Comparison of Regression and Classification Models for Multi-Source Direction of Arrival Estimation with Convolutional Recurrent Neural Networks

Nils Poschadel, Stephan Preihs, Jürgen Peissig
Institute of Communications Technology, Leibniz University Hannover, Germany
{*poschadel, preihs, peissig*}*@ikt.uni-hannover.de*

Abstract
Direction of arrival estimation using deep learning is a well-established research area, usually approached as a regression or classification model. While classification methods have been used preferentially for some time especially for multi-source localization, regression methods have become increasingly popular due to their ability to provide continuous direction estimation. However, multi-source regression is a challenging problem as it requires addressing the issue of permutation invariance. Regression and classification approaches have already been compared in detail for single-source localization. For multi-source localization, there are numerous proposed methods in both categories, but, to the best of our knowledge, there is no systematic comparison between the two options. This study aims to fill this gap by providing a comprehensive analysis of regression and classification models, especially in multi-source localization scenarios.

Introduction
Deep learning based direction of arrival (DOA) estimation has been studied extensively both as regression and classification problem with each approach having its own challenges and advantages. For one-source scenarios, the two methods have already been compared in detail [1], [2], with regression seeming to be more accurate in scenarios with diffuse interference, while classification appeares to be more robust in the presence of localized interference.

When comparing both approaches in a multi-source scenario, it is important to take into account that each method involves its own peculiarities. Classification models usually require a spherical grid definition, which introduces discretization errors and often involves post-processing techniques such as peak-picking to identify unique sound sources in the prediction. In contrast, regression models need to address the permutation invariance problem, since there is no direct assignment to a specific sound source or the predicted order may vary.

In addition, the number of sound sources is often (as in this study) assumed to be prior knowledge. However, a generalization to an arbitrary number can be done in different ways. A straightforward approach (independent of regression or classification) is to preestimate the number of sound sources. In the case of classification, the generalization can also be achieved naturally, e. g., by introducing a threshold value in the classification. There are also approaches for regression, such as the activity-coupled DOA (ACCDOA) estimation [3], which joins localization and detection information by scaling a DOA vector by its probability of belonging to an active source. However, the detection of a sound source should not be part of these investigations. Furthermore, the focus is only on the localization of static sound sources and not on the tracking of moving sources such that challenges like identity switches can be neglected [4]. Therefore, a frame-wise permutation invariant training (fPIT) strategy is used in contrast to utterance-level PIT (uPIT) [5] or sliding PIT (sPIT) [4].

Even though multi-source sound source localization has already been investigated both as classification and regression task, there is a lack of explicit comparison of the two approaches with specific investigation of the different parameters such as the spherical grid or the peak-picking strategy. To address this gap, this work utilizes synthesized first-order Ambisonics speech signals as well as a convolutional recurrent neural network (CRNN) as basic architecture of the deep learning model [6], [7], which will be explained in more detail in the next section.

Model
The deep learning models used in this investigation follow the same basic structure as in [6]–[8]. A detailed overview of the network's architecture is given in Fig. 1. The sole differences between the regression and the classification models appear in the final layer, i. e. the number of output units $\mathrm{dim}_{\mathrm{out}}$, the output activation $\mathrm{act}_{\mathrm{out}}$ as well as the loss function. Details on these parameters will be provided in the next paragraphs.

As evaluation metric we define the localization error as the time average of the frame-wise angular distances between label and prediction. When calculating the angular distance, the permutation of the predicted sound sources resulting in the minimum angular distance is chosen.

All the models were implemented using the TensorFlow framework and optimized using the Nadam optimizer while incorporating early stopping on the validation set.

Classification
In the classification approach, the DOA estimation is interpreted as the task of estimating whether or not each point on a predefined grid corresponds to the direction of an active sound source or not. Here, the following quasi-uniform grid on the unit 2-sphere is used:

$$\begin{aligned} \theta_i &= -90 + \frac{i}{I} \cdot 180 && \text{, with } i \in \{0, \dots, I\}, \\ \phi_j^i &= -180 + \frac{j}{J^i + 1} \cdot 360 && \text{, with } j \in \{0, \dots, J^i\}, \end{aligned} \tag{1}$$

with $I = \lfloor \frac{180}{\alpha} \rfloor$, $J^i = \lfloor \frac{360}{\alpha} \cos(\theta_i) \rfloor$, and a grid resolution parameter α which results in a grid of $n_{\mathrm{grid}} =$

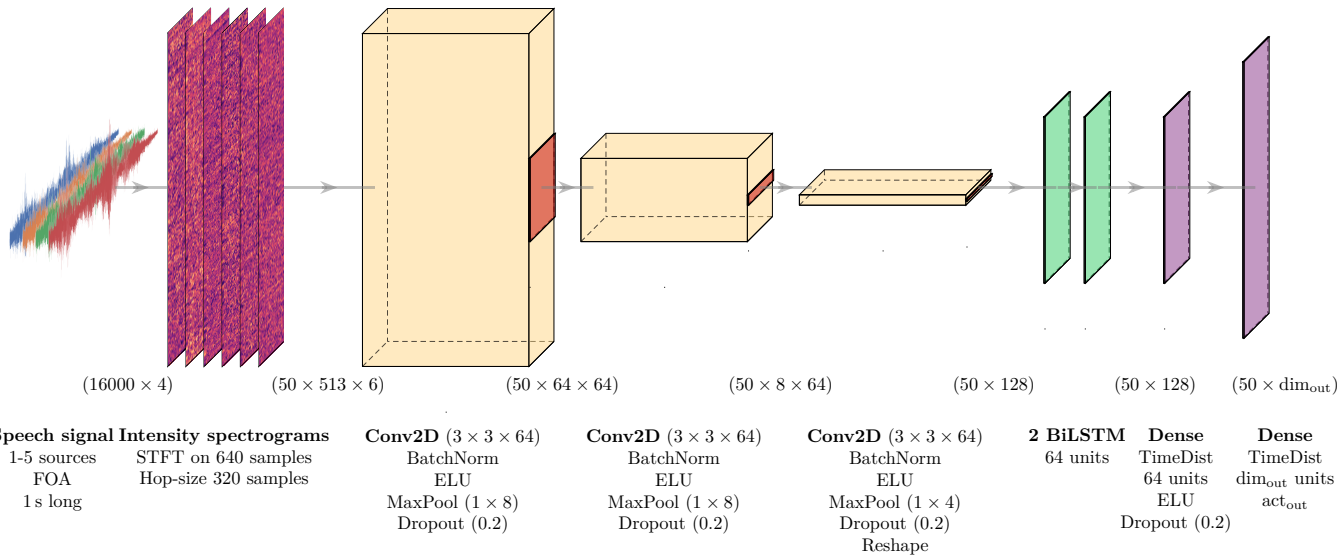

Figure 1: CRNN architecture for DOA estimation used for the regression and classification models.

$\sum_{i=0}^{I} \left(J^i + 1\right)$ points. The angle ϕ is the azimuth, which is zero at the frontal direction and increasing counterclockwise; θ is the elevation, which is zero at the horizontal plane and positive above. In this paper, five different grids as listed in Tab. 1 are compared.

According to this classification setting, the target of the CRNNs is a multi-hot-encoded vector of size $\text{dim}_{\text{out}} = n_{\text{grid}}$ for each time frame, with each index corresponding to a DOA according (1). From this vector, the n_{sources} largest values that correspond to the respective DOAs are chosen. When doing so, it may happen that neighboring grid bins are assigned high values during prediction and therefore picked as different sound sources, while actually belonging to the same sound source. In order to avoid this kind of misclassification, a so-called peak-picking strategy is applied. There are different methods, e. g. a spatial smoothing over neighboring directions [8]. In this study, a sound source is only assigned to a grid bin if this bin has a minimum angular distance of $\text{min}_{\text{dist}} = 0°, 5°, 10°, 15°$ to all previously assigned sound sources.

The final activation function act_{out} are a softmax and a sigmoid activation for the one- and multi-source classification case, together with a categorical and binary cross-entropy loss function, respectively.

Table 1: Approximated maximum possible discretization error induced by the different spherical grids.

Name	α	n_{grid}	Discretization error in °
grid_{1669}	5	1669	3.5
grid_{749}	7.5	749	5.7
grid_{425}	10	425	7.0
grid_{345}	11	345	7.8
grid_{263}	12.5	263	8.9

Regression

In the regression approach, the Cartesian coordinates of the DOA vector of every sound source are directly estimated by the model for each time frame, resulting in an output dimension $\text{dim}_{\text{out}} = 3 \cdot n_{\text{sources}}$. As mentioned above, permutations of the sound sources have to be considered. Therefore, the minimum mean squared error (MSE) over all sound source permutations is applied as loss function implementing a fPIT strategy. The final activation function act_{out} of the model is a linear activation followed by a normalization to unit-norm.

Data

The procedure for generating the training, validation, and testing data is described below. For further details the reader is referred to [6]. The data used for this study was generated from a set of SRIRs simulated with the MCRoomSim toolbox [9] as spherical harmonics signals. The acoustic properties of the walls were set to plausible, randomly chosen surfaces of the GRAP database [10]. For the multi-source case, additional SRIRs were selected belonging to the same room but to a different source and having an angular distance of at least 15° from each other. These SRIRs were then each convolved with a different speech sample from the TIMIT database. The spherical harmonics speech signals were added at a random signal-to-interference-ratio (SIR) between 0 and 10 dB relative to the first source. The signals were cut to the minimum length of the respective individual speech signals, such that the respective target number of sound sources is active during the entire length of the signal.

In the single-source case, ambient babble noise was added to the speech signal at a signal-to-noise ratio (SNR) between 0 and 20 dB, whereas in the multi-source case, ambient babble noise was added to the speech signals at a constant SNR of 20 dB. Finally, these sentences were cut to one-second-sequences and resampled to 16 kHz.

For further evaluations of the DOA estimation performance in a more realistic scenario, additional testing data based on SRIRs measured in the Immersive Media Lab [11] of our institute was synthesized according to the pro-

cedure mentioned above. In the following section, both results on data synthesized from simulated and measured SRIRs are reported.

As input feature for our models, first-order Ambisonics (FOA) pseudointensity spectrograms [6]–[8] were used. In the implementation, a short-time Fourier transform (STFT) is performed on 640 samples using a Hann window along with zero-padding to the FFT size of 1024 samples and a hop size of 320 samples, resulting in 50 time frames and 513 frequency bins.

Results

The localization performance of the regression and classification models for the different numbers of sound sources are shown in Fig. 2. In the single-source case, the regression model slightly outperforms all classification models with a median localization error of 2.3° and 4.5° for the simulated and measured data, respectively, which is consistent with the results in [1], [2]. For the classification models, a finer grid resolution always improves the localization result which is likely due to the discretization errors associated with the different grids.

In the multi-source case, the regression model and the best classification model achieve comparable results for all numbers of sound sources as well as for both simulated and measured data, with the classification being slightly more accurate. However, this is mainly the case if the prior knowledge that all sound sources have a minimum distance of 15° is fully used for identifying unique sound sources. If the minimum distance between the predicted sources is reduced, the misclassifications increase considerably, especially for the grids with a finer resolution. This pattern is mainly expressed in an increased variance of the localization error and less in the median error.

For two or three sources, the comparatively fine grids grid_{749} and grid_{1669} achieve the best results with a median localization error of about 5.9° and 7.3° on the measured data, respectively. For four or five sources, the coarser grid grid_{425} always achieves the best results with a median localization error of 8.4° as well as 9.7°.

Conclusion

All in all, the classification approaches achieved very accurate localization results when prior information about the distribution of the sound sources was used during the identification of unique sound sources during postprocessing. If less prior information is available or can be assumed, more sophisticated peak-picking strategies may have to be considered or a coarser grid resolution has to be chosen (which implicitly corresponds to peak-picking). In general, the more sources there are, the coarser the grid should be chosen. The spherical grid with 425 directional bins has turned out to be a good compromise for the classification models for all numbers of sound sources. Regression approaches provide almost equivalent performances compared to the best classification approaches while containing fewer adjustable hyperparameters.

References

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

[2] L. Perotin, A. Defossez, E. Vincent, R. Serizel, and A. Guerin, "Regression Versus Classification for Neural Network Based Audio Source Localization," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 343–347.

[3] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-Coupled Cartesian Direction of Arrival Representation for Sound Event Localization And Detection," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 915–919.

[4] D. Diaz-Guerra, A. Politis, and T. Virtanen, *Position tracking of a varying number of sound sources with sliding permutation invariant training*, 2022. [Online]. Available: http://arxiv.org/pdf/2210.14536v1.

[5] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[6] N. Poschadel, R. Hupke, S. Preihs, and J. Peissig, "Direction of Arrival Estimation of Noisy Speech using Convolutional Recurrent Neural Networks with Higher-Order Ambisonics Signals," in *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, pp. 211–215.

[7] N. Poschadel, S. Preihs, and J. Peissig, "Multi-Source Direction of Arrival Estimation of Noisy Speech using Convolutional Recurrent Neural Networks with Higher-Order Ambisonics Signals," in *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, pp. 1015–1019.

[8] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-Based Multiple DoA Estimation Using Acoustic Intensity Features for Ambisonics Recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[9] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik, "Room acoustics simulation for multichannel microphone arrays," *International Symposium on Room Acoustics (ISRA) 2010*, 2010.

[10] D. Ackermann, M. Ilse, D. Grigoriev, *et al.*, *A Ground Truth on Room Acoustical Analysis and Perception (GRAP)*, 2018.

[11] R. Hupke, M. Nophut, S. Li, R. Schlieper, S. Preihs, and J. Peissig, "The Immersive Media Laboratory: Installation of a Novel Multichannel Audio Laboratory for Immersive Media Applications," in *Audio Engineering Society Convention 144*, 2018.

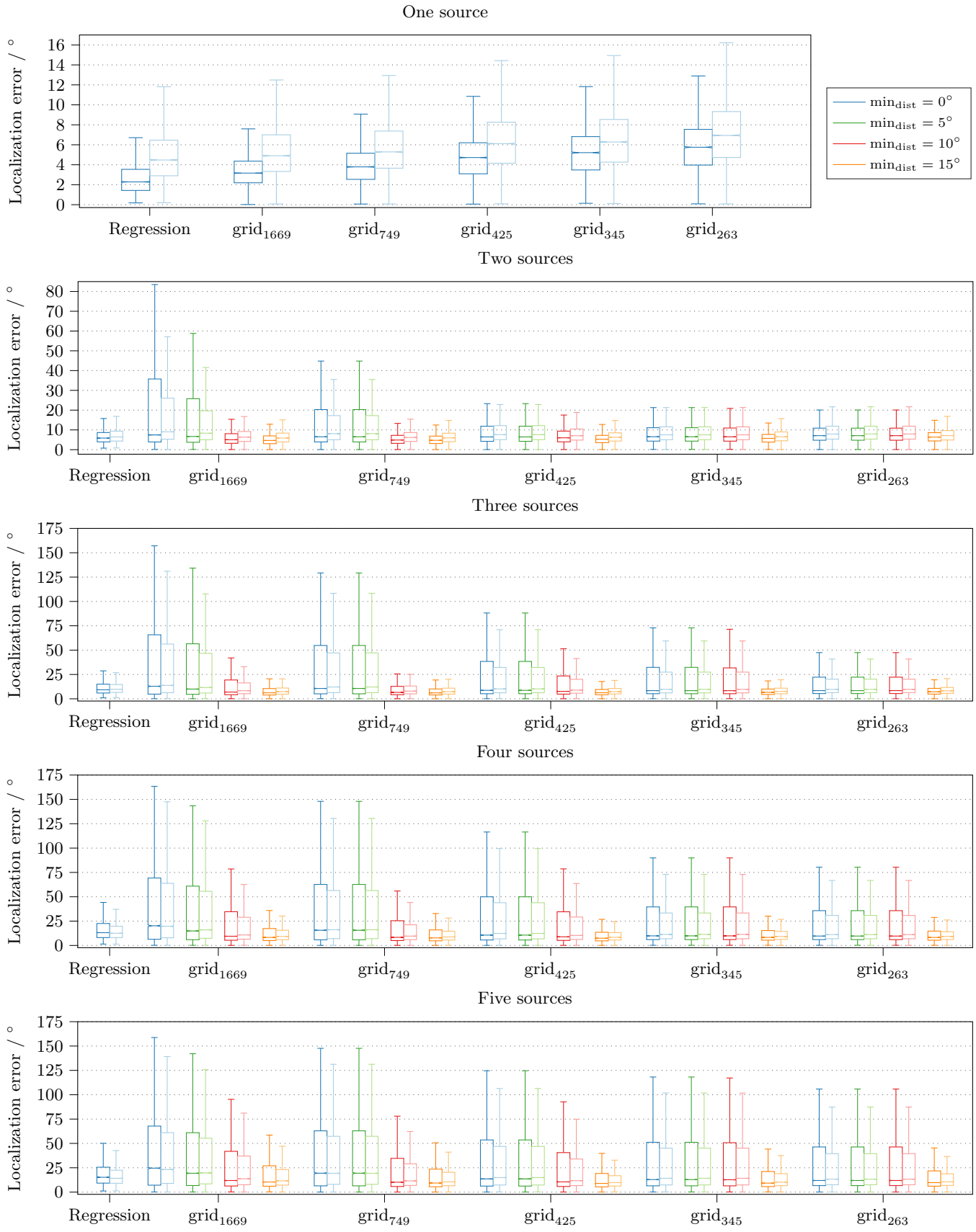

Figure 2: Box plots of the localization errors for the regression and classification models on different grids and with different minimum distances between predicted sound sources for one to five sources. The boxes are drawn from the first to the third quartile and the horizontal line depcits the median. The dark and light boxes represent the results on the data synthesized from simulated and measured SRIRs, respectively. Please also note the different scalings of the vertical axes.