A study on the relative accuracy and robustness of the convolutional recurrent neural network based approach to binaural sound source localisation

Jago T. Reed-Jones[1], Paul Fergus[1], David L. Ellis[1], and Karl O. Jones[1]

[1]*Liverpool John Moores University, UK*

Correspondence should be addressed to Jago Reed-Jones (j.t.reedjones@2019.ljmu.ac.uk)

ABSTRACT

Binaural sound source localization is the task of finding the location of a sound source using binaural audio as affected by the head-related transfer functions (HRTFs) of a binaural array. The most common approach to this is to train a convolutional neural network directly on the magnitude and phase of the binaural audio. Recurrent layers can then also be introduced to allow for consideration of the temporal context of the binaural data, as to create a convolutional recurrent neural network (CRNN).

This work compares the relative performance of this approach for speech localization on the horizontal plane using four different CRNN models based on different types of recurrent layers; Conv-GRU, Conv-BiGRU, Conv-LSTM, and Conv-BiLSTM, as well as a baseline system of a more conventional CNN with no recurrent layers. These systems were trained and tested on datasets of binaural audio created by convolution of speech samples with BRIRs of 120 rooms, for 50 azimuthal directions. Additive noise created from additional sound sources on the horizontal plane were also added to the signal.

Results show a clear preference for use of CRNN over CNN, with overall localization error and front-back confusion being reduced, with it additionally being seen that such systems are less effected by increasing reverb time and reduced signal to noise ratio. Comparing the recurrent layers also reveals that LSTM based layers see the best overall localisation performance, while layers with bidirectionality are more robust, and so overall finding a preference for Conv-BiLSTM for the task.

1 Introduction

Classical methods of sound source localisation such as generalised cross-correlation [1], beamforming methods [2], and subspace methods such as MUSIC [3] and ESPRIT [4] tend to be improved by increasing numbers of transducers, leading to the dominance of many element microphone array for this task.

In nature, however, most hearing species of animal have evolved to have only two ears, despite the importance of sound localisation ability for survival [5]. It follow then that 2 sensors, if used correctly, must be adequate for a useful sound source localisation system.

Binaural Sound Source Localisation (BSSL) is the task of locating a sound source based upon measurement of the sound field by binaural array; typically a physical full head simulator as used in binaural auralisation, such that the same cues used in human sound localisation are present in the binaural signal.

A successful BSSL system, being one that is accurate and robust in real-world scenarios, has clear applications in robotics [6, 7], hearing aids [8, 9] and beyond. Humans identify the Direction of Arrival (DoA) of a sound source based upon the binaural cues Interaural Time Difference (ITD) and Interaural Level Difference (ILD), being the differences in level and time of arrival of a sound between the ears. This can be used for finding DoA in the frontal horizontal plane, however on the full sphere of directions a given ITD and ILD pair will have a range of possible directions; leading to the cone of confusion.

The filtering inflicted upon a sound source by the pinna and other parts of the head and torso, the Head Related Transfer Function (HRTF), differs somewhat uniquely for different DoAs. These differences provide strong context for resolving the cone of confusion. In addition to this, humans also assess how these cues change in relation to head pose and location changes, as well as other contextual cues, to better further resolve a sound source's location in 3d space.

A primitive BSSL approach is to estimate DoA based only upon ITD, by applying the generalised cross-correlation to estimate ITD, and to map this to an azimuth either trigonemtrically, or ideally through a look-up table of measured or tuned values. Early BSSL systems expanded this to combine this approach also with analysis of ILD [10]. Estimating the HRTF of a binaural signal and correlating this to HRTFs in memory provided an approach to full sphere binaural sound source localisation [11].

Most interest in this task has come since the popularisation of machine learning. Earlier approaches consist of creating hand-crafted features representing ITD and ILD and perhaps monaural cues, and using this to train at first probablistic models [12, 13] and then deep neural networks [14, 15, 16, 17].

Following this, a considerable shift towards application of Convolutional Neural Networks (CNNs) occurred [18, 19, 20, 21, 22], wherein the network is trained on a more raw form of the binaural audio, typically vectors or matrices representing the magnitude and phase of the signals.

Concurrently in the field Sound Source Localisation by microphone array, introduction of recurrence into CNN architectures were becoming the popular approach [23], in models known as Convolution Recurrent Neural Networks (CRNNs). Their use is shown to improve performance over CNN for BSSL [24], and this approach has also begun to be used in the field of BSSL [25, 26].

Based upon the continued popularity of CRNN in the wider field of sound source localisation, it is supposed this trend will continue. In aid of this, this work examines the use of recurrence, comparing possible choices of recurrent layer which can be used in systems, assessing the performance based both on overall accuracy, but also robustness to adverse acoustic conditions. To the test the relative performance of different recurrent layers a comparative analysis of the performance of four different CRNN architectures is made, with the models containg four different types of recurrent layer:

- Gated Recurrent Unit (GRU)
- Bidirectional Gated Recurrent Unit (BiGRU)
- Long Short Term Memory (LSTM)
- Bidirectional Long Short Term Memory (BiLSTM)

To create models referred to as Conv-GRU, Conv-BiGRU, Conv-LSTM and Conv-BiLSTM. These are chosen due to their general popularity in recurrent neural networks. With the field of BSSL, BiGRU has previously seen application [25, 26] in CRNN. Beyond this, LSTM has been used for DoA estimation [27] but not alongside convolutional layers, and BiLSTM and GRU have both been used in a CRNN type configuration, but not directly for DoA estimation, but for binaural signal enhancement within a larger BSSL system [28, 29].

These four models are compared also to a baseline, being an equivalent system with no recurrent layer, so as to represent a CNN. Alongside the other modes, this is simply referred to as Conv.

All of these systems are trained upon the the same training datasets using an identical optimiser, and evaluated with the same testing datasets. These models and datasets are described fully in the following chapters.

2 Models

To minimise possible differences occurring to randomness between the models, A CRNN architecture is used where the convolutional layers and the recurrent layers are trained separately, so that the convolutional layers can be trained identically across all systems.

This is done by training two CNNs upon the training dataset of magnitude and phase representations, and then the CNNs used to classify the training dataset but it is the activations at the final convolutional layers that are saved instead of the output of the system. This creates a new dataset of activations, upon which the five systems representing four recurrent layers alongside the baseline are trained.

The CNN consists of only three convolution layers, filters of consistent size throughout the system, as seen in Table 1. It is at the final Relu activation function that the activations are extracted.

These activations are then used to train four shallow RNNs consisting only of an Input layer, a recurrent layer, and a dense layer, and a softmax layer. The recurrent layers, in all cases, have 200 units. This architecture is described in Table 2. The baseline, meanwhile, classifies the activations by use of a Multilayer Perceptron (MLP), as seen in Table 3. It is possible to see that the MLP described in Table 3 is exactly equivalent to the final layers of the CNN in Table 1, hence why this is deemed representative of a CNN despite an unusual abstraction in the training process.

Layer Type	Parameters
Input Layer	
2D Convolution	([6,6], 18)
Batch Normalisation	
Relu	
Max Pooling	(2,2)
2D Convolution	([6,6], 18)
Batch Normalisation	
Relu	
Max Pooling	(2,2)
2D Convolution	([6,6], 18)
Batch Normalisation	
Relu	
Dense	[50]
Softmax	

Table 1: CNN to be adapted into Convolutional Layers

Layer Type	Parameters
Input Layer	
Recurrent Layer	200 Units
Dense	50
Softmax	

Table 2: Generic RNN architecture

Layer Type	Parameters
Input Layer	
Dense	50
Softmax	

Table 3: MLP architecture

3 Binaural Dataset

Binaural datasets are synthesised through binaural auralisation. For these, we create a signal model which considers the effects of reverberation and additive noise, such that for a single sound source $x(t)$ the received pressure is described as:

$$p_{L,R} = x(t) * \text{brir}_{L,R}(t,\varphi) + \eta(t) \quad (1)$$

where p is the sound pressure at the ears, L,R refer to left and right channels, brir is the binaural room impulse response, φ is azimuth and η is additive noise. Almost all previous literature on BSSL focuses on the localisation of human speech. Due to this, for $x(t)$ we use speech utterances taken from the TSP corpus [30]. This corpus is chosen due to it's 48kHz sample rate, higher than that of other commonly used speech corpi. BRIRs can be further modelled as made up of of a head related impulse response (HRIR) and the reflections making up the room impulse response

$$\text{brir}(t,\varphi) = \text{hrir}(t,\varphi) + \sum_{n=1}^{N} A_n hrir(t-\tau_n, \varphi_n, \theta_n) \quad (2)$$

where N is the number of reflections, A_n represents the gain of the reflection, τ represents the time delay of the reflection, and φ_n and θ_n represent the azimuth and elevation of the angle of arrival of the reflection.

To create the BRIRs, HRIRs are taken from the CIPIC dataset [31] for the KEMAR mannequin subject. The CIPIC dataset contains HRIRs for fifty doa on the horizontal plane. Sources at these fifty positions are considered for this test, encompassing the full plane.

The reflection parameters are created for shoebox rooms by image source method (ISM) [32]. This is done for 90 rooms in the training dataset, another 10 rooms in the validation dataset, and 30 rooms in the testing dataset.

All of these rooms have randomised dimensions between 1-10 metres. They then have their absorption coefficients controlled such as to achieve target RT60 values. For the training and validation datasets, these target values are randomised between 0.5-1.5 seconds. For the testing dataset, 10 room dimensions are randomised, and then for each of these randomised rooms the absorption coefficients are generated to make target reverb times of [0.5, 1, 1.5] seconds. This is done to create statistical significance at these sampled reverb times for plotting.

The additive noise is modelled as being the sum of

some interefering sound sources, that is sound sources different to the speaker, but which are also binauralised.

$$\eta(t) = A_0 \sum_{k=1}^{K} A_k(n(t) * \text{brir}_{L,R}(t, \varphi_k)) \qquad (3)$$

where A_0 is an overall gain of the noise mixture, A_k is relative gain of each sound source, K is the number interfering noise sources.
The number interfering sources is determined randomly in the range $1 \leq 10$. The relative level A_k, and azimuth φ_k each interfering sound source are randomly generated. The overall gain A_0 is randomly generated in the training dataset but such that a signal-to-noise ratio (SNR) in range -12dB to -36dB is achieved. For the testing dataset, for every sound source at every room three signal to noise ratios $[-12, -24, -36]dB$ are targeted.
For the testing dataset, 1000 1 second utterances are convolved with BRIRs of all 50 source directions, with noise added, to create 50,000 training files. For the validation dataset, another 100 files were used to create 5,000 binauralised files.
For the testing dataset, another 100 one second utterances were combined with all 9 combinations of RT60 and SNR, for all fifty source directions.

4 Audio to Network

The one second audio files are treated as sequences to classify, each of which will have one DoA estimate created, but are broken into multiple steps in a sequnce. This is done by splitting the 1 second file into 10 100mS files.
The models use 2D convolutional layers, and so a two dimensional form is required; this is achieved differently for the two feature representations. Firstly, a matrix representing interaural phase difference (IPD) is found by applying STFT to the signal, and finding the difference between the two resulting phase matrices. The STFT was done with a hanning window of 425 samples with an overlap of 256 samples, with the end result being a matrix $[147, 19]$
A matrix representing magnitude of the spectrum is found through gammatone descomposition by gammatone filterbank with 147 bands. The resulting frequency banded signals are equivalently windowed, and the average energy in each window is calculated resulting in a matrix of the size $[147, 19, 2]$. The logarithm of this matrix is then taking.

5 Training

The two CNNs, four RNNs and the MLP making up the CRNN and CNN systems are all trained using an Adam optimizer with equivalent training parameters, of an initial training rate of 1e-4, a decay rate of 0.9. The CNNs were trained over 100 epochs, while the RNNs and MLP were trained over 200 epochs.
The validation loss is monitored for signs of overfit, however this was not encountered in any case so early stopping was not required.

6 Evaluation

Performance is evaluted by four metrics:

Classification Rate The rate of succesful classifications, such that $Y_{pred} = Y_{test}$

Front-Back Confusion Rate The rate at which the system falsely classifies within $\pm 10^\circ$ of the front-back reversal of Y_{test}

Root Mean Square Localisation Error (RMSLE) The root mean square, but algorithmically accounting angle wrap-around

RMSLE (Mirrored) The RMSLE, but with both Y_{pred} and Y_{test} such that they are on the frontal horizontal plane.

These metrics are shown averaged over the entire testing dataset in Table 4, while the the systems are shown plotted against RT60 and SNR in figures 1 and 2 respectively.

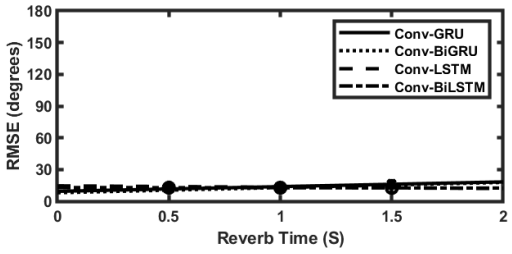

Fig. 1: RMSLE with respect to RT60

	Classification Rate	Front-Back Confusion Rate	RMSLE	RMSLE (Mirrored)
Conv	76.49%	1.7%	22.33°	9.55°
Conv-GRU	79.96%	0.8%	13.96°	6.3°
Conv-BiGRU	80.03%	0.68%	13.02°	6.44°
Conv-LSTM	80.6%	0.62%	13.39°	6.03°
Conv-BiLSTM	81.11%	0%	12.8°	6.65°

Table 4: Accuracy, Front-back Confusion, RMSLE and RMSLE (Mirrored) of all four CRNN architectures

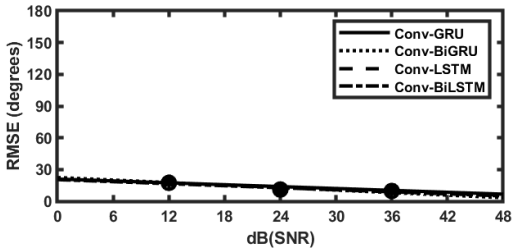

Fig. 2: RMSLE with respect to SNR

7 Discussion

The results shown in Table 4 corroborate the previous finding [24] that use of CRNN leads to an increase in accuracy over CNN. Here, however, it is found that this is advantage is true of a range of recurrent layers. The difference between different layers does not lead to substantial performance changes, with the best and worst performing only being seperated by 1°. Based upon the results seen, however, it is concluded that BiLSTM is the most preferable choice of recurrent layer for performing BSSL with CRNN.

In terms of robustness, the difference is also small, however it is also Conv-BiLSTM which seems most resilient to increases in reveberation, further supporting the case for the of BiLSTM layers.

While performance certainly suffered when looking at the full horizontal plane, rather than just the frontal, none of the systems had large difficulty with front-back reversals. However, it is notable that the rate of front-back reversal is exceptionally low for Conv-BiLSTM, with the rate rounding to 0%.

8 Conclusion

This work has performed a comparative analysis of four CRNN architectures containing different types of recurrent layer, by training the four systems with identical optimisers upon identical datasets, with identical parameters in the convolution layers for the four systems The four systems, and a baseline, were evaluated upon a testing dataset of binaural audio with controlled levels of additive noise and reverberation.

Based upon this testing, it is recommended not just that CRNN should be favoured over CNN for BSSL, but specifically that BiLSTM is the optimal choice of recurrent layer for the application.

References

[1] Knapp, C. and Carter, G., "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4), pp. 320–327, 1976, doi: 10.1109/TASSP.1976.1162830.

[2] DiBiase, J. H., Silverman, H. F., and Brandstein, M. S., "Robust localization in reverberant rooms," in *Microphone arrays: signal processing techniques and applications*, pp. 157–180, Springer, 2001.

[3] Schmidt, R., "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, 34(3), pp. 276–280, 1986, doi:10.1109/TAP.1986.1143830.

[4] Roy, R. and Kailath, T., "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on acoustics, speech, and signal processing*, 37(7), pp. 984–995, 1989.

[5] Schnupp, J. and Carr, C., "On hearing with more than one ear: lessons from evolution," *Nature Neuroscience*, 12(6), pp. 692–697, 2009, doi:10.1038/nn.2325.

[6] Nguyen, Q., Girin, L., Bailly, G., Elisei, F., and Nguyen, D.-C., "Autonomous sensorimotor learning for sound source localization by a humanoid

robot," in *IROS 2018-Workshop on Crossmodal Learning for Intelligent Robotics in conjunction with IEEE/RSJ IROS*, 2018.

[7] Dávila-Chacón, J., Liu, J., and Wermter, S., "Enhanced robot speech recognition using biomimetic binaural sound source localization," *IEEE transactions on neural networks and learning systems*, 30(1), pp. 138–150, 2018.

[8] Froehlich, M., Freels, K., and Powers, T. A., "Speech recognition benefit obtained from binaural beamforming hearing aids: Comparison to omnidirectional and individuals with normal hearing," *Audiology Online*, 14338, pp. 1–8, 2015.

[9] Goli, P. and van de Par, S., "Deep Learning-Based Speech Specific Source Localization by Using Binaural and Monaural Microphone Arrays in Hearing Aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, pp. 1652–1666, 2023, doi:10.1109/TASLP.2023.3268734.

[10] Macpherson, E. A., "A computer model of binaural localization for stereo imaging measurement," *Journal of the Audio Engineering Society*, 39(9), pp. 604–622, 1991.

[11] Keyrouz, F., "Advanced Binaural Sound Localization in 3-D for Humanoid Robots," *IEEE Transactions on Instrumentation and Measurement*, 63(9), pp. 2098–2107, 2014, doi:10.1109/TIM.2014.2308051.

[12] May, T., van de Par, S., and Kohlrausch, A., "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1), pp. 1–13, 2011, doi:10.1109/TASL.2010.2042128.

[13] May, T., Ma, N., and Brown, G. J., "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2679–2683, 2015, doi:10.1109/ICASSP.2015.7178457.

[14] Ma, N., May, T., and Brown, G. J., "Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12), pp. 2444–2453, 2017, doi:10.1109/TASLP.2017.2750760.

[15] Ma, N., Gonzalez, J. A., and Brown, G. J., "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11), pp. 2122–2131, 2018.

[16] Lovedee-Turner, M. and Murphy, D., "Application of Machine Learning for the Spatial Analysis of Binaural Room Impulse Responses," *Applied Sciences*, 8(1), 2018, ISSN 2076-3417, doi:10.3390/app8010105.

[17] O'Dwyer, H., Bates, E., and Boland, F. M., "A Machine Learning Approach to Detecting Sound-Source Elevation in Adverse Environments," in *Audio Engineering Society Convention 144*, 2018.

[18] Vecchiotti, P., Ma, N., Squartini, S., and Brown, G. J., "End-to-end binaural sound localisation from the raw waveform," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 451–455, IEEE, 2019.

[19] Xu, Y., Afshar, S., Singh, R. K., Wang, R., van Schaik, A., and Hamilton, T. J., "A Binaural Sound Localization System using Deep Convolutional Neural Networks," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2019, doi:10.1109/ISCAS.2019.8702345.

[20] Zhou, L., Ma, K., Wang, L., Chen, Y., and Tang, Y., "Binaural Sound Source Localization Based on Convolutional Neural Network." *Computers, Materials & Continua*, 60(2), 2019.

[21] Jiang, S., Wu, L., Yuan, P., Sun, Y., and Liu, H., "Deep and CNN fusion method for binaural sound source localisation," *The Journal of Engineering*, 2020(13), pp. 511–516, 2020, doi:https://doi.org/10.1049/joe.2019.1207.

[22] Reed-Jones, J. T., Jones, K. O., Fergus, P., Marsland, J., and Ellis, D. L., "Comparison of Performance in Binaural Sound Source Localisation

using Convolutional Neural Networks for differing Feature Representations," in *Audio Engineering Society Convention 154*, Audio Engineering Society, 2023.

[23] Adavanne, S., Politis, A., and Virtanen, T., "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1462–1466, IEEE, 2018.

[24] Reed-Jones, J. T., Fergus, P., Ellis, D. L., Marsland, J., and Jones, K. O., "Improving Full Horizontal Plane Binaural Sound Localization by use of BiLSTM," in *2024 International Conference on Information Technologies (InfoTech)*, IEEE, 2024.

[25] García-Barrios, G., Krause, D. A., Politis, A., Mesaros, A., Gutiérrez-Arriola, J. M., and Fraile, R., "Binaural source localization using deep learning and head rotation information," in *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 36–40, IEEE, 2022.

[26] Krause, D. A., García-Barrios, G., Politis, A., and Mesaros, A., "Binaural Sound Source Distance Estimation and Localization for a Moving Listener," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, pp. 996–1011, 2024, doi:10.1109/TASLP.2023.3346297.

[27] Massicotte, P., Chaoui, H., and Ouameur, M. A., "LSTM with Scattering Decomposition-Based Feature Extraction for Binaural Sound Source Localization," in *2022 20th IEEE Interregional NEWCAS Conference (NEWCAS)*, pp. 436–440, 2022, doi:10.1109/NEWCAS52662.2022.9841963.

[28] Yang, B., Li, X., and Liu, H., "Supervised Direct-Path Relative Transfer Function Learning for Binaural Sound Source Localization," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 825–829, 2021, doi:10.1109/ICASSP39728.2021.9413923.

[29] Yang, B., Liu, H., and Li, X., "Learning Deep Direct-Path Relative Transfer Function for Binaural Sound Source Localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, pp. 3491–3503, 2021, doi:10.1109/TASLP.2021.3120641.

[30] Kabal, P., "TSP speech database," *McGill University, Database Version*, 1(0), pp. 09–02, 2002.

[31] Algazi, V., Duda, R., Thompson, D., and Avendano, C., "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pp. 99–102, 2001, doi:10.1109/ASPAA.2001.969552.

[32] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 65(4), pp. 943–950, 1979.