
Proceedings Paper:
Hu, Q., Ma, N. and Brown, G.J. orcid.org/0000-0001-8565-5476 (2023) Robust binaural sound localisation with temporal attention. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Proceedings. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 04-10 Jun 2023, Rhodes Island, Greece. Institute of Electrical and Electronics Engineers (IEEE) . ISBN 9781728163284

https://doi.org/10.1109/ICASSP49357.2023.10096640

ROBUST BINAURAL SOUND LOCALISATION WITH TEMPORAL ATTENTION

Qi Hu[1*] *Ning Ma*[2] *and Guy J. Brown*[2]

[1]Key Laboratory of Speech Acoustics and Content Understanding
Institute of Acoustics, Chinese Academy of Sciences, Beijing, CHINA
[2]Department of Computer Science, University of Sheffield, Sheffield, UK

ABSTRACT

Despite there being clear evidence for attentional effects in biological spatial hearing, relatively few machine hearing systems exploit attention in binaural sound localisation. This paper addresses this issue by proposing a novel binaural machine hearing system with temporal attention for robust localisation of sound sources in noisy and reverberant conditions. A convolutional neural network is employed to extract noise-robust localisation features, which are similar to interaural phase difference, directly from phase spectra of the left and right ears for each frame. A temporal attention layer operates on top of these frame-level features by incorporating outputs of a temporal mask estimation module that indicate target dominance within each frame. The combined features are then exploited by fully connected layers, which map them to the corresponding source azimuth. Both the temporal mask estimation module and the sound localisation module are trained jointly in a multi-task learning manner. Our evaluation shows that the proposed system is able to accurately estimate the azimuth of a sound source in various reverberant and noisy conditions.

Index Terms— temporal attention, sound source localisation, temporal mask estimation, multi-task learning, phase spectrum

1. INTRODUCTION

Sound source localisation is a fundamental issue in signal processing and forms an integral part of numerous acoustic signal processing tasks, including sound event detection [1], noise reduction [2], and sound source separation [3, 4]. A variety of approaches have previously been proposed to address this problem, including generalised cross-correlation with phase transform (GCC-PHAT) [5], the steered-response-power (SRP) [6–8], subspace methods [9], and deep-learning based methods [10–21]. Many of these methods, such as GCC-PHAT, SRP and the subspace-based methods, originate from narrow-band antenna signal processing. They are agnostic with respect to array geometry and directional properties, and can handle multiple simultaneously active narrow-band sources. However, their localisation performance declines in the presence of reverberation and noise, because the summation of GCC coefficients in GCC-PHAT or SRP exhibits spurious or broadened peaks, and the constructed noise space as in MUSIC [9] may not correspond to the true one (e.g., it may not be orthogonal to the signal subspace).

Deep neural networks (DNNs) have also been widely used in sound source localisation [10–21]. There are broadly three different types of approaches. 1) DNNs are employed to enhance spatial features, e.g, direct-path relative transfer function [19], interaural phase difference (IPDs) [13], sound intensity vectors (IVs) [17], and steering-response-power [22], which are then fed into an independent back-end localisation system. 2) Time-frequency (T-F) masks related to a target source are estimated by a DNN and used to weight noisy spatial features for the subsequent localisation task [11, 12, 14, 20]. 3) Noisy features are directly fed into a DNN localisation model that incorporates some robust strategies, such as head movements [10] or multicondition training for back-end DNNs [10, 23]. Generally, the front- and back-end processes in the above DNN-based approaches are decoupled, which may not provide the best localisation performance. It has been demonstrated that a joint end-to-end optimisation of the front- and back-end processes can boost the performance of a downstream task [24].

This paper proposes a temporal attention-based binaural sound source localisation system which robustly estimates the azimuth of a speech source by jointly training a temporal attention mask estimator and a sound localisation module in a multi-task learning fashion. Instead of explicitly extracting binaural cues, the system uses a convolutional neural network (CNN) framework with 2-dimensional (2-D) kernels that operate directly on the phase spectrum of the left and right ear signals. Features derived from the magnitude spectrum are fed into a temporal mask estimator (TME) to estimate masks which are then used by an attention layer to combine the CNN-derived deep localisation features across the time domain. These two modules are jointly trained using multi-task learning to alleviate the mismatch problem. Our evaluation shows that the proposed system is able to accurately estimate the azimuth of a speech source in challenging noisy and reverberant conditions.

The rest of this paper is organised as follows. Section 2 describes the proposed DNN-based binaural localisation system. The experimental settings and the evaluation framework are introduced in Section 3. Section 4 presents and discusses the experimental results and conclusions are given in Section 5.

2. SYSTEM DESCRIPTION

2.1. Binaural sound source localisation

The baseline system for binaural sound source localisation, illustrated in Fig. 1 (without the temporal attention (TAttn) layer), consists of two stages. The first stage extracts localisation features from phase spectra with four convolutional layers. The extracted features are then passed to the second stage which uses three fully connected layers to perform azimuth estimation as a classification task.

The input feature is the phase component from the short-term Fourier transform (STFT) coefficients of both ear signals. The left- and right-ear signals, indicated by 'L' and 'R' in Fig. 1, are directly

*This work was partly supported by the National Natural Science Foundation of China (11774380), and China Scholarship Council (201904910080). Most of the work was done while Dr. Hu was on an academic visit to the University of Sheffield. Electronic mail: qhu@mail.ioa.ac.cn, {n.ma, g.j.brown}@sheffield.ac.uk

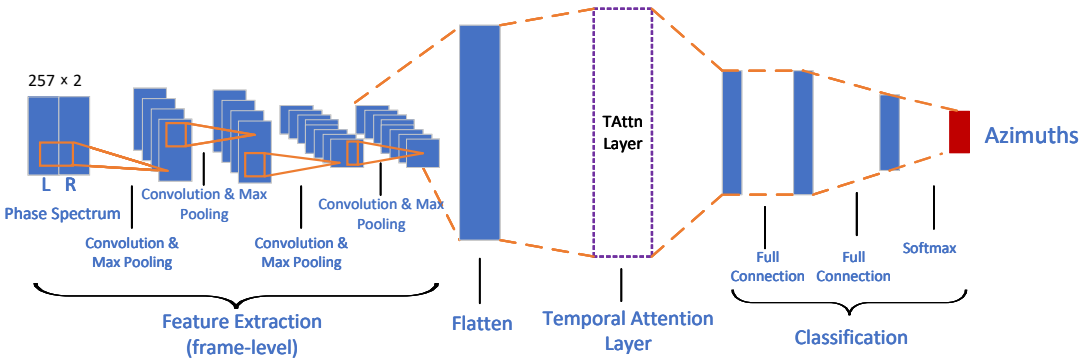

Fig. 1. The architecture of CNN using phase spectrum for binaural sound azimuth estimation. 'L' and 'R' denote the left and right channels respectively. The temporal attention layer may or may not be included in the network depending on different localisation systems.

used as inputs to the CNNs. The signals are sampled at 16 kHz and framed with 20 ms window size and 10 ms overlap. In each frame, the left and right channels are stacked together to form an input matrix of size 2×257 (512-point STFT). All phase values are wrapped to $[-\pi, \pi]$ and then normalised to $[-1, 1]$ as inputs to the CNN. A convolutional layer with 16 2-D kernels of size 2×9 is firstly employed to extract IPD-like features from input features, where 2 corresponds to the binaural channels and 9 corresponds to 9 frequency bins. Next, the outputs from the first convolution layer are down-sampled by a 1×2 max pooling layer to reduce over-fitting and also the computational cost. The down-sampled features are further processed by the following three convolutional layers with kernel sizes of 1×3, 1×3, 1×3 and channels of 16, 32, 32 respectively. The outputs of each layer is followed by 1×2 max pooling and the rectified linear unit (ReLU) activation. After the convolutional layers, each frame-level features are flattened to one vector which is then either combined with other feature vectors along time using a TAttn layer or fed directly into three fully connected layers to perform azimuth estimation. Each of the three dense layers consists of 512 hidden units with ReLU activation and a dropout rate of 0.5.

2.2. Sound localisation with temporal attention

We propose to utilise temporal attention to integrate context information embedded in temporal masks to improve the performance of the baseline system. One intuitive idea is to weight the estimated azimuth probabilities of each frame according to temporal masks which represent the frame-level dominance of the target signal across time. The final probability vector is the average of these weighed probabilities. This method is referred to as *shallow integration*, since it only uses temporal information to combine outputs of the localisation system [25]. We propose a novel method for integrating the temporal information, referred to as *deep integration* (Fig. 1), by inserting an intermediate layer into the CNN localisation system that weights the deep features from the CNN.

A weighted average pooling layer is employed as a temporal attention layer to combine frame-based deep features extracted by the CNN feature extraction stage. This attention layer first uses the softmax function to normalise temporal masks over all frames as follows:

$$\alpha_t = \frac{exp(e_t)}{\sum_{\tau=1}^{T} exp(e_\tau)}, \tag{1}$$

where e_t is the temporal mask value at frame t, and T is the total number of frames. The normalised attention score α_t represents the importance of each frame and is used to calculate the weighted statistics of deep features. For each utterance, the weighted mean vector is estimated as:

$$\tilde{\mu} = \sum_{t}^{T} \alpha_t \cdot h_t, \tag{2}$$

where h_t is the output features from the feature extraction CNN at frame t. The final output of the pooling layer is given by the vectors of the weighted mean $\tilde{\mu}$.

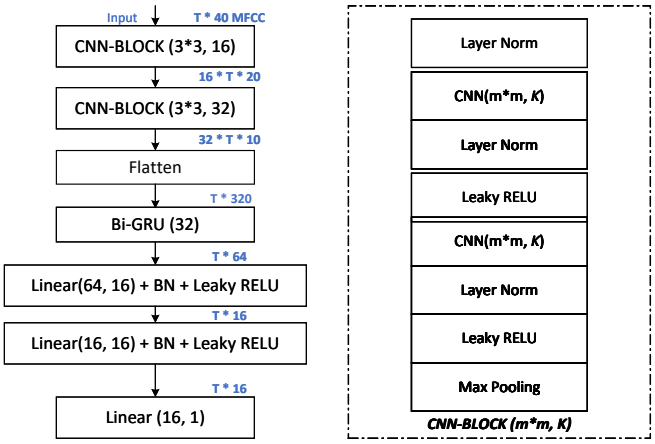

Fig. 2. The TME network topology. The dotted line box on the right is the architecture of 'CNN-Block' used in the TME. T represents the number of frames.

The TME network (Fig. 2) is adapted from a Voice Activity Detector (VAD) with the architecture proposed in SpeechBrain [26]. We discard the sigmoid output layer in the original VAD model so that the TME works as a regression model that maps noisy acoustic features to the corresponding oracle soft masks.

$$IRM(t) = \frac{S^2(t)}{S^2(t) + N^2(t)} \tag{3}$$

where $S(t)$ and $N(t)$ denote the magnitude spectrum of target and interfering signals at the t-th time frame. $IRM(t)$ is the oracle soft mask of the t-th frame, which indicates the dominance of the target signal in the frame. The mean squared error (MSE) loss function L_{mse} is used for training the TME.

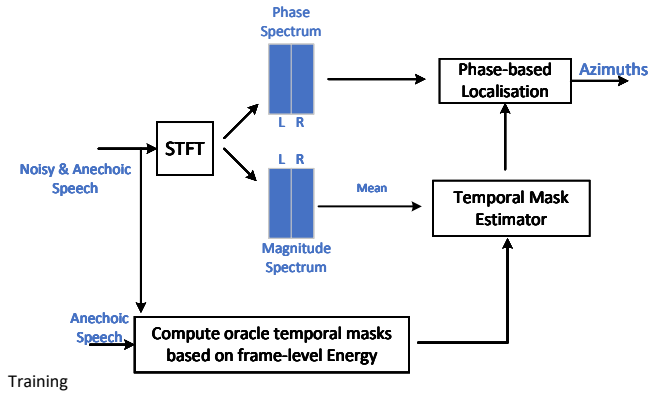

Fig. 3. The framework of the proposed phase-based system using temporal attention for binaural sound localisation. 'L' and 'R' represent speech spectrum of left and right ears respectively. 'Mean' is the average operator across two-ear signals.

2.3. Multi-task learning for azimuth estimation

The estimated mask from the TME module is designed for source separation, which is not necessarily the optimal mask for source localisation. To alleviate the potential mismatch between estimated masks and the source localisation module, we propose a multi-task learning strategy to jointly train the two modules, which encourages the TME to output an optimal mask for the source localisation task.

The complete system is illustrated in Fig. 3. The TME and azimuth estimation network are first pre-trained separately. The training loss function for the joint-training system is a combination of L_{mse} and L_{ce}, corresponding to the TME and the azimuth estimation network respectively. The final loss function is as follows:

$$\mathcal{L}_{multi} = \beta \cdot \mathcal{L}_{mse} + \mathcal{L}_{ce} \tag{4}$$

where β is set to 0.005 heuristically through experiments on the validation set.

3. EVALUATION

3.1. Datasets

The target speech signals were selected from the TIMIT Corpus [27]. All binaural signals were created using head-related impulse responses (HRIRs) or binaural room impulse responses (BRIRs). The speech signals were spatialised across the full 180° azimuth range in steps of 5°. 30 sentences were randomly selected for each of the 37 azimuth locations from the TIMIT train and test subsets for creating the training and test sets, respectively. For training, the KEMAR anechoic HRIRs [28] were used to simulate the free field condition. Azimuths only in the front horizontal plane were considered. For evaluation the Surrey BRIRs [29] were used to simulate different reverberant room conditions. The Surrey database was recorded using a Cortex head and torso simulator (HATS) and includes four room conditions with various amounts of reverberation. Table 1 lists the reverberation time (T_{60}), the direct-to-reverberant ratio (DRR) and initial time delay gap (ITDG) of each room.

Since the training and testing set were created using the impulse responses recorded from different dummy heads, multicondition training (MCT) was applied in the training phase to increase the robustness of the localisation systems. Previous studies [10, 23, 30] have shown that MCT improves robustness by introducing uncertainty into the statistical models of binaural cues. Noise was also added to the test data to evaluate the systems in different noisy conditions. Noisy binaural signals were created by mixing a target signal at a specified azimuth with diffuse noise at *randomly* selected signal-to-noise ratios (SNRs) within $[0, 20]$ dB for training and at *fixed* SNRs (0, 5, 10, and 20 dB) for testing.

Table 1. Room properties of the Surrey BRIRs Dataset [29]

	Room A	Room B	Room C	Room D
T_{60}(s)	0.32	0.47	0.68	0.89
DRR (dB)	6.09	5.31	8.82	6.12
$ITDG$ (ms)	8.72	9.66	11.9	21.6

3.2. Localisation systems

The baseline system was a state-of-the-art DNN-based localisation system using GCC-PHAT features [15]. GCC-PHAT features were computed as the inverse transform of the frequency domain cross-correlation of two audio signals captured by a microphone pair. MCT was also employed to the GCC-PHAT system to improve the robustness to noise and reverberation.

Four different models were evaluated using the proposed framework. The baseline model is *shallow integration* (Section 2.1) where the frame-level output probabilities of the localisation system are integrated without the attention layer. The other three models were *deep integration* systems which employed the temporal attention layer. TAttn–E made use of temporal masks estimated by an independently trained TME. TAttn–J was the joint optimisation network where the TME and the azimuth estimation network were jointly trained using the multi-task learning loss function (Eq. 4). Finally, TAttn–O used normalised *oracle* temporal masks in the attention layer to combine deep features. This system demonstrated the ceiling performance of all the attention systems.

3.3. Experimental setup

The oracle masks for both training and testing were computed according to Eq. 3. For training, the anechoic clean signals were used as the 'target' signal, and the diffuse noise was considered as the 'interferer'. For testing, since all four rooms contained reverberations, the *pseudo-anechoic* clean signals were used as the 'target', created by truncating the original BRIRs before the first reflection according to ITDGs (see Table 1). The 'interferer' was defined as the remaining reflections and diffuse noise.

The CRNN-based TME network was pre-trained on the LibriParty dataset [26] (a synthetic cocktail-party scenario dataset derived from LibriSpeech [31]) and heavily relied on data augmentation to improve its robustness. The pre-trained model was then fine-tuned on the *training* set according to oracle masks computed from anechoic and clean signals.

The *Adam* optimiser with a learning rate of 0.001 and a batch size of 32 was employed. Training with a decreasing learning rate was stopped after 50 epochs and early stopping was applied if no improvement was observed on the validation set for 7 epochs.

The predictions made for each frame in a 1-sec chunk were averaged to report a single azimuth for each chunk. Chunk-based evaluation was adopted in order to avoid the issue that a speech signal typically includes short pauses where there is no directional sound source. The azimuth corresponding to the largest posterior probability was selected as the estimated azimuth. The performance of

Table 2. Localisation RMSE in degrees for different systems in various conditions. Average is computed across rooms and SNRs.

	Room A				Room B				Room C				Room D				Avg.
SNR (dB)	20	10	5	0	20	10	5	0	20	10	5	0	20	10	5	0	
GCC-PHAT	4.9	36.1	56.3	60.1	15.4	45.7	55.7	57.7	10.8	40.5	55.4	60.4	15.8	45.0	57.9	64.3	42.6
+ MCT	2.0	5.9	7.0	9.2	1.6	5.4	8.7	13.3	3.2	5.9	7.1	20.3	2.6	5.1	6.3	13.3	7.3
Shallow	3.3	6.1	8.2	13.6	2.7	4.6	7.4	16.1	2.9	4.9	7.2	19.9	3.3	5.4	8.0	19.6	8.3
TAttn–E	**1.6**	**1.8**	5.5	15.3	**1.0**	5.2	**4.8**	15.2	2.2	2.2	3.2	**9.0**	**1.8**	**2.1**	5.1	19.0	5.9
TAttn–J	**1.6**	**1.8**	**2.9**	**7.9**	1.1	**1.6**	5.1	**12.7**	**2.1**	**2.1**	**2.9**	11.8	1.9	**2.1**	**3.8**	**9.0**	**4.4**
TAttn–O	1.6	1.8	2.5	13.0	1.0	1.4	3.2	10.9	2.2	2.2	2.7	6.0	1.8	2.0	2.8	13.3	4.3

Table 3. Localisation accuracy (%) for different systems in various conditions. Average is computed across rooms and SNRs.

	Room A				Room B				Room C				Room D				Avg.
SNR (dB)	20	10	5	0	20	10	5	0	20	10	5	0	20	10	5	0	
GCC-PHAT	99.4	74.3	41.1	20.6	96.3	59.4	32.7	19.0	97.2	62.9	34.2	17.9	96.0	64.5	35.6	19.9	54.4
+ MCT	99.8	97.8	93.6	85.3	99.5	95.7	92.8	83.3	99.8	97.8	92.2	80.9	99.6	94.9	91.5	81.6	92.9
Shallow	99.7	96.9	90.8	80.4	99.8	96.0	90.3	78.6	99.9	98.3	94.2	75.3	99.8	97.6	90.7	72.1	91.3
TAttn–E	**100**	99.8	97.9	86.4	100	99.8	96.5	82.4	**100**	99.5	97.8	83.2	100	99.5	97.7	78.0	94.9
TAttn–J	**100**	**100**	**98.6**	**89.3**	**100**	**99.9**	**97.7**	**88.7**	**100**	**99.7**	**97.9**	**90.7**	**100**	**99.6**	**98.4**	**90.4**	**96.9**
TAttn–O	100	99.9	98.9	91.3	100	99.9	98.0	88.7	100	99.8	98.5	90.6	100	99.6	98.7	87.0	96.9

the models was reported using two metrics: root mean square error (RMSE) in degrees and localisation accuracy (LocACC). The LocACC was measured by computing the absolute distance between the ground-truth source azimuth and the estimated azimuth with a threshold of 5°.

4. RESULTS AND DISCUSSION

Tables 2 and 3 list the localisation RMSE and LocACC results across different SNRs and reverberant conditions, respectively. In general, the performances of all systems decreased as the SNR decreased. Across different room conditions, the systems performed worse in the more reverberand Room B (lowest DRR) and Room D (longest reverberation time).

The GCC-PHAT baseline without MCT performed poorly in conditions where the SNR was lower than 20 dB or the reverberation was strong (e.g. room B). When trained with MCT, the GCC-PHAT system's performance greatly improved and in most conditions it achieved a similar accuracy to the Shallow system. This is expected as both the GCC-PHAT and the Shallow systems employed phase-based features. However, as the level of noise and reverberation increased, their performance started to decrease and the benefit of the temporal attention became more apparent. All the *deep integration* systems achieved significantly higher localisation accuracy than the *shallow integration* systems, especially in low SNR conditions ($\leq$10 dB).

The TAttn–E system, which employed estimated masks in the attention layer, performed reasonably well at high SNRs when compared to the TAttn–O system, which employed oracle masks. Both systems performed well at SNRs above 10 dB, achieving close to 100% accuracy. At high SNRs, since the speech source dominated in most frames the errors in estimation of temporal masks may not have a significant impact on the localisation accuracy. However, in more adverse conditions (SNR $\leq$ 10 dB, especially in the more reverberant rooms B and D), the performance of TAttn–E became worse. In these conditions, the mask estimation errors started to have a more negative impact on the localisation performance and a better error-tolerance strategy was needed.

By jointly training the mask estimation network and the sound localisation network, the TAttn–J system was able to significantly reduce the localisation errors over the TAttn–E system. The biggest error reduction was again achieved at lower SNRs and in the more reverberant rooms B and D. On average, the TAttn–J system achieved results very close to the TAtten–O system, which employed oracle masks, both in RMSE (4.4° vs 4.3°) and in LocACC (96.9% vs 96.9%). This is likely due to alleviation of the mismatching issue between the TME mask and the sound localisation task during the joint training, which could learn to produce more optimised masks for the sound localisation task.

5. CONCLUSIONS

This paper has proposed a novel binaural machine hearing system that employs temporal attention for robust sound localisation in noisy and reverberant conditions. Instead of enhancing the signal or weighting the output probabilities, the temporal attention layer operates on frame-level deep features within the localisation DNN. By jointly optimising the localisation process and the temporal mask estimation process in a multi-task learning fashion, the proposed TAtten–J system has the opportunities to reduce the mismatch in the two processes and employs masks that are more suitable for the sound localisation task.

Our evaluation in different SNRs and room conditions using the Surrey database has shown that the jointly optimised attention-based system is more accurate in localising sound sources than the attention-based localisation system with separately estimated masks, especially in more reverberation and noisy conditions. The performance is also significantly better than the GCC-PHAT baseline and the shallow-integration system.

Future work will focus on extending the system to employ spectro-temporal attention, which would be useful particularly for narrow-band intrusions where the localisation network may need to disregard a band of frequencies. We will also explore a more integrated approach to mask estimation and sound localisation by exploiting azimuth-based embedding in mask estimation.

6. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 34–48, 2018.

[2] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*, Wiely, 2018.

[3] C. Han, Y. Luo, and N. Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *IEEE ICASSP*, 2020, pp. 6404–6408.

[4] K. Tan, B. Xu, A. Kumar, E. Nachmani, and Y. Adi, "Sagrnn: Self-attentive gated rnn for binaural speaker separation with interaural cue preservation," *IEEE Signal Processing Letters*, vol. 28, pp. 26–30, 2020.

[5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE TASSP*, vol. 24, pp. 320–327, 1976.

[6] M. Brandstein and H. Silverman, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," in *IEEE ICASSP*, 1997.

[7] V. Krishnaveni, T. Kesavamurthy, and B. Aparna, "Beamforming for direction-of-arrival (doa) estimation - a survey," *International Journal of Computer Applications*, vol. 26, 2013.

[8] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated in an adaptive beamformer for hearing aids," *IEEE/ACM TASLP*, vol. 26, pp. 515–528, 2018.

[9] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas and Propagation*, vol. 34, pp. 276–280, 1986.

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

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

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

[13] J. Pak and J. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM TASLP*, vol. 27, pp. 1335–1345, 2019.

[14] W. Zhang, Y. Zhou, and Y. Qian, "Robust DOA estimation based on convolutional neural network and time-frequency masking," in *INTERSPEECH*, Graz, Austria, 2019.

[15] P. Vecchiotti, N. Ma, S. Squartini, and G. Brown, "End-to-end binaural sound localisation from the raw waveform," in *IEEE ICASSP*, 2019, pp. 451–455.

[16] W. Mack, U. Bharadwaj, S. Chakrabarty, and E. Habets, "Signal-aware broadband DOA estimation using attention mechanisms," in *IEEE ICASSP*, 2020, vol. 77, pp. 4930–4934.

[17] M. Yasuda, Y. Koizumi, S. Saito, H. Uematsu, and K. Imoto, "Sound event localization based on sound intensity vector refined by dnn based denoising and source separation," in *IEEE ICASSP*, 2020, pp. 651–655.

[18] T. Jenrungrot, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman, "The cone of silence: Speech separation by localization," in *the 34-th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, 2020.

[19] B. Yang, H. Liu, and X. Li, "Learning deep direct-path relative transfer function for binaural sound source localization," *IEEE/ACM TASLP*, vol. 29, pp. 3491–3503, 2021.

[20] I. Ornolfsson, T. Dau, N. Ma, and T. May, "Exploiting non-negative matrix factorization for binaural sound localization in the presence of directional interference," in *IEEE ICASSP*, Toronto, ON, Canada, 2021, pp. 6125–6129.

[21] P. Grumiaux, S. Kitic, L. Girin, and A. Guerin, "A survey of sound source localization with deep learning methods," *JASA*, vol. 152, no. 1, pp. 107–151, 2022.

[22] P. Pertila and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *IEEE ICASSP*, 2017, pp. 6125–6129.

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

[24] T. Ochiai, S. Watanabe, T. Hori, and J. Hershey, "Multichannel end-to-end speech recognition," *arXiv:1703.04783v1*, 2017.

[25] S. Chakrabarty and E. 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, 2019.

[26] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J. Chou, S. Yeh, S. Fu, C. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Yan Gao, R. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," *arXiv:2106.04624v1*, 2021.

[27] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Technique Report*, vol. 11, 1992.

[28] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A free database of head-related impulse response measurements in the horizontal plane with multiple distances.," in *The 130th AES Convention*, 2011.

[29] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modelling for source separation in reverberant environments," *IEEE TASLP*, vol. 18, no. 7, pp. 1867–1871, 2010.

[30] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE TASLP*, vol. 20, no. 5, pp. 1503–1512, 2012.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE ICASSP*, 2015.