TIME DIFFERENCE OF ARRIVAL ESTIMATION OF SPEECH SIGNALS USING DEEP NEURAL NETWORKS WITH INTEGRATED TIME-FREQUENCY MASKING

Pasi Pertilä, Mikko Parviainen

Faculty of Information Technology and Communication Sciences, Tampere University, Finland

ABSTRACT

The Time Difference of Arrival (TDoA) of a sound wavefront impinging on a microphone pair carries spatial information about the source. However, captured speech typically contains dynamic non-speech interference sources and noise. Therefore, the TDoA estimates fluctuate between speech and interference. Deep Neural Networks (DNNs) have been applied for Time-Frequency (TF) masking for Acoustic Source Localization (ASL) to filter out non-speech components from a speaker location likelihood function. However, the type of TF mask for this task is not obvious. Secondly, the DNN should estimate the TDoA values, but existing solutions estimate the TF mask instead. To overcome these issues, a direct formulation of the TF masking as a part of a DNN-based ASL structure is proposed. Furthermore, the proposed network operates in an online manner, i.e., producing estimates frame-by-frame. Combined with the use of recurrent layers it exploits the sequential progression of speaker related TDoAs. Training with different microphone spacings allows model re-use for different microphone pair geometries in inference. Real-data experiments with smartphone recordings of speech in interference demonstrate the network's generalization capability.

Index Terms— Acoustic Source Localization, Microphone Arrays, Recurrent Neural Networks, Time-Frequency Masking

1. INTRODUCTION

The extraction of spatial information, such as Direction of Arrival (DoA) or Time Difference of Arrival (TDoA), from a sound source emitted wavefront is important for several applications from automatic camera steering [1] to beamforming [2]. A traditional approach to Acoustic Source Localization (ASL) is to estimate the maximum likelihood of the source position given the multichannel audio. Such a likelihood function is referred to as an acoustic map [3]. The Steered Response Power (SRP) with phase transform (SRP-PHAT) is considered as a robust tool for ASL and it builds the acoustic map as the sum of the Generalized Cross-Correlation (GCC) with phase transform (GCC-PHAT) values steered with delays related to the propagation model [4].

Several works that utilize Deep Neural Networks (DNNs) in ASL have appeared in recent years. DOA estimation is treated as a classification problem in many approaches [5, 6, 7, 8, 9]. The first end-to-end DNN based ASL method was presented in [10], which also summarizes recent works. Approaches that utilize DNNs in conjunction with ASL derived features include [11], which uses a Convolutional Neural Network (CNN) to provide a DoA estimate with a Minimum Variance Distortionless Response (MVDR) beamformer output power values as the input feature. In [12] a Voice Activity Detection (VAD) system is used to select frames from which GCC-PHAT values are obtained as the input features for a CNN to then further model the speaker position. Approaches to combine the DNN method with the traditional ASL frameworks include [13], where a CNN was used to predict a Time-Frequency (TF) mask to cancel non-speech values from GCC-PHAT in order to obtain an acoustic map related to the speech source. Recently, [14] proposed Recurrent Neural Networks (RNNs) to predict a TF mask, which was applied to GCC-PHAT, beamforming, and subspace beamforming methods to increase the accuracy of the speaker's TDoA estimate. However, both approaches rely on specifying a target TF mask, and the type of the suitable mask for such a task is not self-evident. In [15], TF masking in conjunction with MVDR beamforming to output a signal for Automatic Speech Recognition (ASR) was proposed. The authors further propose to minimize the ASR error rate by fine-tuning the TF mask prediction.

This work proposes a two-part DNN structure for the estimation of TDoA values of a single speaker. The first part learns to produce a TF mask. The second part then applies the predicted TF mask to the GCC-PHAT to remove the contribution of non-speech interference and then estimates the TDoA value related to the speaker's position. The proposed DNN structure allows the examination of different learning strategies for TF masking -based TDoA estimation to find a suitable masking strategy. In contrast to other TF masking -based ASL approaches [14, 13], to the knowledge of the authors, the proposed method is the first that integrates the TF masking into a DNN that predicts TDoA values. In contrast to offline approaches, e.g. [14], where a TF mask is predicted at the end of each sentence, the proposed method operates in an online fashion and produces a TDoA (and a TF mask) estimate for every input frame. The TDoA estimation uses regression, which can produce continuous values in contrast to classification with a discrete set of output values.

The proposed DNN structure utilizes recurrent layers to exploit the temporal structure of the data. Combined with the online processing, this allows the network to follow a source that is in motion. A single model trained with a range of microphone spacings allows the utilization of different microphone distances during inference without model retraining. The method was trained and tested with static sources using simulated impulse responses and further tested on a set of dynamic smartphone recordings of speech.

This paper is organized as follows. Section 2 describes the signal model and TDoA estimation using TF masking. Section 3 presents the proposed model. Section 4 describes the data, and Section 5 presents the results. Section 6 concludes the paper.

2. SIGNAL MODEL AND TDOA ESTIMATION

The i^{th} microphone signal is denoted as $x_i(t,k)$ in the time-frequency domain, where $k = 0,\ldots,K-1$ is discrete frequency index and t is processing frame index. In the case of a single speaker, the signal is modeled as the sum of the reverberated speech signal $s_{n=0}(t,k)$ in the presence of interference sources $s_{n>0}(t,k)$ and noise $e_i(t,k)$

$$x_i(t,k) = \textstyle\sum_n h_{i,n}(t,k) \cdot s_n(t,k) + e_i(t,k), \qquad (1)$$

where $h_{i,n}(t,k)$ is the Room Impulse Response (RIR) between the nth source and the ith microphone, here $i = 0, 1$.

The direct path delay from nth source position $\mathbf{p}_n$ to the ith microphone position $\mathbf{m}_i$ is $\tau_i^n = \|\mathbf{m}_i - \mathbf{p}_n\| \cdot c^{-1}$, where c is the speed of sound. While the direct delay is not measurable without knowledge of the source signal, the TDoA between two microphones i and i', defined as $\tau_{ii'}^n = \tau_i^n - \tau_{i'}^n$, is measurable. The GCC-PHAT [16] is a popular method for TDoA estimation

$$R_{ii'}(\tau,t) = \sum_{k=0}^{K-1} R_{ii'}(\tau,t,k) = \sum_{k=0}^{K-1} \frac{x_i(t,k)\cdot x_{i'}^*(t,k)}{|x_i(t,k)||x_{i'}^*(t,k)|} e^{\jmath\cdot\tau\cdot\omega_k},$$
$$= 2\sum_{k=0}^{K/2+1} \cos\left(\angle x_i(t,k) - \angle x_{i'}(t,k) + \tau\omega_k\right), \quad (2)$$

where $\jmath$ is the imaginary unit, $(\cdot)^*$ denotes complex conjugate, τ denotes the pairwise delay value, and $\omega_k = 2\pi k/K$ is the angular frequency. The TF mask $\eta_{ii'}(t,k)$ can be applied to GCC-PHAT (2) with multiplication

$$R_{ii'}^m(\tau,t) = \sum_{k=0}^{K-1} \eta_{ii'}(t,k) \cdot R_{ii'}(\tau,t,k). \quad (3)$$

The estimate for (masked) TDoA is obtained at time frame t

$$\hat{\tau}_{ii'}^m(t) = \arg\max_{\tau} R_{ii'}^m(\tau,t). \quad (4)$$

3. PROPOSED METHOD

The proposed method uses spatial measurements (GCC-PHAT values) and magnitude spectrum as DNN input features to estimate the speaker's TDoA. The practical implementation uses frequency bands instead of Discrete Fourier Transform (DFT) resolution to reduce memory consumption.

3.1. Input Features

Based on (2), the positive frequencies ($\omega_k \in [0, \pi]$) of the spectrum are sufficient for modeling GCC-PHAT values for each frequency bin k, and are used as input features for a range of values of τ

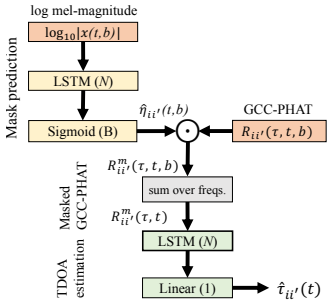
$$R_{ii'}(\tau,t,k) = \cos\left(\angle x_i(t,k) - \angle x_{i'}(t,k) + \tau\omega_k\right). \quad (5)$$

Here, the mel-frequency resolution spatial input feature $R_{ii'}(\tau,t,b)$ was obtained by multiplying $R_{ii'}(\tau,t,k)$ with a weight matrix $W(k,b)$ that defines a mel-filterbank, consisting of equally spaced triangular filters overlapping with adjacent bands (refer e.g. to [17]).

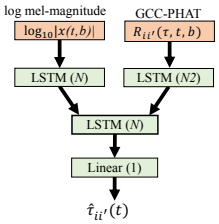
The magnitude spectrum feature ($\log_{10}|x(t,b)|$) consisted of mel-frequency band log-magnitude values averaged over the microphone pair.

3.2. DNN Achitecture for TDoA Estimation Using Masking

The proposed DNN model predicts the TDoA (and optionally also the TF mask) using the mel-frequency resolution input features i) GCC-PHAT $R(\tau,t,b)$ and ii) log-magnitude $\log_{10}|x(t,b)|$. To implement this, the proposed DNN uses the input magnitude spectrum to predict a mel-frequency resolution TF mask $\eta(t,b)$ that is multiplied with the mel-frequency resolution GCC-PHAT $R(\tau,t,b)$ (for each value of τ separately) before integrating over the frequency bands to produce the masked GCC-PHAT $R_{ii'}^m(\tau,t)$. The masked GCC-PHAT is then fed as the input of a TDoA estimation subnetwork, which predicts the final TDoA value. The network utilizes recurrent Long Short-Term Memory (LSTM) cells in both the mask prediction stage and in the TDoA prediction stage. The motivation is as follows: the recurrent LSTM cells can retain long history information [18], which is generally desired when processing sequential data such as speech. Having the previous TDoA output available while predicting the next value helps to avoid spurious noise and interference generated peaks while maintaining a smooth trajectory of the speaker's TDoA values. Figure 1a illustrates the proposed model architecture.



(a) The DNN architecture for TF masking -based TDoA estimation.



(b) The DNN architecture for direct TDoA estimation.

Fig. 1: Panel a) illustrates proposed TF masking -based approach. The sigmoid -type activation layer has B output values that represent the TF mask $\eta_{ii'}(t,b)$, which is multiplied with each delay value τ of $R_{ii'}(\tau,t,b)$, and then summed over frequency bands before TDoA estimation. Panel b) depicts the direct approach, which produces TDoA without masking. This model's last LSTM layer takes $N + N2$ inputs, where N, and $N2$ denote the number of neurons.

3.3. DNN Training Approaches

The below approaches (A)-(D) to train the proposed DNN architecture were experimented with and compared to direct approach (E).

(A) **Implicit mask training**: Train using only the TDoA output. Output: TDoA.

(B) **Joint training**: Train mask prediction and the TDoA prediction simultaneously. Output: TF mask and TDoA.

(C) **Explicit mask training**: First train the TF mask prediction layers and then freeze their weights while training the TDoA estimation layers. Output: TF mask and TDoA.

(D) **No masking**: Omit the masking process ($\eta_{ii}(t,b) \equiv 1$ in Fig. 1a), and train to predict the TDoA from GCC-PHAT values integrated over the frequency range. Output: TDoA.

(E) **Direct approach**: The masking stage is omitted, while the model inputs are kept the same. See Fig. 1b. Output: TDoA.

The approach (E) tests the benefit of imposing the TF mask on the GCC-PHAT and then integrating over the frequency range instead of using a direct approach with the same inputs. Training was stopped if the output error[1] of the validation data did not decrease in 40 consecutive epochs or reached 550 epochs with the adagrad optimizer [19]. The approach (C) used 50 % of the training data to train the mask, and the rest to predict the TDoA with the frozen TF mask layers. This was made to avoid memorizing the training data in the mask prediction stage. The Mean Absolute Error (MAE) of TDoA was used as the optimization criteria, since it was observed to reduce the impact of outlier TDoA values in contrast to Mean Squared Error (MSE).

3.4. TF Mask for Training

The oracle mask is obtained by utilizing the Phase Sensitive Filter (PSF) [20], which is also used as the target for the TF mask in models (B), and (C):

$$\eta_i(t,k) = \frac{|x_{\text{i,dp}}(t,k)|}{|x_i(t,k)|} \cos(\theta_i), \quad (6)$$

where θ_i is the phase difference between the direct path signal component $x_{\text{i,dp}}(t,k)$ and the noisy and reverberant observation $x_i(t,k)$,

[1] For DNN (B) the sum of TDoA and TF-mask errors

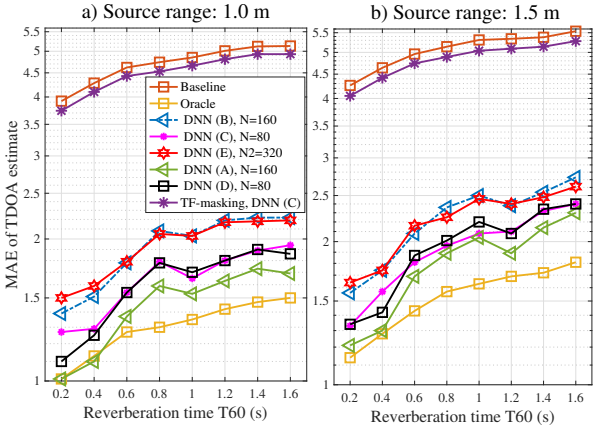


Fig. 2: Simulation test data MAE values (samples, 16 kHz) for different DNN variants at 1 and 1.5 meter speaker distances. The baseline is obtained using GCC-PHAT. The number of DNN parameters is (A): 83151, (B):268671, (C):83151, (D):45201, (E):2750801.

and the mask used in (3) is $\eta_{ii'}(t,k) = \eta_i(t,k) \cdot \eta_{i'}(t,k)$. The mask is finally converted to mel-frequency resolution by multiplying with the matrix $W(k,b)$. Oracle masked TDoA is also reported and it is obtained using (3), (4), and (6) in the mel-frequency resolution.

4. SIMULATED AND REAL-DATA

This section describes the production of the simulation and real-data.

The simulated signals consisted of speech sentences from TIMIT [21] database convolved with synthetic RIRs produced with the image source method [22] in a cuboid shaped room with width, depth, and height $[7, 6.8, 3]$ m, where each dimension was scaled with a uniform random variable in range $[0.5, 1.5]$ to obtain different room sizes. A stereo signal with six microphone spacings between [25–30] cm in 1 cm steps was simulated for training, and for the validation and testing sets five values between [10–26] cm in 4 cm steps were used, with an additional 2.5 mm and 5 mm bias added for validation and testing set microphone spacings, respectively. The horizontal source distance was set to 1 m from the array center, with 10 cm random range and height fluctuation for each sentence to avoid consistent echoes. The reverberation time T_{60} values $150, 300, 450$, and 600 ms were used in the training data, 250 and 500 ms for the validation, and test data contained values between $[200, 1600]$ ms in 200 ms steps. The desired T_{60} was obtained by iteratively solving a single absorption coefficient value common to all surfaces with the Eyering's reverberation formula [23]. A 360° range of horizontal source angles was simulated using 180 angles in the training and validation data, and by using 90 angles in the test data. Small biases were added to validation and test angles to avoid using exactly the same source angles. Training data therefore consisted of 4320 sentences (3 hours of data), validation 1800 sentences (1.2 hours), and test data 7200 sentences (5.2 hours). Different speaker IDs were present in the training, validation, and testing data.

The speech data is then mixed with ambient recordings from both indoor and outdoor environments from the DEMAND database [24]. The database [24] contains array recordings with different microphone spacings between 5 cm and 22 cm, and the stereo pair with nearest microphone spacing to the simulated pair was used as the interference. The Signal to Interference Ratio (SIR) was randomly drawn between $[-5, +5]$ dB for each sentence and

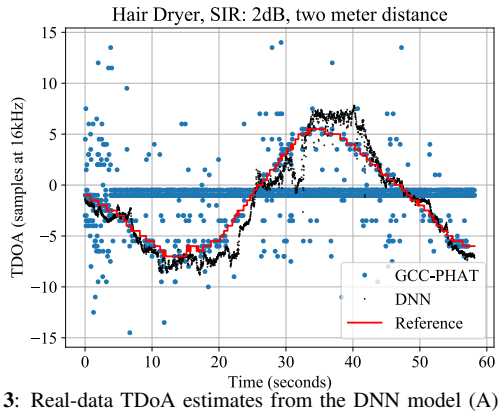


Fig. 3: Real-data TDoA estimates from the DNN model (A) for a speaker (distance 2 m) in presence of a loud interference source (hair dryer). TDoA values near zero belong to the interference source.

White Gaussian Noise (WGN) was added to result in +6 dB Signal to Noise Ratio (SNR). The level of the observed reverberant speech signal was used to report SIR and SNR. Non-overlapping parts of the interference recordings was used in training, validation, and testing data sets. The simulation was performed at 48 kHz sampling rate, and finally all audio was downsampled to 16 kHz for processing.

4.1. Real-Data Description

Gathering of real-data was performed in an office with dimensions $4.1 \times 4.2 \times 3.2$ m, and a reverberation time of 410 ms. A smartphone[2] with reported height of 154.2 mm and microphones in its both ends was used to capture a speech signal from Librispeech [25] played back from a loudspeaker[3] at 1 m distance. The recording was made in three different positions of the room, during which the smartphone was turned slowly to change the source angle. The experiment was repeated at 2 m distance. The height of the loudspeaker and the approximate smartphone height was 1.6 m. Each recording lasted one minute, and a total of six minutes of data was collected.

To capture interference, three different types of everyday interference signals from BBC sound effects library[4] (hospital corridor, interior background, and hair dryer) were played back from the loudspeaker and recorded with the phone mounted to a stand at 1.6 m height located approximately 2 m in front of the loudspeaker. A fourth interference recording contained an active drip coffee brewing machine at the same horizontal distance on a desk at 75 cm height. Each of the captured speech signals was mixed with each of the four types of interference to create 24 test signals. Each test signals was mixed in multiple SIR conditions between -10 dB and $+40$ dB. The SIR is reported with respect to the captured speech signals. The reference TDoAs were obtained by applying a median filter (order 55) to the TDoAs of interference free speech recordings, obtained with GCC-PHAT. Sampling rate of audio was 16 kHz.

5. RESULTS

All data was processed in short windows of 20 ms with 10 ms hop length. The spatial feature tensor $R_{ii'}(\tau, t, b)$ was pre-calculated for possible τ values with the maximum physical microphone spacing of 31 cm with half samples precision, i.e., the TDoA range was

[2]Huawei Mate 10 Pro, `https://consumer.huawei.com/en/phones/mate10-pro/specs/`

[3]Genelec 8010A, `https://www.genelec.com/8010`

[4]obtained from *Stockmusic*, `www.stockmusic.com`

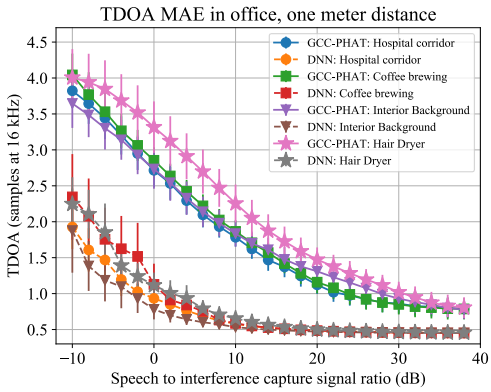


Fig. 4: Real-data performance in the office at 1 meter speaker distance for Baseline (GCC-PHAT) and DNN (A).

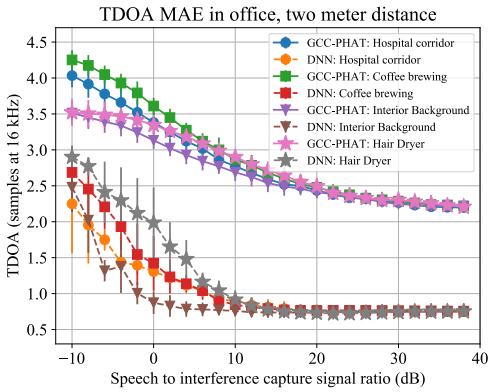


Fig. 5: Real-data performance in the office at 2 meter speaker distance for Baseline (GCC-PHAT) and DNN (A).

$\tau = [-15.0, -14.5, \ldots, 14.5, 15.0]$ samples at 16 kHz. Number of DFT bins was $K = 512$, and the number of mel-frequency bands was chosen as $B = 30$. Different hyper-parameter values for neurons $N = 40, 80, 160, 320$, were tried out, and the best test-set performance was reported with the corresponding number of network parameters (weights and biases). A sequence length of 160 frames and a mini-batch size of 20 samples were selected empirically. The model (E) was trained by fixing the common number of neurons (N) for the TDoA and magnitude processing LSTMs to 80, while trying out 40,80,160, and 320 neurons for the processing of GCC-PHAT input $(N2)$.

5.1. Simulation results

The obtained TDoA estimate was compared with the ground truth TDoA value, and the MAE is reported as the average value over all sentences with different microphone spacings and different source angles for each reverberation level. Figure 2 depicts the MAE of TDoA estimation for source distance a) 1.0 m, and b) 1.5 m, and the number of network parameters is listed.

The TDoA obtained from the mel-frequency resolution GCC-PHAT (Baseline) has the highest MAE values in different reverberation levels and at both source distances. The TDoA of mel-frequency resolution oracle masked GCC-PHAT (Oracle) has the lowest error. The DNN (A) achieves the smallest MAE in all reverberation levels of the compared methods, and even reaches (Oracle) in low-reverberation (T60$\leq$400 ms). By first learning the TF mask, and then learning to predict the TDoA value DNN (C) results in larger MAE than model (A). Interestingly, the MAE of using only the GCC-PHAT for TDoA estimation, i.e., DNN (D), has comparable results. The joined learning of TF mask and TDoA estimation, i.e. DNN (B), has increased error. Finally, the direct approach of DNN (E) results in similar performance as DNN (B), but requires ten times more parameters. All DNN variants outperformed the baseline. Using only the predicted TF mask of model (C) and extracting the TDoA using (4) (Fig. 2, "TF-masking, DNN (C)") resulted in slight improvement over the Baseline.

Based on results of DNN (D) and the results of the method that used only the predicted TF mask, it is apparent that the sequential processing capability offered by the LSTM layers converting the (masked) GCC-PHAT values into TDoA contributes the majority of the TDoA improvement. Secondly, the results of DNN (A) suggest that the TDoA is marginally better when the TF mask is inferred from data than to have been learned using PSF as the target mask.

5.2. Real-data results

The best performing architecture and training approach, i.e. DNN (A), is contrasted with the GCC-PHAT -based TDoA (Baseline). Figure 3 illustrates the resulting TDoAs of speech recorded at two meter distance in strong interference (SIR $+2$ dB). Note, that the DNN (A) is able to naturally follow the speaker related TDoA values, even when the TDoA trail is sparse and it crosses the interference source three times (0 s, 25 s, and 47 s). Tracking such TDoA data would be a difficult task for a traditional target tracking approach, such as Kalman filtering [26], when relying on source dynamics alone. Figure 4 illustrates the MAE of TDoAs from GCC-PHAT and DNN (A) output averaged over the three recordings as a function of the SIR for each type of interference. The vertical bars illustrate the error's standard deviation. The DNN (A) outperforms the Baseline, and its MAE converges to half a sample in SIR above 15 dB, which is the TDoA resolution of the GCC-PHAT input feature. Figure 5 reports DNN (A)'s capability to generalize to distant sources. Again, DNN (A) outperforms the Baseline. To conclude the real-data experiments, the DNN (A) was successful in outperforming the baseline method with actual recorded smartphone data in an office containing a new speaker, more distant sources than in training, and in the presence of different types of interference.

6. CONCLUSIONS

DNN-based TF masking has been previously used in conjunction with acoustic speaker localization. Designing the type of mask for this task is not obvious, and the desired output is the spatial measurement (here the TDoA value) and not the TF mask itself. To address these issues, this paper proposed a DNN architecture that used magnitude spectrum information to derive a TF mask that was then applied to the spatial features (GCC-PHAT). The masked GCC-PHAT was then integrated over the frequency range and used to predict the TDoA value of the microphone pair. A DNN architecture and training variant that learned to only predict the TDoA using implicit TF masking outperformed other variants that additionally learned to predict the oracle mask. The approach also outperformed a direct DNN for TDoA estimation without the TF masking stage and frequency integration, with 30 times less parameters. The proposed architecture used LSTM cells to process the sequential signal, which was observed to contribute largely to the performance of the approach. The proposed solution showed capacity to improve the TDoA estimation over the GCC-PHAT baseline with smartphone recorded speech in the presence of different types of interference. The solution's online capability improves applicability in real-time systems.

7. REFERENCES

[1] B. Dahlan, W. Mansoor, M. Abbasi, and P. Honarbakhsh, "Sound source localization for automatic camera steering," in *The 7th International Conference on Networked Computing and Advanced Information Management*, 2011.

[2] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4 – 24, 1988.

[3] A. Brutti, M. Omologo, and P. Svaizer, "Comparison Between Different Sound Source Localization Techniques Based on a Real Data Collection," in *Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, 2008, pp. 69–72.

[4] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust Localization in Reverberant Rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., chapter 8, pp. 157–180. Springer-Verlag, 2001.

[5] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.

[6] S. Chakrabarty and E. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 136–140.

[7] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[8] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1462–1466.

[9] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 74–79.

[10] J.M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," *Sensors*, vol. 18, no. 10, 2018.

[11] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 103–116, 2018.

[12] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Localizing speakers in multiple rooms by using deep neural networks," *Computer Speech & Language*, vol. 49, pp. 83 – 106, 2018.

[13] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 6125–6129.

[14] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *Proc. Interspeech*, 2018, pp. 322–326.

[15] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 3246–3250.

[16] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 24, no. 4, pp. 320 – 327, Aug 1976.

[17] M. Woelfel and J. McDonough, *Distant Speech Recognition*, John Wiley & Sons, 2009.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[20] H. Erdogan, J. R Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993, Linguistic Data Consortium, Philadelphia.

[22] J. Allen and D. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943 – 950, 1979.

[23] H. Kuttruff, *Room Acoustics*, Taylor & Francis, 5th edition edition, 2009.

[24] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *21st International Congress on Acoustics*, 2013.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[26] S. Särkkä, *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013.