LOCALIZATION OF SOUND SOURCES WITH KNOWN STATISTICS IN THE PRESENCE OF INTERFERERS

Kainan Chen, Jürgen T. Geiger, Karim Helwani, and Mohammad J. Taghizadeh

Huawei European Research Center, Munich, Germany

ABSTRACT

Methods are available for simultaneous localization of multiple (unknown) audio sources using microphone arrays. Typical algorithms aim at localizing all active sources. They moreover require that the number of sources is known and is less than or equal the number of microphones. This constraint cannot be satisfied in many real-life situations and noisy environments. We present an algorithm for localizing an audio source with known statistics in a multi-source environment. The proposed method circumvents the mentioned problems by using a phase-preserving signal extraction method on the input signal. A binary mask is estimated and used to retain only the information of the target source in the original microphone signals. The masked signals are fed to a modified version of a conventional localization algorithm, which now localizes only the target source. Experimental results obtained from real recordings show that the proposed method can successfully detect and localize the target source.

Index Terms— Sound Source Localization, DOA, NMF, Speaker Recognition

1. INTRODUCTION

Sound source localization is commonly done using microphone array(s), and there are many high-accuracy algorithms which can localize multiple simultaneously active sources, typically, done either by an estimation of either Time Difference Of Arrival (TDOA) or Direction Of Arrival (DOA). For DOA estimation, typically, subspace approaches are applied, such as MUltiple SIgnal Classification (MUSIC) [1], Estimation of Signal Parameter via Rotational Invariance Technique (ESPRIT) [2]. To cope with axis-symmetric array geometries, several algorithms have been proposed such as Eigenbeam Processing [3]. These are approaches that extend the subspace approaches to be applied in suitable transform-domains. In TDOA-based localization, multi-channel filtering can be used to estimate the time difference. Prominent examples for TDOA estimation are the generalized cross-correlation (GCC) method [4], adaptive eigenvalue decomposition (AED) [5], and TRINICON [6]. In order to improve the robustness of the GCC method, a weighting scheme of the GCC functions was proposed. Maximum likelihood estimation of the weights has been considered in the presence of uncorrelated noise, while the phase transform is an efficient and reliable approach to overcome effect of reverberation [7, 8]. Further advancements rely on identification of the speaker-microphone acoustic channel for reverberant speech localization [9, 10]. Other strategies have been sought for multiple-source localization and tracking, such as in [11]. A related scenario is a reverberant environment with background noise, where a Minimum Variance Distortionless Response (MVDR) beamformer can be used to localize the source and enhance the signal [12, 13].

The research leading to these results has received funding from the European Commision Union Seventh Framework Programme (FP7/2007/2013) under grant agreement 607480 LASIE.

1.1. Related work

To problem we address in this work is to detect and localize a specified source with known statistics, for example for forensic applications, where training data are available for a speaker, and the goal is to detect and localize the speaker in a crowded environment. For multi-source localization, the framework of TRINICON [6, 14] can be used. However, in addition to the requirement of having less sources to be localized than microphones it doesn't offer a method to distinguish the localized sources. In all briefly reviewed algorithms so far, the target is to localize all active sound sources regardless whether it is of interest or no. This fact, in combination with the typical limitation to have less or equal active sources than microphones makes sound localization in public spaces rather challenging. Approaches to cope with the limitation of the number of sources such as [15] generally require to learn the acoustic environment which is inapplicable in the scenario of a public space where the acoustic properties of the environment can change drastically in uncontrolled manner. A verified speaker localization method is proposed in [16]. This technique relies on recovery of the harmonics of a desired speech signal in noisy and reverberant condition. In [17], a related problem is addressed where a noise-free version of the target signal is available. The system presented in [18] addresses the problem of unsupervised speech extraction and localization. In [19], a spectral mask was estimated with a deep neural network, as preprocessing to a localization algorithm to reduce the noise. In general, source extraction methods are not sensitive in phase separation, such as [20], which delivers good extraction in source spectrum magnitude, but the phase is different to the original. As both DOA or TDOA algorithms are very sensitive in phase differences between the channels, using such sound source extraction method as preprocessing for localization shouldn't work if the phase is not guaranteed.

1.2. Contribution

The idea of the proposed method is to use blind source separation, such as Non-negative Matrix Factorization (NMF), as a preprocessing step before the actual localization algorithm. Monaural source separation is applied to the mixture of all microphone signals from the array. From the separated signals, the target source (a speaker, in our case), is identified using simple speaker identification techniques. A binary mask is estimated using the separated signals and the identification result, in order to retain only the information of the target source in the original microphone signals. The filtered signals are processed with the ESPRIT localization algorithm, which is

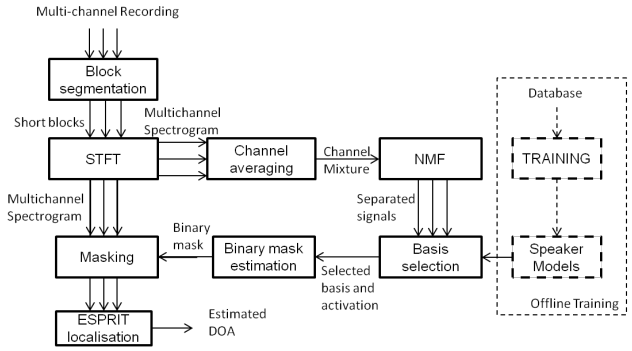

Fig. 1. System overview

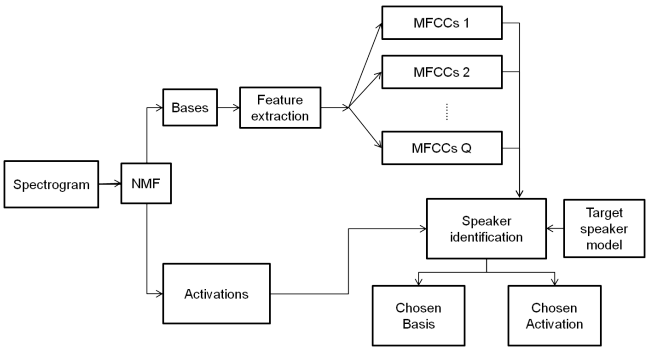

Fig. 2. Details of the source separation and basis selection process

subject to a small modification, such that only the target source is localized. Experiments were performed with real recordings, and the results show that the proposed method successfully manages to localize only the target source, while the other sources are suppressed.

2. PROPOSED APPROACH

The overview of our approach is shown in Fig. 1. In an offline part, speaker models of potential target speakers are trained from a training database. We model the speaker signals with Gaussian Mixture Models (GMMs) using MFCCs as features. The online part of the system starts with a short-time Fourier transform of all microphone channels. Single-channel blind source separation using NMF is applied to the magnitude spectrogram mixture (simple averaging) of all channels. The separated signals are classified using the trained models in order to find target speaker among them. Then, the multichannel microphone signals are multiplied with a binary mask which is derived from the results of source separation and speaker identification. The masked signals are used as input to a modified version of ESPRIT localization, which results in a DOA estimate for the target speaker.

It is assumed that the number of sources is known, although, due to the source separation approach, this number is not critical for the performance. While the number of sources is generally not limited, in our experiments, we tested the method with two and three active sources from different DOAs.

2.1. Target source training

Since we want to localize an audio source with known statistics, these statistics need to be learned in an offline training phase. A more difficult, but also more useful scenario is when both the target source and the other sources are different persons speaking. Thus, in this work, we focus only on speech sources and use speaker recognition methods to train and identify the target speakers.

An established method in speaker recognition is to use MFCCs as features and model these with GMMs [21]. In our scenario, the speaker recognition task is rather simple – identifying one specified target speaker from only a few sources, and therefore we use this approach.

2.2. Non-negative matrix factorization

In the localization step of our method, we first convert each channel of a signal into a magnitude matrix $N \times F$ in time-frequency domain by STFT, where N is the length of Discrete Fourier Transform (DFT), F is the number of frames. We define the spectrogram of the p_{th} channel as $\mathbf{V}_p$, and the averaged spectrogram as $\mathbf{V}$. We factorize the matrix $\mathbf{V}$ using NMF [22]. The NMF factorizes the nonnegative matrix $\mathbf{V}$ into two nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$,

$$\mathbf{V} = \mathbf{W}\mathbf{H}. \tag{1}$$

Assume $\mathbf{V}$ is an $N \times F$ matrix, the dimension of the factorization is r. After the factorization, $\mathbf{W}$ is an $N \times r$ basis matrix, and $\mathbf{H}$ is an $r \times F$ activation matrix. We use the implementation described in [22]. To approximate the factorization in Eq. (1), the method defines a cost function to minimize the Kullback-Leibler divergence between $\mathbf{V}$ and $\mathbf{WH}$:

$$\mathrm{D}(\mathbf{A}||\mathbf{B}) := \sum_{i,j} \mathbf{A}_{ij} \log \frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}} - \mathbf{A}_{ij} + \mathbf{B}_{ij} \tag{2}$$

where i, j denote the i-th row and j-th column of the matrices $\mathbf{A}$ and $\mathbf{B}$. The dimension r of this factorization is equal to the number of active sources Q. The rows of the factorized matrix $\mathbf{W}$ are the spectral features of each source, and the columns of matrix $\mathbf{H}$ are their activations. $\mathbf{W}$ and $\mathbf{H}$ are computed with iterative update rules, which ensure that the distance D is minimized. We suggest using the multiplicative update rules for divergence distance, which are proposed in [22].

The details of the basis selection step from Fig. 1 are shown in Fig. 2. MFCC features are extracted for each of the separated bases, and, using the trained model of the target speaker, the basis corresponding to the target speaker s is found with maximum likelihood classification. As each basis maps to an activation vector, we choose the activation vector that belongs to our chosen basis. We take $\mathbf{W}_s$ (column vector) and $\mathbf{H}_s$ (row vector) as the chosen basis and activation vector.

2.3. Binary mask

The signal extraction step of our system is performed on a monaural mixture of all microphone signals. In this section, we describe an approach to estimate a binary mask that is applied on the original microphone signals. The main purpose is to find when and at which frequencies the specified target sound source is active, so we choose elements from the spectrogram based on the chosen basis and activations, using a binary mask.

First, we define the reconstruction of the target source as $\mathbf{\Psi}$,

$$\mathbf{\Psi} = \mathbf{W}_s \times \mathbf{H}_s, \tag{3}$$

which is an $N \times F$ matrix, in time-frequency domain, and its entry Ψ_{ij} in i_{th} row and j_{th} column denotes the energy of the basis in i_{th} frequency band and j_{th} time frame.

The binary mask is determined as

$$V_{ij}^{'} = \begin{cases} 1 & \Psi_{ij} > \tau V_{ij}, \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathbf{V}^{'}$ denotes mask matrix of the selection result and $V_{ij}^{'}$ denotes its entry in i_{th} row and j_{th} column. The variable $\tau\ (0 < \tau \leq 1)$ is a threshold. We suggest setting τ between 0.3 and 0.5. In block-based processing, such as real-time processing, only a few frames are analyzed together. The choice of $\tau > 0.3$ guarantees that there is enough information for localization. It is not necessary for the specified source to be predominant (the localizaton method is introduced below), and therefore τ is not set higher than 0.5, otherwise there would be too much useful information wasted. The selected spectrogram in the p_{th} channel is $\mathbf{V}_p^{'}$,

$$\mathbf{V}_p^{'} = \mathbf{V}_p \circ \mathbf{V}^{'}, \tag{5}$$

where $\circ$ denotes Schur Product.

There are two reasons we use binary mask instead of soft mask in Eq. (4). First, we factorize the averaged spectrogram, and the energy of the specified source is not uniformly distributed in each channel. Therefore we can't apply the same soft mask to all channels. Second, applying a soft mask or a Wiener filter, e.g. as proposed in [23, 24], corresponds to an amplitude modification. This leaves the phase unchanged and thus does not improve the eigenvalue energy distribution in the subsequent localization step.

2.4. DOA estimation based on time-frequency information

We prefer to use ESPRIT as the DOA estimation algorithm if the microphone array fits its requirement. The original algorithm is introduced in [2]. It was designed for narrow-band signals and can estimate DOA from the multi-channel signal matrix directly in time domain. Our signals are broadband speech signals, and therefore we have to transform the signal into frequency domain and run ESPRIT in each narrow band. We use short-time Fourier analysis of overlapping frames (parametrisation details are given in Section 4.2). The covariance matrix $\mathbf{R}_{f_i}$ is computed for narrow bands in frequency domain,

$$\mathbf{R}_{f_i} = \mathbf{V}_{f_i}^{'} \cdot \mathbf{V}_{f_i}^{'\text{H}}, \tag{6}$$

where $\mathbf{V}_{f_i}^{'}$ denotes the selected narrow band spectrogram with all the channels together, defined as

$$\mathbf{V}_{f_i}^{'} = \left[\mathbf{V}_{1f_i}^{'}, \ldots, \mathbf{V}_{Pf_i}^{'}\right]^{\text{T}}, \tag{7}$$

where f_i denotes the i_{th} frequency band, and P denotes the number of microphones in the array.

In the conventional ESPRIT algorithm, the second step is to decompose the covariance matrix $\mathbf{R}_{f_i}$ into its eigenvalues and eigenvectors,

$$\mathbf{R}_{f_i}\mathbf{U} = \boldsymbol{\lambda}\mathbf{U}, \tag{8}$$

where $\mathbf{U}$ denotes eigenvector matrix, and $\boldsymbol{\lambda}$ denotes the eigenvalue diagonal matrix. Then, the eigenvectors with first highest Q (source number) eigenvalues are taken into further processing. In our method, however, as we use a threshold parameter τ for the binary mask, the eigenvector corresponding to the specified sound source might not always have the largest eigenvalue. Thus we take only one eigenvector with the highest relationship to the specified sound source, as we describe in the following. We first estimate an SNR in this narrow band signal,

$$\delta_{f_i} = \overline{\sum_{j=1}^{F} \frac{\Psi_{ij}}{V_{ij}}}, \tag{9}$$

where δ_{f_i} denotes the averaged SNR through frames in the i_{th} frequency band of the averaged spectrogram $\mathbf{V}$. The SNR is defined as the ratio of the energy of the specified speaker to the full energy. Then we compute the energy E_m in the m_{th} eigenvector,

$$E_m = |\lambda_{mm}|, m = 1, \ldots, P. \tag{10}$$

The sum of the eigenvectors energy is E,

$$E = \sum_{m=1}^{P} E_m \tag{11}$$

and $\delta_m^{'} = \frac{E_m}{E}$ is the energy ratio for the m_{th} eigenvector: The eigenvector $\mathbf{U}_s$ is chosen by the eigenvalue which has the similar energy ratio ($\delta_m^{'}$) to the SNR δ_{f_i}:

$$\mathbf{U}_s = \left\{ \mathbf{U}_m \middle| \min_m |\delta_m^{'} - \delta_{f_i}| \right\}. \tag{12}$$

With the chosen $\mathbf{U}_s$, the subsequent steps in ESPRIT are exactly the same as in [2], and we achieve an estimated DOA for each frequency band. In order to obtain one localization result per frame, the results over all frequencies are taken and a peak detection method is applied on the histogram.

3. EVALUATION

3.1. Recordings

The proposed method is evaluated with real recordings. Recordings were taken in an acoustically treated room with the dimensions $5\text{m} \times 10\text{m} \times 3\text{m}$. The room has sound absorbing material hanging on each wall to make sure the reverberation time RT_{60} attains values between 0.5 and 0.7 s.

The linear microphone array consists of five uniformly distributed AKG C562 CM microphones (4 cm distance from each other). We recorded samples with two or three speakers, where one of them is the specified target sound source. The sources were placed at different angles, at a distance of 3 m from the microphone array.

Speech samples were taken from the TIMIT database [25], with leading and trailing silence cut off. Target source and interference sources were always taken from different speakers, and multiple samples were connected to have 10 s long recordings. For training the target speaker model, 35 s recordings of the specified target speaker were used (clean and disjunct from the test recordings).

3.2. Parametrisation

During the localization phase, the recording is segmented into blocks with a length of 1 s, with a hop size of 0.25 s. Localization results are computed for each of these blocks. If the length of the blocks is set too small, the quality of NMF in blind source separation will be low. If it is to too large, the basis would be too general to describe the spectrum in each frame.

In each block, the STFT is computed for smaller frames, with a length of 4 096 samples and a hop size of 2 048 samples, using Hann windows. The sampling rate is 44 100 Hz. NMF is performed on magnitude spectrograms, and the number of bases r is set to the number of active sources. The number of iterations is set to 200 (normally it converges faster than this). The threshold τ for the mask estimation is set to $\tau = 0.4$.

For speaker modelling, we use the standard configuration of 13 MFCC coefficients in the frequency band of 0-8 000 Hz, together with delta and delta-delta coefficients, resulting in 39 features per

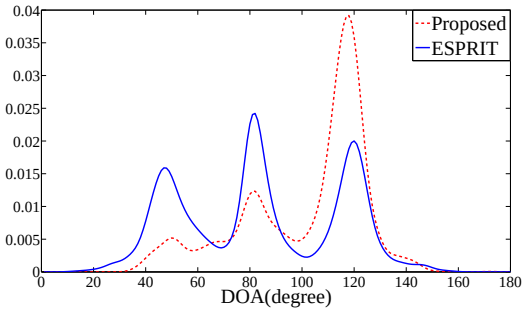

Fig. 3. Normalised histogram of DOA estimates for one block, comparing the proposed method with ESPRIT

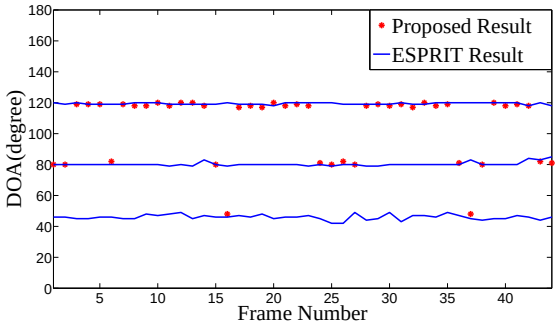

Fig. 4. Block-wise results for DOA estimation

block. The window size and hop size are the same as in the STFT for localization processing. Possible target speakers are modelled with GMMs with diagonal covariance matrices. We employed 32 mixture components, which represents a good trade-off, considering a scenario with low amounts of training data and short blocks of testing data.

3.3. Results

Fig. 3 shows the results for one block of a recording with three active sources as a normalised histogram of DOA estimates. The proposed method is compared with the conventional ESPRIT algorithm. The specified target source is played from $120°$, and two other sources are played from $45°$ and $80°$, respectively. Results for ESPRIT are represented by the solid line, that we can't specify our target source's DOA from it. The dashed line shows the result of the proposed algorithm. It can be seen that the peak of the target source is clearly highlighted, while the other two peaks are attenuated. Therefore, the target source can be found successfully with a simple peak detection method. Fig. 4 shows the detailed block-wise results in this experiment after peak detection. In general, the proposed method is quite robust in detecting only the target speaker. Some systematic errors are made, for example in low-energy regions of the target speaker.

We have done experiments with 3 active sources like the one shown above, and with 2 active sources, from $[45°, 90°]$ and $[60°, 90°]$, where the specified sound source is from $90°$ in both cases. Summarized results are shown in table 1. The second row "DOA error" shows the average DOA error of the proposed method, suggesting that the DOA results of the proposed method are robust. In the third row "Tgt detection with spkID", the average block-wise accuracy of detecting the correct source is shown, including the GMM speaker recognition. Better results are achieved with only two active sources. In order to analyze the performance independently of the speaker recognition method, the second last row shows results using an oracle speaker recognition that doesn't make mistakes. For reference, the last row of the table shows the GMM speaker identification accuracy. Because the NMF does a very rough unsupervised blind source separation, it leads to a higher GMM speaker recognition error rate as compared to single source speaker recognition.

Source angle	$45°$ $80°$ $120°$	$60°$ $90°$	$45°$ $90°$
DOA error	$1.52°$	$0.43°$	$0.54°$
Tgt detection with spkID	66.7%	84.4%	84.4%
Tgt detection ora. spkID	77.8%	88.9%	86.7%
GMM accuracy	75.6%	86.7%	84.4%

Table 1. Summarised results in terms of DOA error, target detection with real and oracle speaker identification, and GMM accuracy

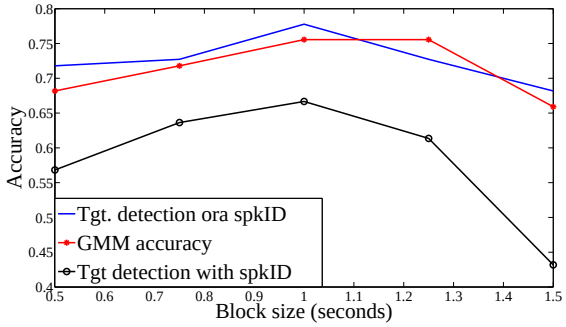

Fig. 5. Results for three sources, depending on the block size

We explored the influence of the block size for the recording with 3 sources. For the results in Table 1, an analysis block size of 1 s was used. Results for different block sizes (keeping the hop size constant at 0.25 s) are shown in Fig. 5. The choice of 1 s delivers the best results, while for larger block sizes, the NMF result becomes inaccurate, and for smaller block sizes, there is not enough information included for the NMF separation process.

4. CONCLUSIONS

We proposed a method for audio source localization with known source statistics. The method uses unsupervised source separation as a preprocessing step to a conventional localization algorithm. From the separated sources, a specified target source is identified and a spectral mask is estimated. The original microphone array signals are processed with this mask and used for source localization. Our results with real recordings show that the method is successful at detecting and localizing the target source in a multi-source environment.

The result of the source separation step is important for both speaker recognition and localization. Currently we use the averaged spectrum across all channels as input to the NMF. Using the relation between channels is a way to improve the quality, such as [26], which delivers better blind source separation. Since we only take the magnitude spectrum in the result of BSS, one direction of future work is towards a less complex multi-channel NMF version to improve the accuracy of our algorithm.

5. REFERENCES

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

[2] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 37(7), pp. 984–995, 1989.

[3] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 89–92.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 24(4), pp. 320–327, 1976.

[5] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *the Journal of the Acoustical Society of America*, vol. 107(1), pp. 384–391, 1999.

[6] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: a versatile framework for multichannel blind signal processing," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, pp. 889–92.

[7] M. Omologo and P. Svaizer, "Acoustic source localization in noisy and reverberant environments using csp analysis," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996, pp. 921–924.

[8] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Elsevier Signal Processing*, vol. 92(8), pp. 1950–1960, 2012.

[9] F. Ribeir, C. Zhang, D. Florencio, and D. Ba, "Using reverberation to improve range and elevation discrimination in sound source localization," in *IEEE Transactions on Acoustics, Speech, Signal Processing*, 2010, pp. 731–736.

[10] S. Nam and R. Gribonval, "Physics-driven structured cosparse modeling for source localization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 5397–5400.

[11] F. Nesta, P. Svaizer, and M. Omologo, "Cumulative state coherence transform for a robust two-channel multiple source localization," in *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 290–297.

[12] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10 (3), pp. 538–548, 2008.

[13] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 18(1), pp. 158–170, 2009.

[14] A. Lombard, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using blind adaptive MIMO system identification," in *Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006, pp. 7–12.

[15] R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011, pp. 245–248.

[16] A. Asaei, M. J. Taghizadeh, M. Bahrololum, and M. Ghanbari, "Verified speaker localization utilizing voicing level in split-bands," *Elsevier Signal Processing*, vol. 89(6), pp. 1038–1049, 2009.

[17] M. Farmani, M. S. Pedersen, Z. Tan, and J. Jensen, "Maximum likelihood approach to informed sound source localization for hearing aid applications," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 16–20.

[18] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 19(7), pp. 1900–1912, 2011.

[19] W. Q. Zheng, Y. X. Zou, and C. Ritz, "Spectral mask estimation using deep neural networks for inter-sensor data ratio model based robust DOA estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 325–329.

[20] T. Otsuka, K. Ishiguro, T. Yoshioka, H. Sawada, and H. Okuno, "Multichannel sound source dereverberation and separation for arbitrary number of sources based on bayesian nonparametrics," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 22(12), pp. 2218–2232, 2014.

[21] D. Reynolds, "An overview of automatic speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. 4072–4075.

[22] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds., pp. 556–562. MIT Press, 2001.

[23] A. Spriet, M. Moonen, and J. Wouters, "Spatially preprocessed speech distortion weighted multi-channel wiener filtering for noise reduction," *Elsevier Signal Processing*, vol. 84(12), pp. 2367–2387, 2004.

[24] H. Cho, J. Choi, and H. Ko, "Robust sound source localization using a wiener filter," in *Proc. IEEE Conference on Emerging Technologies & Factory Automation (ETFA)*, 2013, pp. 1–6.

[25] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM," Philadelphia: Linguistic Data Consortium, 1993.

[26] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 18(3), pp. 550–563, 2010.