ACOUSTIC LOCALIZATION USING SPATIAL PROBABILITY IN NOISY AND REVERBERANT ENVIRONMENTS

Sebastian Braun, Ivan Tashev

Microsoft Research,
Redmond, WA, USA

ABSTRACT

In realistic acoustic sound source localization scenarios, we often encounter not only the presence of multiple simultaneous sound sources, but also reverberation and noise. We propose a novel multi-source localization method based on the spatial sound presence probability (SSPP). The SSPP can be computed using prior knowledge of the anechoic relative transfer functions (RTFs), which incorporate magnitude and phase information, and makes the approach general for any device and geometry. From the SSPP we can not only obtain multiple simultaneous sound source direction estimates, but also their spatial presence probability. The SSPP can be used for a probabilistic update of the estimated directions, and can further be used to determine the dominant sound source. We demonstrate the robustness of our method in challenging non-stationary scenarios for single- and multi-speaker localization in noisy and reverberant conditions. The proposed method still localizes a sound source at 8 m with an average error below 7°.

Index Terms— Acoustic localization, direction-of-arrival, spatial probability

1. INTRODUCTION

Acoustic sound source localization receives a lot of recent research attention [1], and is an important task for various applications, among them speech enhancement using beamforming [2], sound source separation [3], spatial acoustic scene analysis or spatial sound object detection, and spatial encoding techniques [4]. The increasing popularity of farfield communication and voice control demands reliable localization of sound sources at several meters distance from the microphones in any reverberant and noisy environment, in presence of potentially multiple directional sound sources.

Sound source localization methods using the time difference of arrival based on the generalized cross-correlation (GCC) [5] often fail in more complex scenarios with multiple sources and reverberation due to violation of the signal model. Most localization methods model the signal in the time-frequency domain as the direct sound wave plus some noise or interference. Due to the spectral sparsity of speech, assuming only a single directional sound wave per time-frequency bin can hold well even in presence of multiple active speech sources. Narrowband direction-of-arrival (DOA) estimators such as MUSIC [6] and ESPRIT [7] can even estimate multiple DOAs per time frequency. However, it is not straightforward to estimate the number of sources, and the combination of the narrowband DOA estimates to one or multiple broadband direction estimates often creates an additional permutation problem. Furthermore, ESPRIT, more efficient MUSIC solutions such as root-MUSIC [8], or spherical harmonic domain methods [9] impose constraints on the microphone geometry and microphone directivity, which prohibits generalization to arbitrary microphone arrays. Neural networks have been used for sound localization [10, 11]. However, creating enough training data and re-training for each specific array can be often impractical. In [12, 13], source separation methods are used to obtain time-frequency masks to obtain multiple GCC-based direction estimates.

In this paper, we propose a sound source localization method that i) is general for any microphone array, ii) can locate multiple simultaneously active, possibly moving sound sources, iii) increases farfield localization accuracy in noisy and reverberant environments, iv) provides a presence probability per estimated source direction, and v) has low requirements on computational complexity, memory and no processing delay. In [14], we proposed a method to compute the probability of a narrowband sound wave arriving from a single spatial location using a relative transfer function (RTF) feature. In this paper, we generalize this method to multiple directions to create a spatial sound presence probability (SSPP) map, which indicates regions of spatial source presence. From the SSPP, we propose to continuously track a maximum number of sound sources, which can be simultaneously active. This way, we obtain estimates of direction and presence of a source at each direction. In contrast to many other multi-source localization methods, the proposed method does not require prior knowledge of the number of active sources, which is often challenging to determine in practice. Compared to the probabilistic source localization method proposed in [15] using GCC features with support vector machines, our proposed approach uses time-frequency sparsity to generalize better for multi-source scenarios. The associated SSPPs for each direction estimate allow us to determine the dominant active sound source in each time frame. We evaluate the proposed sound localization method in challenging scenarios with strong reverberation and large source distances, non-stationary and spatially inhomogenous noise, and changing positions of single and multiple speakers.

2. SIGNAL MODEL

We assume that the sound captured at the microphone with index $m \in \{1, \dots, M\}$ is given in the short-time Fourier transform (STFT) domain by

$$Y_m(k,n) = \sum_{i=1}^{I} H_{m,i}(k)\, S_i(k,n) + V_m(k,n), \qquad (1)$$

where $S_i(k,n)$ are I speech source signals, $H_{m,i}(k)$ are the direct path acoustic transfer functions of source i to microphone m, $V_m(k,n)$ models noise and reverberation, and k and n are the frequency and time frame indices, respectively. Given that the speech

sources $S_i(k,n)$ are spectrally sparse, we assume that there is only a single dominant speech source per time-frequency bin. Hence, we model the microphone signals by

$$Y_m(k,n) = A_{m,1}(k,\mathbf{r}_\mathrm{d})\, X_\mathrm{d}(k,n) + U_m(k,n) \tag{2}$$

where $A_{m,1}(k,\mathbf{r}_\mathrm{d}) = H_{m,\mathrm{d}}(k)/H_{1,\mathrm{d}}(k)$ is the relative direct transfer function of the dominant source $S_\mathrm{d}(k,n)$ between the m-th and 1st microphone, $X_\mathrm{d}(k,n) = H_{m,1}(k)S_\mathrm{d}(k,n)$ is the dominant speech source signal from location $\mathbf{r}_\mathrm{d}$ at the fist microphone, and $U_m(k,n) = \sum_{i\setminus\mathrm{d}} H_{m,i}(k)\, S_i(k,n) + V_m(k,n)$ models the noise, reverberation, and all residual components, such as the non-dominant speech sources. Note that due to the sparsity assumption, the dominant source location $\mathbf{r}_\mathrm{d} \in \{\mathbf{r}_1,\ldots,\mathbf{r}_I\}$ can vary across frequency within the same time frame n. Although we denote the source locations as absolute cartesian vectors $\mathbf{r} = [r_\mathrm{x}, r_\mathrm{y}, r_\mathrm{z}]$, in farfield localization, $\mathbf{r}$ are usually distance-independent unit vectors, and will be referred as *direction* in the remainder of the paper.

3. STATE-OF-THE-ART ACOUSTIC LOCALIZATION

A well-known robust and generally applicable localization method is the steered response power with phase transform (SRP-PHAT) [16]. The direction estimate is obtained by the maximum of the normalized cross-power spectral density (CPSD), steered in all possible directions $\mathbf{r}_\ell \in \{\mathbf{r}_1,\ldots,\mathbf{r}_L\}$ i. e.

$$\widehat{\mathbf{r}}_\mathrm{SRP}(n) = \arg\max_\ell \sum_k \left| \sum_{m=1}^{M} A_m^*(k,\mathbf{r}_\ell)\, \frac{\Phi_{m,1}(k,n)}{|\Phi_{m,1}(k,n)|} \right|^2 \tag{3}$$

where $\Phi_{m,1}(k,n) = \mathrm{E}\{Y_m(k,n)Y_1^*(k,n)\}$ is the cross-power spectral density between the m-th and first microphone signals. The expectation operator $\mathrm{E}\{\cdot\}$ can be approximated by first-order recursive smoothing with a small time constant.

A widely used narrowband localization method designed for uniform circular array (UCA) geometries is beamspace root-MUSIC (BS-RM) [17]. While BS-RM yields similar accuracy but lower computational complexity than the traditional MUSIC algorithm [6, 18], its complexity is still rather high. To obtain potentially multiple robust broadband direction estimates, k-means clustering has been successfully employed to the estimated narrowband directions or spatial features [19, 20]. In our baseline algorithm, we use the slightly more robust k-medians clustering [21] with recursive initialization to obtain multiple broadband direction estimates. Preliminary experiments showed that additional robustness checks such as the coherence test or onset detection as proposed in [20] deteriorated the results in adverse conditions. We assume that the dominant direction is given by the cluster centroid with the largest amount of data points.

4. SPATIAL PROBABILITY BASED LOCALIZATION USING THE RTF INPRODUCT

In [14], we have proposed a method to compute the SSPP of a directional sound source with respect to a single direction based on the inproduct between the estimated and given anechoic RTF. In the following, we formulate the SSPP for multiple directions, such that the maximum SSPP per direction is an indicator for the most probable source direction. The narrowband spatial probabilities are combined into a global broadband SSPP, which can be used to obtain multiple source directions and their presence probabilities. Furthermore, the spatial presence probabilities of the estimated source directions can be used for a probabilistic update to track the source direction estimates.

4.1. Spatial sound presence probability

By neglecting the noise term in (2), the RTF can be estimated from the microphone signals in the least-squares sense by

$$\widehat{A}_{m,1}(k,n) = \frac{\Phi_{m,1}(k,n)}{\Phi_{1,1}(k,n)}. \tag{4}$$

Note that in the presence of noise, the RTF estimate given by (4) is biased. Although there exists a variety of more sophisticated and unbiased RTF estimators [22, 23, 24], we use (4) to keep the computational complexity low.

Let us define the estimated RTF vector and the anechoic RTF vectors for the potential source directions $\mathbf{r}_\ell \in \{\mathbf{r}_1,\ldots,\mathbf{r}_L\}$ as

$$\widehat{\mathbf{a}}(k,n) = \begin{bmatrix}\widehat{A}_{2,1}(k,n) & \ldots & \widehat{A}_{M,1}(k,n)\end{bmatrix}^T, \tag{5}$$

$$\mathbf{a}_\ell(k) = \begin{bmatrix}A_{2,1}(k,\mathbf{r}_\ell) & \ldots & A_{M,1}(k,\mathbf{r}_\ell)\end{bmatrix}^T, \tag{6}$$

which are both vectors of length $M-1$. As a distance measure between the potential and observed RTF vectors, we utilize the normalized vector inproduct, which can also be interpreted as the cosine of the hermitian angle [25, 26, 14]

$$\Delta_{\ell,k,n} = \cos\langle \mathbf{a}_\ell(k), \widehat{\mathbf{a}}(k,n)\rangle = \frac{\Re\left\{\mathbf{a}_\ell^H(k)\,\widehat{\mathbf{a}}(k,n)\right\}}{\|\mathbf{a}_\ell(k)\|\,\|\widehat{\mathbf{a}}(k,n)\|}, \tag{7}$$

where $\Re\{\cdot\}$ is the real part operator. Note that $-1 \le \Delta_{\ell,k,n} \le 1$ is bounded. The feature $\Delta_{\ell,k,n}$ becomes one, when the estimated RTF is close to an anechoic source from $\mathbf{r}_\ell$, otherwise we expect the cosine angle to be smaller than one, or even negative.

Following the concept of RTF-based spatial probabilities [14], we can compute the conditional probability $P(H_\ell|\Delta_{\ell,k,n})$ that the observed time-frequency bin originates from direction $\mathbf{r}_\ell$ by

$$P(H_\ell|\Delta_{\ell,k,n}) = \frac{P(H_\ell)p(\Delta_{\ell,k,n}|H_\ell)}{P(\bar{H}_\ell)p(\Delta_{\ell,k,n}|\bar{H}_\ell) + P(H_\ell)p(\Delta_{\ell,k,n}|H_\ell)} \tag{8}$$

where $P(\bar{H}_\ell) = 1 - P(H_\ell)$. In [14] it was proposed to model the likelihood function for spatial speech presence $p(\Delta_{\ell,k,n}|H_\ell)$ by an exponential function, and the likelihood for speech absence $p(\Delta_{\ell,k,n}|\bar{H}_\ell)$ by a raised cosine function, given in [14, Eqs. (9), (11)], respectively. We assume an equal *a priori* probability ratio of $P(H_\ell) = P(\bar{H}_\ell) = 0.5$.

The global broadband SSPP $P(\mathbf{r}_\ell, n)$ is obtained by the arithmetic average of the narrowband probabilities within the frequency range of interest as

$$P(\mathbf{r}_\ell, n) = \frac{1}{k_\mathrm{max} - k_\mathrm{min} + 1} \sum_{k_\mathrm{min}}^{k_\mathrm{max}} P(H_\ell|\Delta_{\ell,k,n}), \tag{9}$$

where the lower and upper frequency bin limits k_min and k_max should be chosen according to the array aperture and the spatial aliasing frequency, respectively. We propose to apply a temporal smoothing to the global SSPP function $P(\mathbf{r}_\ell, n)$ with fast attack and slow release time constants.

An example of the global SSPP function is shown in Fig. 1 on top for a single speech source with road noise. The ground truth direction of the speech source is shown as black dashed line for

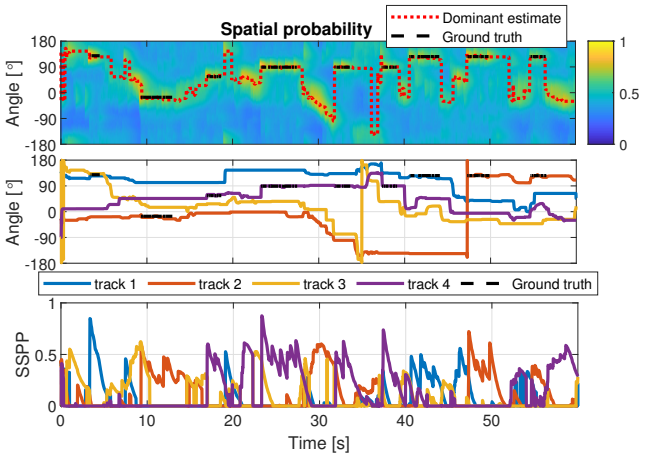

Figure 1: From top to bottom: SSPP, probability per estimated source track, and estimated directions. Example for road noise with SNR $= 10$ dB, $T_{60} = 0.8$ s and distance 4 m.

frames, where the speech source is active. The presence of the speech source can be well observed as regions with high SSPP, while also other highly directional noise components e. g. sound of passing by cars can be observed.

4.2. Probabilistic tracking of directional sources

As we are uncertain how many sound sources are active at frame n, we constantly track a maximum of $I_{\max} \geq I$ source directions $\mathbf{r}_i$ along with their presence probabilities $P(\mathbf{r}_i, n)$ for $i \in \{1, \ldots, I_{\max}\}$.

From the SSPP function $P(\mathbf{r}, n)$, we extract the $I_n \leq I_{\max}$ highest spatial peaks at $\widehat{\mathbf{r}}_{i'}$, $i' \in \{1, \ldots, I_n\}$ at each time step n. Note that determining the peaks of $P(\mathbf{r}, n)$ becomes a 2-dimensional problem, if we consider varying elevation angles. Only spatial peaks above the spatial noise floor are considered. The SSPP noise floor can be tracked by

$$\mu_P(n) = \alpha_P \mu_P(n-1) + (1-\alpha_P)\frac{1}{L}\sum_{\ell=1}^{L} P(\mathbf{r}_\ell, n), \qquad (10)$$

where α_P is a very slow time smoothing constant.

The I_n instantaneous spatial peaks need to be assigned to the existing direction estimate tracks. From the previous time step, we have $I_{\max}$ source direction estimates $\widehat{\mathbf{r}}_i(n-1)$. The I_n instantaneous source directions are mapped uniquely to the closest previous source direction $\widehat{\mathbf{r}}_{i'}(n) \rightarrow \widetilde{\mathbf{r}}_i(n)$ by

$$\widetilde{\mathbf{r}}_i(n) = \arg\min_i \|\widehat{\mathbf{r}}_{i'}(n) - \widehat{\mathbf{r}}_i(n-1)\|_2^2 \quad \forall i' \in \{1, \ldots, I_n\}. \qquad (11)$$

Note that direction book-keeping using cartesian unit vectors avoids the angular wrap-around problem. After the mapping, the $I_{\max}$ direction estimates can be updated using the SSPPs by

$$\widehat{\mathbf{r}}_i(n) = (1 - P_i(n-1))\widehat{\mathbf{r}}_i(n-1) + P_i(n)\widetilde{\mathbf{r}}_i(n), \qquad (12)$$

where the probabilities $P_i(n)$ are compensated for the spatial noise floor by

$$P_i(n) = \frac{P(\widetilde{\mathbf{r}}_i, n) - \mu_P(n)}{1 - \mu_P(n)}. \qquad (13)$$

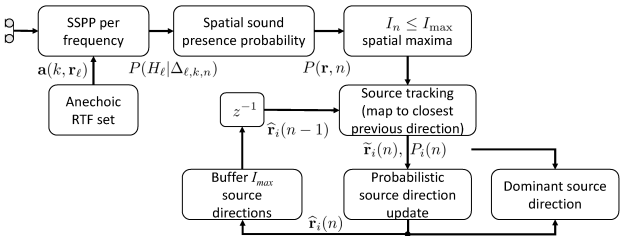

Figure 2: Proposed SSPP-based localization system.

The probabilistic update ensures that direction tracks are only updated, if a source in the respective direction is present, and remain constant during speech absence.

The number of estimated directions $I_{\max}$ should be chosen larger than the maximum number of expected sound sources to avoid distraction of the direction estimate tracks. An example for a single speech source in a road noise scenario is shown in Fig. 1 using $I_{\max} = 4$. The Fig. 1 top shows the estimated SSPP and the true source location as black dashed line. The SSPP map clearly indicates the spatial presence of the speech source, while also other directional noise components are visible. Note that $P(\mathbf{r}_\ell, n)$ is naturally smooth across direction due to smearing by reverberation, which eases peak extraction. Fig. 1 center shows the four estimated direction tracks, while the noise floor compensated SSPPs (13) are shown on Fig. 1 bottom in the same colors. We can observe that the SSPP is only high, when a directional source is active, while the direction tracks of inactive sources are not updated.

For a multi-source estimation task, the final result are the $I_{\max}$ direction tracks along with their SSPPs. For acoustic beamforming, the goal is often to focus on a single speech source. By assuming that the dominant source is the desired one, the dominant source direction per frame can simply be obtained as the direction estimate with the maximum SSPP $\forall i$, which is shown as red dotted line on the top figure in Fig. 1. An overview of the proposed system is shown in Fig. 2.

5. EVALUATION

5.1. Dataset and evaluation criteria

We created two different datasets, a single speech source scenario and a multi speech source scenario. Male and female speech utterances of length between 2 to 5 s from an internal database were concatenated to files of 60 s length by inserting random pauses of 0 to 4 s length between each utterance. The room impulse responses were generated using the image method [27] for a UCA with 6 omnidirectional microphones on a radius of 4 cm by simulating shoebox rooms with the reverberation times $\{0.3, 0.5, 0.8\}$ s. The source position changed randomly after 1 to 4 utterances. For the single-source scenario, the source angle could take any value between $[-179, 180]^\circ$ in 1° steps, while the source-array distance was constant per file with $\{1, 2, 4, 6, 8\}$ m. For the multi-source scenario, the source angle changed randomly after 1 to 4 utterances on an angle grid between $[-179, 180]^\circ$ with a 21° increment, with a randomly selected source distance from $\{1, 2, 4, 6, 8\}$ m per utterance and source. In the multi-source scenario, 3 speech sources were active simultaneously. After convolving the room impulse responses with the utterances and summing the speech signals in the multi-source case, spatial noise recordings from a *bar*, *road*,

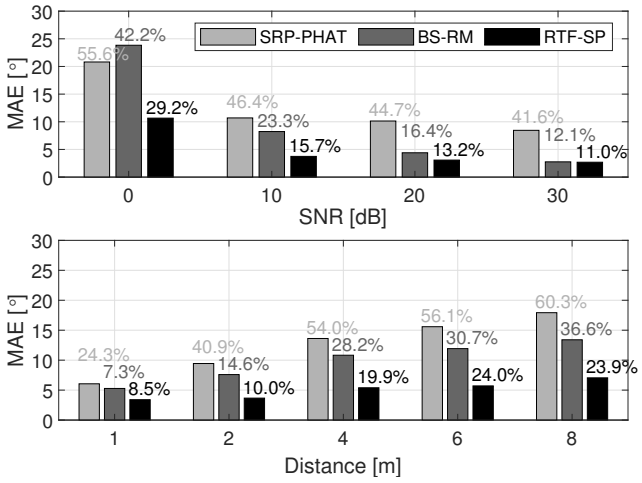

Figure 3: Mean absolute localization error (bars) and miss rate with 5° tolerance (numbers on top) for a single active speech source with changing position over SNR and distance.

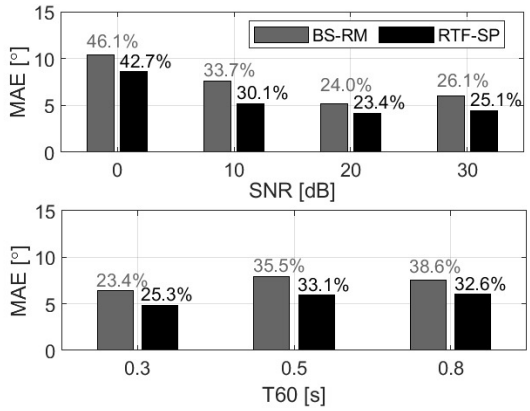

Figure 4: Mean absolute localization error (bars) and miss rate with 5° tolerance (numbers on top) for 3 simultaneous speech sources.

or *office* scenario were added with a signal-to-noise ratio (SNR) of $\{0, 10, 20, 30\}$ dB. The spatial noise was recorded in the Ambisonics format and rendered to the microphone setup [28]. This resulted in a single- and multi-source dataset of 180 and 36 files.

As evaluation criterion, we use the mean absolute error (MAE) between the true and estimated source direction, only computed for frames where the sources are active. Additionally, we compute the localization miss rate within 5° as the ratio of the number of direction estimates, where the absolute localization error is above 5°, related to the total number of active frames for each source. For the multi-source experiment, these error measures are computed between the closest true and estimated directions. This criterion disregards consistency of the source tracks, which is beyond the scope, and not the goal of this paper.

5.2. Implementation

In our experiments, the audio data sampled at 16 kHz was processed using a STFT with square-root Hann windows of 32 ms length and $\Delta t = 16$ ms frame shift. The parameters to compute the narrowband spatial probabilites (8) were chosen as proposed in [14]. The anechoic RTFs $\mathbf{a}_\ell(k)$ were computed from the microphone geometry using the omnidirectional far-field microphone model, where the possible direction set $\{\mathbf{r}_1, \ldots, \mathbf{r}_L\}$ were the azimuth angles in the horizontal plane between $[-175, 180]^\circ$ in 5°. Note that the source angles of the dataset are chosen so that only very few source directions lie exactly on the discrete RTF direction set used in the implementation. The recursive smoothing time constants for estimating the CPSDs were 0.025 s, for the SSPP smoothing of (9) we used 0.002 s attack and 1.2 s release time constants, and for the spatial noise floor tracking (10) α_P had 0.22 s attack and 11.2 s release[1]. We estimated $I_{\max} = 4$ source direction tracks in both the single- and multi-source experiment.

5.3. Results

Figure 3 shows the results for the single-source experiment. The bars indicate the localization MAE in degrees, and the miss rate is shown as numbers in % on top of each bar. The results per SNR averaged over all T_{60} and distance conditions are shown on top, while the results per distance averaged over T_{60} and SNR are shown below. We can observe that the proposed method, denoted as *RTF-SP*, yields the lowest MAE in all conditions, and the lowest miss rate in all conditions except at 1 m distance, where *BS-RM* is slightly better. Note that the computational complexity of *BS-RM* exceeds the complexity of the proposed method by a factor larger than 10. *SRP-PHAT* shows a large discrepancy between MAE and miss rate at low SNR as its direction estimate has a large variance. While the performance of *SRP-PHAT* and *BS-RM* constantly decreases towards lower SNR and larger distances, the performance of *RTF-SP* drops much less significant.

Figure 4 shows the results for the multi-source experiment, depending on SNR (top) and T_{60} (bottom). The proposed method outperforms *BS-RM* except for the miss rate at $T_{60} = 0.3$ s. While the MAE and miss rates at high SNR approximately double compared to the single source scenario, the performance becomes similar at lower SNR. Note that the multi-source evaluation disregards the source assignment problem, and the SSPP is not used in contrast to the single-source experiment. Therefore, we observe smaller errors at low SNRs in Fig. 4. The performance of both *BS-RM* and *RTF-SP* does not decrease significantly for the highest T_{60}.

6. CONCLUSION

We have proposed a noise- and reverberation-robust source localization method for multiple sources. The localization uses spatial probabilities based on a RTF correlation feature incorporating knowledge of the anechoic RTFs in the directions of interest. The method is able to localize multiple simultaneously active sound sources in adverse environments using the spatial sound presence probability. In addition to each direction estimate, the method provides also the presence probability associated to each direction estimate, which indicates source activity and estimation confidence. In a further step, the most likely dominant active source can be determined as the most probable active source. The proposed method was evaluated in realistic and very challenging scenarios with reverberation times up to 0.8 s, distances up to 8 m, and non-stationary, spatially inhomogenous noise.

[1] *Attack* and *release* refers to rising and falling signals, respectively.

7. REFERENCES

[1] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "LOCATA challenge-evaluation tasks and measures," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept. 2018, pp. 565–569.

[2] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[3] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[4] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, June 2007.

[5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[6] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[7] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 984–995, 1989.

[8] B. D. Rao and K. V. S. Hari, "Performance analysis of root-MUSIC," in *Signals, Systems and Computers, 1988. Twenty-Second Asilomar Conference on*, vol. 2, 1988, pp. 578–582.

[9] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 178–192, Jan. 2017.

[10] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 2814–2818.

[11] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, March 2019.

[12] V. G. Reju, R. S. Rashobh, A. H. T. Nguyen, and A. W. H. Khong, "An efficient multi-source DOA estimation algorithm for underdetermined system," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept. 2018, pp. 86–90.

[13] Z. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 178–188, Jan. 2019.

[14] S. Braun and I. Tashev, "Directional interference suppression using a spatial relative transfer function feature," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[15] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sep. 2014, pp. 99–103.

[16] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Apr. 1997, pp. 375–378.

[17] M. D. Zoltowski and C. P. Mathews, "Direction finding with uniform circular arrays via phase mode excitation and beamspace root-MUSIC," vol. 5, pp. 245–248, 1992.

[18] J. P. Dmochowski and J. Benesty, "Microphone arrays: Fundamental concepts," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer, Jan. 2010, ch. 11.

[19] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[20] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust DOA estimation of multiple speech sources," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2287–2291.

[21] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[22] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sept. 2004.

[23] M. Schwab, P. Noll, and T. Sikora, "Noise robust relative transfer function estimation," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sept 2006, pp. 1–5.

[24] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, March 2017, pp. 11–15.

[25] J. L. Coolidge, "Hermitian metrics," *Annals of Mathematics*, vol. 22, no. 1, 1920.

[26] D. Y. Levin, E. A. Habets, and S. Gannot, "Near-field signal acquisition for smartglasses using two acoustic vector-sensors," *Speech Communication*, vol. 83, pp. 42 – 53, 2016.

[27] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[28] H. Gamper, L. Corbin, D. Johnston, and I. J. Tashev, "Synthesis of device-independent noise corpora for speech quality assessment," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept. 2016, pp. 1–5.