Deep Learning for Sound Source Localisation based on Data from the DCASE Challenge 2020

Pablo Marhoff, Adam Kujawski, Ennes Sarradj
Technische Universität Berlin, 10587 Berlin, Deutschland, p.marhoff@tu-berlin.de

Introduction

Sound contains a large amount of information about our everyday environment. The development of signal processing methods to detect the presence of a specific sound event and to determine their spatial location is of major importance in a variety of application areas, including context-sensitive mobile devices, industrial robots, hearing aids or autonomous vehicles. A successful implementation of acoustic analysis systems can enable an automatic, camera-free description of the environment. This can help machines to act more seamlessly in complex environments [1]. Nowadays, microphone arrays are used to obtain rich information from an existing sound field. Traditional methods like beamforming are commonly used to process the microphone array's channel data and to solve the problem of localising sound sources or to obtain spatially filtered source signals.

In recent years, research development has tended to move away from physical model-based methods towards data-driven algorithms [2]. The latter are capable of extracting information from raw data by means of artificial intelligence. Specifically in localisation tasks, the use of Deep Learning and Deep Neural Networks has been proven to yield fast and accurate localisation results for a broad range of source signal types, including speech [3], real-world sound events [4] or stationary noise [5]. One appealing approach is to use advanced image recognition models to extract the spatial information from an acoustic source mapping, as done by Kujawski et al. [5]. So far, this concept has only been verified on single source cases and stationary synthetic data.

The aim of this contribution is to investigate whether a similar approach can be successfully used to localise real world non-stationary sound events in indoor environments from spherical microphone array data. Furthermore, it is examined whether the performance and robustness of the localisation results can be improved compared to model-based methods like beamforming. A new end-to-end model is introduced and applied to the "TAU-NIGENS Spatial Sound Events 2020" dataset [6] from the DCASE2020 Challenge [7].

Methodology

Figure 1 illustrates the concept for the localisation of sound events investigated in this work. It is worth noting that the model focuses exclusively on localisation, which means that no classification of the occurring sound events is performed. Starting from the raw channel data of a four-channel microphone array, the model aims at predicting the direction-of-arrival (DOA) in the horizontal and vertical plane for a single sound event. Therefore, the model consists of three main stages of processing:

1. **Level threshold stage:** determines the presence of a sound event for a certain time frame
2. **Beamforming stage:** processes an active labelled time frame and yields multiple one-third octave source maps
3. **Source map recognition stage:** contains a convolutional neural network (CNN) architecture that processes the beamforming maps and provides an estimate of the DOA

During the first stage, the audio signals are divided into 100 ms frames. The sound pressure level (SPL) of the first chan-

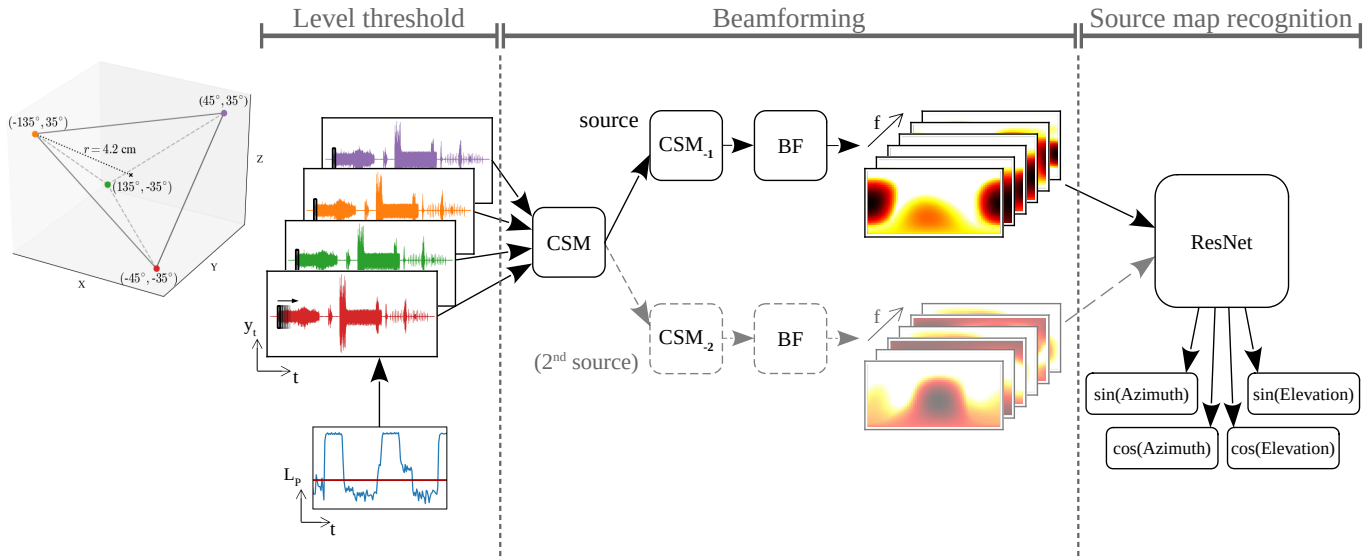

Figure 1: Block diagram of the end-to-end model for source localisation

nel (purple) is calculated over each time frame. A customised SPL-threshold is then used to decide if a frame includes an active source and should be further processed.
The beamforming stage creates a discrete source mapping $\mathbf{B} \in \mathbb{R}^{S_\theta \times T_\phi}$ on a spherical grid. The grid surrounds the microphone array with a radius of 1 m to the array center. S_θ corresponds to the number of discrete elevation angles and T_ϕ represents the total number of azimuth angles. The resolution of the source mapping results from the selected number of grid points. The grid points are arranged equiangularly, resulting in oversampling at the poles and undersampling in the area of the horizontal great circle. A uniform grid structure in the azimuth-elevation plane is obtained, which is suitable as input for the CNN. For all following steps, the spherical grid is rolled out into the two-dimensional representation.
In contrast to [5], eigenvalue beamforming in the frequency domain is used to calculate the source maps. Therefore, an estimate of the cross-spectral matrix (CSM) $\mathbf{C} \in \mathbb{C}^{M \times M}$ for M microphones is calculated by using Welch's method [8], with

$$\mathbf{C} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{p}_k \mathbf{p}_k^{\mathrm{H}}, \tag{1}$$

where H stands for the hermitian transpose. The M microphone signals for a single time frame are first divided by temporal windowing into K blocks consisting of N samples. Afterwards, each individual signal block is transformed into the frequency domain by a fast Fourier transform (FFT). The single frequency coefficients of the k-th signal block and for all sensors are organised in the vector $\mathbf{p} \in \mathbb{C}^M$. In this work, a time frame of 100 ms length corresponds to $K = 5$ with $N = 512$. Subsequently, an eigendecomposition of the CSM is performed. Assuming that the eigenvalues Λ and the eigenvectors $\mathbf{v}$ of the CSM belong to the available sources, a CSM can be resynthesised for each eigenvalue. In combination with subsequent beamforming, a decomposed representation of the source mapping is obtained by

$$b_i(\theta, \phi) = \mathbf{h}^{\mathrm{H}} \mathbf{C}_i \mathbf{h} \quad \text{with} \quad \mathbf{C}_i = \mathbf{v}_i \Lambda_{ii} \mathbf{v}_i^{\mathrm{H}}. \tag{2}$$

In Equation 2, $\mathbf{h}$ is the steering vector that connects each microphone position with a search point of the spherical grid. The use of eigenvalue beamforming enables to filter out specific source components and to suppress unwanted noise by disregarding small eigenvalues. It should be noted that, depending on the source type, the distribution of the spectral energy can vary over different FFT coefficients. For this reason, $F = 9$ one-third octave beamforming maps were calculated in the frequency range of 500 Hz to 3150 Hz. Each associated source map is than considered as a separate image channel.
The first two stages can be declared as a feature extraction section that collects and prepares the data for Deep Learning. The last stage is than used to yield an accurate and robust prediction by means of the advanced ResNet model $\mathcal{F}_{\mathrm{ResNet}}$. The source maps obtained from the Beamforming stage are fed into the ResNet model that outputs a vector containing the expected DOA of the source for a single frame:

$$\mathcal{F}_{\mathrm{ResNet}} : \mathbb{R}^{S_\theta \times T_\phi \times F} \rightarrow \mathbb{R}^4. \tag{3}$$

The ResNet model was developed by He et al. [9] to solve image recognition tasks. It is a convolutional neural network (CNN) in which the activation of a neuron results from a discrete convolution [10]. By using convolution operations instead of matrix-vector operations, significantly fewer weights are needed and local features are recognised regardless of their position or orientation [2]. A fundamental difference of the ResNet to other CNNs are the residual blocks. These blocks determine the residual between the input and the output to the layer.
Since the target values – azimuth (ϕ) and elevation (θ) of the DOA – are circular data, they must be encoded appropriately to avoid the representation-related leap between 2π and 0 and an incorrect error calculation due to its periodic nature. The output vector therefore corresponds to the calculated DOA decomposed into its sine and cosine components

$$\mathcal{F}_{\mathrm{ResNet}}(\mathbf{B}) := [\phi_{sin}, \phi_{cos}, \theta_{sin}, \theta_{cos}]. \tag{4}$$

Model Architecture

Table 1 shows the detailed structure of the ResNet used in this work. The dimension of the input layer – and thus of all further layers – results from the size of the beamforming maps and the number of frequency bands considered.
The first discrete convolution merges the beamforming maps of all frequency bands by means of a three-dimensional kernel dimension. The number of kernels used for the convolution determines the number of feature maps of the following layer. The main component of the ResNet consist of the three residual layers, here called Res I, Res II and Res III. Each of these layers in turn consists of three successively stacked blocks. The optimum structure of the layers has been investigated by He et al. [9]. A residual layer halves the dimension of the feature maps and, in order to maintain the amount of information, doubles their number.

Table 1: Network architecture and no. of trainable parameters of the used ResNet based on He et al. [9]. The upper value in the square brackets corresponds to the number of kernels or their size in the first convolution, the lower value to the second convolution. The subsequent factor stands for the number of these corresponding blocks per layer.

Block	Dim. Input	Dim. Output	No. kernels	Kernel Dim.	No. weights
Conv. Layer	35 x 70 x 9	35 x 70 x 25	25	3 x 3 x 9	2,025
Res I	35 x 70 x 25	18 x 35 x 25	$\begin{bmatrix}25\\25\end{bmatrix}$ x 3	$\begin{bmatrix}3 \text{ x } 3\\3 \text{ x } 3\end{bmatrix}$ x 3	34,675
Res II	18 x 35 x 25	9 x 18 x 50	$\begin{bmatrix}25\\50\end{bmatrix}, \begin{bmatrix}50\\50\end{bmatrix}$ x 2	$\begin{bmatrix}3 \text{ x } 3\\3 \text{ x } 3\end{bmatrix}$ x 3	125,550
Res III	9 x 18 x 50	5 x 9 x 100	$\begin{bmatrix}50\\100\end{bmatrix}, \begin{bmatrix}100\\100\end{bmatrix}$ x 2	$\begin{bmatrix}3 \text{ x } 3\\3 \text{ x } 3\end{bmatrix}$ x 3	501,100
BN & Pooling	5 x 9 x 100	100 x 1	1	5 x 9	200
Regression	100 x 1	400 x 1	*400 nodes*		40,400
Output Layer	400 x 1	4 x 1	*4 nodes*		1,604

In the Batch normalisation (BN) and Pooling layer, the mean value of each feature map is calculated and passed through a regression layer into the output layer, which generates the target values.

Dataset

The dataset [6] used in this work was taken from the Sound Event Localisation and Detection Challenge (DCASE2020) [7]. The "TAU-NIGENS Spatial Sound Events 2020" dataset contains 800 one-minute audio files. Of these, 400 files constitute the training dataset with a total of 120,624 frames with an active sound source. The remaining half forms the test dataset. The semi-synthetic microphone signals are randomly composed from a set of 14 different sound events, convolved with one of 15 impulse responses (IRs) of real rooms from the Tampere University campus (Finland). The IRs were recorded using the microphone array shown in Figure 1. Further, additional ambient noise (e.g. office, canteen/cafeteria, corridor) with different signal-to-noise ratios from 30 dB (noiseless) to a minimum of 6 dB (noisy) exists at the recordings. The spatially moving and non-moving sound events are documented with a resolution of 100 ms. The occurring sound event classes include transient or non-transient events (e.g. knocking on door, piano), as well as broad or narrow band sounds (e.g. running engine, alarm).

Experimental Settings

In this contribution, the dimension of the spherical grid has been set to $S_\theta = 70$ and $T_\phi = 35$, resulting in a spatial resolution of at least 5.1°.

The model optimisation is treated as a supervised learning problem. The training objective is to minimise the mean squared error E between the model estimation $\mathbf{y}_s = [\phi_{sin}, \phi_{cos}, \theta_{sin}, \theta_{cos}]$ and the label $\hat{\mathbf{y}}_s$ for each sample denoted by the lower index s. Here, the Adam optimisation algorithm [11] with a learning rate of 0.001 was used.

A localisation error $\Delta\sigma$ in degrees [°] has been defined to assess the performance of the model. This error is defined as the shortest angular distance between the actual DOA ($\hat{\phi}$, $\hat{\theta}$) and the predicted DOA (ϕ, θ), which corresponds to the great-circle distance

$$\Delta\sigma = \arccos(\sin\hat{\theta} \cdot \sin\theta + \cos\hat{\theta} \cdot \cos\theta \cdot \cos|\hat{\phi} - \phi|). \quad (5)$$

The training of the ResNet lasted 150 epochs with 2000 iterations per epoch. The batch size was set to $|\mathcal{B}| = 8$.

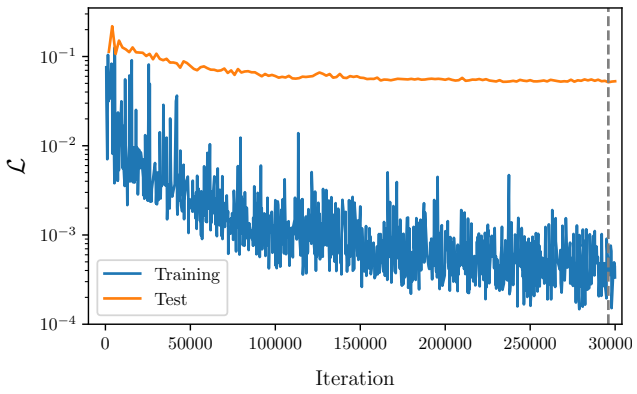

Figure 2: Training and test error development

Results

The development of training (blue) and test error (orange) during training is shown in Figure 2. The vertical dashed line marks the location of the minimum test error. The training was stopped after 300,000 iterations to limit the CPU time. The training step with the lowest loss value over the evaluation data has been achieved at the 296,000th iteration, with a minimum error of 0.05.

The angular location with the maximum energy across the source maps of all frequencies was considered in order to obtain an estimate of a sound events DOA with the eigenvalue beamforming (EB) method. This procedure will be denoted as EB_{max} in the following. Figure 3 compares the localisation performance of the data-driven ResNet model (middle) with the model-based EB_{max} method (upper) and the true reference values (bottom) for a specific audio file from the test dataset. It is clearly visible that the localisation performance of the ResNet is significantly superior to the beamforming method. The results of the ResNet appears to be more robust since only few outliers can be observed. The mean localisation error for the whole file decreases from 32.7° (EB_{max}) to 11.8° by using the ResNet. However, the estimated elevation values appear to have a tendency towards 0° compared to the reference values. This behaviour was observed for all audio files and sound event classes. It is assumed that this systematic bias in elevation estimation can be attributed to an unbalanced training dataset. The probability that a sound

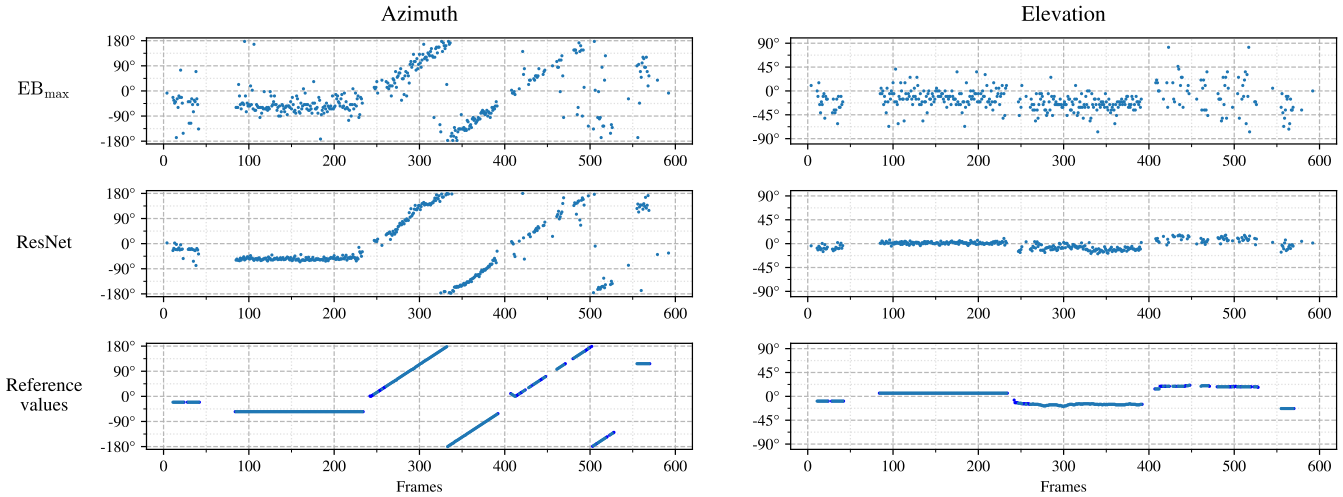

Figure 3: Comparison of the localisation results. Barking dog (Frame 11-42), crash (84-234), running engine (242-392), crying baby (407-528), female speech (555-570)

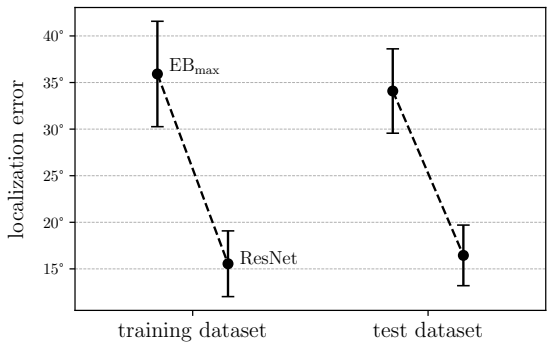

Figure 4: Mean and standard deviation of the two methods, EB_{max} (left) and ResNet (right) for the training and the test dataset

event occurs near the zero elevation plane is particularly high. Considerably large and constant localisation errors were observed for narrowband sources. This was particularly noticeable for the "alarm" sound class.

The localisation error over the training and test dataset is shown in Figure 4. The results obtained with the beamforming method (EB_{max}) are compared against the ResNet model. On average, the localisation error is reduced significantly through the use of the ResNet by 54.4 % (from 35.3° to 16.1°). Furthermore, a decrease of the mean standard deviation from 10.1 ° to 6.5 ° can be observed. Although the mean error from the ResNet over the test dataset is larger than over the training dataset, it can be assumed that no overfitting to the training data has occurred. Comparing the mean error of the single file (illustrated in Figure 3) with the mean error over the test dataset it can also be concluded, that the results observed over the audio file can be generalised to the entire dataset.

Summary

In this contribution, it was demonstrated how Deep Learning can improve the localisation results for non-artificial sound sources based on acoustic source maps compared to beamforming. A new model for the localisation of non-stationary real-world sound events was proposed. The end-to-end model consists of three stages of processing: Detection of active sound sources, feature extraction using eigenvalue beamforming and estimation of the DOA through a ResNet. By the use of eigenvalue beamforming, the essential information regarding the DOA of a temporally non-overlapping sound source has been extracted. The information was then brought into a structure that is suitable as input data for the ResNet. This method improved the localisation error significantly by almost 60 % compared to the beamforming localisation results.

References

[1] Y. Schulz et al., "Hearing What You Cannot See: Acoustic Vehicle Detection Around Corners," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2587-2594, 2021, doi: 10.1109/LRA.2021.3062254.

[2] Y. LeCun et al., "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015, doi: 10.1038/nature14539.

[3] J. Vera-Diaz et al., "Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates," *Sensors*, vol. 18, no. 10, p. 3418, 2018, doi: 10.3390/s18103418.

[4] Y. Cao et al., "Event-Independent Network for Polyphonic Sound Event Localization and Detection," 2020, ArXiv: abs/2010.00140.

[5] A. Kujawski et al., "A deep learning method for grid-free localization and quantification of sound sources," *J. Acoust. Soc. Am.*, vol. 146, no. 3, pp. EL225–EL231, 2019, doi: 10.1121/1.5126020.

[6] A. Politis et al., "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *Proc. Workshop Detection and Classification Acoustic Scenes and Events*, 2020, arXiv: 2006.01919.

[7] S. Adavanne et al., "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, 2019, doi: 10.1109/jstsp.2018.2885636.

[8] P.D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio and Electroacoust.*, vol. 15, no. 2, pp. 70–73, 1967, doi: 10.1109/TAU.1967.1161901.

[9] K. He et al., "Identity mappings in deep residual networks," *Lect. Notes Comput. Sci. (including subseries Lect. Notes Artif. Int. and Lect. Notes Bioinformat.)*, pp. 630-645, 2016, arXiv: 1603.05027.

[10] Y. Lecun et al., "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998, doi: 10.1109/5.726791.

[11] D. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization," *Int. Conf. Learning Representations*, San Diego, CA, USA, May 2015, arXiv: 1412.6980.