forum acusticum 2026
8-12 September · Graz, Austria

TOWARDS THE EVALUATION OF COMPLEX ACOUSTIC ENVIRONMENTS USING SOUND SOURCE SEPARATION METHODS

Sean Kenji Cragg[1], Will Bailey[1], Antonio J Torija[1]

[1] *Acoustics Innovation Institute, University of Salford, United Kingdom*

Abstract

Perception of a soundscape is shaped significantly by the dominant sound sources present in an acoustic environment. Spatial characteristics have also been suggested to influence how a soundscape is perceived. Soundscape recordings are often analysed holistically rather than at the level of individual sources; the ability to separate and analyse sources individually could enable deeper insight into how specific sources drive overall perception.

Spatial audio formats are commonly used in soundscape assessment, yet a source stripped of its directional character cannot be used to investigate spatial contributions to perception, or re-rendered for listening experiments. The preservation of spatial characteristics, however, appears to be relatively unexplored for the task of universal sound separation.

This paper presents a minimal modification of ConvTasNet to accept and output first-order ambisonic (FOA) signals for universal sound separation (separating sources of any class), comparing three loss functions (SNR, SI-SDR, and a multichannel SI-SDR variant) for FOA-to-FOA separation. Results indicate that the multichannel SI-SDR loss substantially improves both separation performance and preservation of source direction compared with conventional SI-SDR and SNR losses. These early-stage findings suggest spatially-preserving source separation may hold promise as a tool for future soundscape studies.

Keywords: *soundscape, source separation, deep learning, ambisonics*

1. Introduction

The individual sources present within a soundscape can influence the perception of the soundscape as a whole. A prior soundscape study by Pérez-Martínez et al. [1] at the Alhambra of Granada, Spain, found that when visitors predominantly found a pleasant sound to be the dominant source (such as birds or water) at a given site, the higher the overall reported soundscape quality. Similarly, an online listening experiment investigating how sound sources affect annoyance in urban soundscapes found that the presence of traffic related sound sources increased annoyance while the presence of nature sounds decreased it. The study also found that the number of identifiable sources had a significant relationship with annoyance ratings [2]. Although this suggests that the specific sources present may play a role in how a soundscape is perceived, the individual sources are treated only as present or absent, rather than as individually quantifiable contributions to the mixture - as such it is difficult to quantify how a given source may contribute to the perceived annoyance. Many soundscape studies have also made use of psychoacoustic indicators to try to explain perceptual attributes [3]. However, without the use of source separation methods such an analysis will be carried out on an acoustic environment as a whole as opposed to analysing individual sources.

Sound source separation, the task of recovering the individual signals that make up an audio mixture, offers a potential route to gaining further understanding on how specific sources effect the perception of a soundscape. Rather than relying on listeners to report which sources are present/dominant, or computing acoustic and psychoacoustic metrics from a recording as a whole, separation would allow each source's own signal to be isolated and analysed independently. Source separation could also enable more direct control over listening experiments than is possible with soundscape recordings alone, for example by allowing specific sources to be removed or added from recordings to test their effect on perceptual attributes, or a single source to be evaluated for its psychoacoustic properties in isolation from the rest

***Corresponding author**: s.k.cragg@edu.salford.ac.uk.*

of the mixture. The ability to separate and analyse sources individually should therefore enable deeper insight into how specific sources drive overall perception, in a way that is not achievable through holistic analysis or subjective labelling alone.

The field of sound source separation has seen substantial progress in recent years, particularly with the recent rise of deep learning-based approaches [4]. While much of this work has often focused on fixed-domain separation, recently the task of *universal sound separation* has been proposed [5], which is the task of separating sources from a mixture irrespective of the types of sound present. Given that a soundscape can contain sounds from a wide range of sound classes, universal sound separation is a promising step towards making source separation applicable to soundscape studies. However, current universal sound separation methods may not yet be fully suited to soundscape studies. One reason for this is that much work in this area focuses on generating single channel outputs. The seminal universal sound separation paper examined the separation of single channel mixtures [5] and while more recent work explored universal sound separation using multichannel input [6], [7], the outputs of these methods however, remain single channel. The preservation of the multichannel format in the separated output could be valuable for soundscape analysis. Prior work has suggested that the spatial characteristics of a soundscape may themselves influence its perception [8], and a spatially-preserved output would allow this influence to be investigated at the level of individual sources rather than only the mixture as a whole. Also, a spatial output could be re-rendered for controlled listening experiments, allowing separated sources to be presented or removed while retaining their original spatial character, rather than being reduced to a single channel stripped of its spatial characteristics. An ambisonic-to-ambisonic separation network, AmbiSep [9], has been proposed, but this work targets speech separation. To the best of our knowledge, spatial output preservation has not yet been explored in the context of universal sound separation.

This paper investigates whether spatially-aware universal sound separation is achievable, by making a simple modification to an existing, single-channel separation architecture, Conv-TasNet [10], to accept and output first-order ambisonic (FOA) signals. Rather than designing a new architecture around spatial audio from the outset, this minimal modification is used to establish a baseline understanding of whether spatially-aware separation is easily achievable, before more complex architectural or dataset extensions are considered. This paper presents early-stage results from an ongoing investigation into sound source separation of ambisonic audio mixtures, with the aim of establishing whether the approach could be viable in future soundscape studies.

2. Method

The aim of this study is to separate sound sources of arbitrary classes from a two-source mixture in FOA, while preserving their spatial characteristics. As part of this, the use of different loss functions to train the model is also compared. The work presented here is an early experiment within a larger ongoing research project. This section describes the data used to train and evaluate the model, the network architecture, and the spatial evaluation methodology used to assess it.

2.1. Dataset

The dataset used to train and evaluate the model consists of synthetic FOA mixtures containing two sound sources of arbitrary sound class, along with the FOA representations of the ground truth audio. The source audio was obtained from the FSD50K dataset [11], two sources were selected at random from the dataset to include in each mixture - ensuring that two sources of the same sound class are not present in the same mixture. A random azimuth was chosen for each source from a set of possible azimuths at 10-degree intervals, with no two sources occupying the same azimuth within a mixture. Elevation angle was fixed at 0 degrees. No reverberation was added; a single ground reflection was included for each source. The Python framework Acoular [12] was used to render the audio, with additional modifications made support encoding to FOA. Each data-point was 4 seconds long with a sample rate of 16kHz. 4500 samples were used for training, 500 for development validation and 1876 samples were used for evaluation.

2.2. Model Architecture

The model was based on the Pytorch [13] implementation of ConvTas-Net [10]. The model was modified to operate on FOA audio by altering the input and output channel count. The number of channels in the encoder and decoder blocks were increased from one to four, no changes were made to the separation module. All hyperparameters were left at PyTorch's defaults. The encoder/decoder convolution kernel had a size of 16, with 512 features passed to the mask generator. The mask generator used a convolution kernel size of 3, an input/output feature dimension of 128 for each convolutional block, an internal feature dimension of 512 within each convolutional block, 8 layers per convolutional block, 3 convolutional blocks, and a sigmoid activation function for the mask output.

2.3. Training

The model was trained for up to 200 epochs with a batch size of 5, using the Adam optimiser with an initial learning rate of 0.001. An adaptive learning rate was used, halving the learning rate after 6 epochs with no improvement in validation loss, along with early stopping after 10 epochs without improvement and gradient clipping at a maximum norm of 5.0.

2.3.1. Loss Functions

The model was trained using 3 different loss functions. Signal-to-noise-ratio (SNR), scale-invariant signal-to-

distortion-ratio (SI-SDR) [14] and a multichannel variant of SI-SDR [9]. Both SI-SDR and the multichannel variant calculates an optimal scaling factor between the target and the source prior to the SDR calculation, the key difference, however, is that the multichannel variant calculates a single optimal scaling factor across all channels, keeping the inter-channel ratios intact, whereas the standard SI-SDR in this case calculates a separate scaling factor for each channel. Permutation invariant training (PIT) [15] was used to account for the arbitrary ordering of sources. For each of these functions the mean is taken across all channels and the negative of this value is used as the loss. TorchMetric's [16] implementation of SNR, SI-SDR and PIT are used.

2.4. Evaluation

Separation performance was evaluated using SDR improvement (SDRi) [17] and SI-SDR improvement (SI-SDRi) [14], calculated per source on the evaluation dataset, using the mean across channels.

To assess whether spatial characteristics were preserved through separation, a weighted angular error metric was used to compare the estimated direction of arrival of each separated source against the ground truth audio. This was done by comparing the intensity vectors of the target source and the estimated source at each time-frequency bin (t, f). The active intensity vector was computed as shown in equation (1) [18],

$$\mathbf{I}_{(t,f)} = -\begin{bmatrix} \mathrm{Re}\{W_{(t,f)} X^*_{(t,f)}\} \\ \mathrm{Re}\{W_{(t,f)} Y^*_{(t,f)}\} \\ \mathrm{Re}\{W_{(t,f)} Z^*_{(t,f)}\} \end{bmatrix} \tag{1}$$

where W, X, Y, Z are the corresponding FOA channels. The angular difference between the direction of arrival at each T-F bin was calculated as shown in equations 2 and 3, where d is the direction of arrival.

$$\mathbf{d}_{(t,f)} = -\frac{\mathbf{I}_{(t,f)}}{\|\mathbf{I}_{(t,f)}\|} \tag{2}$$

$$\theta_{(t,f)} = \arccos\left(\mathbf{d}^{\text{estimate}}_{(t,f)} \cdot \mathbf{d}^{\text{target}}_{(t,f)}\right) \tag{3}$$

To give more importance to bins with more energy, each bin was weighted by the energy of the target signal at that bin, calculated as shown in equation (4) [19],

$$E_{(t,f)} = |W^{\text{target}}_{(t,f)}|^2 + \|\mathbf{V}^{\text{target}}_{(t,f)}\|^2 \tag{4}$$

where

$$\mathbf{V}_{(t,f)} = \begin{bmatrix} X_{(t,f)} \\ Y_{(t,f)} \\ Z_{(t,f)} \end{bmatrix} \tag{5}$$

Giving an overall weighted mean angular error for each source.

$$\bar{\theta} = \frac{\sum_{t,f} E_{(t,f)}\, \theta_{(t,f)}}{\sum_{t,f} E_{(t,f)}} \tag{6}$$

Table 1: Mean separation performance and spatial accuracy on the evaluation set (n = 1876), by training loss function. SI-SDRi and SDRi are reported in dB; spatial error is the weighted mean angular error of each time-frequency bin in degrees. Values in parentheses are standard deviations.

Loss Function	SDRi (dB)	SI-SDRi (dB)	Weighted Angular Error (°)
SNR	-0.2 (1.6)	4.4 (4.0)	42.4 (40.5)
SI-SDR	6.7 (6.5)	8.7 (12.9)	75.2 (37.6)
Multichannel SI-SDR	18.4 (7.8)	23.6 (14.5)	3.1 (9.4)

3. Results

Table 1 reports the mean SI-SDRi, SDRi and weighted angular error across the evaluation set for each loss function. The multichannel SI-SDR loss function substantially outperformed the other loss functions evaluated, both in terms of separation performance and in maintaining the spatial characteristics of the target source.

4. Discussion

The results presented here represent a small part of a larger, ongoing research project, and should be interpreted as an initial test of whether multichannel, spatially-aware separation is feasible in simple mixtures before moving on to more complex scenarios. These results show that even a minimal architectural change to a single-channel separation network was sufficient to achieve good separation performance in the FOA domain for two-source mixtures without reverberation.

As highlighted in the results section, the model trained with the multichannel SI-SDR loss function performed substantially better across all metrics. As the multichannel SI-SDR variant uses a single scaling factor which is shared across all the ambisonic channels, this should guide the model towards preserving the inter-channel ratios between the signals, leading to preservation of the separated source direction. The low value observed in the weighted angular error suggests that this shared scaling-factor across the channels does prove successful in preserving the direction of the estimated sources. In contrast, the SI-SDR loss function which has a separate scaling factor for each channel, inherently treats each channel as a separate entity. While this may lead to sufficient separation of individual channels, it will do this with no regard of its relationship with the other channels, leading to a potential distortion of the estimated source's direction. The high amount of angular error, suggests that source direction is not well preserved with this loss function. As the SNR loss function does not include any scaling factor and compares the estimates directly to the targets, in theory it should aim to implicitly preserve the direction of the source. However, the source direction was not well preserved, though this

could be due to poor separation performance in general. It must be pointed out that the angular error metric used is not fully independent of the separation quality, since it is computed by comparing the intensity vectors of the separated source and the target signal. Where the separation performance is poor, the noise from the interfering source is likely to corrupt the estimated intensity vectors producing a high angular error that may reflect the contribution from the interfering source rather than a failure to preserve the direction of the target source. Using the energy of the target source to weight the average, may mitigate this to some extent as the weighting favours bins where the target signal is high; however, bins where the sources overlap will still be exposed to interference from the other source.

While the multichannel SI-SDR loss function's better preservation of source direction could be expected, given that the shared scaling factor is designed to preserve inter-channel relationships, it was less clear why the multichannel SI-SDR loss function also outperformed the SI-SDR loss function on the SI-SDRi metric. Given that this metric is derived from SI-SDR itself, this result was not necessarily anticipated. One possible, though speculative, explanation for the observed behaviour is that by constraining the model to preserve inter-channel relationships (and therefore the spatial characteristics), the multichannel SI-SDR loss function encouraged the model to more readily exploit the spatial cues for separation. A loss function that discourages the model from discarding the spatial cues early in training may therefore support better separation and not just better preservation of the source direction after separation has occurred. Regarding the separation performance in Table 1, it must be noted that the models trained on these loss functions did not show consistent performance across all channels and sources. The reported value is the mean across all channels for each source, there may be differences between the sources or the channels. For the model trained with SI-SDR, there was a stark drop in performance in the third channel for both sources, this channel had a mean SI-SDRi of -0.02 dB across both sources, whereas the other channels across both sources had a mean SI-SDRi of 11.6 dB. This large disparity suggests that the model did not learn to separate this channel. This channel encodes the vertical component of the soundfield. In the dataset, elevation remains constant and the receiver is placed level with the source; because of this, the channel includes only the added ground reflections, leading to less energy in this channel and the source being heard only from below. These factors may have contributed to the poor separation of this channel. As such, in the future, the dataset should be expanded to include elevation differences or to utilise data augmentation techniques such as channel swapping [20] to see if this alleviates this observed issue. For the SNR-trained model, there were differences between the two separated sources. One source showed a consistently near-zero mean SI-SDRi across all channels of -0.03 dB, indicating this source remained effectively unseparated. The second source achieved a higher mean SI-SDRi (8.84 dB), but this is not corroborated by SDRi, which had a mean of -0.42 dB across the channels. From informal listening over a few of the evaluation samples, it appeared as though this second source was often near silent. Given this, it is possible the SI-SDRi value reported may overstate its true subjective separation quality.

It is also worth noting how minimal the architectural change to the network was. No component of Conv-TasNet's separation module was modified to take advantage of the relationships between the separate ambisonic channels. The only change was increasing the number of channels in the encoder and decoder from one to four. Given this, it is notable that the multichannel SI-SDR loss function was still able to achieve good separation performance while preserving the source direction without any architectural mechanism explicitly designed to support this. At the same time, the poor performance of SNR and SI-SDR loss functions should not be taken as evidence that these loss functions are unsuitable for ambisonic source separation; it is possible that an architecture that is explicitly designed to take advantage of the inter-channel relationships could narrow the performance gap between these loss functions.

From a soundscape perspective, preservation of source direction may be particularly important, as spatial characteristics may contribute to perception [8]. The improved directional preservation achieved by the multichannel SI-SDR loss function therefore suggests potential utility not only for source analysis but also for future perceptual listening experiments involving manipulated soundscape recordings.

While these results are encouraging, several limitations should be considered before this approach can be applied to soundscape analysis. The multichannel SI-SDR loss function, while it constrains the scaling factor to be shared across channels, is still invariant to a single scale factor applied globally to all four channels. In practice, this means the levels of the separated outputs are not consistent with those of the sources in the original mixture. This may be a significant issue for further analysis of separated sources if the level of the estimated source cannot be retained, as this could prevent meaningful comparisons between sources for features such as loudness. The dataset used here is also considerably simpler than a typical soundscape recording. Mixtures contained only two stationary sources, with no reverberation and a fixed elevation. Two-source mixtures in particular are of limited relevance to soundscape analysis, where recordings typically contain a larger and more variable number of concurrent sources. A further limitation concerns the temporal overlap of the mixtures: the source audio clips varied in duration and mixture generation did not explicitly enforce temporal overlap between sources. It was informally observed that separation quality was higher for mixtures where the

sources did not overlap in time compared to those where they did. As a result, the presented metrics may slightly overstate performance compared to datasets which consist of fully overlapping sources. The results presented here should be treated as an initial test of feasibility under simplified conditions, rather than as representative of performance on real soundscape recordings.

5. Conclusions and Future Work

This paper investigated whether class-agnostic ambisonic-to-ambisonic source separation was feasible by making a minor change to a single-channel source separation architecture. Among the three loss functions evaluated, the multichannel SI-SDR loss, which applies a shared amplitude scaling factor across all four ambisonic channels, substantially outperformed both standard SI-SDR and SNR, providing a baseline for future development of spatially-aware source separation methods tailored to soundscape assessment and manipulation.

These results establish feasibility on a constrained problem rather than as a soundscape-ready method. The multichannel SI-SDR loss, while effective at preserving direction, remains invariant to the absolute output level of the separated source, and the dataset used contained only two stationary sources with no reverberation - conditions considerably simpler than a typical soundscape recording. Addressing these gaps is the immediate focus of the next stage of this work: extending the dataset to include reverberation, elevation variation, and a greater and more variable number of concurrent sources. The absolute scale issue will also be addressed, either through a loss function that penalises global scale error directly or a post-hoc calibration step. Another direction worth exploring is developing models that separate sources based on sound class or type, rather than separating a fixed number of discrete sources from a recording. This may be more useful in recordings containing a large number of similar sources (for example, a crowd of people speaking or a group of birds singing), where attempting to separate each individual source is likely impractical, and a class-based grouping may be more meaningful for practical analysis purposes.

References

[1] G. Pérez-Martínez, A. J. Torija, and D. P. Ruiz, "Soundscape assessment of a monumental place: A methodology based on the perception of dominant sounds," *Landscape and Urban Planning*, vol. 169, pp. 12–21, 2018.

[2] A. Mitchell et al., "Effects of Soundscape Complexity on Urban Noise Annoyance Ratings: A Large-Scale Online Listening Experiment," *International Journal of Environmental Research and Public Health*, vol. 19, no. 22, p. 14 872, 2022.

[3] M. S. Engel, A. Fiebig, C. Pfaffenbach, and J. Fels, "A Review of the Use of Psychoacoustic Indicators on Soundscape Studies," *Current Pollution Reports*, vol. 7, no. 3, pp. 359–378, 2021.

[4] S. Araki, N. Ito, R. Haeb-Umbach, G. Wichern, Z.-Q. Wang, and Y. Mitsufuji, "30+ Years of Source Separation Research: Achievements and Future Challenges," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India: IEEE, 2025, pp. 1–5.

[5] I. Kavalerov et al., "Universal Sound Separation," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA: IEEE, 2019, pp. 175–179.

[6] D. Lee and J.-W. Choi, "DeFT-Mamba: Universal Multichannel Sound Separation and Polyphonic Audio Classification," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India: IEEE, 2025, pp. 1–5.

[7] D. Wu, X. Wu, and T. Qu, "Leveraging Sound Source Trajectories for Universal Sound Separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2337–2348, 2025.

[8] L. Hermida and I. Pavón, "Spatial aspects in urban soundscapes: Binaural parameters application in the study of soundscapes from Bogotá-Colombia and Brasília-Brazil," *Applied Acoustics*, vol. 145, pp. 420–430, 2019.

[9] A. Herzog, S. R. Chetupalli, and E. A. P. Habets, "AmbiSep: Joint Ambisonic-to-Ambisonic Speech Separation and Noise Reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3081–3094, 2023.

[10] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[11] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[12] E. Sarradj et al., *Acoular – Acoustic testing and source mapping software*, 2025. [Online]. Available: `https://zenodo.org/records/17425756`.

[13] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[14] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK: IEEE, 2019, pp. 626–630.

[15] D. Yu, M. Kolbaek, 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)*, New Orleans, LA: IEEE, 2017, pp. 241–245.

[16] N. Detlefsen et al., "TorchMetrics - Measuring Reproducibility in PyTorch," *Journal of Open Source Software*, vol. 7, no. 70, p. 4101, 2022.

[17] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[18] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "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.

[19] V. Pulkki, "Directional Audio Coding in Spatial Sound Reproduction and Stereo Upmixing," in *AES Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond*, Journal of the Audio Engineering Society, 2006.

[20] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A Four-Stage Data Augmentation Approach to ResNet-Conformer Based Acoustic Modeling for Sound Event Localization and Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.