, note that the global average DRR is negative for all demonstrated condi-

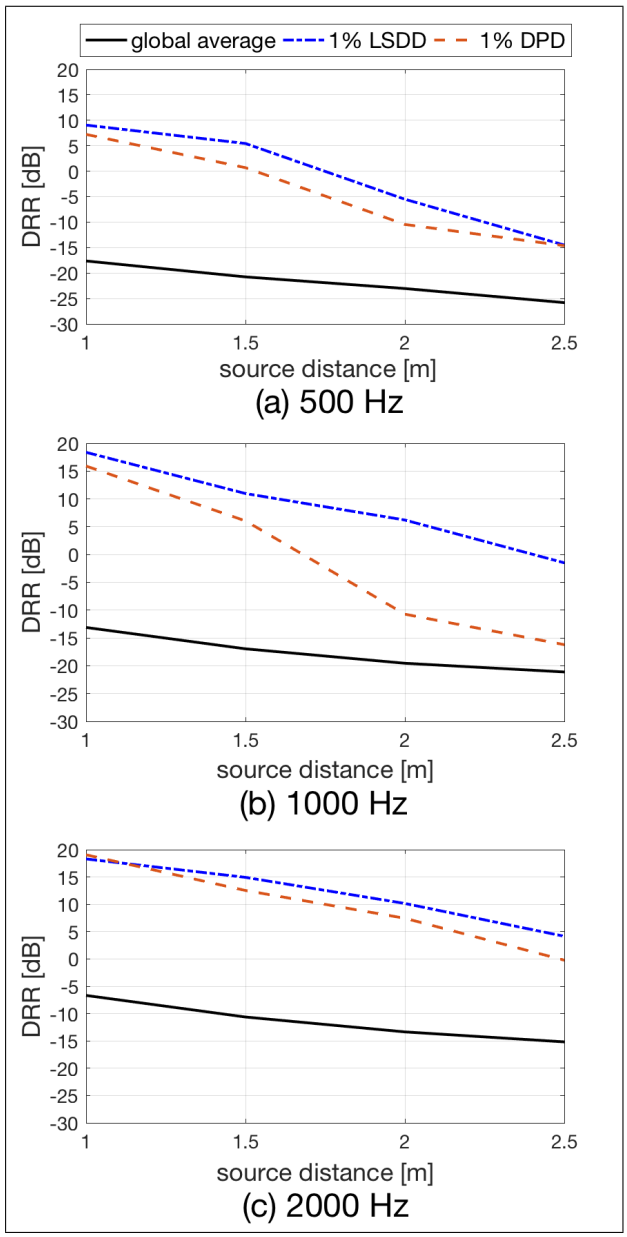


Figure 8. Improvement in the average DRR of the top-rated 1% of the bins as a function of source distance and for three different frequencies. Each frequency is represented by a 250 Hz band, which are 4 adjacent bins of the corresponding spectrograms.

tions, suggesting that the room is highly reverberant. The global average DRR roughly follows the inverse relation to the source distance, as is expected in a diffuse field. Moreover, the average DRR improves towards higher frequencies, which is probably due to a reduction in the corresponding narrowband reverberation time [16]. It is interesting to note that, in general, the higher the global DRR, the more significant is the improvement in DRR obtained using both methods (LSDD and the DPD).

This observation is supported by the relative improvement in DRR for close source locations, which is as high as 30 dB. For farther sources, the improve-

ment in DRR drops to about 20 dB at 1000 and 2000 Hz and drops to even 10 dB at 500 Hz. Overall, it can be seen that the LSDD method is capable of providing an average DRR of above 10 dB, while starting as low as −10 dB or even less. It can also be seen that the LSDD method attains somewhat higher DRR values than the DPD, in many cases the advantage reaches 5 dB or more.

## 6. Future work

There are several open questions that can be suggested for future investigation. One particularly important issue is related to the choice of $R_{th}$, as briefly discussed in Section 4. It should be realized that the threshold value affects two inversely related aspects: (i) the average DRR improvement and (ii) the number of the selected bins. For example, choosing a very high threshold may lead to a great improvement in the DRR, but also result in a small number of bins, which may be insufficient for inference of statistics required by the subsequent DoA estimation algorithm, e.g. for inference of the covariance matrix when using a subspace method [12]. On the other hand, choosing too low a threshold value may simply fail to provide a significant improvement in the average DRR. This implies that an optimal choice may lie somewhere in between and may depend on the specific DoA algorithm and other involved parameters. Hence, getting a better insight into this question is of utmost importance for optimizing the benefits displayed by the proposed approach.

## 7. Conclusion

A new method was proposed to automatically select the STFT bins dominated by the direct sound, while rejecting those contaminated by reflections. The method was described and discussed in the context of improved robustness to reverberation when estimating DoA in realistic environments. The direct-sound selection criterion proposed here is based on the relative peak strength of the local SDD spectrum; it operates entirely in the space domain and, thereby, avoids the limitations related to the SH-domain processing involved in most state-of-the-art approaches. By using experimental data, it was demonstrated that, in many cases, the proposed method is capable of improving the average DRR of the selected bins by as much as 30 dB. It is emphasized that the proposed method should be seen as a preprocessing step that may be exploited by many of the commonly used DoA estimation algorithms in order to increase their robustness to reverberation. The trade-off between the improvement in the average DRR and the total number of the selected bins was explained and its further investigation is suggested for future work.

**References**


[1] 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*, vol. 22, pp. 1494–1505, Oct 2014.

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

[3] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization using estimation consistency in the time-frequency domain," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 516–520, March 2017.

[4] V. Tourbabin and B. Rafaely, "Speaker localization by humanoid robots in reverberant environments," in *2014 IEEE 28th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, pp. 1–5, Dec 2014.

[5] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer, 1st ed., 2008.

[6] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015.

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

[8] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, Inc., New York, USA, 2002.

[9] V. Tourbabin and B. Rafaely, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1803–1814, Dec 2014.

[10] M. Maazaoui, K. Abed-Meraim, and Y. Grenier, "Adaptive blind source separation with hrtfs beamforming preprocessing," in *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 269–272, June 2012.

[11] P. Stoica and K. C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1132–1143, Jul 1990.

[12] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, pp. 276–280, Mar 1986.

[13] K. Harmanci, J. Tabrikian, and J. L. Krolik, "Relationships between adaptive minimum variance beamforming and optimal source localization," *IEEE Transactions on Signal Processing*, vol. 48, pp. 1–12, Jan 2000.

[14] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*, Feb 2000.

[15] "EBU SQAM CD: Sound quality assessment material recordings for subjective tests." published online, Oct. 2008. Companion document: EBU Tech 3253.

[16] J. Backus, *The Acoustical Foundations of Music*. W W Norton, New York, 2 ed., 1977.