# The Effect of Noise Reduction upon Voiceprint Integrity

Oli Harrisson[a,b], Jago T Reed-Jones[a], Kay Morrison[a,b], Colin Robinson[a], Karl Jones[a]

[a] *Applied Forensic Technology Group, School of Engineering, Liverpool John Moores University*
*3 Byrom Street, Liverpool, United Kingdom, L3 3AF*

[b] *Merseyside Police*
*15 Cazneau Street, Liverpool, United Kingdom, L3 3AN*
*oliver.harrisson@merseyside.police.uk*

***Abstract*— Audio evidence is often full of noise and it may be advantageous to apply noise-reduction in order to discern dialogue or other sounds; however, this risks damaging any audio cues used in voiceprint analysis. This paper seeks to assess the impact noise-reduction systems have through the application of noise reduction to a small sample, and analysis of a contained voiceprint.**

**It was found that in adverse conditions and / or by application of extreme parameters, enough damage to the cues can be done to make voiceprint analysis ill-advised, and this technique should be reserved for recordings with favourable signal to noise ratios.**

***Keywords*— Voiceprint, Noise Reduction, Audio Forensics, Multimedia Forensics, Audio Signal Processing**

## I. Introduction

A common form of multimedia evidence is audio recordings of speech which causes the need to identify the speaker(s). This can be achieved through auditory analysis, or through visual analysis of a time-frequency representation of the audio.

Common phonemes are compared between audio recordings to identify unique features [1], or the voiceprint, analogous to a fingerprint.

Audio used for evidence, however, is often recorded under adverse environmental and technical conditions and contains a high amount of noise. In these scenarios, noise reduction could be applied to improve the speech to noise ratio.

Common noise reduction techniques rely upon spectral subtraction, where an estimate of the noise- time frequency composition is made and subtracted from the original signal. This, however, could inadvertently remove part or all of specific cues used by forensic practitioners to identify a voiceprint.

This paper seeks to identify the degree to which this destruction takes place and make a recommendation on whether voiceprint analysis may be carried out on recordings post-noise reduction.

## II. Method

Recordings were made of a spoken sentence, and ambient noises. The spoken sentence consists of "I should tell you that the room is a bit smaller than you imagined, but that does make it cheaper". The words "You" and "room" are sounds formed by the resonances within the nasal cavity and are the hardest to alter at free will. Words such as "little" and "this" are principally formed by the articulators in the oral cavity and the words "should" and "cheap" are the fricatives: sounds issued by the restriction of air passing through the articulators [6].

The noise profiles consisted of recordings of two settings as inspired by [2,3]. The first noise profile is of mechanical (motor) noise which would form a consistent noise profile yet would still contain frequencies which occupy the same spectral space as the voice. It is also recorded outdoors adding random noise bursts in places due to weather conditions. This profile was inspired by the recording featured by Fraser [3] who used a speech sample captured alongside a continuous (amplified) noise profile over a telephone line.

The second noise is the ambience of a coffee shop with additional voices occupying the same spectral space. This was chosen as an analogy to the adverse recording conditions which can be found in police interview rooms. All recordings were created using the same equipment and settings, and stored as 24bit 96kHz PCM Audio. Real world audio sources may vary greatly in standard and recording condition [2].

Each noise profile was scaled to a series of different amplitudes at 6dB increments and summed with the speech to create a total of 12 samples.

Each sample was opened in an audio editing software. A spectral profile of the noise (obtained at a point within the file before any speech signal) was used with the Spectral Denoise 'Learn' function which takes an average spectral profile of the noise to be removed from the recording. The 'de-noise' curve was altered so that the new output noise had a 'flat' spectral response after reduction (as opposed to higher energy within specific frequency band(s)) before subtraction is applied. As suggested by [4], each file was subjected to repeat "iteratives"

of the spectral noise reduction process with the curve realigned to ensure a flat response, until the noise was eliminated whilst the voiceprint is maintained.

Spectrograms from before and after the spectral subtraction, as well as notes on spectral subtraction processes to allow for reproducibility were obtained. Figure 1 shows an example of these spectrograms for one sample, in which degradation of voice characteristics can be seen.

The files could then be visually examined and assessment notes were made similar to those seen in Figure 2. In addition, amplitude measurements of the harmonics on keywords described by [1] were recorded, an example of which is shown in Figure 3.

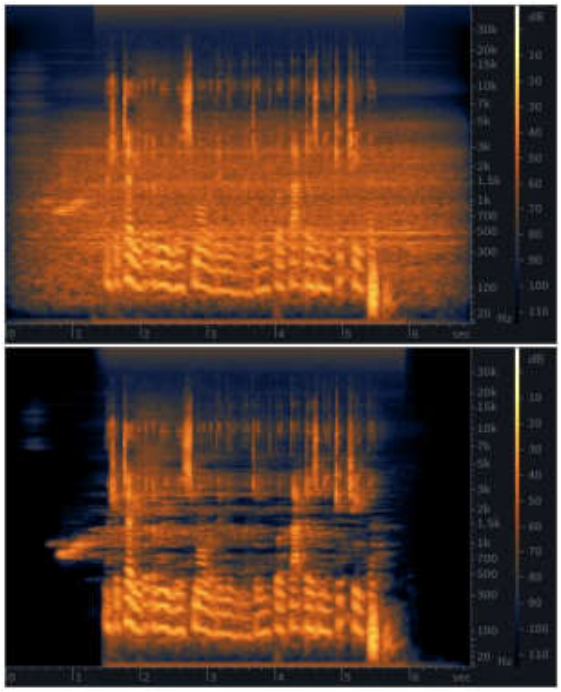

Fig. 1 Spectrogram from before (top) and after (bottom) application of noise reduction from iZotope RX 7

| Spectral Measurements | |
|---|---|
| Initial Inspection Notes | Fundamental frequency and the first four harmonics remain intact, some of the upper harmonics between 500Hz and 2kHz have been degraded by the noise reduction. Some unique vowel-consonant transition harmonics have been lost.<br><br>Still able to compare favourably with the original sample.<br><br>Speech is very much audible and intelligible. |
| Fundamental Frequency Range | 88Hz |
| Notable points of degradation to overall sound. | Between 500Hz and 2kHz, partial degradation. E.g. at the word "room", the 5th, 6th, 7th and 8th harmonics have been masked by the (reduced) noise profile. |
| (New) Noise Profile Details | Peak = -55dBFS, RMS = -64dBFS |

Fig. 2 Example notes from Visual Inspection

| Frequency / Harmonic Amplitude Relationship Measurements | | |
|---|---|---|
| Contour ID | Amplitude | Point / Word of Measurement |
| Fundamental Frequency | -18.1dBFS | "tell" |
| Harmonic 1 | -19.0dBFS | "tell" |
| Harmonic 2 | -18.1dBFS | "tell" |
| Harmonic 3 | -17.6dBFS | "tell" |
| Harmonic 4 | -29.1dBFS | "tell" |
| Harmonic 5 | -33.8dBFS | "tell" |
| Harmonic 6 | -35.4dBFS | "tell" |
| Harmonic 7 | -34.3dBFS | "tell" |
| Harmonic 8 | -36.4dBFS | "tell" |

Fig. 3 Example recording of harmonic measurements

## III. Findings and Discussion

During visual examination of the recordings, it was found that the lower harmonics and fundamental frequencies of the speech were preserved in nearly all recordings whereas the upper harmonics including those in the most prominent vowel sounds and key words were masked by the noise profiles; normally once the noise profile had reached a peak of -18dBFS and above. This was especially true of the coffee shop ambience; the various harmonic contours from different voices masked those of the target voice at nearly all amplitudes.

The mechanical noise was a much more sustained sound though it contained frequencies whose amplitude (at -18dBFS and above) was equal to, or sufficient enough to mask the contours of the lower harmonics. At the highest amplitude, the mechanical noise was able to completely mask the upper harmonics despite their shift in pitch and amplitude.

Both noise profiles had significant energy in the lower frequencies; requiring considerable manipulation of the denoising curve to achieve as flat a spectral response as possible.

The application of the noise reduction would reduce the noise profile by as much as 30dB after two to four processes. (once significant voice degradation appeared on the lower harmonics of the voice print, it was decided to stop the processes). On the mechanical noise samples, the frequencies which masked the harmonics before noise reduction were reduced significantly; yet this also severely degraded entire (narrow) frequency bands which in turn degraded the finer detail of the harmonic pitch variation. Whilst the higher and lower harmonics would remain intact, the harmonics with most amplitude variation were eroded making it difficult to visually inspect and match the contours. Despite the hypothesis, it was the voiceprint summed with mechanical noise which suffered more degradation after the noise reduction processes.

On the coffee shop ambience, noise reduction which targeted the background speech was effective in reducing the overall noise volume; however, this also eroded the target voice in

parts. Compared to samples with the sustained mechanical noise, it was easier to find higher frequency points of the key speech to inspect the contour behaviour; although not on the vowel transition sounds. Such was the overall degradation however, it soon became difficult to differentiate target speech contours and background contours.

Across all samples, the parts of the phrase which appear to have survived degradation are the words "tell", "that" and "cheaper". This is possibly because these words start and end with fricatives and plosives which consists of the most audible energy.

Despite Kersta's research [1] which proposed a set of key words used by students to identify a speaker, none of the listed cues were measurable after the noise reduction processes. However, the word "tell" which features a similar vowel sound to "a" (lower case pronunciation) gave a visually strong representation of the target voice and so was chosen for acquiring the measurements. At a normal speaking rate, this word along with "that" is a combination of fricative energy and sustained vowel. It was one of very few strong cue words in the processed test files (on -18dBFS upward).

The fundamental frequency and the first three harmonics retained their dominance in the spectral space across all the samples; although where the noise profile originally had an amplitude of -12dBFS, -6dBFS and 0dBFS, the reduced profile (across the full frequency spectrum) had a dB RMS value higher than second harmonic upwards. In these circumstances, it remains possible to ascertain the relation between the harmonics; but it remains advisable that the practitioner examine the relationship between the harmonics and the surrounding noise to eliminate any potential doubt. The spectral analyser showed a 20dB difference between surrounding noise and target signal when examining these harmonics.

The efficacy of the noise reduction performed best when the background noise level was at or below -18dBFS. A lower background noise level (such as at -30dBFS) still risked degradation to the voiceprint but such a risk in this experiment did not outweigh any potential value a noise reduction process could bring to the sample. The mid and upper contours were still visible within the voiceprint.

When measuring the amplitude of the harmonics in the word "tell", the relationship of the initial recording showed only 2 to 3dB differences between the first two harmonics before a 12dB drop to the remain harmonics which then had 2 to 3dB differences; the post processed harmonics at the same word either had wider differences (thus requiring a difference tolerance) or were un-measurable because they were too difficult to pick out from the remaining profile.

## IV. Conclusion

Before conducting any noise testing, visual examinations of the test files revealed that when placed in real world conditions, voiceprints have unique yet small character traits which are readily available when recorded in studio conditions. Once subjected to dubbed noise profiles (sourced from real world situations), these nuances are the first to disappear.

From the experiments, the author concludes and would like to propose that any practitioner wishing to use noise reduction to obtain a clearer voiceprint only does so when there is a target signal to noise ratio of 18dBFS. If this ratio is less, the practitioner should ensure they can clearly demonstrate their justifications for continuing with noise reduction and also demonstrate the pitch and amplitude variable behaviours of the mid and upper harmonic contours. The author would also acknowledge that a spectrogram cannot alone provide voice identity and that a voice analysis should still be conducted by experts in linguistics and phonetics; with spectrograms provided as secondary support to aid the phonetician.

Further to this, the author notes that recently introduced FSR guidelines request the over processing be avoided. Great care must be taken not to use excessive noise reduction on an evidential recording [3,5]; as this will erode the audible information making it difficult to eliminate any doubt.

## References

[1] L. Kersta "Voiceprint Identification", Nature vol.196, pp.1253-1257 (1962)

[2] J. Careless "Forensic Audio Analysis" Government Video, vol. 19, no. 7, pp.32-33 (2008)

[3] H. Fraser "Don't believe your ears: 'enhancing' forensic audio can mislead juries in criminal trial" [online] (2019) url: ://theconversation.com/dont-believe-your-ears-enhancing-forensic-audio-can-mislead-juries-in-c riminal-trials-113844

[4] S. Ogata and T. Shimamuma "Reinforced Spectral Subtraction Method to Enhance Speech Signal" Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. TENCON 2001, vol.1, pp.242-245 (2001)

[5] Forensic Science Regulator "Codes of Practice and Conduct. Appendix: Speech and Audio Forensic Services FSR-C-134" [online] (2020) url: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/912393/134_FSR_Speech_and_Audio_Appendix_FSR-C-134_Issue_2.pdf Accessed on 26th October 2020

[6] F. Lorenz (2013) "Basics of Phonetics and Phonology in English" Logos Verlag, Berlin

# 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

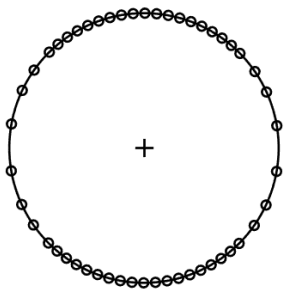

Fig. 1 Source Directions on Azimuthal Plane used in datasets

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 $\alpha$ 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 | ([2,2], 8) |
| Batch Normalisation | |
| ReLu | |
| Max Pooling | (2,2) |
| **Layer 2** | |
| Convolution Layer | ([8,8], 16) |
| Batch Normalisation | |
| ReLu | |
| Max Pooling | (2,2) |
| **Layer 3** | |
| Convolution Layer | ([16,16], 32) |
| Batch Normalisation | |
| ReLu | |
| Max Pooling | (2,2) |
| **Output Layer** | |
| Dense | 50 |
| Softmax | |

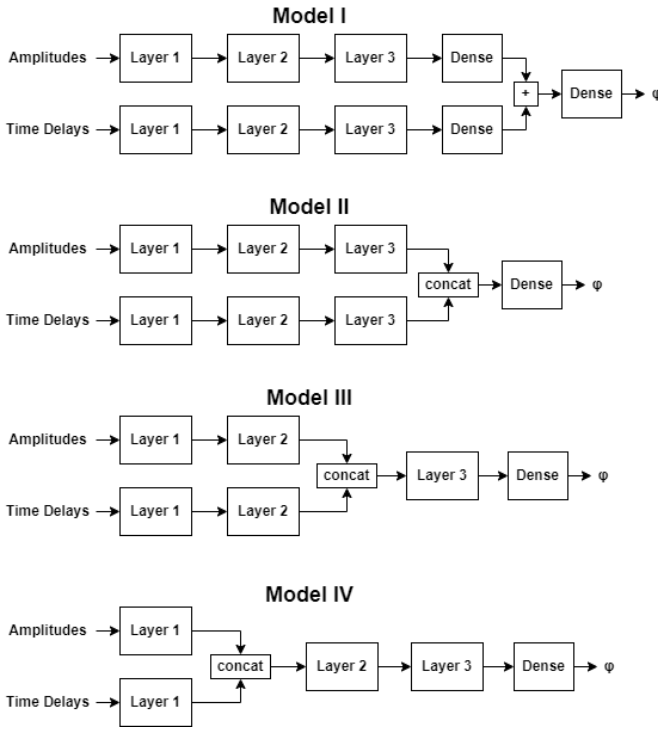


Fig. 2 Framework for all four Models

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.

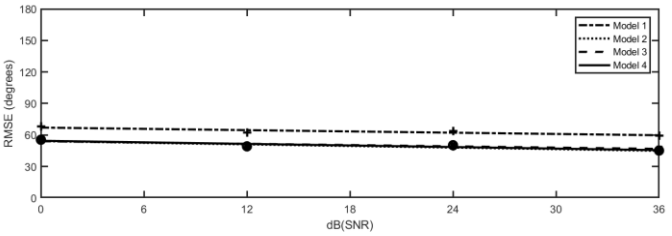


Fig. 3 RMSE with respect to SNR

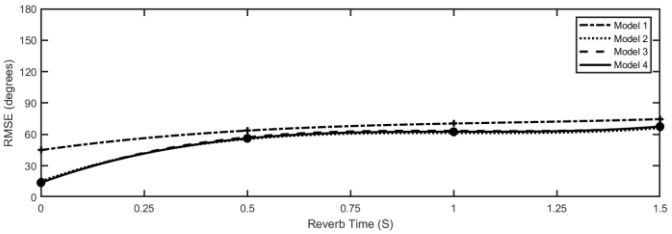


Fig. 4 RMSE with respect to Reverb Time

## 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