A Deep Learning Approach for Angle Specific Source Separation from Raw Ambisonics Signals

Francesc Lluís[1], Nils Meyer-Kahlen[2], Vasileios Chatziioannou[1], Alex Hofmann[1]
[1] *University of Music and Performing Arts Vienna, 1030 Vienna, Austria, Email: lluis-salvado@mdw.ac.at*
[2] *Aalto University, 02150 Espoo, Finland, Email: nils.meyer-kahlen@aalto.fi*

Introduction

Deep neural networks have significantly increased the performance of sound source separation in recent years, consistently outperforming previous methods [1]. At the same time, higher order Ambisonics is more broadly adopted, for example in virtual reality applications. Ambisonics is a scene-based spatial audio format, in which the spatial information is contained in the full set of signals, as opposed to object-based formats, in which each sound source is stored separately together with positional metadata [2]. Therefore, individual sound sources are not readily available when given an Ambisonics scene.

Here, we present a deep learning approach that performs end-to-end source separation from raw Ambisonics signals, conditioned on a specific direction. In practise, the algorithm can be used like a beamformer, but instead of separating sources only in the spatial domain, it also benefits from the ability of source separation that inherently operates in the time-frequency domain. Related research proposes to use multi-channel audio recordings and a neural network to localize and separate speech signals in the horizontal plane [3]. However, this approach is limited to only audio input from rotationally symmetric microphone arrays and uses a pre-shift method which is not applicable in audio formats that are agnostic to the microphone array configuration, such as Ambisonics.

Results on anechoic musical mixtures show that the neural network can extract sound from a specific target direction. The largest gains over a modal beamformer are obtained for low orders or proximate sources. This method may for example be part of a processing chain converting Ambisonics audio recordings, which are nowadays easy to capture using available microphones, to object-based formats.

Dataset generation

We use signals from the MUSDB18 [4] dataset to create training, validation, and testing data. The MUSDB18 dataset contains a 'train' folder with 100 songs and a 'test' folder with 50 songs. For each song, the dataset provides the isolated signals of the *drums*, *bass* and *vocals* sound sources. We use signals from 90 songs in the 'train' folder to generate training data and the remaining 10 songs are used to generate validation data. For training and validation, a single example is created by first selecting six-second long audio segments from the isolated signals at a random time. Therefore, every isolated signal is taken from a random song for each source, so they do not necessarily come from the same piece. In addition, we verify that the particular source is active in the audio segment. Then, to create an Ambisonics mix, a random direction is assigned to each of the three sources, and the audio segments are encoded up to fourth-order Ambisonics [2, p. 71]

$$\boldsymbol{\chi}_N(t) = \sum_{q=1}^{3} \boldsymbol{y}_N^T(\boldsymbol{\theta}_q) s_q(t), \tag{1}$$

where s_q is the audio segment of the source q and $\boldsymbol{y}_N(\boldsymbol{\theta}) = \begin{bmatrix} Y_0^0(\boldsymbol{\theta}) & Y_1^{-1}(\boldsymbol{\theta}) & Y_1^0(\boldsymbol{\theta}) & \dots & Y_N^N(\boldsymbol{\theta}) \end{bmatrix}$ is a vector of spherical harmonics (SH), evaluated at the direction $\boldsymbol{\theta} = [\phi, \vartheta]$, with azimuth angle ϕ and zenith angle ϑ. Note that such an encoding corresponds to placing a sound source in the far field with respect to an ideal Ambisonics receiver. During source placement, in the data generation step, it is assured that all pairs of sources are at least $\varphi_{i,j} = 5°$ apart from each other. The distance is defined as the great circle distance between the encoding direction of two sources

$$\varphi_{i,j} = \arctan \boldsymbol{x}_i^T \boldsymbol{x}_j, \tag{2}$$

where $\boldsymbol{x}_i = \begin{bmatrix} \cos\phi_i \sin\vartheta_i, \sin\phi_i \sin\vartheta_i, \cos\vartheta_i \end{bmatrix}^\top$ is a normalized direction vector. Further, in 30% of all created mixes, we force one source to be silent. For training the presented model, 10000 mixes are generated. Additionally, 1000 mixes are created for validation.

Regarding test data generation, the same encoding is applied to generate a total of 1000 mixes using the 'test' folder. In this case, single examples are created using six-second long audio segments coming from the same song at the same time, as one would expect, and no sources are silenced. As we aim to investigate the effect that proximate sources have on the separation performance, we create two test sets. In the first one, sources are randomly placed on the sphere ensuring that no two sources are within 5° great circle distance. In the second one, sources are randomly placed such that their great circle distance is between 5° and 10°.

Neural Network

Architecture. We adapt the Demucs neural network [5] to extract the signal s from a specific target direction $\boldsymbol{\theta}$, given the raw Ambisonics mixture as input signal $\boldsymbol{\chi}_N$. Specifically, Demucs input channels are modified according to the Ambisonics order and a global conditioning approach [3] is used to guide the separation according to the target direction.

Demucs operates in the waveform domain and is based on a U-Net convolutional architecture, i.e. encoder-decoder

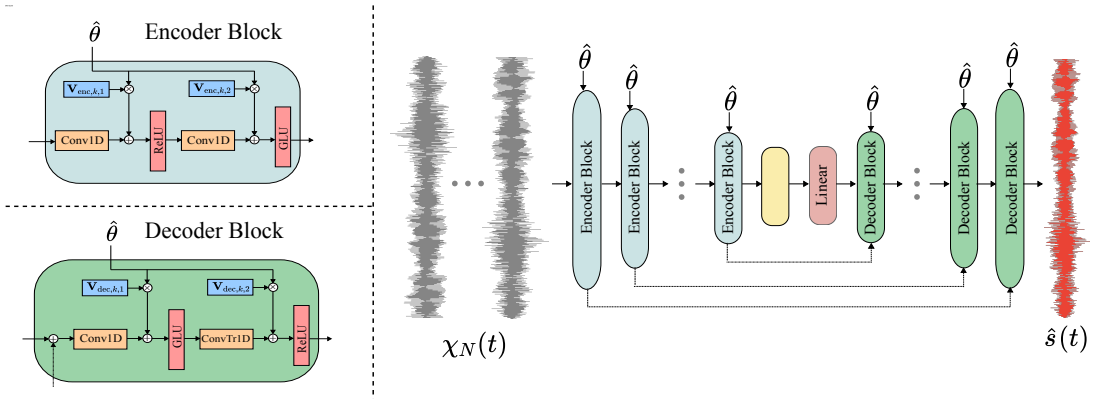

Figure 1: Top left: Encoder block diagram, used in early parts of the neural network. Bottom left: Decoder block diagram, used in late parts of the neural network. Right: Overview diagram of the neural network. $\boldsymbol{\chi}_N(t)$ refers to the Ambisonics mixture, $\hat{\boldsymbol{\theta}}$ is the scaled target direction, and $\hat{s}(t)$ is the estimated separated signal.

architecture with skip connections (See Fig.1). Demucs uses 6 convolutional blocks for both the encoder and the decoder. At each encoder block, the input feature map is downsampled while the feature channels are doubled. The decoder reverses this procedure by upsampling the feature maps and halving the feature channels at each block. A Bidirectional Long-Short Term Memory (BiLSTM) between the encoder and decoder is applied to provide long range context. A key characteristic of U-net-like architectures are the skip connections (represented as dashed arrows in Fig.1) which allow the decoder to access the feature maps computed by the encoder at the same level of hierarchy. This enables the decoder to access low-level information that may be lost when propagated through the network. In the current case, it allows to propagate level and phase differences between the raw Ambisonics channels. This is relevant to learn the correspondence between the spatial information and the conditioning direction for further source separation.

Conditioning. We scale the target angles to condition the neural network. Specifically, given a target direction $\boldsymbol{\theta} = [\phi, \vartheta]$, with azimuth angle $\phi \in [-\pi, \pi]$ and zenith angle $\vartheta \in [0, \pi]$, the scaled target direction $\hat{\boldsymbol{\theta}} = [\hat{\phi}, \hat{\vartheta}]$ is defined as $\hat{\phi} \in [-1, 1]$ and $\hat{\vartheta} \in [-1, 1]$. Then, $\hat{\boldsymbol{\theta}}$ is inserted at each block of the audio network after being multiplied by a learnable linear projection as in [3].

Supervised Training. The network is trained in a supervised manner. For each instance, the network is provided with the Ambisonics mix, the target direction and the ground truth signal, which can be silent as discussed above. The parameters of the network are optimized to reduce the ℓ_1 loss between the estimated signal and the ground truth signal in the target direction. During training, as target direction, we randomly select one of the three sources directions and uniformly perturb it within a 2.5° window. This perturbation determines the spatial selectivity of the network when using it. The network is trained for 350 epochs using the Adam optimizer. Learning rate is set to 1×10^{-4} and it is reduced by a factor 0.1 after 10 epochs with no improvement in the validation set loss. The batch size is set to 16. After the training process, we select the weights with the lowest validation loss for testing purposes. Both training and testing are conducted on a single Titan RTX GPU. The training stage takes about 72 hours while the inference takes 47 milliseconds for a single data sample (value averaged from 300 different separation predictions)

Baseline

For comparison to the neural network separation, we use the output of a linear modal beamformer pointed to the desired target direction. In the SH domain, the shape of an axisymmetric beam is determined by a set of weights, $\boldsymbol{w} = \left[w_0, w_1, w_1, w_1, w_2, ...w_N\right]$, with one unique weight per Ambisonics order

$$s(t) = \boldsymbol{y}_N(\boldsymbol{\theta}_q)\mathrm{diag}(\boldsymbol{w})\boldsymbol{\chi}_N(t). \tag{3}$$

For the following comparison, a max-$\boldsymbol{r}_\mathrm{E}$ beamformer is used, which offers a good compromise between side lobe strength and main lobe width, by minimizing the directional spread (maximal $\boldsymbol{r}_\mathrm{E}$ vector). Max-$\boldsymbol{r}_\mathrm{E}$ weights are approximated by [2]

$$w_n \approx P_n\Big(\frac{\cos\left(137.9°\right)}{N + 1.51}\Big), \tag{4}$$

where P_n is the Legendre function of order n.

Evaluation

We evaluate the audio source separation performance using the scale-invariant source to distortion ratio (SI-SDR) [6]. The SI-SDR between a signal s and its estimate $\hat{s}$ is defined in units of decibels as

$$\mathrm{SI\text{-}SDR}(s, \hat{s}) = 10\log_{10}\left(\frac{\|\alpha s\|^2}{\|\alpha s - \hat{s}\|^2}\right), \tag{5}$$

where $\alpha = \mathrm{argmin}_\alpha\|\alpha s - \hat{s}\|^2 = \hat{s}^T s/\|s\|^2$. Specifically, for each mixture in the test set, we compute the separations using the ground truth direction of each source in case it is not silent, i.e. $\boldsymbol{\theta}_q$, as target direction for both the neural network (NN) and the linear modal beamformer (BF). Then, we assess the separation performance between the estimated signal $\hat{s}_q$ and the ground truth signal s_q.

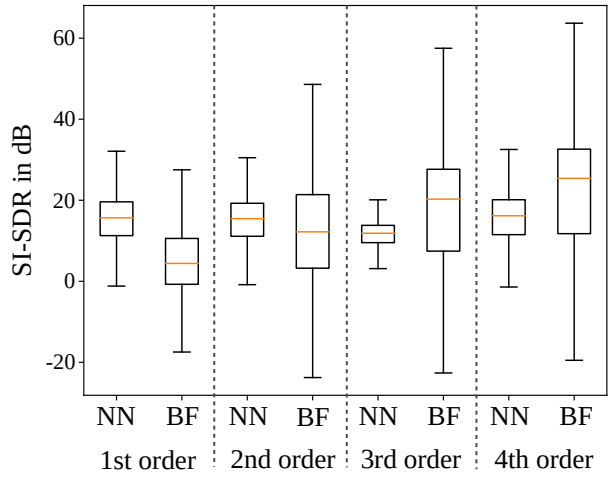

Figure 2: SI-SDR box plot on the test set where sources are randomly placed on the sphere. The box extension shows the first and third quartile of the data with a line at the median. The whiskers extending from the box show the range of the data.

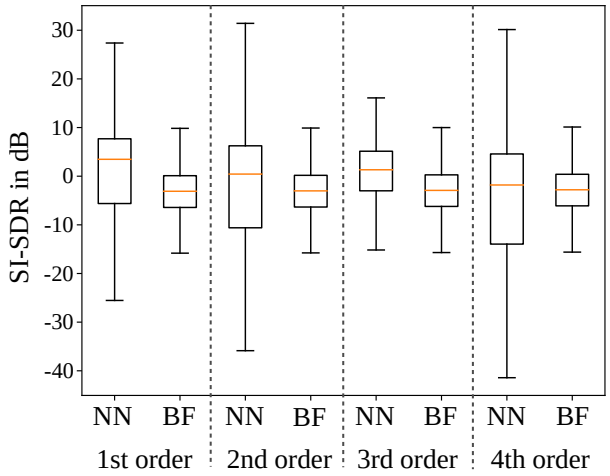

Figure 3: SI-SDR box plot on the test set where sources are randomly placed within distances of $5° \leq \varphi_{i,j} \leq 10°$. The box extension shows the first and third quartile of the data with a line at the median. The whiskers extending from the box show the range of the data.

Ambi. Order	Neural Network	Beamformer
1	**15.66**	4.40
2	**15.44**	12.20
3	11.85	**20.27**
4	16.16	**25.40**

Table 1: SI-SDR median scores in dB calculated for the test set, where sources are randomly placed on the sphere. It is assured that no two sources are within 5° great circle distance on the sphere. As a reference, the median SI-SDR achieved when the omni channel of the Ambisonics mix is used as the separated source is -3.32 dB.

Ambi. Order	Neural Network	Beamformer
1	**3.47**	-3.09
2	**0.43**	-3.01
3	**1.32**	-2.91
4	**-1.80**	-2.79

Table 2: SI-SDR median scores in dB calculated for the test set, where sources are randomly placed within distances of $5° \leq \varphi_{i,j} \leq 10°$. As a reference, the median SI-SDR achieved when the omni channel of the Ambisonics mix is used as the separated source is -3.16 dB.

Results

We are interested in evaluating the source separation performance depending on the Ambisonics order. To this end, we compare the separation performance considering Ambisonics mixes with an order between one and four. Note that for each order, we adapt the input channels of the neural network accordingly and retrain the model.

Results in Table 1 show that when sources are randomly placed, the neural network separation performance is better for low orders (1st and 2nd) while the beamformer separation is better for high orders (3rd and 4th). Interestingly, a similar neural network source separation performance is observed, independently of the encoding order it has been trained and tested. For the third order network, the separation is even worse than for the lower order networks. Hence, low orders seem enough for the neural network to learn the correspondence between the spatial information contained in the Ambisonics signal and the conditioning direction to perform source separation. Figure 2 compares the neural network and the beamformer source separation performance for different encoding orders when sources are randomly placed over the complete sphere.

We observe that in examples where sources are close to each other, the neural network provides a better separation performance compared to the linear beamformer. This is to be expected because the neural network can learn to separate the sources through non-linear operations in time, frequency and space, while the linear beamformer (see Eq. 3) is strictly limited by its spatial selectivity. To further investigate how the proximity of the sources affects the separation performance, we use the same trained models from the previous experiment, but apply them in the test set that we generated ensuring that sources are randomly located 5° to 10° degrees great circle distance close to each other. Results in Table 2 and Figure 3 show that when sources are close, the neural network performs better separation than the beamformer for any encoding order. Also, the network better separates the sources when it has been trained using lower orders.

In order to better understand the behaviour of the network, Figure 4 compares the output of the linear beamformer with max-$\boldsymbol{r}_{\mathrm{E}}$ weights to the network output for a dense grid of target directions in case of a first order Ambisonics mix. In this example, two sources, marked as red dots, are close to each other, while the other one is farther away. First of all, by computing the root-mean-

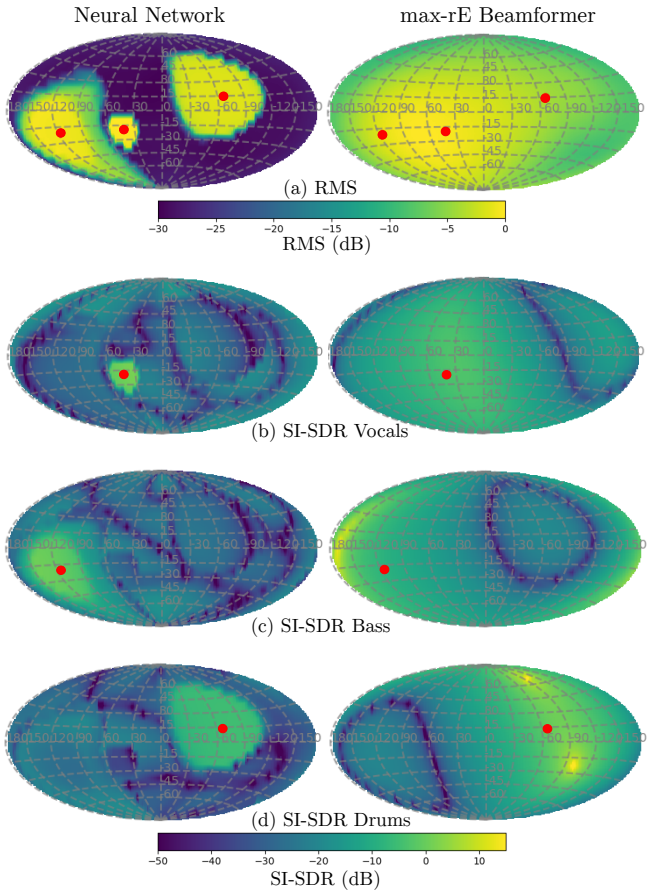

Figure 4: Visualization of Neural Network and Beamformer a)RMS and b),c),d) SI-SDR for audio predicted in a discrete set of directions (100 azi and 50 zen) using a first order Ambisonics mix where two sources are close and the other is far. The red dots correspond to the location of the sources.

square (RMS) of the predicted audio at each direction, it can be seen that the neural network is able to achieve a higher contrast between silent regions and source regions, which is not captured by the metrics shown so far. By computing the SI-SDR at each direction between the predicted source and the ground truth source, it can also be seen that the regions, in which the output of the neural network is similar to the ground truth source, are much more distinct than in the case of the beamformer. Hence, the non-linear source separation processing achieved by the network offers high spatial selectivity, even in the case of first order input.

Conclusion

In this work we presented a deep learning approach that performs source separation from raw Ambisonics signals, conditioned on a target direction. The results suggest that through conditioning with a target angle, a source separation neural network can implicitly learn the correspondence between the spatial information contained in the Ambisonics signal and the conditioning direction. We show that first-order Ambisonics mixtures are sufficient for the neural network to learn the correspondence between the mixture spatial information and the target direction to perform source separation. Interestingly, the largest improvement with respect to a linear beamformer with max-$\boldsymbol{r}_{\mathrm{E}}$ weighting are achieved when the neural network is trained and operated on first order mixtures. Additionally, the neural network performs better separation, for any order, when the sources are close to each other. When sources are randomly placed over the complete sphere, the neural network yields better separation at lower orders, i.e. 1st and 2nd order, while the beamformer performs better separation for higher orders, i.e. 3rd and 4rt. In theory, it should be possible to train a network that achieves at least linear resolution in all cases, for example by optimally combining the output of a linear beamformer and the neural network, which will be the subject of future work. Further, future work will include assessing the performance with real-world recordings from spherical microphone arrays.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812719.

References

[1] Stöter, F., Liutkus, A., Nobutaka, I.: The 2018 signal separation evaluation campaign, in International Conference on Latent Variable Analysis and Signal Separation, Springer, 2018, pp. 293–305

[2] Zotter, F. and Frank, M.: Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality, vol. 19., Springer International Publishing, 2019. `https://doi.org/10.1007/978-3-030-17207-7`

[3] Jenrungrot, T., Jayaram, V., Seitz, S., and Kemelmacher-Shlizerman, I.: The Cone of Silence: Speech Separation by Localization, in Advances in Neural Information Processing Systems, 2020

[4] Rafii, Z., Liutkus, A., Stöter, F., Mimilakis, S. and Bittner, R.: The MUSDB18 corpus for music separation. `https://doi.org/10.5281/zenodo.1117372`

[5] Défossez, A., Usunier, N., Bottou, L., and Bach, F.: Music Source Separation in the Waveform Domain, 2019

[6] Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.: SDR–half-baked or well done?, in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 626–630