Study of speaker localization under dynamic and reverberant environments

Daniel A. Mitchell* and Boaz Rafaely[†]

School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel

ABSTRACT

Speaker localization in a reverberant environment is a fundamental problem in audio signal processing. Many solutions have been developed to tackle this problem. However previous algorithms typically assume a stationary environment in which both the microphone array and the sound sources are not moving. With the emergence of wearable microphone arrays, acoustic scenes have become dynamic with moving sources and arrays. This calls for algorithms that perform well in dynamic environments. In this article, we study the performance of a speaker localization algorithm in such an environment. The study is based on the recently published EasyCom speech dataset recorded in reverberant and noisy environments using a wearable array on glasses. Although the localization algorithm performs well in static environments, its performance fell substantially when used on the EasyCom dataset. The paper presents performance analysis and proposes methods for improvement.

Keywords: Direction of Arrival (DOA) estimation, Direct Path Dominance (DPD) test, wearable microphone arrays, EasyCom dataset

1 INTRODUCTION

Localizing multiple sound sources recorded with a microphone array in an enclosure is an important task used in a wide range of applications such as speech enhancement, source separation and video conferencing [2]. Therefore, many direction-of-arrival (DOA) estimation methods have been developed for this task. These include methods based on beam-forming [11], subspace methods such as multiple signal classification (MUSIC) [9], and time-delay of arrival estimation methods [3]. Many of the algorithms based on these methods were designed assuming a free-field environment. When these algorithms are used in a more common reverberant environment their DOA performance degrades. The reason for this is that in a reverberant environment, room reflections mask the direct sound which carries the DOA information. Recently, however, several methods have been developed for DOA estimation of multiple speakers that are robust to reverberation. One such method processes the microphone signals in the time-frequency domain, and employs a direct-path dominance (DPD) test [7] to identify time-frequency bins that are dominated by the direct sound from the source. Algorithms which use this method have been widely studied assuming a static environment, where both the sound sources and the microphone array are stationary. On the other hand, these algorithms have been less intensely studied in a more realistic, dynamic environment, where the sound sources and/or the microphone array are moving.

In dynamic environments, the motion of the sound sources and/or microphone array may lead to rapid change of DOAs in time. Thus to accurately trace the DOA of speakers requires a short interval between successive DOA estimates. Additionally, the DOA estimates may be smoothed in time using a tracking algorithm. Although DOA estimation and tracking algorithms in dynamic environments have been the subject of several recent studies [5, 12], including the Acoustic Source Localization and Tracking (LOCATA) [6] challenge, none of these have included experiments with wearable microphone arrays which may bring new challenges. Such scenes are becoming increasingly more popular due to the increased interest in applications involving augmented reality.

In this article we address the problem of DOA estimation in a noisy dynamic environment involving a wearable microphone array. The experiments were performed using the Easy Communication (EasyCom) dataset

---

*mitchdan@post.bgu.ac.il   [†]br@bgu.ac.il

[4] which was explicitly designed to represent a realistic cocktail-party environment. The DOA estimates were computed using a computationally efficient algorithm which had been shown previously to have good source localization performance in a static reverberant environment [10]. The algorithm incorporates a DPD test and operates in the time-frequency domain. We study the performance and limitations of this algorithm on the EasyCom dataset under different operating parameters. We also introduce two modifications of the algorithm and study their improvement of performance.

2 MATHEMATICAL MODEL

In this section we first briefly present the model assumed for the recorded signal as captured by the microphone array at each time-frequency bin (t, f). Then, we describe the local space domain distance (LSDD) algorithm [10] for the DOA estimation at each bin (t, f).

2.1 Signal model

Assume a microphone array with M microphones arranged according to a pre-defined geometry. Next, consider a sound field comprised of K far-field sources, arriving from directions $\Psi_k, k \in \{1, 2, \ldots, K\}$. These sources represent the direct sound from the speakers in the scene, as well as reflections (reverberations) due to objects and room boundaries.

In the next step, the recorded microphone signals are transformed into the joint time-frequency domain by applying the short-time Fourier transform (STFT). This is done by first separating the speech signal into short time intervals of length δt. A fast Fourier transform (FFT) is then applied to each time segment. Following these pre-processing steps, the signal received by the microphone array can then be described in the STFT domain as

$$\mathbf{x}(t, f) = \sum_{k=1}^{K} s_k(t, f) \mathbf{v}(f, \Psi_k) + \mathbf{n}(t, f) , \tag{1}$$

where $\mathbf{x}(t, f) = [x_1(t, f), x_2(t, f), \ldots, x_M(t, f)]^T$ is an $M \times 1$ complex vector denoting the signal as measured by the microphones composing the array; $\mathbf{v}(f, \Psi_k) = [v_1(f, \Psi_k), v_2(f, \Psi_k), \ldots, v_M(f, \Psi_k)]^T$ is an $M \times 1$ vector denoting the response of the microphone array to a unit-amplitude plane wave at frequency f arriving from the k'th source in direction Ψ_k; $s_k(t, f)$ is a scalar which represents the amplitude of the k'th sound source signal; and $\mathbf{n}(t, f) = [n_1(t, f), n_2(t, f), \ldots, n_M(t, f)]^T$ is an $M \times 1$ vector denoting the noise in the signal $\mathbf{x}(t, f)$.

2.2 Local Space Domain Distance (LSDD) Algorithm

The LSDD algorithm is a recently developed DOA estimation algorithm characterized by DOA performance that is robust to reverberation. The algorithm was first proposed in [10] and works as follows. The directional spectrum $\mathbf{S}(t, f) = [S_1(t, f), S_2(t, f), \ldots, S_L(t, f)]^T$ computed within this algorithm is an $L \times 1$ vector defined over a grid of DOAs $\Theta_l, l \in \{1, 2, \ldots, L\}$. The l'th component is defined as:

$$S_l(t, f) = d\left(\mathbf{x}(t, f), \mathbf{v}(f, \Theta_l)\right) = d\left([x_1(t, f), \ldots, x_M(t, f)]^T, [v_1(f, \Theta_l), \ldots, v_M(f, \Theta_l)]^T\right) , \tag{2}$$

where $d(\mathbf{a}, \mathbf{b})$ is a function which measures the similarity between two vectors $\mathbf{a}$ and $\mathbf{b}$. In [10], $d(\mathbf{a}, \mathbf{b})$ was defined as

$$d(\mathbf{a}, \mathbf{b}) = \frac{1}{\min_{\beta}\left(\frac{\|\mathbf{a} - \beta \mathbf{b}\|}{\|\mathbf{a}\|}\right)} , \tag{3}$$

where $\|.\|$ is the 2-norm. In this article we shall use, however, the more conventional cosine similarity measure:

$$d(\mathbf{a}, \mathbf{b}) = \frac{|<\mathbf{a}, \mathbf{b}>|}{\|\mathbf{a}\| \|\mathbf{b}\|} , \tag{4}$$

where $<\mathbf{a}, \mathbf{b}>$ denotes the inner product between $\mathbf{a}$ and $\mathbf{b}$.

Given the spectrum vector $\mathbf{S}(t, f)$, the estimated DOA for bin (t, f) is computed by:

$$\hat{\theta}(t, f) = \arg\max_{l} \{S_l(t, f)\} . \tag{5}$$

Some of the bins (t,f) do not, however, contain a valid $\hat{\theta}(t,f)$ value. These are bins in which the direct signal from the speaker is masked by noise and reverberations. We eliminate these bins by calculating a DPD measure value $\chi(t,f)$ for each bin which we then test against a threshold λ. Although there are different methods available for calculating $\chi(t,f)$, we shall, for simplicity, use the following:

$$\chi(t,f) = \max_l \{S_l(t,f)\} . \tag{6}$$

Together, Eqns. (5) and (6) define a (joint) LSDD DOA/DPD algorithm.

2.3 Energy Weighted Local Space Domain Distance (LSDDe) Algorithm

We describe an energy weighted modification for calculating the DPD test value $\chi(t,f)$. In this equation, we weight the DPD test value in Eqn. (6) with its corresponding signal energy. The energy weighted DPD test value is then:

$$\chi(t,f) = \max_l \left\{ S_l(t,f) \cdot MED[|x_1(t,f)|^2, |x_2(t,f)|^2, \ldots, |x_M(t,f)|^2] \right\} , \tag{7}$$

where MED is the median operator. Together Eqns. (5) and (7) define a (joint) LSDDe DOA/DPD algorithm.

3 PROPOSED DOA ESTIMATION ALGORITHM

We propose a new DOA estimation algorithm which was the outcome of investigating the performance of the LSDD algorithm with the EasyCom dataset. It is clear from Eqns. (2), (4) and (5) that the LSDD algorithm does not use any information about the behavior of $\mathbf{S}(t,f)$ with respect to Θ_l. As such information may be useful, it is proposed to incorporate this information using a correlation process, as follows. For each frequency f we define an ïdeal two-dimensional spectrum represented by matrix $\mathbf{W}$, whose elements, $W_{lh} \equiv W(\Theta_l, \Theta_h)$, represent the similarity between the l'th steering vector $\mathbf{v}(f, \Theta_l)$ and the h'th steering vector $\mathbf{v}(f, \Theta_h)$:

$$\begin{aligned} W(\Theta_l, \Theta_h) &= d(\mathbf{v}(f, \Theta_l), \mathbf{v}(f, \Theta_h)) \\ &= d([v_1(f, \Theta_l), \ldots, v_M(f, \Theta_l)]^T, [v_1(f, \Theta_h), \ldots, v_M(f, \Theta_h)]^T) , \forall l, h \in \{1, 2, \ldots, L\} , \end{aligned} \tag{8}$$

where measure d is defined in (4). Now, the similarity between each column vector in $\mathbf{W}$ and the spectrum $\mathbf{S}(t,f)$ is computed, which provides the following indication. Suppose that Θ_h is a DOA of an actual source. Then, we expect the h'th column of $\mathbf{W}$ to be similar to $\mathbf{S}(t,f)$. This, in effect, defines the new directivity based Space Domain Distance (dSDD) DOA estimation algorithm. The corresponding DOA estimate for bin (t,f) is now computed using:

$$\hat{\theta}(t,f) = \arg\max_h \{d(\mathbf{S}(t,f), \mathbf{W}_h)\} , \tag{9}$$

where $\mathbf{W}_h$ denotes the h'th column in $\mathbf{W}$. We follow Eqn. (6) and define a corresponding DPD test measure:

$$\chi(t,f) = \max_h \{d(\mathbf{S}(t,f), \mathbf{W}_h)\} . \tag{10}$$

Together, Eqns. (9) and (10) define a (joint) dSDD DOA/DPD algorithm. It should be noted that under ideal conditions where signal $\mathbf{x}$ in Eqn. (1) is composed of a single plane wave, the two algorithms, LSDD and dSDD should provide the same estimate as they both rely on the same set of steering vectors. However, the motivation for proposing dSDD is the expected robustness against potential noise and reverberation due to the comparison of entire functions, or vectors. This is in contrast to the LSDD where DOA estimates are based on looking for a peak in a function.

As in the case of the LSDD algorithm, we describe an energy weighted dSDD algorithm (dSDDe), in which, for each bin (t,f), we weight the dSDD DPD test measure with the corresponding signal energy. The energy weighted DPD test value is therefore:

$$\chi(t,f) = \max_h \{d(\mathbf{S}(t,f), \mathbf{W}_h)\} \cdot MED \left[|x_1(t,f)|^2, |x_2(t,f)|^2, \ldots, |x_M(t,f)|^2\right] . \tag{11}$$

4 EXPERIMENTS

This section presents an experimental study that aims to investigate the performance of the LSDD and the dSDD with the EasyComm dataset. First, experimental setup and methodology are presented. This is then followed by the evaluation of the results.

4.1 Set-up

The experiments described in this article were performed on the EasyCom dataset [4]. This dataset was designed with the aim of analyzing the cocktail party effect with audio signals captured by augmented reality (AR) glasses equipped with an egocentric six-channel microphone array. Figure 1 shows a schematic drawing of the glasses with locations of the microphones [4].

The dataset contains recordings of natural conversations in a noisy restaurant environment. Participants were equipped with close-talk microphones, a camera and tracking markers. They were asked to engage in conversations during several tasks, including introductions, ordering food, solving puzzles, playing games and reading sentences. The recordings also contain an egocentric video viewpoint of the participants. The pose (position and rotation) of every participant was also recorded. The dataset was additionally labelled with annotators of voice activity.

Figure 1. Illustration of the AR glasses with locations of microphones [4]. Four of the microphones are fixed rigidly to the glasses and two of the microphones are placed in the user's ears.

The signals recorded by the microphones were sampled at a rate of 48 kHz. The recorded data was transformed into the STFT domain using a 1024 samples ($\simeq 20$ msec) Hann window with an overlap of 512 samples. The microphone signals in the STFT domain were employed as an input to the algorithms under study.

4.2 Methodology

Evaluation of the DOA/DPD algorithms incorporated a direction search with a resolution of $5°$ which was limited to the horizontal plane. The ground truth azimuthal DOA (Ψ_k) was obtained from the EasyCom dataset as a function of time. Altogether, a series of three experiments was performed with the EasyCom dataset. The first experiment measured the effective frequency range of the array $[f_{low}, f_{high}]$, and the second and third experiments investigated the effect of frequency smoothing and the length of the time interval ΔT on the performance.

DOA estimation performance was evaluated as follows. For each (t, f) bin, the absolute error

$$\varepsilon(t, f) = |\Psi(t) - \hat{\theta}(t, f)|, \tag{12}$$

between the true DOA $\Psi(t)$ and the estimated DOA $\hat{\theta}(t, f)$ was computed. Note that Eqn. (12) assumes that both $\Psi(t)$ and $\hat{\theta}(t, f)$ are measured with respect to the same axis. In practice, in this dataset, $\Psi(t)$ is measured with respect to an axis defined relative to the room, while $\hat{\theta}(t, f)$ is measured with respect to the orientation of the glasses. Thus, before calculating $\varepsilon(t, f)$, $\hat{\theta}(t, f)$ is transformed to the fixed axis of the room by incorporating head tracking information. In addition, each bin was labeled a "hit" if $\varepsilon(t, f) \leq 10°$; otherwise it was labeled a "miss".

The bins (t, f) were divided into time blocks of length ΔT. For each block, only bins which satisfied the following four conditions are regarded as valid and are used for DOA estimation:

1. The frequency f lies within the effective operating frequency range: $f \in [f_{low}, f_{high}]$.

2. The time t lies within the selected block. Denote T as the middle time of the block then $t \in [T - \Delta T/2, T + \Delta T/2]$ belongs to block T.

3. Voice activity was detected at time t.

4. The DPD test value $\chi(t,f)$ exceeds a threshold λ. For the purpose of this study, the threshold was determined from the percentage p of bins which satisfy the first three conditions with the highest value χ. Note that λ is computed independently for each block T.

The mean value of $\varepsilon(t,f)$ is computed over valid (t,f) bins, i.e. bins which satisfy the above four conditions. By definition, this is the mean absolute DOA estimation error $E(p,T,\Delta T)$ for the block T. Similarly, the mean hit ratio $H(p,T,\Delta T)$ was computed on valid (t,f) bins by dividing the number of valid bins labelled "hit" by the total number of valid bins. Finally the mean absolute error, $\bar{E}(p,\Delta T)$, and the mean hit ratio, $\bar{H}(p,\Delta T)$, for the entire experiment was computed by averaging $E(p,T,\Delta T)$ and $H(p,T,\Delta T)$ over all blocks T.

4.3 Effective operating Frequency Band

The EasyCom dataset involves speech sound which naturally limit the frequency range of interest [8]. This frequency band is reduced in practice by aliasing effects which arise from the microphone array. For a specific steering vector $\mathbf{v}(f,\Theta_h)$ (corresponding to a frequency f and direction Θ_h), the similarity between $\mathbf{v}(f,\Theta_h)$ and the set of steering vectors $\mathbf{v}(f,\Theta_l), l \in \{1,2,...,L\}$ is computed using Eqn. (4). This is repeated for all frequencies f leading to the following measure

$$\Lambda(f,\Theta_l) = d(\mathbf{v}(f,\Theta_h),\mathbf{v}(f,\Theta_l)) . \tag{13}$$

Figure 2 shows Λ for $\Theta_h = 0°$. Visual inspection shows that the preferred frequency band is about $1100 - 2000$ Hz, where at lower frequency the directivity may be too wide, while at higher frequencies significant side lobes may degrade spatial processing. While this is a relatively narrow band of frequencies, in this work it led to best performance. Extending the range of operation for both lower and higher frequencies is proposed for future work.
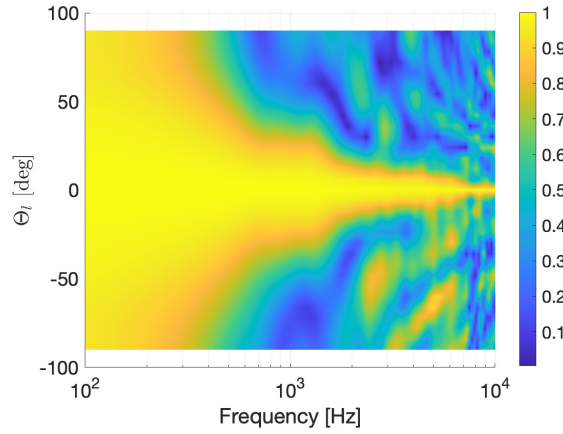

Figure 2. The similarity matrix Λ for $\Theta_h = 0°$

4.4 DOA Error and Frequency Smoothing

The spectrum $\mathbf{S}(t,f)$ plays a key role in the DOA/DPD algorithms. In particular, the authors in [1] have shown that it is beneficial to smooth $\mathbf{S}(t,f)$ in frequency. We investigated the effect of smoothing $\mathbf{S}(t,f)$ over frequency using a moving average filter of length $(2R+1)$. Let $\bar{\mathbf{S}}(t,f) = [\bar{S}_1(t,f),\bar{S}_2(t,f),...,\bar{S}_L(t,f)]^T$ denote the smoothed spectrum, which is computed by

$$\bar{S}_l(t,f) = \sum_{r=-R}^{R} S_l(t,f+r\Delta f)/(2R+1) , \tag{14}$$

where Δf is the STFT frequency bin width and $\mathbf{S}_l(t,f)$ is defined in Eqn. (2).

DOA estimation experiments were carried out on several 1-minute segments extracted from the EasyCom dataset. Altogether $\mathbf{S}(t,f)$ was frequency smoothed using (i) a 9-element filter, (ii) a 3-element filter and (iii) no smoothing. In this experiment, ΔT was fixed to 200 msec. The corresponding results are shown in Fig. 3.
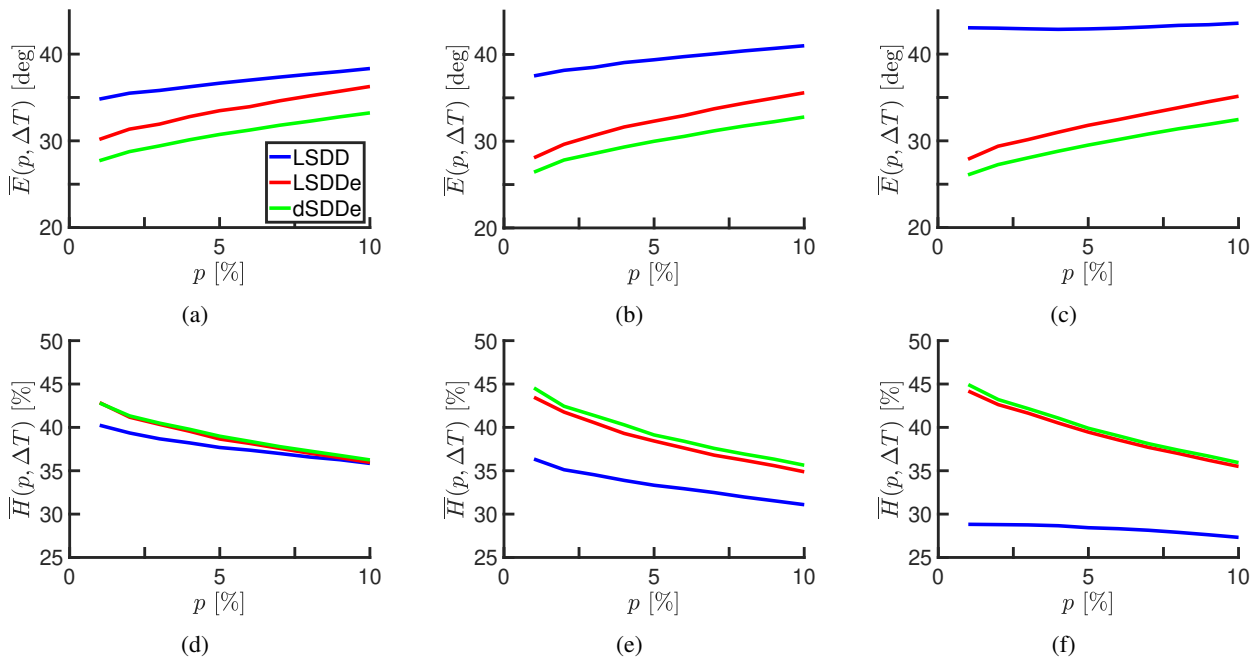

Figure 3. The mean absolute error $\bar{E}(p,\Delta T)$ and the mean hit ratio $\bar{H}(p,\Delta T)$ as a function of p when averaged over ten 1-minute data segments. The time interval was $\Delta T = 200$ msec. (a)-(c) Shows, respectively, $\bar{E}(p,\Delta T)$ with a 9-element smoothing filter, a 3-element smoothing filter and no smoothing. (d)-(f) Show the corresponding $\bar{H}(p,\Delta T)$ curves.

Figure 3 shows that best overall performance in terms of both $\bar{E}(p,\Delta T)$ and $\bar{H}(p,\Delta T)$, for the original LSDD algorithm as described in Sec. 2.2 is obtained with the 9-element filter. On the other hand, for both LSDDe as in Sect. 2.3 and dSDDe as in Sect. 3, the best overall performance is obtained when no smoothing is used. Moreover, for both algorithms the change in performance with respect to smoothing is relatively small.

Finally, we compare the results obtained with the variation of the LSDD algorithm (i.e with a 9-element smoothing filter) and the best variations of the LSDDe and dSDDe algorithms (i.e. no smoothing). At low percentages p the mean absolute error $\bar{E}(p,\Delta T)$ obtained with the dSDDe algorithm is approximately $9°$ lower than that obtained with the LSDD algorithm. Similarly, the mean hit ratio obtained with the dSDDe algorithm is approximately 5% higher than that obtained with the LSDD algorithm.

4.5 DOA Error and Time Interval ΔT

The choice of time interval ΔT may be directly related to the dynamic nature of the dataset. In general, we would like to use a value of ΔT which is small enough such that the environment can be considered spatially stationary within the interval. However if the chosen value of ΔT is too small, DOA performance may degrade. In this experiment, three different values of ΔT were investigated: (a) $\Delta T = 200$ msec; (b) $\Delta T = 300$ msec and (c) $\Delta T = 500$ msec. The DOA/DPD algorithms used the best frequency smoothing as presented in the previous section. Figure 4 shows $\bar{E}(p,\Delta T)$ averaged over ten 1-minute segments for three ΔT values. The figure shows a consistent and significant improvement in performance as ΔT increases. At $\Delta T = 500$ msec. and $p = 1\%$, dSDDe gave a mean absolute DOA error $\bar{E} \approx 20°$. Similar to Fig. 3, dSDDe achieved the best performance, and LSDD the worst.
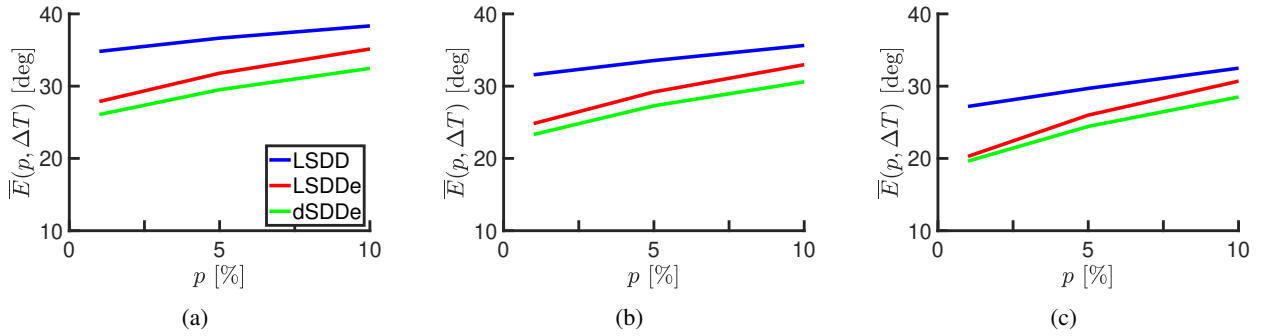
Figure 4. $\bar{E}(p,\Delta T)$ as a function of percentage p for the LSDD, LSDDe and dSDDe algorithms for three different ΔT values, averaged over ten 1-minute segments. (a) $\Delta T = 200$ msec; (b) $\Delta T = 300$ msec; (c) $\Delta T = 500$ msec. Best Frequency smoothing was applied as in Sec. 4.4.

4.6 Summary

Figure 5 illustrates the performance of the dSDDe algorithm together with a timeline of several experimental data values, for an example taken from the EasyCom dataset. In this example there are two active speakers. The time interval ΔT used in the figure is 200 msec.
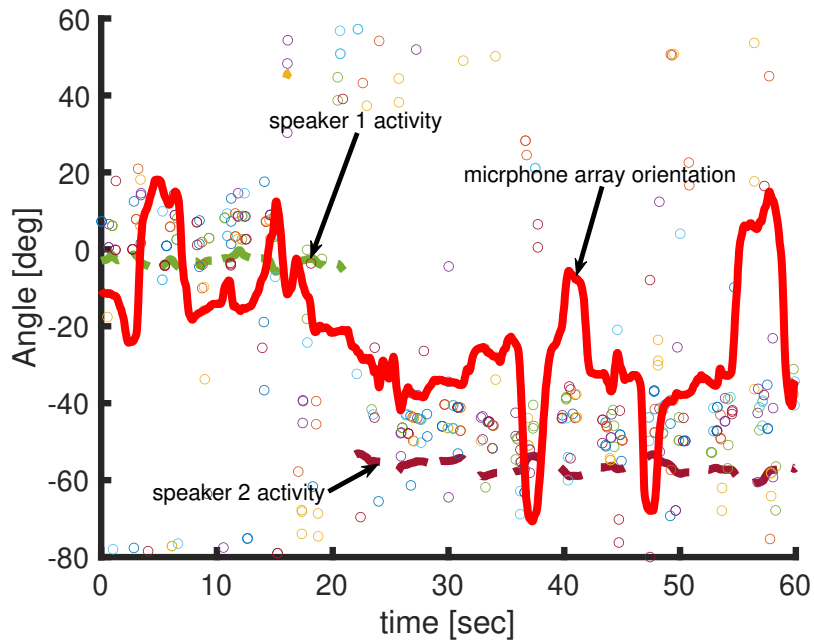

Figure 5. Ground-truth and estimated DOAs for a 1-minute long segment taken from the EasyCom dataset, including two speakers. Ground-truth direction are presented using dashed thick lines. The orientation of the microphone array is marked using thick red line. The DOA estimates $\hat{\theta}(t,f)$ provided by the dSDDe algorithm with $p = 1\%$ and $\Delta T = 200$ msec. are shown as colored circles.

The figure illustrates the dynamic nature of the environment. While much of the DOA estimates fall close to active speakers, some fall very much away from a true DOA. There is also a clear bias in the estimation toward a preferred direction. These findings raise the need for further research to better understand the dataset and the algorithms, and propose improved solutions.

5 CONCLUSION

This work presented three experiments for DOA estimation based on the EasyCom dataset. These preliminary experiments showed:

1. Baseline performance using the original LSDD algorithm showed limited performance.

2. dSDD improved performance by incorporating more detailed spatial information.

3. Energy weighting DPD test values were found to be useful.

4. Both the LSDDe and the dSDDe DOA algorithms do not require frequency smoothing.

REFERENCES

[1] H. Beit-On, V. Tourbabin, and B. Rafaely. The importance of time-frequency averaging for binaural speaker localization in reverberant environments. In *INTERSPEECH*, pages 5071–5075, 2020.

[2] M. Brandstein and D. Ward. *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.

[3] 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*, volume 1, pages 375–378. IEEE, 1997.

[4] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. 2021.

[5] C. Evers and P. A. Naylor. Acoustic slam. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1484–1498, 2018.

[6] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann. The locata challenge data corpus for acoustic source localization and tracking. In *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 410–414. IEEE, 2018.

[7] L. Madmoni and B. Rafaely. Improved direct-path dominance test for speaker localization in reverberant environments. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2424–2428. IEEE, 2018.

[8] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story. The perceptual significance of high-frequency energy in the human voice. *Frontiers in Psychology*, 5:587, 2014.

[9] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[10] V. Tourbabin, D. L. Alon, and R. Mehra. Space domain-based selection of direct-sound bins in the context of improved robustness to reverberation in direction of arrival estimation. In *Proc. 11th European Congress and Exposition on Noise Control Engineering (EURONOISE18)*, pages 2589–2596, 2018.

[11] B. D. Van Veen and K. M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988.

[12] W. Wang and R. Wu. High-dynamic doa estimation based on weighted l 1 minimization. *Progress In Electromagnetics Research C*, 42:253–265, 2013.