CRNN-based Multi-DOA Estimator: Comparing Classification and Regression

Pieter Cooreman, Alexander Bohlender, Nilesh Madhu[1]

IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium
Email: {pieter.cooreman,alexander.bohlender,nilesh.madhu}@ugent.be

Abstract

Deep learning methods have greatly improved the localization of sound sources in adverse conditions. An important consideration in this case is the output representation. Direction of arrival (DOA) estimation can be interpreted as a classification problem, but performing a regression to continuously estimate the DOAs is also possible. Whereas classification and regression were previously compared for particular cases, such as frame-wise DOA estimation and single source conditions, in this paper we study the more general localization of one or two concurrent sources with a convolutional recurrent neural network. Our experiments show that the two approaches perform comparably in single source scenarios. To address the ambiguity in the source-to-output assignment when multiple DOAs are estimated using regression, we consider permutation invariant training and angular sorting of the desired outputs. However, we find that classification is then generally preferred, especially for closely spaced sources.

1 Introduction

In the field of sound source localization (SSL), microphone arrays can be used to estimate the directions of arrival (DOAs) of acoustic sources. Practical applications of SSL include state-of-the-art hearing aids [1], robot audition [2], and drone audition [3]. Traditionally, approaches were based on classical signal processing techniques, of which [4] gives an overview. Although these methods have shown great performance in favorable conditions, realistic scenarios are often more challenging due to low signal-to-noise ratios (SNRs), strong reverberation, and the difficulty of simultaneously localizing multiple sources positioned in close proximity to each other. More recently, deep learning (DL) methods have gained popularity thanks to their data-driven approach improving robustness in such adverse conditions.

A large variety of DL methods have been proposed that differ in the network architecture, the training paradigm, and the input/output representation. A summary is given in [5]. The focus of this paper is on the *output strategy*, in particular. In a broad sense, two types of approaches can be identified: classification and regression. Regression directly returns (multiple) continuous angle estimates, but the number of outputs (defined before training) may deviate from the number of sources that are active at test time. Classification divides the solution space in discrete zones and assigns a probability of source activity to each. However, the definition of these discrete zones limits the achievable resolution. Therefore, the choice of regression or classification is an important factor in the design of a DL-based DOA estimator.

In the literature, there are only few quantitative comparisons between classification and regression. In [6] and [7], regression is compared to classification in the case of a single active source. In both works, a convolutional recurrent neural network (CRNN) is used of which the output layer size is dependent on the output strategy. Whereas [6] observed that regression outperforms classification, no single system stands out in the results of [7]. It is noted that regression is a justifiable option even though it is less commonly used.

In contrast, [8] proposed a deep neural network (DNN) architecture to localize two sources with regression and compared it to a DNN-based classifier. The array covariance matrix is used as input and fed through a number of hidden layers, resulting in two outputs that directly correspond to the DOA angles of the sources. Generally, the regression approach shows slight improvements over the classification. They also confirmed that, for adverse conditions such as low SNRs and closely spaced sources, DL models perform significantly better than the classical MUSIC algorithm [9]. However, the compared networks only consist of feedforward layers, which make them incapable of tracking the DOAs over time. Incorporating temporal context causes the regression to be more complicated as the model is expected to generate continuous DOA paths, making the regression-classification comparison again a valuable investigation.

When a multi-source scenario with continuous outputs is considered, regression faces the problem of an ambiguous source-to-output assignment. In [10], this is handled either with permutation invariant training (PIT) or by sorting the angles in ascending order. In this paper, we also consider these two approaches.

To the best of our knowledge, in this paper we conduct the first quantitative analysis directly comparing regression and classification for the localization of multiple sources using a DNN architecture that incorporates temporal context. The employed CRNN model is taken from [11], where a classification-based approach was adopted. Our experiments show that the regression approach performs similarly as the classification (baseline) model for single source localization. For multi-source scenarios, however, the classification achieves a significantly higher accuracy than the regression.

In Sec. 2 we formulate the problem of DOA estimation. Sec. 3 presents the DOA classification baseline, before Sec. 4 describes which adjustments are needed to perform a regression instead. Experimental results are discussed in Sec. 5. Finally, conclusions and further research opportunities are presented in Sec. 6.

2 DOA estimation

We aim to localize J sound sources based on the signals captured by an array of N microphones. We assume the nth microphone signal $x_n(t)$ is the sum of reverberant source signals and a noise term $v_n(t)$. Here, the reverberant signal from the jth source is modeled as the dry source signal $s_j(t)$ convolved with the impulse response (IR), $a_{nj}(t)$, that describes how sound propagates from source j to microphone n. We obtain

$$x_n(t) = \sum_{j=1}^{J} a_{nj}(t) * s_j(t) + v_n(t), \tag{1}$$

where $*$ denotes convolution. In this work, we primarily focus on *speech* signals, such that each source corresponds to one talker.

The K-point short-time Fourier transform (STFT) representation of the microphone signals is expressed as $X_n(k, \lambda)$ where k is the frequency bin and λ the time frame index. For DOA estimation, we are particularly interested in the phase component $\angle X_n(k, \lambda)$ as the interchannel phase differences contain the spatial information needed for the localization. To characterize the DOA of a source, we only consider the azimuth angle φ of the spherical coordinate system of which the microphone array is at the center.

2.1 Classification

For classification, we divide the solution space $0 \leq \varphi < 2\pi$ into N_φ uniform discrete zones, each covering a range of $2\pi/N_\varphi$. The desired output of a DOA classifier is then 1 for all DOA zones with source activity, and 0 otherwise.

[1]This work is supported by the Research Foundation - Flanders (FWO) under grant numbers 11G0721N and G081420N.

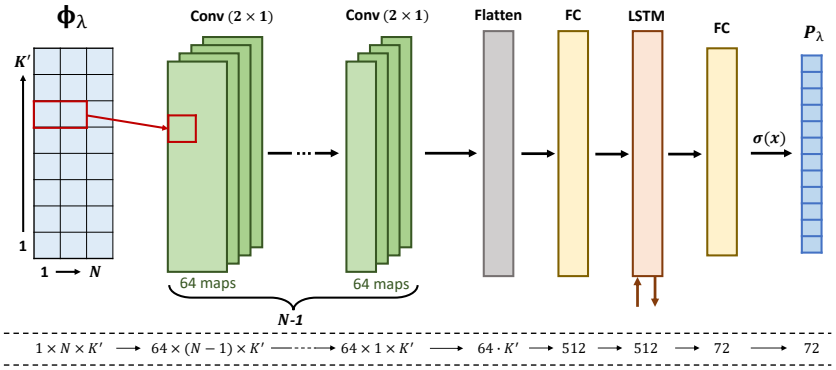


Figure 1: Baseline CRNN-C architecture. Features are extracted from the phase maps Φ_λ through a CNN. The classifier that follows produces a vector of probabilities $\boldsymbol{P}_\lambda$ (one entry per DOA class). The dimensions of the data after each step are listed below the figure.

2.2 Regression

A regression approach requires only one continuous output per source: the azimuth angle $\varphi \in [0, 2\pi)$. A disadvantage of this representation is that the angle wraps around at the interval bounds. To mitigate this problem, sine and cosine of φ can be estimated instead [12].

Intuitively, regression may be seen as the more natural choice due to the unlimited resolution, whereas the finite number of classes introduce a quantization noise in the case of classification. However, regression returns a fixed number of outputs regardless of the (generally unknown) number of active sources, which complicates the interpretation of the results. Furthermore, in the multi-source case, there is an ambiguity concerning which output corresponds to which source (permutation problem).

3 Baseline classification model

The CRNN model proposed in [11] is described here as it is used as a baseline. We will call this DOA classifier CRNN-C.

3.1 Input representation

For a certain time frame λ, there are N microphone signals with K frequency bins each. We only consider frequencies up to the Nyquist frequency: $k \in \{0, \dots, K'-1\}$ with $K' = K/2+1$. From the phase component $\angle X_n(k, \lambda)$, an $N \times K'$ matrix (termed the phase map [13]) Φ_λ is formed that serves as input to the network.

3.2 Neural network architecture

The CRNN architecture is shown in Fig. 1. First, convolutional layers are defined with an output size of 64 feature maps. The kernels of these layers operate only across the microphone (channel) dimension to extract inter-channel features without mixing the information from different frequencies. This is motivated in [13] with the assumption that there is only one dominant speech source in each time-frequency bin (W-disjoint orthogonality [14]), thereby improving robustness when there are concurrent speakers. With a kernel size of (2×1), the information from all channels is combined over the course of $N-1$ convolutional layers. This part of the architecture serves as the feature extractor, as it generates 64 features per frequency bin that contain the spatial information needed for the DOA estimation.

Subsequently, the DOA classifier uses these features to compute probabilities for each of the defined angular zones. This part of the network consists of fully connected (FC) layers, as well as a long short-term memory (LSTM) layer to incorporate temporal information. After the final FC layer, sigmoid activation $\sigma(x)$ is applied. The output is a vector of probabilities $\boldsymbol{P}_\lambda$ for each of the $N_\varphi = 72$ DOA classes.

3.3 Training paradigm

The online data generation of [11] that is used for training the DNNs in this work is briefly described in this section. Microphone signal mixtures are generated according to the signal model of (1). Dry speech signals are convolved with simulated room impulse responses (RIR). Additive mixtures of up to 2 concurrent sources are simulated, the activity of each of which is modeled by a first order Markov chain. Thus, the state of each source randomly switches between two states: *active* and *inactive*, which happens on average each 1.5s. Temporally uncorrelated diffuse noise (spherically isotropic noise field) is added according to an SNR uniformly sampled in the range of 0 dB to 30 dB.

RIRs were simulated for a variety of rooms and microphone array positions to make the training data more diverse. The target output for each class is set to 1 when it corresponds to the DOA of an active source, and 0 otherwise. The binary cross-entropy (BCE) loss function is used to compare target and estimated output. An AdamW optimizer [15] is used with a learning rate of 0.0001 and weight decay of 0.002. Dropout [16] (rate 0.5) and batch normalization [17] are applied for regularization.

4 Proposed regression model

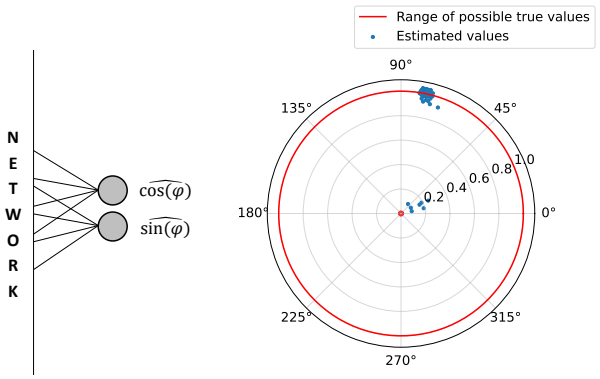
In this section we describe how the baseline model of Sec. 3 is adapted in order to perform regression instead of classification. We distinguish two scenarios: single source (SS) and multi-source (MS) regression.

4.1 Single source

For the single source scenario, we set the number of output nodes to 2 instead of $N_\varphi = 72$. Moreover, linear output activation is used instead of sigmoid. This is illustrated in Fig. 2a, where $\widehat{(\cdot)}$ indicates that sine and cosine of the azimuth angle φ are estimated. Combined, the outputs $(\widehat{\cos\varphi}, \widehat{\sin\varphi})$ may be seen as a point in the xy-plane (not necessarily on the unit circle), from which the DOA estimate $\widehat{\varphi}$ can be extracted. Note that these outputs are only meaningful when they correspond to an *active* source. During periods of source inactivity, the target values are undefined. We propose two solutions for this.

Straightforwardly, periods of inactivity can be ignored during training: we define a "masked" mean square error (MSE) loss function that disregards inactive frames, i. e. these do not contribute to the loss. As a result, the network is optimized entirely for the task of estimating the DOA when a source is active, but not to (additionally) perform source activity detection.

Alternatively, the target values for both cosine and sine can be explicitly set to 0 in frames with inactivity. This is illustrated in Fig. 2b, where the blue dots represent a possible output of the model. The possible positions for target values are indicated in red. The network is then trained with an MSE loss, where the underlying error may be interpreted as the distance in the xy-plane between estimate and target.



(a) Network outputs for a single source: estimated sine and cosine of the azimuth angle.

(b) Example regression output on xy-plane for a source at $\varphi = 80°$, where $x = \widehat{\cos\varphi}$ and $y = \widehat{\sin\varphi}$. Each blue dot is the output at a certain time frame.

Figure 2: Output strategy for single source regression.

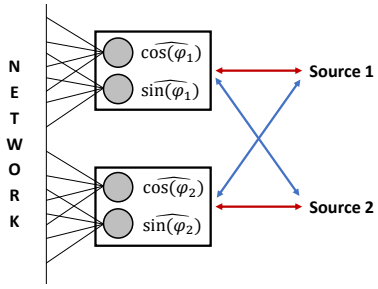


Figure 3: Output strategy for multi-source regression. For two sources, there are two possible source-to-output assignments.

The resulting two regression models for SS localization are referred to as CRNN-R-SS-1 (ignore frames with inactivity) and CRNN-R-SS-2 (set the target to 0 in these frames).

4.2 Multi-source

The regression-based multi-source localization is complicated by the source permutation problem. For simplicity, we only consider two sources here, but the extension to more sources is straightforward. In this case we have 4 output nodes as shown in Fig. 3, where the source permutation problem is illustrated as well. Here, we need to consider 2 possible source-to-output assignments as indicated by the red and blue arrows.

To address the source permutation problem, two approaches known from the closely related task of speaker separation are considered [18, 19]. The first approach uses angle sorting, which was termed location-based training (LBT) in [19]. The outputs of the network are then expected to be sorted according to the DOA angles e. g. in ascending order. By imposing this sorting convention in the labels, the model must also learn how to sort its outputs. This can be beneficial for training as the output is then uniquely defined. A drawback, however, is that the permutation can suddenly change when an already active source becomes inactive or when a source newly becomes active.

An alternative is given by PIT [18]. In this case, we consider all *frame-level* permutations during training, and the one with the lowest loss is used to update the network.

For simplicity, we only consider one setup where the loss only includes frames in which both sources are active at the same time. The resulting two regression models for MS localization are referred to as CRNN-R-MS-LBT and CRNN-R-MS-PIT. Putting outputs to zero in case sources are inactive, similar to the single source case, is an alternative approach, but out of scope for this paper.

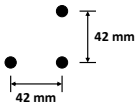


Figure 4: 3-mic subarray of the UMA-16 mic array [20].

5 Evaluation

Signals are sampled at $f_s = 16\,\text{kHz}$. Transformation to the STFT domain is done with a frame length of $M = 512$ (square-root Hann analysis window) and hop size 160, which corresponds to a step of 10 ms. A subset of the UMA-16 microphone array [20] is used, of which the geometry is shown in Fig. 4 and kept the same for both training and evaluation. We use the training setup described in Sec. 3.3.

5.1 Evaluation setup

For the evaluation, too, signals are mixed according to the signal model (1) to generate realistic microphone signals. Clean speech signals are taken from the TSP speech database [21]. RIRs were recorded in a meeting room with approximate dimensions $7.50\,\text{m} \times 5.00\,\text{m} \times 2.65\,\text{m}$ for angles $\varphi \in \{0°, 20°, \ldots, 180°\}$ and source-array distances of 1 m and 2 m. For the background noise, the pub noise from the ETSI database [22] was rerecorded under relatively diffuse conditions.

For every experiment, we simulate up to 2 sources each consisting of 5 concatenated utterances. At the end of an utterance, the DOA angle is changed to a new random angle with a probability of 50%. The DOAs of two different sources are different at all times. We aggregate the results of 50 experiments to obtain reliable results. The resulting total signal duration is 565.51 s.

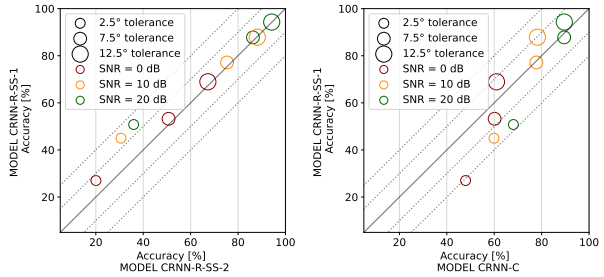
5.2 Single source comparison

First we consider the single source scenario. From the classification (CRNN-C) output, the class with the highest probability is selected as the DOA estimate. In the case of regression (CRNN-R), the DOA is computed based on the sine and cosine estimates using the arctan2 function. Performance is measured with the localization accuracy, which indicates the percentage of frames (during source activity) where the DOA estimate is "correct", i. e. where the absolute error does not exceed a defined threshold. We consider tolerated errors of $\pm 2.5°$, $\pm 7.5°$, and $\pm 12.5°$. This is to allow a fair comparison between regression and classification, since the classes have a resolution of 5°. The experiments are performed for three different SNRs: 0 dB, 10 dB, and 20 dB.

Fig. 5a compares the two SS regression models. The x-axis represents the accuracy of the CRNN-R-SS-2 model, whereas the y-axis shows the accuracy of the CRNN-R-SS-1 model. This allows the performance of the two models to be compared directly. Each data point represents one set of experimental conditions, where the *color* of the marker indicates the SNR, and the *size* indicates the tolerated error. If a point lies on the solid diagonal line ($x = y$), both models perform equally well. Points above this line indicate an improvement of CRNN-R-SS-1 compared to CRNN-R-SS-2.

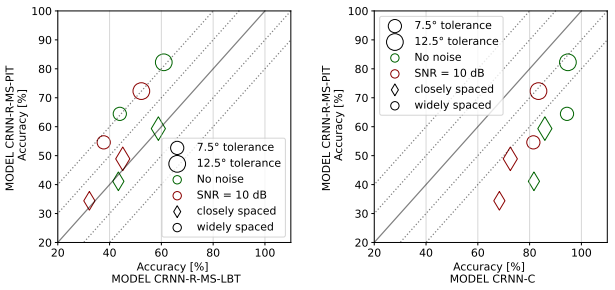
First, we observe that CRNN-R-SS-1 generally achieves higher localization accuracy scores than CRNN-R-SS-2, especially for the smallest tolerated error of $\pm 2.5°$. This suggests that including source activity detection in the training adversely affects the DOA estimation.

Next, we compare our best performing regression model (y-axis: CRNN-R-SS-1) with the classifier (x-axis: CRNN-C) in Fig. 5b. Considering a specific tolerance, the results for the different SNRs lie approximately on a diagonal line. This suggests that the robustness against diffuse noise is rather similar for regression and classification. Whereas the classifier performs better for small tolerated errors, the regression model outperforms the classification model when the threshold is increased. For instance, at an SNR of 10 dB and a tolerance of $\pm 12.5°$, the regression model achieves an accuracy of 87.9%, as compared to only 78.2% using



(a) Comparison regression models. (b) Regression vs classification.

Figure 5: Results for single source scenario.



(a) Comparison of PIT and LBT regression models. (b) Comparison of PIT regression and classification.

Figure 6: Results for multi-source scenario.

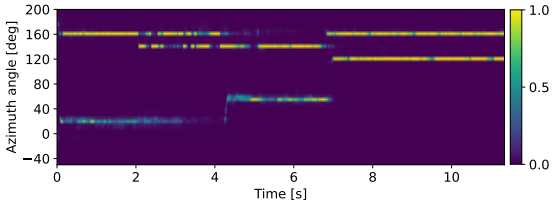
classification. This may be a result of the multi-label classification approach, which does not account for the proximity between different classes in its loss function. A high probability in an incorrect DOA class is penalized equally regardless of how large the error is.

5.3 Multi-source comparison

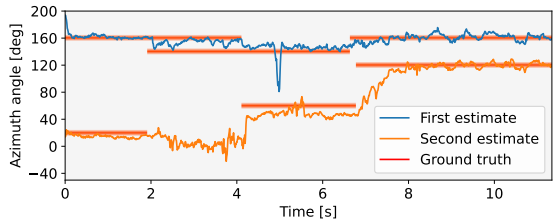
We now investigate the performance for an experimental setup with 2 concurrent sources. To evaluate the classification approach, we can use the same (unchanged) network in this case. The 2 highest peaks (local maxima) in each frame are then selected as the DOA estimates. These are compared to the target values by considering all possible permutations. We report the results for two different tolerated errors ($\pm 7.5°$or $\pm 12.5°$) and two different SNRs (10 dB or noiseless). Further, we also make a distinction based on the difference between the DOAs of the two sources. Due to the setup used to record RIRs, only multiples of 20° are possible. We therefore consider one set of results for "closely spaced" sources (differences of either 20° or 40°) and one set of results for "widely spaced" sources (differences $60°, 80°, \ldots, 180°$).

First we compare the two regression models CRNN-R-MS-LBT and CRNN-R-MS-PIT that solve the source permutation problem by angle sorting and permutation invariant training, respectively. The results in Fig. 6a clearly show that PIT (y-axis) outperforms LBT (x-axis). This could be because additionally requiring the cosine and sine outputs to be sorted based on the underlying angles is detrimental to the core task of estimating the DOAs. Mainly, the PIT and LBT results differ for the setup with widely spaced sources (circular marker ◯), whereas both perform comparably for the localization of closely spaced sources.

Again, we select the best performing regression model (PIT) for the comparison with the classification-based approach. The results are shown in Fig. 6b. It is immediately apparent that classification generally outperforms regression in this case. For the most challenging conditions ($\pm 7.5°$ tolerance, 10 dB SNR, closely spaced), the classifier achieves an accuracy of 68.3%, as opposed to only 34.4% using regression. Regarding the effect of tolerated error and SNR, we observe similar trends as in the SS scenario. We also see that the setup with closely spaced sources



(a) Classifier output for two simultaneously active sources.



(b) Regression output for two simultaneously active sources.

Figure 7: Comparison of outputs for single example.

causes the performance of the regression model to degrade more strongly than that of the classifier. To better understand this observation, we now consider one example more closely.

Fig. 7 shows the outputs of both models for the same experiment (no additive noise). In both cases, the x-axis indicates time and the y-axis the DOA. The upper plot shows the output *probabilities* of the classifier (represented by different colors). The lower plot shows the estimated *angles* computed based on the output of the regression model (blue and orange lines), along with the true angles (red line). The most difficult scenario occurs between 2 s and 4 s when the difference between the two source DOAs is only 20°. Then, it appears that the regression model is unable to distinguish between the 2 sources. Instead, the output for the first source (blue line) lies somewhere between both true angles (140° and 160°), presumably because the squared error can thereby be limited in both directions. At the same time, the output for the second source (orange line) remains at an angle of around 20°, where the speaker was located during the first utterance (up to around 2 s). The classifier, in contrast, is able to simultaneously localize the 2 sources quite accurately.

6 Conclusions

The choice between a classification-based and a regression-based approach is an important factor in the design of a DNN for SSL. Previous works comparing the two did not consider multi-source scenarios, or used networks that cannot take advantage of temporal context. In this paper, therefore, we perform a comparison of the two output strategies in both single and multi-source scenarios based on a CRNN architecture.

In the SS case, regression and classification perform comparably. Whereas the regression model more often localizes sources correctly when a certain error can be tolerated, classification outperforms regression when the acceptable error threshold is small. The more common occurrence of estimates that are considerably off when using the classification approach could be related to the multi-label classification approach.

For the MS localization, we find that PIT should be preferred over LBT to address the permutation problem when using the regression approach, possibly because additionally requiring the outputs to be sorted takes away from the performance on the core task of estimating the DOAs. However, a significantly higher performance is achieved with the classification approach in this case, whereby the permutation problem can be avoided altogether.

In future work, the inclusion of magnitude information could be considered so that (especially closely spaced) sources can be distinguished more easily. This should increase the accuracy of both the classification and regression based models.

References

[1] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2015.

[2] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5610–5614, IEEE, 2015.

[3] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone Audition: Sound Source Localization Using On-Board Microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 508–519, 2022.

[4] N. Madhu and R. Martin, "Acoustic source localization with microphone arrays," in *Advances in Digital Speech Transmission* (R. Martin, U. Heute, and C. Antweiler, eds.), pp. 135–170, Wiley New York, 2008.

[5] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[6] Z. Tang, J. D. Kanu, K. Hogan, and D. Manocha, "Regression and Classification for Direction-of-Arrival Estimation with Convolutional Recurrent Neural Networks," in *Proc. Interspeech 2019*, pp. 654–658, 2019.

[7] L. Perotin, A. Défossez, E. Vincent, R. Serizel, and A. Guérin, "Regression versus classification for neural network based audio source localization," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 343–347, 2019.

[8] Y. Xiong, A. Liu, X. Gao, and A. Yauhen, "DOA Estimation Using Deep Neural Network with Regression," in *2022 5th International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 1–5, IEEE, 2022.

[9] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[10] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Computer Speech & Language*, vol. 75, p. 101360, 2022.

[11] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in CNN based multisource DOA estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1594–1608, 2021.

[12] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Improvement of doa estimation by using quaternion output in sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pp. 244–247, 01 2019.

[13] S. Chakrabarty and E. A. 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.

[14] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–529, IEEE, 2002.

[15] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–19, 2019.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 448–456, PMLR, 07–09 Jul 2015.

[18] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, 2017.

[19] H. Taherian, K. Tan, and D. Wang, "Location-based training for multi-channel talker-independent speaker separation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700, 2022.

[20] miniDSP, "UMA-16 USB microphone array." `https://www.minidsp.com/products/usb-audio-interface/uma-16-microphone-array`.

[21] P. Kabal, "TSP speech database," tech. rep., McGill University, Montreal, Quebec, Canada, 2002.

[22] European Telecommunications Standards Institute, "Speech Processing, Transmission and Quality Aspects (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database," Tech. Rep. ETSI EG 202 396-1, ETSI, 2008.