Early vs Late Fusion in Binaural Sound Source Localisation using CNN

Jago T. Reed-Jones[a], John Marsland[a], David L. Ellis[a], Paul Fergus[b], Karl O. Jones[a]

[a]School of Engineering, Liverpool John Moores University
3 Byrom Street, Liverpool, United Kingdom, L3 3AF
j.t.reedjones@2019.ljmu.ac.uk

[b]School of Computer Science and Mathematics, Liverpool John Moores University
3 Byrom Street, Liverpool, United Kingdom, L3 3AF

Abstract— **In Binaural Sound Source Localisation there are two representations of the signals which contain useful cues for localisation: the time/phase frequency spectrum and the magnitude frequency spectrum. This typically leads to two branch CNN architectures being employed achieve localisation.**

This paper compares the difference in performance between models which employ early and later fusion of these two branches, finding only negligible differences and thus concluding that this is an unimportant consideration in the design of such systems.

Keywords— **Binaural Sound Localization, Sound Source Localization, Convolutional Neural Networks, Audio Signal Processing, Machine Learning**

I. Introduction

Binaural Sound Source Localisation (BSSL) is the task of estimating of the Direction of Arrival of Sound Source using recordings of a sound field made with a binaural array.

This approach differs from traditional methods of Sound Source Localisation (SSL) in that a binaural array contains only two sensors, as opposed to the large arrays of sensors used in other methods.

This can be achieved through means of Binaural Cues: the Interaural Level Difference (ILD) and the Interaural Time Difference (ITD). Only using Binaural cues, however, is not adequate for localisation in the full azimuthal range, as there are two solutions for a given ILD & ITD: a position in front of the head, and the mirror position behind the head. This ambiguity can be resolved through analysis of the frequency response, as at different source positions the filtering of the signal of the head is unique. This is the head related transfer function (HRTF).

While only some works have dealt with localising in the full azimuthal range [1], a common approach for this task is utilising Convolutional Neural Networks (CNNs) [1-4]. CNNs are ideal for this task as they are capable of taking frequency domain representations of the audio signal and extracting relevant features.

Typically this will involve some combination of representations of the magnitude differences and phase or time differences of the sound arriving at the ears, leading to two branch architectures.

This work will look at the effect changing the point of fusion of such a model has on the localisation performance, the point of fusion being at which point the two branches are concatenated into a single branch.

To do such, four CNN models of differing points of fusion are trained and tested on identical datasets of magnitude and time-delay representations of sound.

II. Training & Testing Datasets

A. Audio Datasets

Audio datasets for training and testing were created from which the Time-Frequency (TF) matrices could be created. For such, speech samples were taken from the Librispeech corpus [5], a collection of English language spoken media. For the training dataset, ten 100mS samples were taken from 200 different files in the corpus, making a total of 2000 unique speech samples. For testing, one 100mS sample was taken from 100 different files.

These speech samples were then convolved with Binaural Room Impulse Responses (BRIRs) of ten different rooms and fifty different source directions. The source directions were all on the azimuthal plane, being the source directions available in the CIPIC HRTF dataset [6], from which the HRTFs of a KEMAR mannequin were used to create the BRIRs. These source directions can be seen in Fig. 1

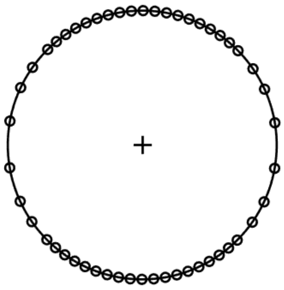

Fig. 1 Source Directions on Azimuthal Plane used in datasets

BRIRs are the impulse responses at the ears for a given source direction, but in a diffuse field rather than the anechoic condition of measured Head Related Impulse Reponses (HRIRs).

BRIRs were simulated using the image source method, for a target reverb times of $T_R = \{0.5,1,1,5\}$seconds where T_R is the time taken for the impulse response to attenuate by 60dB. This was done by randomly generating room dimensions for a rectilinear room with boundaries between 1-10m, and then altering the absorption coefficients of the boundaries to achieve the target reverb times according to the Sabine equation:

$$T_R = \frac{0.161V}{S\alpha}$$

where V is the room volume, S is the surface area, and α is the absorption coefficient.

Creation of BRIRs for three rooms, however, is likely to lead to severe overfit and so multiple room dimensions were created. Five sets of room dimension for three target reverb times were used for training, for a total of 15 unique rooms. Additionally, another five room dimensions were used for the testing dataset leading to another 15 unique rooms.

The training and testing datasets were then convolved with BRIRs and HRIRs, evenly distributed according to the reverb times, as according to Table I.

TABLE I
DISTRIBUTION OF FILES ACROSS REVERB TIMES

T_R	0s	0.5s	1s	1.5s
%	25	25	25	25

This leads to 25% of the files representing the anechoic condition, and 75% of the files representing diffuse fields, with each room being used for 5% of the total number of files.

The other acoustic condition the system is trained and tested under is the addition of noise. This was done by creating noise mixtures which consisted of noise sources convolved with HRIRs and BRIRs matching the room used for the speech sample. For the training dataset, the noise source was pink noise, and for the testing dataset it was a recording of background room noise. A random number of noise sources between 1-10 were used for each audio file, and for each noise source a random azimuth was chosen. The entire noise mixture was then normalised as to achieve target signal-to-noise (SNR) ratios of $dB(SNR) = \{0, 12, 24, 36\}$. This noise mixture was summed to the speech source.

B. *Magnitude Matrices Dataset*

The first branch of the CNN would interpret a magnitude TF-matrix created from the audio dataset. This was created by decomposing the audio into frequency bands using a gammatone filterbank. The filterbank contained 300 filters distributed between 100Hz and 8kHz. Upon decomposition, the resulting band limited signals were then windowed using a hamming window with a length of 465 samples and an overlap of 256 samples. This lead to a 6 windows, from which the average level of energy was taken.

Upon applying this to both left and right channels, the result is a matrix of the size $[300,6,2]$. The values in this matrix were then scaled into deciBels.

C. *Time Delay Matrices Dataset*

To create a matrix of values relating to time-delay, the same bank of 300 band limited signals from gammatone decomposition were used.

For each of these stereo signals, a cross correlation curve was calculated using generalised cross-correlation phase transform (GCC-Phat) algorithm [7,8]. These correlation curves were then truncated to represent the section of the curve relevant to the time delays which can be encountered between the ears, being the central-most 11 samples.

Under perfect conditions, this matrix would look like one vertical line of high values representing the correct time delay, however under reverberant conditions this can be reduced, and so a trained CNN is useful as it can learn to discard useless information in the curves based on the information at other frequency bands.

III. MODELS

To assess the effect fusion has on results, four CNNs were created. Each of these had two input layers to take in magnitude and time-delay matrices, and processed these through the same layers but the point at which the branches were concatenated was change in each instance.

The layers which were present in all models can be found in Table II, and the way in which these were combined can be found in Fig 2.

Notably Model I slightly differs from the other three. In order to test the effect having concatenating after the dense layers has, the layers are instead summed into each other, and then another dense layer is found after the fusion.

TABLE III
LAYERS FOUND IN ALL CNN MODELS

Layer 1	
Convolution Layer Batch Normalisation ReLu Max Pooling	([2,2], 8) (2,2)
Layer 2	
Convolution Layer Batch Normalisation ReLu Max Pooling	([8,8], 16) (2,2)
Layer 3	
Convolution Layer Batch Normalisation ReLu Max Pooling	([16,16], 32) (2,2)
Output Layer	
Dense Softmax	50

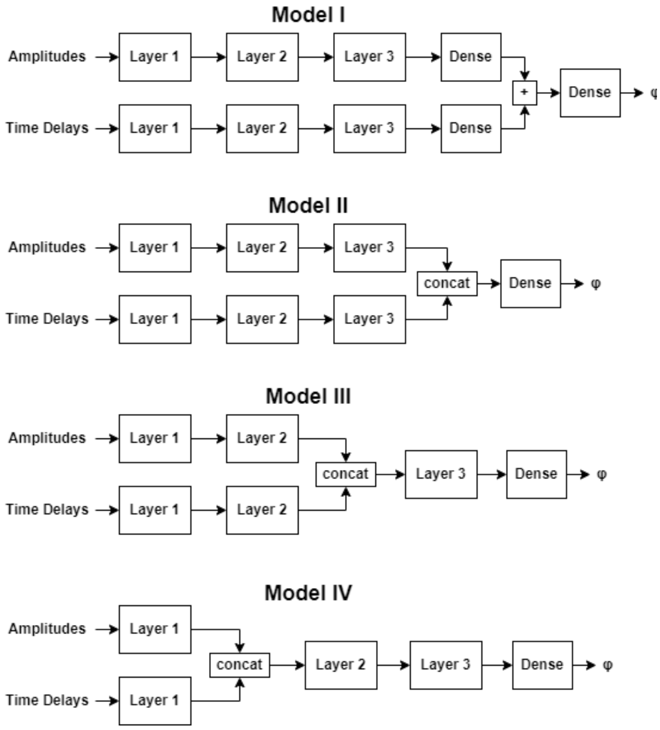

Fig. 2 Framework for all four Models

The models were trained using an Adam optimiser, with a learning rate of 0.001, and a mini batch size of 16. The models were all trained for a period of 200 epochs.

IV. Results

Results are presented in terms of three metrics: Classification accuracy, which is simply the rate at which the network correctly classifies so that predicted azimuth = true azimuth. Root Mean Square Error (RMSE), which is calculated from the difference between predicted azimuth and true azimuth, and finally the Front-Back Confusion Rate, which is the rate at which the network predicts azimuth to be in the front-back mirror position from the true azimuth within a tolerance of ±10°, except for cases where the true azimuth is within ±10° of its mirror position.

TABLE IIIII
Localisation performance metrics for all four models

	Model I	**Model II**	**Model III**	**Model IV**
Classification Accuracy	40.4%	45.8%	46.1%	42.7%
RMSE	64.66°	53.96°	55.2°	54.79°
Front-Back Confusion Rate	4.66%	1.59%	1.79%	1.92%

Additionally, performance was recorded with respect to changing reverb time and SNR in the testing datasets. The RMSE of these is plotted against these two variables in Figs 3 & 4.

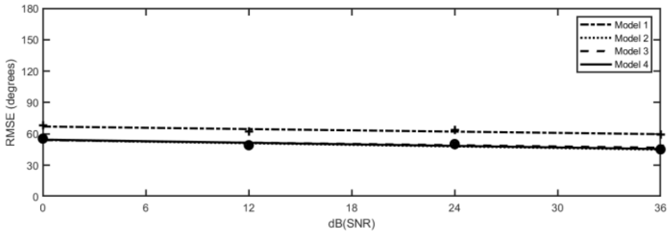

Fig. 3 RMSE with respect to SNR

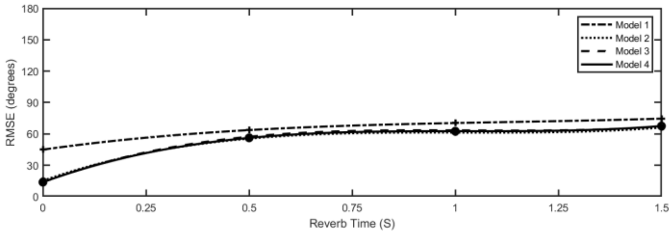

Fig. 4 RMSE with respect to Reverb Time

V. Conclusions

From the results in Table III and Figs 3 & 4, it can be seen that Models II, III and IV perform almost identically in all metrics. This strongly suggests that point of fusion is not a large concern in for the task of BSSL with CNN.

The slightly differing results seen in Model I likely manifest due to the previously mentioned differences in this model to the others. It is likely that the flow of operations now including an extra dense layer has altered the performance, possibly causing a higher degree of overfit.

The level of performance seen in all models is not high, this is likely due to the model heavily overfitting to the BRIRs known to the training set, an idea supported by the much better performance seen when $T_R = 0$, as the problem of generalization between rooms does not exist in the anechoic condition, and so the HRIRs of the testing dataset are the same as the training dataset's.

Given these results, point of fusion is not deemed to be a significant factor in the design of CNNs for this task, and preference is given to early fusion as this can reduce the number of operations required in training and running of the model.

Acknowledgment

This research was supported by funding from the Faculty of Engineering and Technology, Liverpool John Moores University

References

[1] Y. Yang, J. Xi, W. Zhang and L. Zhang, "Full-Sphere Binaural Sound Source Localization Using Multi-task Neural Network," 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 2020, pp. 432-436.

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

[3] C. Pang, H. Liu and X. Li, "Multitask Learning of Time-Frequency CNN for Sound Source Localization," in IEEE Access, vol. 7, pp. 40725-40737, 2019, doi: 10.1109/ACCESS.2019.2905617.

[4] P. Vecchiotti, N. Ma, S. Squartini and G. J. Brown, "End-to-end Binaural Sound Localisation from the Raw Waveform," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 451-455, doi: 10.1109/ICASSP.2019.8683732.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210, 2015, doi:10.1109/ICASSP.2015.7178964.

[6] V. Algazi, R. Duda, D. Thompson, and C. Avendano, “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.

[7] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay”, IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-24(4), pp. 320-327, 1976.

[8] M. S. Brandstein and H. F. Silverman, “A robust method for speech signal time-delay estimation in reverberant rooms”, in Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Munich, Germany, 1997