DIRECTION OF ARRIVAL ESTIMATION IN HIGHLY REVERBERANT ENVIRONMENTS USING SOFT TIME-FREQUENCY MASK

Vladimir Tourbabin, Jacob Donley, Boaz Rafaely, Ravish Mehra

Facebook Reality Labs

ABSTRACT

A recent approach to improving the robustness of sound localization in reverberant environments is based on pre-selection of time-frequency pixels that are dominated by direct sound. This approach is equivalent to applying a binary time-frequency mask prior to the localization stage. Although the binary mask approach was shown to be effective, it may not exploit the information available in the captured signal to its full extent. In an attempt to overcome this limitation, it is hereby proposed to employ a soft mask instead of the binary mask. The proposed weighting scheme is based directly on a metric of the direct-to-reverberant sound ratio in each individual time-frequency pixel. Evaluation using simulated reverberant speech recordings indicates substantial improvement in the localization performance when using the proposed soft mask weighting.

Index Terms— Direction of arrival estimation, Reverberation, Microphone-array processing

1. INTRODUCTION

Direction of Arrival (DoA) estimation is a fundamental microphone array processing method. It is employed for sound source localization with numerous applications in video-conferencing systems, consumer electronics, humanoid robots, security systems, in-car audio, and more. In many cases, the DoA estimation is carried out in reverberant environments, such as living spaces or meeting rooms. High reverberation has been shown to have a detrimental effect on the performance of most DoA estimation algorithms, motivating research and development of robust approaches in the presence of reverberation [1–3].

One relatively recent approach that effectively improves robustness to reverberation is based on an additional pre-processing step, which is executed immediately prior to the DoA estimation. The purpose of this step is to pre-select time-frequency pixels (a.k.a time-frequency bins) that contain sound with high Direct-to-Reverberant Ratio (DRR). Subsequently, a DoA estimator is applied to the selected pixels only, which inherently reduces sensitivity to reverberation [2,4–8]. The preprocessing step is equivalent to applying a binary mask to the signal spectrogram. The binary mask may be non optimal from a signal processing standpoint due to the following reasons: (1) the selected pixels are weighted evenly, although each may contain varying amounts of useful DoA information, (2) the process completely discards pixels whose DRR is deemed to be below a certain threshold, thereby potentially missing useful residual DoA information.

In a recent publication, [9], the authors proposed to cluster the pixels that have passed the selection threshold. Then, the pixels are weighted in accordance with their distance from a corresponding centroid. This approach partially overcomes the even weighting shortcoming mentioned above, while still discarding most of the pixels. The weighting procedure used in [9] is sometimes referred to as the *soft mask* as opposed to the hard binary mask. In the current work, we aim to address both the above shortcomings of the binary mask by exploiting the soft mask approach and generalizing it in the two following ways: (1) we propose to apply the soft mask to all time-frequency pixels without a selection step, (2) we propose to weight the different pixels based directly on their corresponding DRR metric. This way, none of the pixels are discarded and all of the pixels are (presumably) weighted in accordance with the amount of useful DoA information that they contain.

The remainder of the paper is organized as follows. Section 2 introduces notation and provides additional details on selected previous work. Next, in Section 3, the proposed soft mask approach is introduced. Section 4 summarizes the proposed algorithm. Section 5 evaluates the DoA estimation performance when using the soft mask and compares it to the binary mask approach. Conclusions follow.

2. BACKGROUND

The current section presents the reverberant signal model used throughout the paper and briefly introduces some of the relevant previous work.

2.1. Signal model and notation

Consider an arbitrary array of M microphones located either in free-field or configured around a reflective surface. The Short Time Fourier Transform (STFT) pixel with the time frame t and frequency bin f is an $M \times 1$ complex-valued vector; it is assumed to include two major components:

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

where $\mathbf{x}_d(t,f) \in \mathbb{C}^{M\times 1}$ and $\mathbf{x}_r(t,f) \in \mathbb{C}^{M\times 1}$ denote the direct and the reverberant portions of the captured sound, respectively. Vectors $\mathbf{v}(f,\Omega_i)$ stand for the Array Transfer Function (ATF) (a.k.a array steering vector) at frequency f and arrival directions Ω_i, with $i = 0$ indicating the direct sound and $i = 1, ..., I$ denoting the reflected waves. Finally, scalars s_i represent amplitudes of the waves arriving from the corresponding directions and account for the attenuation and the phase shift induced by sound propagation and reflection from the various surfaces. Note that when using the model in (1), the number of the reflected components, I, may grow with increased reverberation time in order to accurately represent the measured signal.

Lastly, in order to avoid potential ambiguities, we explicitly define the local DRR of a given time-frequency pixel as

$$\mathrm{DRR}(t,f) = \frac{\|\mathbf{x}_d(t,f)\|_2}{\|\mathbf{x}_r(t,f)\|_2}. \quad (2)$$

The above definition is assumed henceforth throughout this paper.

2.2. Previous approaches

As mentioned above, several recent publications were concerned with improving DoA estimation robustness to reverberation by applying a binary mask. The purpose of the mask was to discriminate between time frequency pixels with high DRR and the rest of the pixels that contain a significant amount of reverberant energy.

As the local DRR of individual pixels is not readily available, several different methods were proposed to compute DRR-related metrics from the array input signals [2, 4–8]. Some of these methods attempt to assess the amount of local variation of the DoA estimates [4,5], while arguing that high DRR pixels are expected to display higher consistency in their immediate time-frequency vicinity. Another recently proposed hybrid method combines signal power, sound field diffuseness estimations, and speech presence probability in a single selection criterion [7]. Additional highly effective methods operate in the spherical harmonics domain by exploiting frequency smoothing [2] or local sound field directivity [6].

An effective method operating in the element-space domain was recently proposed for an arbitrary array geometry [8]. The method is called the Local Space Domain Distance (LSDD) and is based on a measure of similarity between the ATF and a given time-frequency pixel. In particular, the method computes a spatial spectrum for each of the pixels as follows:

$$S_{t,f}(\Omega_j) = \cos^{-1} \frac{|\mathbf{v}^H(f,\Omega_j) \cdot \mathbf{x}(t,f)|}{\|\mathbf{v}(f,\Omega_j)\|_2 \cdot \|\mathbf{x}(t,f)\|_2}, \quad j = 1, ..., J, \quad (3)$$

where $\{\Omega_j\}_{j=1}^J$ is a set of DoAs that can be chosen in accordance with the desired resolution and the available resources. It is emphasized that the expression in (3) is the Hermitian angle between the two complex-valued vectors $\mathbf{v}^H(f,\Omega_j)$ and $\mathbf{x}(t,f)$ [10]. Put in plain words, the Hermitian angle measures the angle between the two lines defined by the complex vectors. The angle is limited to the range $[0, 90]$ deg, with 0 deg denoting parallel lines and 90 deg denoting perpendicular lines. The most computationally demanding term in (3), which cannot be pre-computed, is probably the inner product $\mathbf{v}^H(f,\Omega_j) \cdot \mathbf{x}(t,f)$. Nevertheless, the term can be computed for all arrival directions $\{\Omega_j\}_{j=1}^J$ in parallel using a single matrix-vector multiplication.

Using the spectrum defined in (3), the LSDD metric for a given time-frequency pixel, $\mathbf{x}(t,f)$, is defined as

$$\mathrm{LSDD}(t,f) = \min_j \left(S_{t,f}(\Omega_j)\right) \quad [\mathrm{deg}], \quad (4)$$

which is simply the minimal distance in degrees between the pixel and the ATF. Hence, a small $\mathrm{LSDD}(t,f)$ indicates that the pixel (t,f) is similar to the transfer function in a certain direction, implying that it is dominated by a single wave arriving from that direction. This wave is likely to be the direct sound wave because it usually arrives before the reflections. On the other hand, a larger value of $\mathrm{LSDD}(t,f)$ indicates that the pixel is farther away from the ATF and, therefore, is likely to contain a significant amount of reverberant energy. In [8], the metric was shown to be strongly correlated to a ground-truth local DRR; it was used to compute a binary mask for selecting the high DRR pixels as follows:

$$\mathrm{BM}(t,f) = \begin{cases} 1, & \mathrm{LSDD}(t,f) < \theta \\ 0, & else \end{cases}, \quad (5)$$

where θ is an arbitrary threshold that usually can be set by a trial and error process to obtain desired performance in a given scenario. Finally, the DoA can be estimated by selecting the direction that minimizes the average spatial spectrum of the selected pixels, i.e

$$\Omega^\star = \underset{\Omega_j}{\mathrm{argmin}} \sum_{t,f} \mathrm{BM}(t,f) \cdot S_{t,f}(\Omega_j). \quad (6)$$

It is emphasized that a single source case is discussed here for simplicity, while an extension to the multiple source case can be obtained by selecting several directions corresponding to anti-peaks (dips) of the average spectrum.

In the following section, we exploit the LSDD metric in order to introduce the *soft mask*.

3. SOFT MASK

The binary mask approach divides all the time-frequency pixels into two groups, those which contain the direct sound component only, and those which are contaminated by reflected sound. This is a crude approximation to reality, because the local DRR of a given pixel is a continuous quantity [8]. This, in turn, implies that the amount of useful DoA information contained in a given pixel may be a continuous quantity, as well. This idea is demonstrated here by assessing the DoA performance as a function of the LSDD metric. For that purpose, 26 min of reverberant speech were generated in three different rooms using the image method [11]. The reverberation time ranged from 763 ms to 1115 ms. See Section 5.1 for additional details. Using the STFT of the reverberant speech recordings, a DoA estimate and the LSDD metric were computed for each pixel individually. The DoA estimate was obtained by finding the minimum of the pixel's spatial spectrum in (3), while the LSDD metric was computed directly as outlined in (4). Then, the pixels were grouped in accordance with their LSDD value and an average DoA error was computed for each group. The result is plotted in Fig. 1. It can be seen that, as expected, the average DoA error increases for pixels with higher LSDD values (reduced DRR). There are two important points in relation to the DoA error behavior that should be noted. First, in line with the above-mentioned hypothesis, the DoA error appears to be a continuously-valued and monotonic function of the LSDD metric, as opposed to the crude binary approximation. Second, we recall that in the cases where no DoA information is available, the average error is expected to be 90 deg and note that here, this error level is only reached for the pixels with the highest LSDD values. This suggests that most of the pixels contain some amount of information about the DoA, which could,

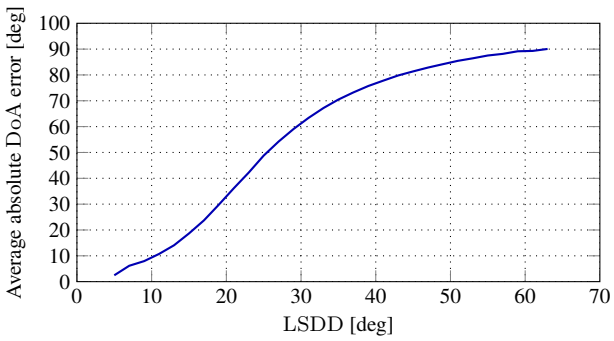

Figure 1: Average DoA error vs. the LSDD metric. Based on simulated reverberant audio recording of total length 26 min, using STFT frame of 256 samples, 50% overlap, and sampling frequency of 16 kHz. The analysis uses frequencies in the range $1800 - 3600$ Hz, resulting in a total of over 6 million time-frequency pixels. See Section 5.1 for additional details.

potentially, become useful in the estimation process.

In order to exploit the potentially useful information in all the time-frequency pixels and in order to weight the different pixels in accordance with the amount of information contained within them, it is hereby proposed to use soft mask weighting as a replacement for the binary mask. Furthermore, in the light of the correlation between the LSDD metric and the DoA performance, it seems natural to exploit the metric for establishing the soft mask weights. In particular, we propose the following soft mask weights:

$$\mathrm{SM}(t,f) = \mathrm{LSDD}(t,f)^{-\alpha}. \tag{7}$$

The parameter α is introduced here to serve as a selectivity factor, i.e. the higher α is, the more weight is steered towards the low LSDD pixels. Note that the expression in (7) is undefined for $\mathrm{LSDD}(t,f) = 0$. In practice, this is a very rare case and can be dealt with by, for example, setting the corresponding weight to the highest weight in the mask. An example of the proposed soft mask weight function is illustrated in Fig. 2 along with the binary mask for comparison.

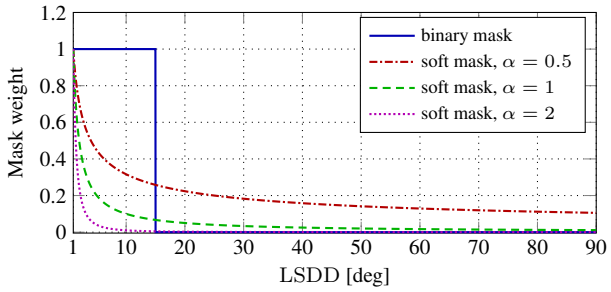

Figure 2: An illustration of the soft mask weighting function with three different α values. The binary weight function is also provided for comparison. The binary mask threshold was arbitrarily set to 15 deg.

Finally, the soft mask-based DoA estimator can be obtained by simply swapping $\mathrm{BM}(t,f)$ with $\mathrm{SM}(t,f)$ in (6):

$$\Omega^\star = \underset{\Omega_j}{\operatorname{argmin}} \sum_{t,f} \mathrm{SM}(t,f) \cdot S_{t,f}(\Omega_j). \tag{8}$$

In the following section, the proposed algorithm is formally outlined for the reader's convenience. Then, in Section 5, the effect of the proposed soft mask weighting scheme on the DoA estimation performance is analyzed and compared to the performance obtained with the previously employed binary mask.

4. ALGORITHM SUMMARY

The current section summarizes the DoA estimation algorithm that uses the soft mask weighting scheme as proposed above. It is assumed that the algorithm receives at its input a precomputed STFT of an audio buffer, i.e. $\mathbf{x}(t,f)$, $t = 1,...,T$, $f = 1,...,F$. Then, the algorithm proceeds as is summarized in Algorithm 1. Note that in (3), the algorithm implicitly uses $\mathbf{v}(f,\Omega_j)$. In practice, the ATF of a given array can be obtained in multiple ways including analytic modeling [8], numerical simulation [12], and/or direct measurement [13].

5. SIMULATION STUDY

The current section summarizes a simulation study that has been carried out in order to assess the effect of the proposed soft mask weighting scheme and to compare its performance to the previously employed binary mask approach.

Algorithm 1 LSDD-based DoA estimation with soft mask

1: input: $\mathbf{x}(t,f)$, $t = 1,...,T$, $f = 1,...,F$
2: init: $\mathrm{SM}(t,f) = 0$, $t = 1,...,T$, $f = 1,...,F$
3: **for all** $t \in \{1,...,T\}$ and $f \in \{1,...,F\}$ **do**
4: compute: $S_{t,f}(\Omega_j)$, $j = 1,...,J$ [use Eq. (3)]
5: compute: $\mathrm{LSDD}(t,f)$ [use Eq. (4)]
6: compute: $\mathrm{SM}(t,f)$ [use Eq. (7)]
7: **end**
8: compute: $\Omega^\star$ [use Eq. (8)]
9: output: $\Omega^\star$

5.1. Setup

Reverberant recordings of human speech were simulated in three different rooms. The dimensions of the rooms and their broadband reverberation time, T_{60}, are summarized in Table 1. The dimensions were chosen in accordance with real dimensions of three actual existing meeting rooms.

Table 1: Physical dimensions and selected acoustic properties of the three rooms used in the simulation study.

#	x [m]	y [m]	z [m]	T_{60} [ms]	d_{cr} [m]
1	3.55	3.62	5.00	786	0.52
2	5.45	3.97	2.89	763	0.52
3	6.48	9.35	3.66	1115	0.80

Reverberant impulse responses were computed for the microphone array configuration illustrated in Fig. 3. The room impulse responses were obtained using the image method [11].

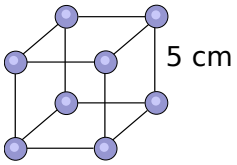

Figure 3: An illustration of the array configuration with 8 microphones located at the corners of a cube with an edge length of 5cm.

In each room, 32 different source and receiver (microphone-array center) locations were simulated. The locations were randomly drawn from a uniform distribution over the entire volume of the room with the two following constraints: (1) no receiver/source was allowed within 0.5m of any of the walls, (2) receiver-source distance is within an expected conversant range of $1-2$ m, which is greater than the critical distance in all the simulated rooms (see Table 1). In each scenario, dry speech recordings from the TIMIT corpus [14] with average length of approximately 16 sec, were convolved with the simulated room impulse responses. Total length of the simulated recordings was 26 minutes.

The sampling rate of the simulated recordings was 16 kHz. The recordings were transformed into the STFT domain using frames of 256 samples and 50 % overlap. In a preliminary investigation it was found that the array configuration used here performs best in the frequency range of $1800-3600$ Hz. Hence, the frequency bins

outside this range were discarded. The remaining data was divided into short buffers of duration L_{buf} [sec]. The buffer length that was used in the different experiments is detailed below. Using the binary and the soft mask approaches, the DoA was estimated for each of the buffers separately. An average of the absolute angular error was computed, and served as the measure of the DoA estimation performance.

Lastly, the reader may recall that the binary mask approach has an inherent LSDD threshold, as described in (5). Hence, when using the binary mask approach, the DoA estimation is only defined on those buffers that contain at least one pixel that passes the threshold. Therefore, the analysis and comparison between the binary and soft masks in the following subsection is carried out only using the buffers that have passed the appropriate LSDD threshold.

5.2. Results and discussion

First, the DoA estimation was carried out using the soft mask with different values of the selectivity parameter, α, in order to study its effect on the performance of the method. The results are plotted in Fig. 4. It can be seen that the soft mask performs better than the binary mask over most of the range of α. The DoA error is minimal around $\alpha = 3$ increasing from that point in both directions. This behavior can be explained as follows: when α is low, the soft mask simply uses all of the pixels and, hence, performs less well than the binary mask; when α is high, the error increases because the soft mask becomes too selective, effectively discarding the DoA information contained in most of the pixels.

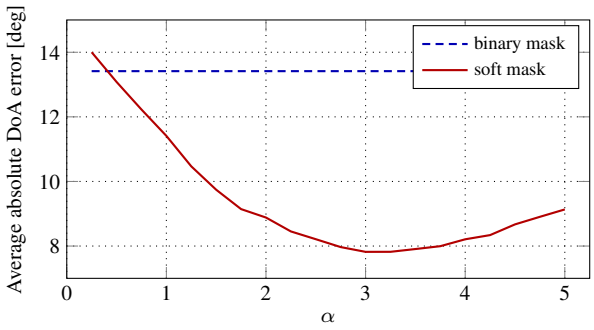

Figure 4: Average absolute DoA error as a function of the selectivity parameter, α (see Eq. (7)). The audio buffer length is $L_{buf} = 360$ ms, the LSDD threshold used with the binary mask is $\theta = 12$ deg.

When performing DoA estimation, it is usually desirable to keep the audio buffer duration, L_{buf}, as small as possible. This is especially important when a fast estimation of moving sound sources is required. Hence, the second experiment reported here aimed to compare the DoA performance of the binary and the soft masks as a function of the audio buffer duration. The results are plotted in Fig. 5. It can be seen that the soft mask leads to a significant reduction of the DoA error in all of the computed range of the buffer duration. In particular, note that a 10 deg average error can be obtained when applying the soft mask to buffers of as low as 50 ms. At the same time, in order to obtain similar performance when using the binary mask, the audio buffer duration would need to be increased all the way to 1000 ms.

Recall that the performance of the DoA estimator with the binary mask is strongly tied to the choice of the LSDD threshold, θ. Hence, the last experiment described here compares the performance of the binary and the soft masks for three different values of the LSDD threshold, $\theta = 5, 10$ and 15 deg. The binary mask approach is only capable of producing a DoA estimate from audio

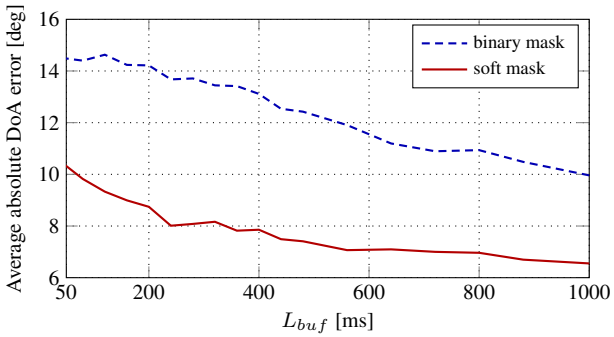

Figure 5: Average absolute DoA error as a function of the audio buffer duration used to obtain a single DoA estimate. The selectivity parameter with the soft mask is $\alpha = 3$, the LSDD threshold used with the binary mask is $\theta = 12$ deg.

buffers that contain at least one pixel satisfying the corresponding LSDD threshold, otherwise all of the pixels are discarded. Hence, here, three sets of the audio buffers were first identified by selecting those which contain at least one pixel that passes the appropriate LSDD threshold. Note that, by definition, the set of buffers for $\theta = 5$ deg is contained within the set corresponding to $\theta = 10$ deg, which, in turn, is contained in the set of $\theta = 15$ deg. The DoA estimation was carried out on the three sets using both the binary and the soft mask methods. The average DoA error obtained in the three cases is shown in Fig. 6. First, note that in agreement with previously reported results [8], the DoA error obtained with the binary mask approach increases for larger values of the LSDD threshold. Second, in all three cases of the LSDD threshold, applying the soft mask instead of the binary mask leads to a reduction in the average DoA estimation error.

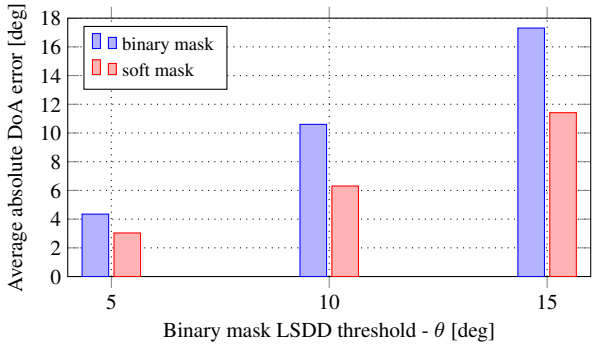

Figure 6: Average absolute DoA error for different values of the LSDD threshold used in the binary mask method. The audio buffer length is $L_{buf} = 360$ ms, the selectivity is $\alpha = 3$.

6. CONCLUSION

The current work has proposed a way to further improve the reverberation-robust LSDD algorithm for DoA estimation by introducing a soft time-frequency mask. It was pointed out that the soft mask has the potential to exploit the available DoA information to a greater extent, as compared to the previously employed binary mask. Using the proposed method, a substantial improvement in the DoA accuracy has been demonstrated. Future work may focus on further optimization of the soft mask weighting scheme.

7. REFERENCES

[1] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, April 1997, pp. 375–378.

[2] 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, no. 10, pp. 1494–1505, Oct 2014.

[3] O. Schwartz, Y. Dorfan, E. A. P. Habets, and S. Gannot, "Multi-speaker DOA estimation in reverberation conditions using expectation-maximization," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2016, pp. 1–5.

[4] 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)*, March 2017, pp. 516–520.

[5] 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.

[6] 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.

[7] A. Brendel, C. Huang, and W. Kellermann, "STFT bin selection for localization algorithms based on the sparcity of speech signal spectra," *Euronoise 2018, Crete*, 2018.

[8] V. Tourbabin, D. 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," *Euronoise 2018, Crete*, 2018.

[9] K. Wu, V. G. Reju, and A. W. H. Khong, "Multisource DOA estimation in a reverberant environment using a single acoustic vector sensor," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1848–1859, Oct 2018.

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

[11] J. B. Allen and D. 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.

[12] 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, no. 12, pp. 1803–1814, Dec 2014.

[13] 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)*, June 2012, pp. 269–272.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. S. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus," CD-ROM, 1993.