Deep Learning based Sound Source Localization

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

Introduction

Deep learning (DL) methods have demonstrated their effectiveness in spatial and immersive audio and are extensively utilized in various fields. One notable application is sound source localization (SSL), also known as direction of arrival (DOA) estimation, where DL models can compete with or even surpass classical approaches, especially in challenging acoustic conditions. However, it has been shown that DL approaches primarily achieve good and reliable results when the models are not used as black boxes that process vast amounts of raw data and autonomously learn implicit feature representations, but when application, model, and input data are finely tuned to each other.

DL based SSL can typically be divided into different components. An overview of a general pipeline is shown in Fig. 1. The following sections provide brief insights into the different aspects, in particular those investigated by the Institute of Communications Technology at Leibniz University Hannover in recent years. In addition, the current research project "Hooray" will be introduced in the last section, in which the possibilities and performance boundaries of deep learning based SSL using a head-mounted microphone array are being investigated.

The description of the SSL components is explicitly focused on selected topics and does not claim to be exhaustive. Further and more comprehensive information on DL based SSL can be found in the systematic review in [1].

Input Feature Extraction

A key aspect in the design of DL based SSL is the choice of input features. One approach involves using unprocessed inputs, allowing the neural network to autonomously discover the most effective representations for the task, potentially leading to innovative methods of interpreting audio data for localization tasks. This strategy involves feeding the network with multi-channel waveforms directly in the time domain, providing a raw, unfiltered signal. Alternatively, in the time-frequency domain, despite undergoing initial filter-bank processing, magnitude and phase spectrograms could also be regarded as some kind of unprocessed input feature.

However, applying task-specific preprocessing, which typically incorporates the principles of traditional SSL methods, often leads to improved convergence and performance. For raw microphone signals, features such as the generalized cross correlation with phase transform (GCC-PHAT) [2], spatial cue-augmented log-spectrogram (SALSA) [3], or SALSA-lite [4] features are commonly extracted. When dealing with binaural signals, attributes like the interaural level difference (ILD), interaural time difference (ITD), and interaural phase difference (IPD) are typically utilized [5]. Similarly, for first-order Ambisonics signals, the estimation of the (pseudo-) intensity vector has proven highly effective and has therefore been established as a common feature extraction method [6].

When extending to higher-order Ambisonics, there are considerably fewer detailed studies and established standards are not as prevalent as in first-order scenarios. This situation underscores the need for additional investigation into how higher orders might enhance localization performance and which specific features could be most effective for this purpose. Initial investigations using amplitude and phase spectrograms have indicated that higher-order features generally allow for more accurate SSL, especially in multi-speaker scenarios [7], [8]. When investigating more sophisticated higher-order features, e. g. generalizations of the first-order intensity vector to higher orders, preliminary results show considerable differences between the results; however, the use of more sophisticated and complex input features is sometimes associated with additional challenges such as overfitting and thus reduced generalizability. These results are undergoing more detailed analysis and are being prepared for publication.

Input Feature Scaling

In general, the extracted features should not be used as they are, but should be scaled, as this has been shown to improve the performance and convergence of the models. In this paper, a distinction is made between an individual normalization of the audio data (e. g. peak or rms normalization) to remove unwanted and informative bias and the dataset-wide scaling of the extracted features (here in the time-frequency domain). While there are various methods, the so-called standardization or z-score normalization to zero mean and unit variance has proven itself. With time-frequency data, there are generally three degrees of freedom (time, frequency, channel) along which scaling can be performed jointly or individually.

In a systematic investigation with amplitude and phase channels as well as first-order intensity vectors, it has been shown that, especially with magnitude and phase spectrograms it is important to align the different scales of magnitude and phase channels by an individual scaling [9]. On the other hand, relative dependencies between the different channels should be preserved by respective common scaling. The performance differences when using different scaling variants of the intensity vectors were considerably smaller. Both the joint and the individual scaling of real and imaginary parts achieved good results [9].

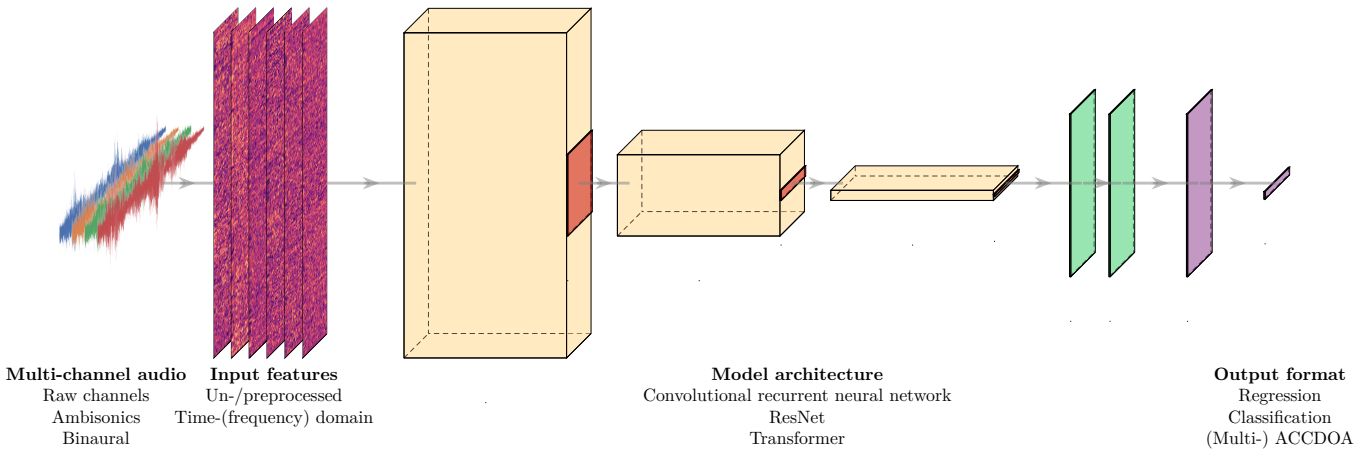

Figure 1: Basic DL pipeline for sound source source localization tasks.

Model and Problem Formulation

Model Architectures

This section provides a brief overview of established neural network architectures and output formats for the task of SSL. For a more comprehensive discussion on these architectures and their specific connections to SSL, please refer to [1]. Additionally, fundamental concepts and principles of deep learning are detailed in [10]–[12].

An established class of neural networks used in SSL are convolutional recurrent neural networks (CRNN). As the name suggests, these networks combine one or more convolutional layers with one or more recurrent layers and thus also provide the usually associated advantages of both methods. Convolutional neural networks, which emerged from the study of the brain's visual cortex, have proven to be very effective in local pattern recognition tasks, while also being resource-efficient due to using partially connected layers and weight sharing [11].

Recurrent neural networks work especially well on sequential data like audio signals [11] by processing sequences of inputs one time step at a time while maintaining a so-called state that reflects the accumulated information from observed data, effectively capturing temporal dependencies [12]. In practice, the advanced RNN variants long short-term memory (LSTM) cell or the gated recurrent unit (GRU) cell are usually deployed, which are designed to overcome the limitations of traditional RNNs in handling long-term sequences. In CRNN architectures, the combined use of convolutional layers and recurrent layers is followed by a final mapping of these integrated features to outputs via a fully-connected layer.

The (self-) attention mechanism is another key element that is increasingly used in SSL models, especially in transformer architectures, where it serves as an alternative to traditional RNNs. This approach allows the neural network to selectively focus on specific segments of the input sequence, assigning weights to different input vectors by evaluating the correlations between them [1].

Output Format

SSL tasks are typically categorized into classification and regression approaches. In the classification paradigm, SSL involves estimating whether each point on a predefined spatial grid corresponds to the direction of an active sound source or not. Therefore, the target of the model is a multi-hot-encoded vector of the size of the grid for each time frame, with each index corresponding to a certain DOA. During inference, the largest values of this vector corresponding to the respective DOAs need to be chosen, though adjacent grid bins may be mistakenly identified as different sources due to their high values. To avoid such misclassifications, a so-called peak-picking strategy is applied. There are different methods, e. g. a spatial smoothing over neighboring directions [6] or a minimum considered angular distance between different sound sources [13].

On the other hand, the regression approach aims to directly predict the Cartesian coordinates of each sound source's DOA vector for each time frame, which requires consideration of source permutations due to lack of direct source assignment or varying prediction order. To implement permutation-invariant training, minimum mean squared error over all permutations is usually used as the loss function. While each method has its own intricacies – classification requires the definition of a spherical grid, which introduces a trade-off between discretization error and task complexity, and regression models typically require a priori information on the number of sound sources – an appropriate choice of hyperparameters can lead to comparable performance between the two [13]–[15].

An alternative method combines detection and localization into an end-to-end task using the activity-coupled Cartesian DOA (ACCDOA) formulation. This approach associates sound event activity with the length of a corresponding Cartesian DOA vector [16], i. e. inactive sound sources are represented by null-vectors while the vectors corresponding to active sound sources have unit norm. During inference, a sound source is then said to be active, if the activity exceeds a certain threshold, e. g. 0.5. This ACCDOA approach outperformed two-branch architectures in the SELD task of the DCASE 2020 challenge, establishing itself as the new baseline method in the following editions [17].

Figure 2: The KEMAR HATS equipped with a Microsoft HoloLens and the 16-channel MEMS microphone array extension.

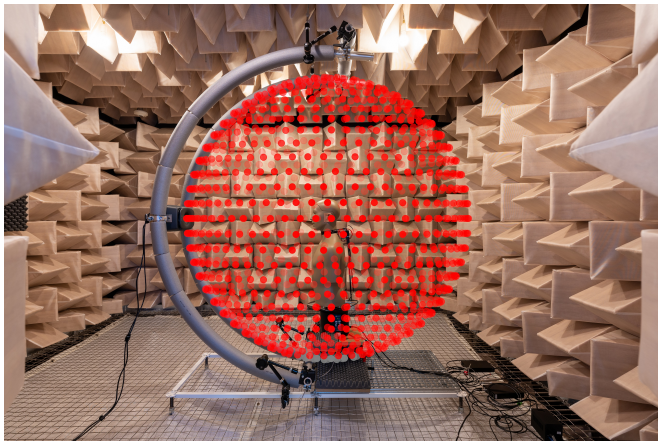

Figure 3: Impulse response measurement setup with the loudspeaker positioned at 1 m distance to the acoustic center of the KEMAR head. The red dots indicate the measurement directions; the grey dots at the bottom show source positions below −72° elevation which could not be measured.

The project Hooray

The research project Hooray – Exploring the Performance Boundaries of a Head-Worn Microphone Array for Deep Learning based Dynamic Acoustic Scene Analysis is about the evaluation of a head-worn 2D microphone array for DL based DOA estimation. Special focus is given to the assessment of the influence of different parameters such as a correct head-above torso orientation (HATO), the number and positioning of the microphones or inclusion of prior information. For example, first studies have shown that head rotations have a positive effect on the localization accuracy of DL models [18], [19]. However, only binaural signals were used and no correct representation of the HATO was taken into account.

For these investigations, two different designs of 16-channel MEMS microphone arrays were developed as extension kits for the Microsoft HoloLens mixed reality headset. The basic setup of the project, which consists of the microphone array and the HoloLens mounted on the head of a KEMAR head and torso simulator (HATS), is depicted in Fig. 2.

In order to be able to consider different types of head movements or HATOs across various scenarios involving multiple sound sources and distinct room acoustics while ensuring high diversity and real-world representativeness in the data, the training and test datasets were chosen to be simulated and synthesized from measured anechoic impulse responses.

For detailed and correct room simulations, full spherical impulse responses were needed for all HATOs. To generate such datasets, motorization of the head rotation was essential for both timing and precision reasons. Therefore, a low-cost open-source head motorization kit for the KEMAR HATS, called the LoCOMo kit was developed [20], [21]. The goal was to build a non-invasive and easy to assemble as well as simple to use motorization kit using off-the-shelf components. Therefore, the design incorporates an external toothed belt drive and an Arduino-controlled stepper motor, accompanied by 3D printed parts, a basic electronic circuit, and a UDP-based MATLAB interface for control.

Given the externally mounted construction of the LoCOMo kit, potential acoustic influence on measured HRTFs was investigated in order to assess the suitability of the kit for different conceivable applications. The acoustic influence of the motorization on the measured HRTFs was identified by a detailed quantitative comparison of broadband binaural cues and the fine spectral structure for five different elevations and at a 10°-resolution of azimuth and HATO. In general, it can be said that the LoCOMo kit has minor acoustic influence, which is mainly constrained to contralateral constellations. Based on the rather small differences identified, especially to the HRTFs of a KEMAR with equivalent neck extension, the use of the motorization should be feasible for automated acquisition of HRTF datasets of high quality and comparability.

Using the LoCOMo kit, nearly full-spherical datasets of impulse responses with variable HATO from −90° to 90° were measured with both the built-in KEMAR microphones and the head-worn microphone array. Fig. 3 shows the measurement setup featuring the KEMAR equipped with the LoCOMo kit and the loudspeaker positioned at an elevation angle of 0°.

Using these impulse response datasets, training and testing datasets were simulated and synthesized incorporating different room acoustics, speaker scenarios and head rotations. The next steps involve the training of the models on the different datasets, followed by a detailed evaluation identifying the influence of all the parameters mentioned above.

Besides common evaluation, benchmarking against state-of-the-art systems and microphone (sub-) arrays and validating the generalizability with real recordings, the models will also be analyzed using interpretability techniques like e. g. Layer-Wise Relevance Propagation (LRP) to allow a more substantial statement on the models' robustness and decision making [22].

References

[1] 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, p. 107, 2022. DOI: 10.1121/10.0011809.

[2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976. DOI: 10.1109/TASSP.1976.1162830.

[3] T. N. Tho Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial Cue-Augmented Log-Spectrogram Features for Polyphonic Sound Event Localization and Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022. DOI: 10.1109/TASLP.2022.3173054.

[4] T. N. Tho Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "SALSA-Lite: A Fast and Effective Feature for Polyphonic Sound Event Localization and Detection with Microphone Arrays," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 716–720. DOI: 10.1109/ICASSP43922.2022.9746132.

[5] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks," in *Fortschritte der Akustik - DAGA 2015*, Nürnberg, Germany, 2015.

[6] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 1, pp. 22–33, 2019. DOI: 10.1109/JSTSP.2019.2900164.

[7] 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 *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, 2021, pp. 211–215. DOI: 10.23919/EUSIPCO54536.2021.9616204.

[8] 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 *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, 2021, pp. 1015–1019. DOI: 10.23919/EUSIPCO54536.2021.9616002.

[9] N. Poschadel, R. Kiyan, S. Preihs, and J. Peissig, "On the Impact of Input Scaling Strategies for Deep Learning based DOA Estimation from Ambisonics Signals," in *Proceedings of the 24th International Congress on Acoustics (ICA 2022)*, Gyeongju, Korea, 2022.

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[11] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019, ISBN: 978-1492032649.

[12] F. Chollet, *Deep Learning with Python*, 2nd ed. New York: Manning Publications Co. LLC, 2021.

[13] N. Poschadel, S. Preihs, and J. Peissig, "Comparison of Regression and Classification Models for Multi-Source Direction of Arrival Estimation with Convolutional Recurrent Neural Networks," in *Fortschritte der Akustik - DAGA 2023*, Hamburg, Germany, 2023.

[14] 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. DOI: 10.21437/Interspeech.2019-1111.

[15] 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. DOI: 10.1109/WASPAA.2019.8937277.

[16] 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 *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 915–919. DOI: 10.1109/ICASSP39728.2021.9413609.

[17] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A Dataset of Dynamic Reverberant Sound Scenes with Directional Interferers for Sound Event Localization and Detection," *arXiv preprint arXiv:2106.06999v2*, 2021.

[18] N. Ma, T. May, and G. J. Brown, "Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localisation of Multiple Sources in Reverberant Environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, 2017, ISSN: 2329-9290. DOI: 10.1109/TASLP.2017.2750760.

[19] G. Garcia-Barrios, D. A. Krause, A. Politis, A. Mesaros, J. M. Gutierrez-Arriola, and R. Fraile, "Binaural source localization using deep learning and head rotation information," in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 36–40. DOI: 10.23919/EUSIPCO55093.2022.9909764.

[20] N. Poschadel, S. Preihs, and J. Peissig, "LoCOMo: A Low-Cost Open-Source Head Motorization Kit," in *155th Convention of the Audio Engineering Society*, New York City, NY, USA, 2023.

[21] N. Poschadel, *Locomo: A low-cost open-source head motorization kit for the kemar head and torso simulator*, GitLab, https://go.lu-h.de/locomo, 2023.

[22] R. Kiyan, N. Poschadel, S. Preihs, and J. Peissig, "Adaption of Layerwise Relevance Propagation for Audio Applications," in *Fortschritte der Akustik - DAGA 2022*, Stuttgart, 2022.