Exploiting Non-Negative Matrix Factorization for Binaural Sound Localization in the Presence of Directional Interference

Ornolfsson, Ingvi; Dau, Torsten; Ma, Ning; May, Tobias

Published in:
Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing

Link to article, DOI:
10.1109/ICASSP39728.2021.9414233

Publication date:
2021

Document Version
Peer reviewed version

Citation (APA):
Ornolfsson, I., Dau, T., Ma, N., & May, T. (2021). Exploiting Non-Negative Matrix Factorization for Binaural Sound Localization in the Presence of Directional Interference. In *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 221-225). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414233

EXPLOITING NON-NEGATIVE MATRIX FACTORIZATION FOR BINAURAL SOUND LOCALIZATION IN THE PRESENCE OF DIRECTIONAL INTERFERENCE

Ingvi Örnolfsson**, Torsten Dau***, Ning Ma* † *and Tobias May* *

†Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
* Department of Health Technology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

ABSTRACT

This study presents a novel solution to the problem of binaural localization of a speaker in the presence of interfering directional noise and reverberation. Using a state-of-the-art binaural localization algorithm based on a deep neural network (DNN), we propose adding a source separation stage based on non-negative matrix factorization (NMF) to improve the localization performance in conditions with interfering sources. The separation stage is coupled with the localization stage and is optimized with respect to a broad range of different acoustic conditions, emphasizing a robust and generalizable solution. The machine listening system is shown to greatly benefit from the NMF-based separation stage at low target-to-masker ratios (TMRs) for a variety of noise types, especially for non-stationary noise. It is also demonstrated that training the NMF algorithm on anechoic speech provides better performance than using reverberant speech, and that optimizing the source separation stage using a localization metric rather than a source separation metric substantially increases the system performance.

***Index Terms*—** Binaural sound source localization, non-negative matrix factorization, source separation, directional interference

1. INTRODUCTION

Automatic localization of sound sources is relevant for many audio applications, such as the steering of a beamformer for use in hearing aids [1]. A beamformer can be used to enhance a particular target direction and reduce directional interferences from other directions. For binaural machine listening, localization is a fairly trivial task to solve when only one speaker is present. However, when the task is to localize one specific sound in the presence of interfering directional sources, it becomes more difficult for machines to achieve human performance. Humans use interaural time differences (ITDs) and interaural level differences (ILDs) to localize sounds [2], and these cues have been shown to be useful for sound localization by machines as well [3, 4]. This study is concerned with the task of localizing a speech source in the presence of directional masking noise. In this situation, the target speaker and the masker provide competing ITDs and ILDs, and those alone will not be sufficient to consistently localize the target speaker.

Only a few studies have directly addressed the problem of localizing sounds in the presence of directional masking noise. One approach has been to introduce a speech separation stage into the localization algorithm, which allows one to selectively weight the evidence from individual time-frequency units based on the probability that the target source is dominant in that particular unit. A measure of this probability can be obtained by means of various source separation methods, each of which have their own advantages and drawbacks. An approach based on a Gaussian mixture model has been shown to be useful in the same task as is considered in this study, but it either requires *a priori* knowledge about the nature of the masking sound, or assumes that the masker can be modelled by a universal background model based on a few distinct noise types [5]. DNN-based approaches have also shown promising results [6, 7], but the complexity of these models makes them potentially unsuitable for applications where computational power is limited, for example in hearing aids. Classical noise reduction techniques can be used as learning-free speech separation approaches, but these are typically unable to deal with non-stationary noise, since they assume that the noise varies more slowly than the speech [8, 9, 10]. Common to all source separation methods is that their hyperparameters are typically optimized with respect to minimizing the distortion and artifacts of the target signal [11]. However, these metrics do not necessarily provide the best signal separation for a localization task, and it may be advantageous to optimize the separation stage with respect to a localization metric instead.

The aim of this study was twofold. The first goal was to investigate whether a source separation method based on non-negative matrix factorization (NMF) can aid sound localization. The NMF-based approach has a number of potential advantages over the previously mentioned ones. It allows one to train a model ahead of time based only on the target source and learn the representation of the interfering sources during the actual separation, thus making the model robust to unseen source types [12]. Moreover, an NMF-based approach does not assume stationarity of the noise source, and is likely to be superior to classical noise reduction techniques when the interfering noise is non-stationary. The second goal of the study was to investigate whether optimizing the separation stage with respect to a localization metric would improve performance compared to optimizing it with respect to classical separation metrics.

2. SYSTEM DESCRIPTION

2.1. System overview

A schematic of the combined separation-localization system is shown in Fig. 1. A binaural noisy mixture is fed separately to two parallel stages, the NMF-based separation stage and the DNN-based localization stage.

The separation stage divided the noisy mixture sampled at a rate of 16 kHz into overlapping frames of 20 ms duration with a shift of 10 ms. Each frame was Hann-windowed and zero-padded to a length of 512 samples, and a short-time discrete Fourier transform (STFT) was computed. To further reduce the dimensionality of this representation, a gammatone filterbank was used to integrate the STFT representation into 32 frequency bands spaced according to the equivalent rectangular band (ERB) scale, with center frequencies from 80 Hz to 8 kHz. In the following, the matrix $\mathbf{X}$, defined in Eq. (1), will be

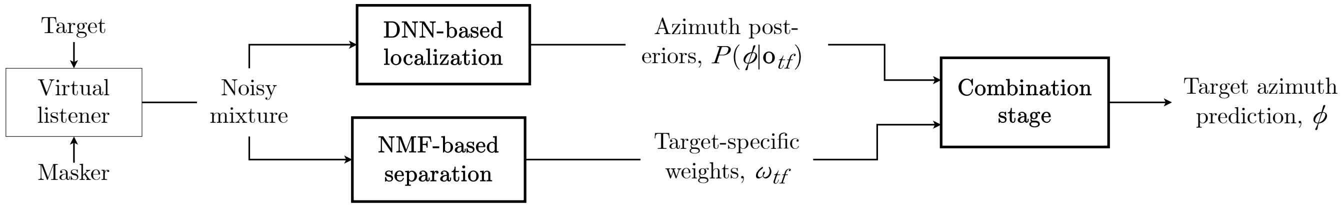
Target
Virtual listener
Masker
Noisy mixture
DNN-based localization
NMF-based separation
Azimuth posteriors, $P(\phi|\mathrm{o}_{tf})$
Target-specific weights, ω_{tf}
Combination stage
Target azimuth prediction, ϕ

Fig. 1. Schematic overview of the combined separation-localization system.

used to denote the magnitude response of this filterbank given the STFT of some noisy mixture. Likewise, matrices $\mathbf{T}$ and $\mathbf{M}$ will be used to denote the magnitude response of the filterbank given the individual target and masker STFTs, respectively. The matrices $\mathbf{T}$ and $\mathbf{M}$ were not made available to the NMF-based separation stage.

The DNN-based localization stage decomposed the noisy mixture into 32 frequency bands using a gammatone filterbank with the same parameters as the separation stage. This allowed the localization stage to extract the cross-correlation function (CCF) feature while retaining a time-frequency representation that was compatible with the weights obtained from the separation stage. In the combination stage, the weights from the separation stage were applied to the azimuth posteriors from the localization stage, and the weighted posteriors were integrated across time and frequency, resulting in a final, target-specific azimuth prediction.

2.2. NMF-based separation stage

The NMF-based source separation stage was based on the algorithm proposed in [12]. In this type of NMF, sometimes termed semi-supervised NMF, a subset of the dictionary matrix is trained before the separation is performed. The pre-trained subset of the dictionary was trained on clean speech, whereas the remaining part was learned during the separation process. The amount of dictionary items used for training on clean speech was determined by the hyperparameter K_T. The algorithm had three other hyperparameters: the rank of the masker dictionaries, K_M, and a sparsity parameter for the target and masker dictionaries, λ_T and λ_M. For any given noisy mixture, the NMF algorithm sought to factorize $\mathbf{X}$, representing the sum of matrices $\mathbf{T}$ and $\mathbf{M}$, into a dictionary $\mathbf{W}$ and an activation matrix $\mathbf{H}$:

$$\mathbf{X} = \mathbf{T} + \mathbf{M} \approx \mathbf{W}\mathbf{H} = [\mathbf{W}_T \mathbf{W}_M] \begin{bmatrix} \mathbf{H}_T \\ \mathbf{H}_M \end{bmatrix}, \tag{1}$$

where $\mathbf{W}_T$ is the fixed target dictionary, $\mathbf{W}_M$ is the flexible masker dictionary and $\mathbf{H}_T$ and $\mathbf{H}_M$ denote the activation matrices for the target and masker, respectively. $\mathbf{W}_M$, $\mathbf{H}_T$ and $\mathbf{H}_M$ were learned by the NMF-algorithm using a gradient descent scheme. Estimates of the target and masker time-frequency representations, $\hat{\mathbf{T}}$ and $\hat{\mathbf{M}}$, were then found as:

$$\begin{aligned} \hat{\mathbf{T}} &= \mathbf{W}_T \mathbf{H}_T \\ \hat{\mathbf{M}} &= \mathbf{W}_M \mathbf{H}_M \end{aligned} \tag{2}$$

These estimates were then passed to the combination stage, where time-frequency specific weights were determined.

2.3. DNN-based localization stage

The DNN-based localization system was the same as the one used in several previous studies [3, 5, 13]. A separate DNN was trained for each of the 32 frequency bands. Although simultaneous sources overlap in time, each time-frequency unit is mostly dominated by a single source. Hence, employing frequency-dependent DNNs was found to be effective for localizing simultaneous sound sources and allows training to be done using single-source data. Each DNN was composed of an input layer with 34 nodes, two hidden layers with 128 nodes each, and 37 output nodes. The 34 input nodes of each DNN corresponded to the 34 features extracted from the corresponding frequency band of the noisy mixture: 33 features from the CCF between the left and right ears using a lag range of ±1 ms, and one interaural level difference (ILD) feature, namely the energy ratio between the left and right ears. The hidden layers used a sigmoid activation function, and the 37 output nodes of the DNN corresponded to the 37 possible angles ranging from -90° to 90° in 5°-steps. The DNN was taken from [5] and was trained without the separation stage. The DNN stage as a whole was considered a fixed building block of the system, and was not subject to the validation procedure presented in this study. In principle, any localization system based on a time-frequency representation of the input signal could benefit from the separation stage proposed in this study.

2.4. Combination stage

The localization stage of the combined system shown in Fig. 1 returned the posterior probability $P(\phi|\mathbf{o}_{tf})$ of a sound source being present at azimuth ϕ, given the binaural feature vector $\mathbf{o}_{tf}$ extracted from frequency band f at time frame t. In order to be able to weight distinct time-frequency bins, a weighting factor ω_{tf} was introduced. Posterior probabilities were then integrated across the time-frequency domain to obtain a final azimuth estimate [3, 5]:

$$P(\phi|\mathbf{o}) = \frac{1}{T} \sum_{t=1}^{T} \frac{\prod_f P(\phi|\mathbf{o}_{tf})^{\omega_{tf}}}{\sum_\phi \prod_f P(\phi|\mathbf{o}_{tf})^{\omega_{tf}}} \tag{3}$$

$P(\phi|\mathbf{o})$ is the probability of the target source being present at azimuth ϕ, given some target weights ω_{tf}. The weights were obtained from the separated target and masker signals as [14]:

$$\omega_{tf} = \left(\frac{\mathrm{TMR}_{tf}}{\mathrm{TMR}_{tf} + 1} \right)^{\beta} \tag{4}$$

where

$$\mathrm{TMR}_{tf} = \left(\frac{\hat{\mathbf{T}}_{tf,L} + \hat{\mathbf{T}}_{tf,R}}{\hat{\mathbf{M}}_{tf,L} + \hat{\mathbf{M}}_{tf,R}} \right)^2 \tag{5}$$

denotes the target-to-masker ratio (TMR) in individual time-frequency bins (t, f). In Eq. (5), TMR is obtained by averaging the target and masker magnitude estimates $\hat{\mathbf{T}}$ and $\hat{\mathbf{M}}$ across the left (L) and right (R) input channels. The ratio of the target and masker estimates is then squared to obtain a power representation. β in Eq. (4) was considered a system hyperparameter, and was optimized in the validation stage.

3. EVALUATION PROCEDURE

The target speech material was selected from the TIMIT Acoustic-Phonetic Continuous Speech Corpus [15], while the masking sounds were taken from the Kaggle Audio Tagging Challenge 2018 database [16] and the ICRA noise database [17]. The TIMIT database contains recordings of ten sentences spoken by each of 630 speakers, and covers eight major dialects of American English. Eight categories of natural sounds from the Kaggle database were used as maskers: applause, the laughter of a toddler, the clattering of a computer keyboard and a variety of musical instruments. In addition, three noises from the ICRA database were selected: ICRA01 (unmodulated speech shaped noise, male), ICRA05 (3-band speech modulated noise, male) and ICRA07 (3-band speech modulated noise, six-person babble).

In each presentation, one target speech recording and one masker were filtered with a binaural room impulse response (BRIR) from the Surrey database [18]. The rooms used had different reverberation times (T_{60}) varying between $0\,\mathrm{s}$ and $0.89\,\mathrm{s}$, and were recorded in an anechoic chamber, an office, a classroom, a cinema-style lecture theatre and a large presentation space with a very high ceiling. The target sentence, BRIR, target and masker azimuth angles and long-term TMR of the mixture were all independently randomized. In the validation stage, the masking noise types were also randomized. In the test stage, the same conditions were used for each noise type (NatNoises-TEST, ICRA01, ICRA05 and ICRA07). Conditions where the target and masker were collocated were not of interest to this study, as in these conditions, the binaural cues from the masker also pointed to the correct location of the target. To avoid these conditions, a minimum separation of at least 10 degrees between target and masker was enforced in both stages and in all conditions.

The standard TRAIN set of the TIMIT database was used for the validation stage, and the TEST set for the test stage. The Kaggle natural sounds noise corpus was also split into a validation set (NatNoises-VAL) and a test set (NatNoises-TEST). From the ICRA database, ICRA01, ICRA05 and ICRA07 were used for testing, but not for validation. To reduce computation time, the validation procedure was split into two phases. First, the model was validated with respect to λ_T, λ_M, K_T and K_M. β was set to 0.5 in this first phase. For the selected model, β was then subsequently optimized in a second validation phase. TIMIT sentences were also used to train the target-specific dictionaries $\mathbf{W}_T$. In both the validation and test stages, the target matrix for training the dictionary was obtained by concatenating 500 sentences from the relevant TIMIT subset (TRAIN or TEST), applying the aforementioned STFT and filterbank processing, and decimating the resulting time-frequency representation in time by a factor of 5. The decimation ensured that the phonetic content in the training set was diverse while keeping the dimensionality of the training matrix low. The 500 sentences used for dictionary training were excluded from the subsequent validation/test procedure. The dataset splits, validation and test stage settings and conditions considered are summarized in Tab. 1.

Table 1. Dataset splits and experimental condition parameters

	Validation stage	*Test stage*
Targets	TIMIT-TRAIN	TIMIT-TEST
Maskers	NatNoises-VAL	NatNoises-TEST, ICRA
Rooms	Surrey anechoic	All five Surrey rooms
Azimuths	$-90°$ to $90°$, $5°$ steps	
Long-term TMRs	-15dBA to 15dBA, 5dBA step	

4. MODEL VALIDATION AND SELECTION

Two distinct experiments were performed, each with their own validation and test stages. Unless otherwise stated, all validation and test stages used the percentage of correctly predicted azimuths (percent correct, PC) as the performance metric. Experiment 1 compared the influence of validating the NMF-based separation stage using either a localization metric or a separation metric, as well as the influence of using either anechoic or reverberant speech for training and validation. For this purpose, two variants of the NMF algorithm were trained. The first NMF variant (NMF-REV) used the most reverberant room (room D) of the Surrey database for training and validation instead of the anechoic condition. The second NMF variant (NMF-SDR) used anechoic speech for training and validation, but used the signal-to-distortion ratio (SDR) metric proposed in [11] for validation instead of the PC. In experiment 1, 250 conditions were used for phase one of the validation, 1,000 for phase two, and 40,000 for testing.

Experiment 2 compared the localization performance of the best NMF system from experiment 1 with three baseline systems. In the first baseline system, the separation stage was omitted. This was achieved by simply setting ω_{tf} was to 1 for all (t, f). In the second baseline system, oracle information about the presence of the target was used by replacing $\hat{\mathbf{T}}$ and $\hat{\mathbf{M}}$ in Eq. (5) with $\mathbf{T}$ and $\mathbf{M}$. The third and final baseline system was a learning-free system inspired by [8]. This system used two independent single-channel noise reduction systems consisting of the minimum mean-square error (MMSE)-based estimator [9] and the adaptive second-order noise power spectral density estimator [10]. The smoothing factor α used by the decision-directed approach for estimating the *a priori* signal-to-noise ratio corresponded to a time constant of $0.396\,\mathrm{s}$ [9]. The respective outputs of the two noise reduction systems were then used to design the adaptive post-filter for arbitrary beamformer (APAB) [19] which was then applied to both the left and the right ear signals. In experiment 2, 750 conditions were used for phase one of the validation, 5,000 for phase two, and 100,000 for testing.

The parameter values resulting from the validation procedure are shown in Tab. 2, along with a label indicating which experiment the different algorithms were used in.

Table 2. Optimal hyperparameter values

Algorithm	*Exp. #*	$\log_2 \beta$	K_T	K_M	λ_T	λ_M
NMF	1,2	-0.5	64	16	0.015	0
NMF-REV	1	-1.5	64	16	0.035	0.010
NMF-SDR	1	10.5	8	8	0.025	0.020
APAB	2	7	-	-	-	-
Oracle	2	4	-	-	-	-
Loc. only	2	-	-	-	-	-

5. PERFORMANCE EVALUATION

Figure 2 shows the results of experiment 1, where the algorithms NMF-REV and NMF-SDR were tested against the standard NMF algorithm. Both variations generally performed worse than the standard NMF algorithm. The NMF-SDR algorithm showed a particularly poor performance, indicating that a combined separation-localization system optimized for SDR does not simultaneously optimize localization performance. Interestingly, the NMF-REV algorithm performed worse than the regular NMF algorithm for low

TMRs even in the room that the NMF-REV algorithm was specifically trained on. This indicates that including reverberation in the training set is not a good choice when using NMF for source separation in low TMR conditions. At high TMRs, there seems to be a slight advantage of using reverberation in the training and validation stages. This may be because reverberation that stems from the target source is more detrimental for the NMF algorithm when the target dictionary is not trained on this condition, whereas reverberation from uncorrelated sources is more easily modelled by the masker dictionary.

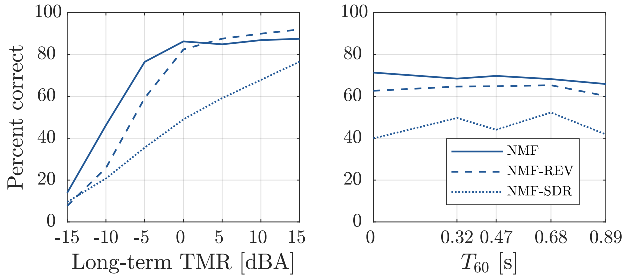

Fig. 2. Results of experiment 1. Localization performance of the three NMF algorithms in the presence of different noise types as a function of the long-term TMR (left panel) and as a function of T_{60}(right panel).

Figure 3 shows the results of experiment 2 for the individual noise types. The APAB algorithm had about 62% correct predictions for the three nonstationary noise types, and around 80% for ICRA01 (stationary speech-shaped noise). This difference was expected, as this algorithm was specifically designed to deal with stationary noise. The NMF algorithm, on the other hand, showed better performance than the APAB baseline algorithm for the three nonstationary noise types, but slightly poorer for the stationary ICRA01 noise.

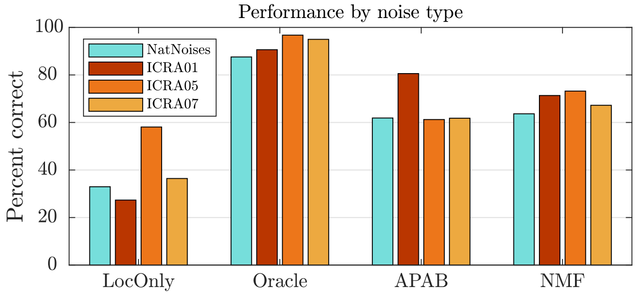

Fig. 3. Results of experiment 2. Localization performance of the four localization algorithms in the presence of different noise types. Results shown are averaged across experimental parameters other than noise type.

Figure 4 shows the results of experiment 2 as a function of long-term TMR and of T_{60}. The NMF algorithm outperformed the APAB baseline algorithm at low TMRs, but did so at the cost of reduced performance at high TMRs. The performance of the NMF algorithm seems to plateau at around 85%. An informal analysis indicated that this effect could be alleviated by using higher ranks for the target and masker dictionaries than those considered in this study. No substantial drop in performance was observed when moving from anechoic to reverberant conditions.

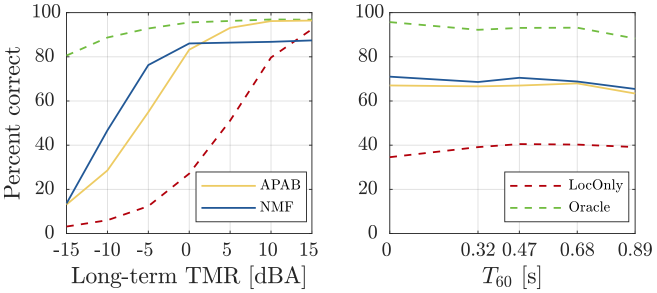

Fig. 4. Results of experiment 2. Localization performance of the four localization algorithms in the presence of different noise types as a function of the long-term TMR (left panel) and as a function of T_{60} (right panel).

6. DISCUSSION AND CONCLUSION

The present study demonstrated that an NMF-based source separation stage can improve the localization performance of a binaural machine listening system in a wide variety of different acoustic conditions. The NMF algorithm outperformed a baseline algorithm based on spectral subtraction in all five rooms considered in this study. The NMF-based separation stage was shown to be especially useful in conditions with low TMR conditions and nonstationary noise. The NMF algorithm performed worse than the baseline in conditions with high TMRs and stationary noise. This is likely attributable to the inherent difference between the separation approaches, the NMF being a learning-based approach. The APAB algorithm is particularly effective in the presence of stationary noise, while the conditions in which the NMF algorithm is likely to be useful are largely defined by what data it is trained on and in which conditions it is validated. It is thus likely that an NMF-based algorithm trained specifically on stationary noise at high TMRs would perform better than the baseline in these conditions, at the potential cost of reducing the general applicability to a wide range of acoustic conditions. With regards to the considered room conditions, training and validating on anechoic speech provided the best performance for all acoustic conditions. This indicates that training on anechoic speech provides a level of robustness towards different acoustic conditions. Training and validating on reverberant conditions, on the other hand, reduced the performance for all room types and across TMRs. This observation is inconsistent with the idea that a more specific training set improves performance in the narrower set of conditions included, as the performance of NMF-REV was lower even in the specific room condition that it was trained on. Optimizing the source separation algorithm using the sound localization performance as the objective was found to be of great importance. This shows that when localizing in the presence of directional interference, the best signal separation does not necessarily lead to the best target source localization.

In summary, it has been demonstrated that an NMF-based approach to source separation is a promising addition to the state-of-the-art localization system. A basic NMF implementation was considered here which can be improved by using convolutional NMF [20] or non-negative tensor factorization [21]. Moreover, a low-latency implementation for real-time processing [6, 7] could be considered in future investigations.

7. REFERENCES

[1] T. Rohdenburg, S. Goetze, V. Hohmann, K. Kammeyer, and B. Kollmeier, "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 2449–2452.

[2] J. Blauert, *Spatial Hearing — The Psychophysics of Human Sound Localization*, Cambridge, MA: MIT Press, 1997.

[3] N. Ma and G. Brown, "Speech localisation in a multitalker mixture by humans and machines," in *Interspeech*, 2016, pp. 3359–3363.

[4] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.

[5] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122–2131, Nov 2018.

[6] G. Naithani, G. Parascandolo, T. Barker, N. H. Pontoppidan, and T. Virtanen, "Low-latency sound source separation using deep neural networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2016.

[7] T. Barker, T. Virtanen, and N. H. Pontoppidan, "Low-latency sound-source-separation using non-negative matrix factorisation with coupled analysis and synthesis dictionaries," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 241–245.

[8] M. Dörbecker and S. Ernst, "Combination of two-channel spectral subtraction and adaptive wiener post-filtering for noise reduction and dereverberation," in *8th European Signal Processing Conference (EUSIPCO)*, 1996, pp. 1–4.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,," IEEE, 1984, vol. 32, pp. 1109–1121.

[10] C. Ris and S. Dupont, "Assessing local noise level estimation methods: Application to noise robust ASR," 2001, vol. 34(1-2), pp. 141–158.

[11] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1462 – 1469, 08 2006.

[12] Schmidt M. N., Larsen J., and Hsiao F., "Wind noise reduction using non-negative sparse coding," in *Proceedings of the IEEE Signal Processing Society Workshop, MLSP*, United States, 2007, pp. 431–436, IEEE.

[13] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, 2017.

[14] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[15] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[16] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," 2018.

[17] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001, PMID: 11465297.

[18] C. Hummersone, "Binaural room impulse response measurements," August 5, 2011.

[19] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., Berlin, Germany, 2001, pp. 39–57, Springer.

[20] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[21] E. L. Benaroya, N. Obin, M. Liuni, A. Roebel, W. Raumel, and S. Argentieri, "Binaural localization of multiple sound sources by non-negative tensor factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1072–1082, 2018.