DOI: 10.61782/fa.2023.0677

forum acusticum 2023

FREQUENCY SMOOTHING WITH OPTIMAL WEIGHTING FOR IMPROVED SPEAKER LOCALIZATION

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

[1] School of electrical engineering, Ben-Gurion University, Israel
[2] Reality Labs Research @ Meta, Redmond, Seatle, USA

ABSTRACT

Abstract Many speaker localization methods apply frequency smoothing to decorrelate the direct-path signal and coherent reflections. However, in practice, only modest decorrelation is accomplished, which may result in performance loss. In this paper, frequency smoothing weights for improved decorrelation are derived and shown to be inversely proportional to the source signal power at the specified frequency. An experimental study demonstrates the added performance when incorporating the proposed weights in a direct path dominance test-based speaker localization method.

Keywords: *Speaker localization, coherent signal subspace, frequency smoothing, direct path dominance test*

1. INTRODUCTION

Direction-of-arrival (DOA) estimation of speakers from microphone array signals is essential for many audio signal processing applications, including speech enhancement, acoustic scene analysis, and spatial audio rendering. However, accurate DOA estimation in reverberant environments can be challenging due to room reflections that can mask the directional information of the speaker.

The direct path dominance (DPD) test is a family of methods that operate in the time-frequency domain and can provide robustness to reverberation by identifying time-frequency (TF) bins where the direct path is dominant and using only these bins for estimating speaker directions. Popular DPD tests, such as [1–4], employ frequency smoothing of the array's spatial correlation matrix (SCM) to decorrelate the reflections. These methods necessitate a high degree of decorrelation to correctly identify the direct path bins. However, only moderate decorrelation may be obtained in practice, which may degrade performance.

In a recent study [5], the PHALCOR algorithm was proposed for localizing early reflections of a single source in a room using a spherical microphone array. One important aspect of the PHALCOR algorithm is the phase alignment transform of the SCM, which can be seen as a generalization of frequency smoothing. In PHALCOR, it is recommended to normalize each SCM by its trace.

In this paper, we propose a similar approach for frequency smoothing and investigate its performance for localizing multiple simultaneous sources using a general array configuration. We show that choosing frequency smoothing weights to be inversely proportional to the source power at the given frequency is optimal in terms of decorrelation. The proposed weighting is incorporated in the DPD test presented in [1, 3] and tested on real recordings from the LOCATA challenge [6]. The results show that the proposed weighting significantly improves DOA estimation performance.

2. SIGNAL MODEL

Consider a sound field consisting of a single source located in a room. The source emits a frequency domain signal $s(f)$ and has a direction of arrival (DOA) Ω_0, measured relative to a point in the room. The sound waves

*__Corresponding author__: hananb@post.bgu.ac.il.

propagate in the room and reflect from the walls, creating multiple reflections. Each reflection is considered as a separate source with a DOA Ω_k and signal $s_k(f)$, which is a delayed and scaled version of the original source signal [7]:

$$s_k(f) = \alpha_k e^{-i2\pi f \tau_k} s(f) \tag{1}$$

Here, τ_k refers to the delay relative to the direct sound, and α_k represents the scaling factor. Additionally, τ_0 and α_0 are normalized to 0 and 1, respectively.

Assuming that the sources are in the far field, the frequency domain microphone signals vector $\mathbf{x}(f) \in \mathbb{C}^Q$ can be represented using the following model:

$$\mathbf{x}(f) = \mathbf{V}\mathbf{s}(f) + \mathbf{n}(f) \tag{2}$$

Here, $\mathbf{s}(f) \triangleq [s_0(f), \ldots, s_K(f)]^{\mathrm{T}}$ is the source signals vector containing the direct sound and the first K reflections, $\mathbf{n}(f)$ represents noise and late reverberation terms, $\mathbf{V} \triangleq [\mathbf{v}(\Omega_0), \ldots, \mathbf{v}(\Omega_K)]$, where $\mathbf{v}(\Omega)$ refers to the steering vector of the array at the direction Ω. Although the steering vectors are usually frequency-dependent, we will assume that the microphone signals have been transformed by focusing matrices that eliminate the frequency dependence of the steering vectors. For more information on focusing matrices, refer to [3].

Assuming that $\mathbf{n}(f)$ and $s(f)$ are uncorrelated, the SCM, $\mathbf{R}_\mathbf{x}(f) \triangleq \mathbb{E}\left[\mathbf{x}(f)\mathbf{x}(f)^{\mathrm{H}}\right]$, at frequency f is given by:

$$\mathbf{R}_\mathbf{x}(f) = \mathbf{V}\mathbf{R}_\mathbf{s}(f)\mathbf{V}^{\mathrm{H}} + \mathbf{R}_\mathbf{n}(f) \tag{3}$$

where $\mathbf{R}_\mathbf{n}(f) \triangleq \mathbb{E}\left[\mathbf{n}(f)\mathbf{n}(f)^{\mathrm{H}}\right]$, and $\mathbf{R}_\mathbf{s}(f) \triangleq \mathbb{E}\left[\mathbf{s}(f)\mathbf{s}(f)^{\mathrm{H}}\right]$. The coherence of the sources implies that $\mathbf{R}_\mathbf{s}(f)$ is a rank-1 matrix, which causes subspace-based methods such as MUSIC to fail [8].

3. FREQUENCY SMOOTHING

Frequency smoothing of the SCM is applied to decorrelate the sources and increase the rank of $\mathbf{R}_\mathbf{s}(f)$ enabling the application of subspace localization methods [8]. The frequency smoothed SCM, $\bar{\mathbf{R}}_\mathbf{x}(f)$, around center frequency f, is calculated as follows:

$$\bar{\mathbf{R}}_\mathbf{x}(f) \triangleq \sum_{j=-J}^{J} w_j \mathbf{R}_\mathbf{x}(f + j\Delta f) \tag{4}$$

where J controls the smoothing bandwidth, Δf is the frequency resolution, and $w_{-J}, \ldots, w_J$ are the weights.

Substituting Eq. (3) in (4), leads to:

$$\bar{\mathbf{R}}_\mathbf{x}(f) = \mathbf{V}\bar{\mathbf{R}}_\mathbf{s}(f)\mathbf{V}^{\mathrm{H}} + \bar{\mathbf{R}}_\mathbf{n}(f) \tag{5}$$

where $\bar{\mathbf{R}}_\mathbf{s}(f) \triangleq \sum_{j=-J}^{J} w_j \mathbf{R}_\mathbf{s}(f + j\Delta f)$, $\bar{\mathbf{R}}_\mathbf{n}(f) \triangleq \sum_{j=-J}^{J} w_j \mathbf{R}_\mathbf{n}(f + j\Delta f)$. When multiple rank-1 matrices are summed up, the resulting matrix, denoted by $\bar{\mathbf{R}}_\mathbf{s}(f)$, may possess a higher effective rank than the original $\mathbf{R}_\mathbf{s}(f)$. To construct $\bar{\mathbf{R}}_\mathbf{s}(f)$, the weights are selected to either be uniform or proportional to the SNR [9]. Nonetheless, these weights were not designed to optimize decorrelation, which is the primary objective of frequency smoothing. Therefore, better weight selection strategies could potentially enhance decorrelation and consequently improve localization accuracy.

4. OPTIMAL WEIGHTS FOR IMPROVED DECORRELATION

Next, we derive weights that optimize decorrelation. For concision, we omit explicit reference to the central frequency f in the forthcoming discussions. The k, k' element of $\bar{\mathbf{R}}_\mathbf{s}$ can be expressed as:

$$\begin{aligned}\left[\bar{\mathbf{R}}_\mathbf{s}\right]_{k,k'} &= \sum_{j=-J}^{J} w_j \sigma_s^2(f_j)\alpha_k\alpha_{k'}^* e^{-i2\pi f_j(\tau_k - \tau_{k'})} \\ &= \alpha_k\alpha_k^* \mathbf{a}_{\tau_k - \tau_{k'}}^{\mathrm{H}} \mathbf{w}\end{aligned} \tag{6}$$

where $f_j \triangleq f + j\Delta f$, $\mathbf{a}_\tau = [\sigma_s^2(f_{-J})e^{i2\pi\tau f_{-J}}, \ldots, \sigma_s^2(f_J)e^{i2\pi\tau f_J}]^{\mathrm{T}}$, and $\mathbf{w} = [w_{-J}, \ldots, w_J]^{\mathrm{T}}$. We aim to determine the weights $w_{-J}, \ldots, w_J$ that minimize the off-diagonal entries of the matrix $\bar{\mathbf{R}}_\mathbf{s}$, relative to the diagonal entries, using an approach that does not require the unknown reflection delays. To achieve this, we propose to minimize $\left|\mathbf{a}_\tau^{\mathrm{H}}\mathbf{w}\right|^2$, averaged over all possible reflection delays τ:

$$\begin{aligned}\underset{\mathbf{w}}{\text{minimize}} \quad & \Delta f \int_0^{\Delta f^{-1}} \left|\mathbf{a}_\tau^{\mathrm{H}}\mathbf{w}\right|^2 d\tau \\ \text{subject to} \quad & \mathbf{a}_0^{\mathrm{H}}\mathbf{w} = C\end{aligned} \tag{7}$$

The purpose of the equality constraint is to avoid minimizing the diagonal elements of $\bar{\mathbf{R}}_\mathbf{s}$, and the precise numerical value of the constant C is immaterial (so long as it is a positive quantity). Because $\left|\mathbf{a}_\tau^{\mathrm{H}}\mathbf{w}\right|^2$ is periodic in τ with period Δf^{-1}, the range of integration is from zero to

Δf^{-1}. The solution to (7) is given by [10]:

$$\mathbf{w} = \frac{\operatorname{diag}(\mathbf{a}_0)^{-2}\mathbf{a}_0 C}{\mathbf{a}_0^{\mathrm{H}}\operatorname{diag}(\mathbf{a}_0)^{-2}\mathbf{a}_0} \tag{8}$$

which implies:

$$w_j = \frac{C}{\sigma_s^2(f_j)} \tag{9}$$

Equation (9) reveals that the optimal decorrelation weights, in the average sense and in the absence of delay knowledge, exhibit an inverse relationship with the power of the source signal. Given the usual unavailability of the knowledge of $\sigma_s^2(f)$, we resort to its estimation based on the trace of $\mathbf{R}_{\mathbf{x}}(f)$.

5. EXPERIMENTAL RESULTS

This section demonstrates the advantage of the proposed weighting when combined in a DPD test based method for multiple speaker localization [1, 3]. The method employs frequency smoothing and necessitates a high degree of decorrelation to correctly identify the direct path bins.

The DPD test is applied with various frequency smoothing weights to real-world recordings of multiple static speakers in a room with a 12-microphone array mounted on a robot's head from the LOCATA challenge [11]. The data was recorded in a laboratory with an approximate reverberation time of $T_{60} = 0.55$ s and includes thirteen scenarios involving two, three or four static speakers.

The initial recordings underwent a down-sampling process from 48 kHz to 16 kHz and were then transformed by the STFT using a Hamming window of 512 samples with a 50% overlap and FFT length of 512 samples. Next, a WINGS focusing transformation was implemented, using a spherical harmonics order of $N = 12$. The smoothed SCM, denoted as $\bar{\mathbf{R}}_{\mathbf{x}}(i,j)$ with time and frequency indices, was computed at various center frequencies using Eq. (4). The weights used for this computation included uniform weights ($w_j = 1$), SNR weights ($w_j = \sigma_s^2(f_j)$) [9], and decorrelation weights ($w_j = \frac{1}{\sigma_s^2(f_j)}$), where $\sigma_s^2(f_j)$ was estimated using $\operatorname{tr}(\mathbf{R}_{\mathbf{x}}(i,j))$ for both SNR and decorrelation weights. Finally, the DPD test was used to select TF bins in the range of $0.5 - 4$ kHz based on whether the ratio of the first two singular values of $\bar{\mathbf{R}}_{\mathbf{x}}(i,j)$ exceeded a certain threshold. The threshold was set such that a given percentage of the TF bins in that range will pass the test. The root mean square error

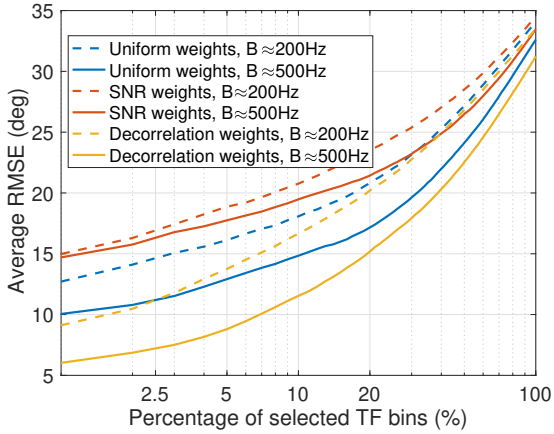

Figure 1. Average (over recordings) RMSE of DOA estimates from selected bins as a function of the percentage of TF bins selected by the DPD test for the various frequency smoothing weights and for various smoothing bandwidths $B = \Delta f\,(2J+1)$ Hz

(RMSE) is defined as:

$$\mathrm{RMSE} \triangleq \sqrt{\frac{1}{|\mathcal{A}|}\sum_{(i,j)\in\mathcal{A}} e(i,j)^2} \tag{10}$$

where $\mathcal{A}$ denotes the set of TF bins selected by the test, and $e(i,j)$ is the angular distance between the DOA estimate at the (i,j) bin and the DOA of closest source to it. The average RMSE as a function of the DPD test threshold is shown in Figure 1, where the results are averaged over multiple recordings. The findings demonstrate that the DPD test incorporating decorrelation weights outperforms the other weighting methods, and that performance improve as the smoothing bandwidth increases. The superior decorrelation with the decorrelation weights, and for wider bandwidths, allows the appropriate selection of direct path bins. Conversely, poor decorrelation, with the uniform and SNR weights, results in TF bins with multiple coherent reflections exhibiting a high singular values ratio (arising from a deficient rank of $\bar{\mathbf{R}}_{\mathbf{s}}(i,j)$), leading to their selection by the test and consequently, degraded localization accuracy.

6. CONCLUSIONS

In conclusion, this paper proposed a novel approach for frequency smoothing in speaker localization methods to improve decorrelation of coherent reflections. The frequency smoothing weights were derived and found to be inversely proportional to the source power at the specified frequency. The proposed weighting was incorporated into a direct path dominance test-based method for speaker localization in reverberant environments, and experiments were conducted on real recordings from the LOCATA challenge. The results showed a significant improvement in DOA estimation performance. The proposed method has the potential to enhance the accuracy of other frequency smoothing based methods localization in various audio signal processing applications.

7. ACKNOWLEDGMENTS

This work has been supported by Reality Labs research @ Meta.

8. 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, no. 10, pp. 1494–1505, 2014.

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

[3] H. Beit-On and B. Rafaely, "Focusing and frequency smoothing for arbitrary arrays with application to speaker localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[4] H. Beit-On, V. Tourbabin, and B. Rafaely, "The importance of time-frequency averaging for binaural speaker localization in reverberant environments," *Proc. Interspeech 2020*, pp. 5071–5075, 2020.

[5] T. Shlomo and B. Rafaely, "Blind localization of early room reflections using phase aligned spatial correlation," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1213–1225, 2021.

[6] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The locata challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[7] 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.

[8] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, 1985.

[9] H. Hung and M. Kaveh, "Focussing matrices for coherent signal-subspace processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 8, pp. 1272–1281, 1988.

[10] H. Beit-On, T. Shlomo, and B. Rafaely, "Weighted frequency smoothing for enhanced speaker localization," Submitted to IEEE TASLP.

[11] "LOCATA website." `www.locata-challenge.org`. Accessed: 2019-10-15.