Audio Engineering Society

Conference Paper 18

Presented at the AES International Conference on Machine Learning and Artificial Intelligence for Audio
2025 September 8–10, London, UK

Predicting binaural colouration using VGGish embeddings

Thomas McKenzie[1], Alec Wright[1], Daniel Turner[2], and Pedro Lladó[3]

[1]*Acoustics and Audio Group, Reid School of Music, University of Edinburgh, United Kingdom*
[2]*Vision, Learning, and Control Group, School of Electronics and Computer Science, University of Southampton, United Kingdom*
[3]*Institute of Sound Recording, University of Surrey, United Kingdom*

Correspondence should be addressed to Thomas McKenzie (thomas.mckenzie@ed.ac.uk)

ABSTRACT

An initial feasibility study is presented exploring the use of a pre-trained feature extractor, designed for large-scale audio classification, applied to the task of predicting the colouration between binaural signals. A multilayer perceptron (MLP) is trained to predict binaural colouration using feature embeddings obtained from the VGGish network and data from five previously conducted listening tests. The evaluation compares seven versions of the network, each trained using different data augmentation methods, to three existing signal processing methods: basic spectral difference (BSD), log. spectral distance (LSD) and an auditory model for predicting binaural colouration (PBC-2). Results show that while the MLP networks are comparable to BSD and LSD, further work is needed to compete with the more accurate PBC-2; such as using specific audio features relevant for colouration.

1 Introduction

Colouration is a perceptual attribute that is central to spatial audio quality evaluation, and often considered more important than localisation [1]. As with any perceptual attribute, the most accurate way to measure it is through controlled listening experiments. However, numerical assessment methods that correctly estimate the perceived colouration are desirable, since they enable quick monitoring of the introduced colouration during development phases of spatial audio systems [2, 3, 4].

A recent study [5] proposed an auditory model that showed strong correlation between listening test data and the model predictions. This model follows a traditional signal processing approach combining signal rectification, high frequency smoothing, equivalent rectangular bandwidth frequency weighting, and signal-dependent weighting of the left and right colouration values to produce a single binaural colouration prediction. A Matlab implementation of the resulting auditory model was made available as part of the Auditory Modeling Toolbox under the model name *mckenzie2025* [5].

To the best of the authors' knowledge, no data-driven model of colouration has yet been explored. Recently, data-driven models have been proposed for functions of auditory processing, often with promising results. These models are either trained using listening experiment data [6] or with true objective data [7, 8, 9]. In many cases, the models can capture nuances of auditory processing that are not fully understood from a psychophysical or neurophysiological point of view,

generalising to unseen conditions and outperforming traditional models in certain tasks. A third strategy is to train data-driven models with estimated data using existing auditory models [10, 11]. However, while synthetic training data is easy to obtain in these cases, and while the resulting models may run significantly faster, their performance is not expected to exceed that of the original models.

This study presents an initial exploration of a data-driven method as an alternative to the auditory model introduced in [5]. The method presented here adopts a pre-trained feature extractor [12] designed for large-scale audio classification, whose feature embeddings are used to train a multilayer perceptron (MLP) regression model. Since the long-term goal is to obtain better results than the ones obtained using [5], the model is trained with listening experiment data as labels. Within the context of the limited colouration data from listening experiments and aiming to optimise the amount of information obtained from them, feature extraction and data augmentation techniques are evaluated for their effect on the model predictions. The data-driven methods explored here are compared against existing colouration metrics obtained using signal processing approaches.

The paper is organised as follows. Section 2 presents a methodology for predicting binaural colouration, whereby a pre-trained feature extractor is used to train MLP regression models and the data augmentation methods are detailed. Section 3 then presents an evaluation of the proposed methods against existing models on their prediction of colouration using a previously unseen subset of signals from the listening tests on perceived spectral similarity. The results of the evaluation are discussed in Section 4, identifying the models that perform the strongest and suggesting why, as well as discussing limitations of the current methodology. Finally, the paper is concluded in Section 5 along with future work.

2 Methods

This section describes the methodology of this study, including the pre-trained VGGish feature extraction model, the multilayer perceptron (MLP) regression model and the training data, along with details of the tested data augmentation methods. All audio signals used in this study are binaural with a sample rate of 48 kHz and 16-bit depth. Fig. 1 presents a block diagram illustrating the training pipeline of the MLP model.

2.1 Listening test data

The dataset used consists of 252 binaural signals, featuring different audio contents and colouration types from five previously conducted listening tests [13, 14, 15, 4, 16]. The number of signals, their audio contents and their colouration types, along with the number of participants in each listening test, are summarised in Table 1.

All listening tests followed a paradigm similar to the multiple stimulus test with hidden reference and anchor (MUSHRA [17]), whereby the similarity between a reference and multiple test signals was rated for the quality of *colouration* on a scale from 0 to 100, whereby a rating of 100 means entirely the same (no colouration). In this study, perceived colouration is therefore derived as the reversed ratings of perceived similarity (where 0 indicates entirely the same, and 100 means as coloured as the low anchor). A single result for each stimulus is taken as the mean of all participants individual ratings.

For each trial in all five listening tests, the reference stimulus was a convolution of the monophonic audio content (see again Table 1) with head-related impulse responses (HRIRs), and anchors were a low-pass filtered version of the reference with a cut-off frequency $f_o = 3.5$ kHz. Test stimuli were presented statically (with a fixed head orientation) over open-back headphones in quiet listening rooms. While the majority of stimuli are simple scenes (single stationary sound sources at a single location), some were complex scenes utilising a mix of multiple sound sources located at different locations on the sphere [15, 16]. The signals with pink noise audio content were around one second in duration, whereas some of the complex scenes (e.g. the pink noise bursts with a train station recording and the percussion recordings) were up to 8 seconds in duration.

The listening test data was randomised prior to being divided into subsets for training, validation and testing, to try and ensure an equal representation of colouration scores within each set. The randomisation used a fixed seed to ensure reproducibility between experiments. The data was then split into training, validation and test subsets, with a split of $60\% : 20\% : 20\%$, meaning a total of 152 binaural signals reserved for training, 50 for validation and 50 for testing.

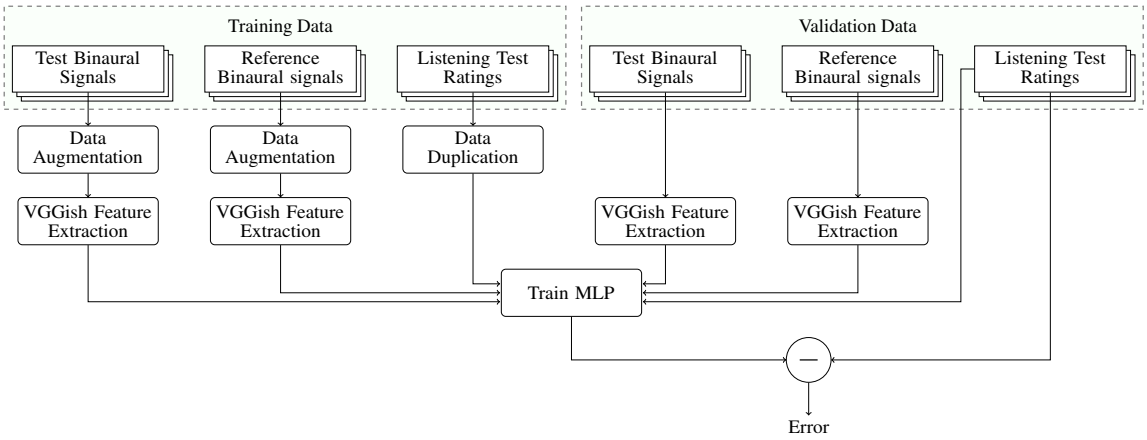

Fig. 1: Block diagram summary of the training of the multilayer perceptron (MLP) models.

Table 1: Brief details of the listening test data used in this study.

Test	# Stimuli	Audio content	Colouration type	# Participants
Mc18 [13]	48	Pink noise	Equalised Binaural Ambisonic HRIRs	20
Mc19a 1 [14]	42	Pink noise	Equalised Binaural Ambisonic HRIRs	20
Mc19a 2 [14]	21	Pink noise bursts and train station recording	Equalised Binaural Ambisonic HRIRs	20
Mc19b 1 [15]	15	Pink noise	Pre-processed Binaural Ambisonic HRIRs	20
Mc19b 2 [15]	15	Percussion recordings	Pre-processed Binaural Ambisonic HRIRs	20
Mc22 [4]	56	Pink noise	10 band equalised HRIRs	9
Ll22 1 [16]	25	Pink noise	HRIRs made wearing different headphones	15
Ll22 2 [16]	25	Speech recordings	HRIRs made wearing different headphones	15
Ll22 3 [16]	5	Rainfall recordings	HRIRs made wearing different headphones	15

2.2 Feature extraction

Given the limited available data, a pre-trained feature extractor was employed to leverage the compact representations learned from a much larger corpus of general audio data. This allows the MLP regression model to be trained on a meaningful, but lower dimensionality, representation of the dataset. The feature extraction mechanism aims at reducing the number of input features into the MLP while maintaining as much relevant information as possible.

In this study, the neural network model detailed in Section 2.3 was trained using features extracted using the VGGish network [12]. VGGish, a modified version of VGG16 architecture, is an 11-layer convolutional neural network (CNN) that has been pre-trained on audio data from the YouTube-8M dataset [18], which incorporates log-mel spectrogram-based inputs for feature extraction. The VGGish model produces 128 features per test signal for each second of audio in a single audio channel. For audio signals that are longer than 1 second, only the first 128 features are used in this study. This is equivalent to truncating the signal to one second, and is done to ensure all signal durations produce results with the same dimensions. It should be adequate, as colouration in the used listening test data is assumed to be approximately constant over the entire signal. As all audio signals are binaural, the left and right channel features are concatenated, resulting in 256 features per binaural signal.

2.3 Model

An MLP was trained to predict binaural colouration using the features extracted using the VGGish extractor as input. A block diagram illustration of the model is

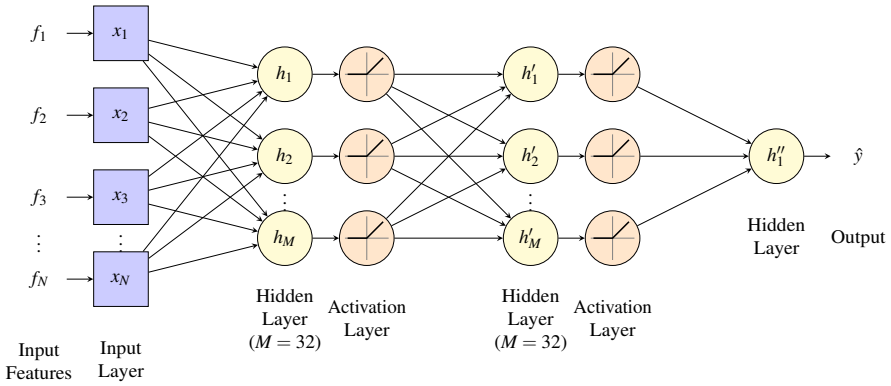

Fig. 2: Structure of the multilayer perceptron (MLP) regression model used in this study to predict binaural colouration.

presented in Fig. 2. The model consists of a feature input layer, the size of which is equal to the number of features outputted by the feature extractor, which in this study was $N = 256$. The feature input layer is followed by two fully connected hidden layers, both of size $M = 32$, and the final output layer, which consists of a single unit to predict the final colouration value. A Rectified Linear Unit (ReLU) activation layer is also applied to the output of each hidden layer.

The model is optimised using mini-batch Stochastic Gradient Descent with Momentum (SGDM), with the mean-squared-error (MSE) between the predicted and ground truth colouration values as the loss function to be minimised. The initial learning rate is 5×10^{-4}, the maximum number of epochs set to 200, and the batch size is set to 32. The validation loss is calculated every 20 epochs, with a validation patience of 10 to allow early stopping. The model with the best validation loss is the one taken forward to be evaluated on the test set.

2.4 Data augmentation

Due to the limited available training data, two data augmentation methods were explored to investigate their potential to improve the performance of the model through effectively increasing the amount of training data. The results and effects on the validation set are reported here.

Many established methods for audio data augmentation either directly or indirectly alter the spectral content of the data [19]. As such, these methods are not suitable for this study as they would likely cause a change in perceived colouration and thus invalidate the associated ground truth value. Therefore, the options available were limited and only temporal changes were employed. The training data was duplicated and augmented, whereas the listening test results were simply duplicated.

The first augmentation method is a time-reversal of the signals (herein referred to as 'reverse'). The discreet time signal $x(n)$, where n denotes the sample, is reversed by

$$x_{\text{rev}}(n) = x(N - n) \tag{1}$$

where N is the total number of samples in the signal.

The second augmentation method is a time-rotation of the signals (herein referred to as 'circshift', whereby the beginning of the signal is moved to the end and the end is brought forward. A signal $x(n)$ processed with circshift[M] becomes

$$x_{\text{circshift}[m]}(n) = x(\frac{mN}{M+1} + n \mod (N)) \tag{2}$$

where m is the index of the duplication and M is the total number of duplications. For example, one duplication of data (denoted here as circshift[1]) would mean the first half of the signal is swapped in position with the second half of the signal. For two duplications (circshift[2]), the first duplication moves first third to the end while the second two thirds are moved to the start; the second duplication moves the first two thirds to the end while the last third is moved to the start. While 'reverse' can only augment the data once, 'circshift' allows for many more augmentations of the data.

Table 2: Results of initial model training using different data augmentation methods. $\overline{\text{RMSE}}$ and STD denote the mean and standard deviation of the root-mean-square error between the ten repeats, respectively.

	Training		Validation	
Augmentation	$\overline{\text{RMSE}}$	STD	$\overline{\text{RMSE}}$	STD
none	8.51	3.9	13	0.867
reverse	8.66	2.78	13.3	0.849
circshift[1]	8.69	3.21	12.5	1.15
circshift[2]	10.1	2.97	12.2	0.586
circshift[3]	8.03	1.26	11	0.528
circshift[4]	8.77	3	11.5	0.994
circshift[5]	7.77	1.42	10.5	0.56

The configurations of augmentation methods employed in this study are summarised as follows:

- none: No data augmentation. Training data dimensions [152,256]
- reverse: Training data signals duplicated with signals time-reversed. Training data dimensions [304,256]
- circshift[1-5]: Training data signals duplicated with signals time-shifted. Training data dimensions [304,256], [456,256], [608,256], [760,256] and [912,256], respectively

To evaluate the effects of the different data augmentation methods, an MLP was trained ten times for each. To ensure repeatability, the random seed was fixed for each iteration, such that the first repeat with each augmentation method would use the same seed, and the next repeat would use a different common seed. Table 2 presents the initial results on the training and validation data for the models trained using seven training datasets: one of which is the original without data augmentation ('none'), and six with data augmentation. Two metrics are reported: the mean ($\overline{\text{RMSE}}$) and the standard deviation (STD) between the root-mean-square error (RMSE) of the ten repeats.

The first observation of these initial results is that overfitting appears evident from the lower RMSE values for training than validation (see Table 2). This is likely due to the small amount of training data. Secondly, the standard deviation between RMSE for each repeat of the training is far higher than that for validation. A likely explanation for this was found by observing that in some training repeats, the loss would briefly jump, and never recover to the values found in other repeats. It may also be a result of there being less data in the validation set.

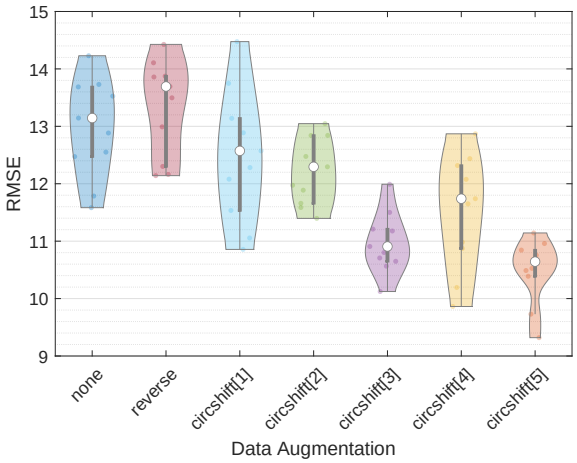

Fig. 3: Root-mean-square error (RMSE) between the validation results with different data augmentation options pre-feature extraction, for 10 repeats of training the model with different random seeds.

To look in more detail at the effect of augmentation methods on the validation set, Fig. 3 presents violin plots of the validation RMSE. Here it seems that the first augmentation method, reverse, does not generally appear to improve the validation loss. The more promising results seem to be as the number of signals in the training dataset increases significantly in size (e.g. for circshift[3-5]): here the validation loss is significantly lower than for no data augmentation. These results helped to inform the evaluation in the following section.

3 Evaluation

To evaluate the final models, the mean listening test ratings (perceived colouration) were compared to predicted colouration values made using the trained neural networks with the as-before-unseen test subset of the data. For an additional comparison to the current state-of-the-art, the results from the proposed neural

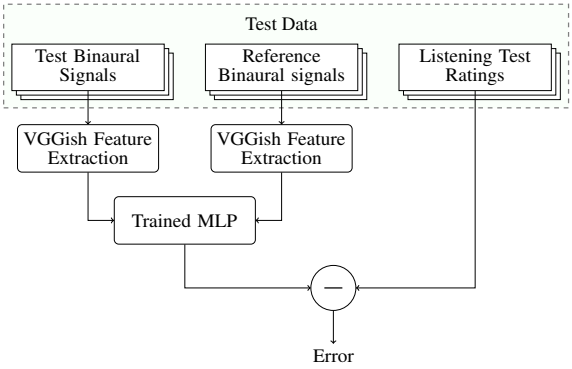

Fig. 4: Block diagram summary of the testing of the trained MLPs.

networks are also compared to three signal processing-based binaural colouration methods. These are basic spectral difference (BSD), log spectral distance (LSD) and PBC-2 (*mckenzie2025* in the auditory modelling toolbox):

- BSD: mean absolute difference in dB between the reference and test spectra
- LSD: Root-mean-square difference in dB between the reference and test spectra
- PBC-2: Auditory model with features including signal rectification, high frequency smoothing, equivalent rectangular bandwidth frequency weighting, binaural weighting

Fig. 4 and Fig. 5 present block diagrams for the evaluation stage using the trained MLP and the PBC-2 auditory model, respectively. This illustrates the differences between the two methodologies.

While the MLP networks have been trained to produce colouration predictions within the same range as the listening test results, the raw BSD and LSD predictions require transformation to fit a range of 0-100. This was done using the linear regression coefficients as calculated in [5], with respective slope and intercept values of $a = 4.803$ and $b = 13.379$ for BSD, and $a = 4.101$ and $b = 10.252$ for LSD. The PBC-2 model already transforms outputs to fit within this range.

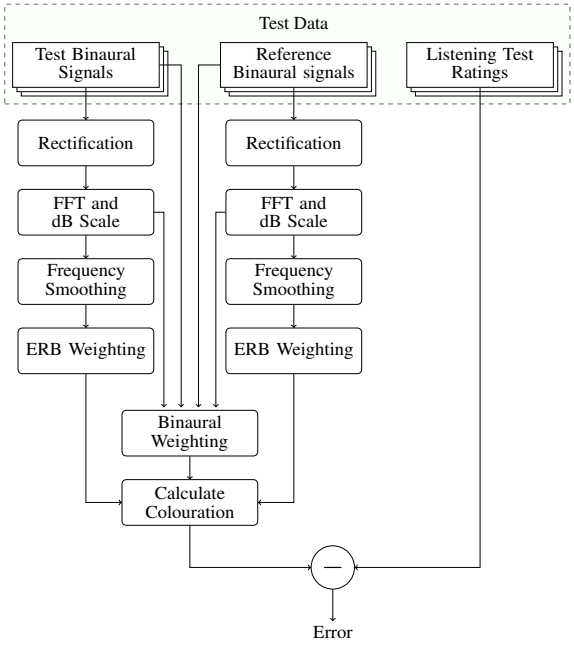

Fig. 5: Block diagram summary of the PBC-2 auditory model [5] and its use in this study.

3.1 Results

The predictions are analysed using a linear regression fitting; the results of which are presented in Table 3. Three measures are reported: R^2_{adj}, the proportion of explained variance (the squared correlation coefficient but corrected for the number of predictors, which is one in this case); MxAE, the maximum absolute error between the ratings and predictions; and RMSE, the root-mean-square error between the ratings and predictions. Note: more accurate predictions of the listening test results are indicated by higher R^2_{adj} values, and lower MxAE and RMSE values.

Firstly, all tested models produce a positive correlation between the predicted and perceived colouration values, suggesting that VGGish features can be used to predict binaural colouration. The range in R^2_{adj} values show that all variants of the MLP networks perform somewhat comparably to the simple signal-processing colouration methods BSD and LSD. However, they are all significantly worse than the more tailored PBC-2. The best performing network was the one with circ-shift[4] training data augmentation, which achieved an $R^2_{\text{adj}} = 0.639$. This demonstrates a notable improvement on BSD and LSD.

Table 3: Results from the linear regression between the predicted and perceived colouration, of the neural networks (top) and signal processing colouration methods (bottom). $R^2_{\text{adj.}}$, MxAE and RMSE denote adjusted explained variance, maximum absolute error and root-mean-square error, respectively.

Processing	$R^2_{\text{adj.}}$	MxAE	RMSE
none	0.549	28.87	12.9
reverse	0.535	36.39	13.1
circshift[1]	0.502	39.83	13.88
circshift[2]	0.531	33.88	13.3
circshift[3]	0.445	37.62	14.38
circshift[4]	0.639	29.59	11.45
circshift[5]	0.541	33.82	13.08
BSD	0.557	34	12.92
LSD	0.521	37.3	13.34
PBC-2	0.901	16.3	6.164

To further understand the results of the regression analysis and how the different model iterations compare, Fig. 6 illustrates the predicted versus perceived colouration and the prediction error. An interesting observation can be made by comparing the predicted error plots. While BSD and LSD seem to significantly overestimate colouration values below 20, the MLP networks and PBC-2 appear somewhat less susceptible to this trend.

4 Discussion

The results show that feature embeddings from VGGish demonstrate a certain degree of generalisability and can be applied to specific tasks, such as predicting binaural colouration. Its performance was comparable to baseline signal processing methods, such as BSD and LSD, with some of the tested feature augmentation and distillation techniques enabling it to outperform these approaches. Generally, the models that better correlated with the listening test results also had lower maximum and average errors. However, even the best-performing MLP configurations performed significantly poorer than the auditory model PBC-2.

A key limitation observed in the results is the likely presence of significant overfitting. This is evident from the pattern of RMSE values across training, validation, and test sets, which typically increased from the training to validation to test set. Notably, models that exhibited consistently lower training and validation loss, such as those employing circshift[3] and circshift[5], produced higher errors for the unseen test set. To address this, future development should employ procedures such as K-fold cross-validation to obtain a more accurate measure of general model performance.

One of the primary constraints affecting model performance is the limited availability of data. The small dataset likely contributed to overfitting and played a significant role in the diminished performance observed on the test set. This suggests that the models were not generalising well beyond the training and validation data. Addressing this limitation would require access to a more extensive dataset, which would be costly and time consuming to acquire.

Additionally, what data there is is still somewhat limited in terms of content and colouration type. Almost 75% of sounds use pink noise as the stimulus. A greater variety is needed with more music and speech, and a greater range of both transient and periodic sounds. Over 50% are binaural renders of Ambisonic signals. Whilst the processing on these signals differs, from equalisation to different order-dependent colouration to HRIR pre-processing steps, more variation in types of colouration should be explored such as between HRIRs from different heads and degradations from perceptual coding algorithms (e.g. lossy data compression). More than 80% are single-source stationary anechoic signals. Reverberant signals, multiple-source complex scenes, moving sources, and varying source widths should be better represented in the data. Finally, the listening tests used in this study primarily featured male Caucasian participants between the ages 20 and 40. This is not representative of the general population and should be addressed, as models trained on the current data will be biased.

With regard to data augmentation, the current approach involved duplicating listening test results. A potential refinement would be to replace the duplication of the mean across all participants and repetitions with a leave-one-out strategy, introducing variation into the duplicated mean perceptual values. Other audio augmentation methods that do not affect colouration should continue to be explored too, as it is likely that even a larger dataset of listening experiment data will be insufficient in many cases.

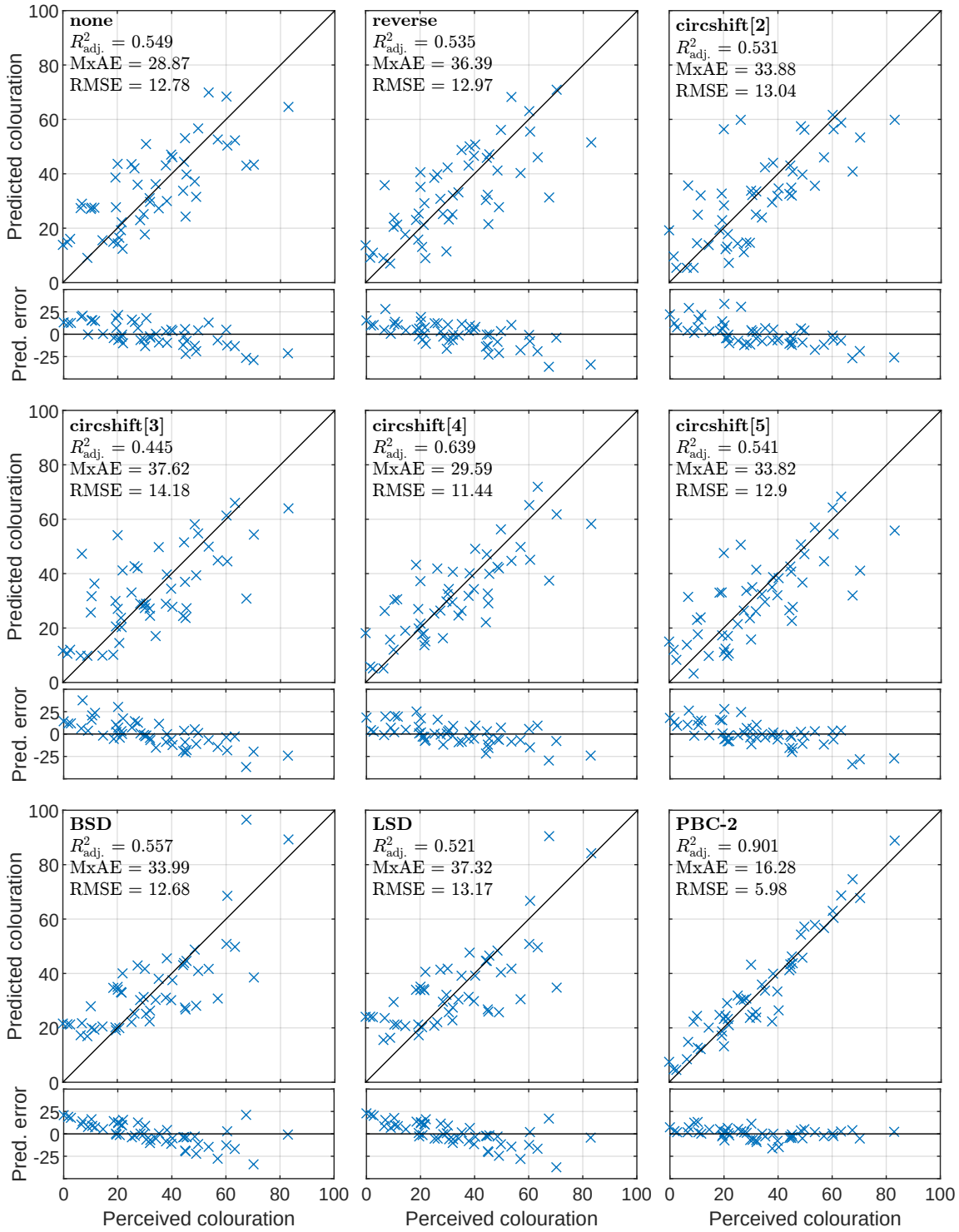

Fig. 6: Model performance by means of the predicted versus rated colouration and the prediction error (pred. error) for each model. $R^2_{\text{adj.}}$, MxAE and RMSE denote adjusted explained variance, maximum absolute error and root-mean-square error, respectively. Note: circshift[1] omitted for readability.

Another point of discussion is the suitability of the used feature extractor. The VGGish network used in this study was originally trained at a sample rate of 16 kHz on a dataset that may not be directly relevant to the task at hand (YouTube audio clips). As many of the important spectral cues in binaural signals are present at high frequencies (above 5 kHz), and a 16 kHz sample rate will have a nyquist frequency of 8 kHz, it is possible that many important spectral features will not be represented in the extracted VGGish features. Given these constraints, it is somewhat surprising that VGGish was able to generalise at all. Future developments will explore alternative feature extraction methods that are more specifically suited to binaural colouration. At a minimum, incorporating some form of auditory front-end modeling prior to feature extraction could be beneficial. Alternatively, leveraging the PBC-2 model in conjunction with a neural network could be investigated.

Finally, other avenues for improving the current methodology include hyperparameter optimisation in the MLP training process. Fine-tuning the model parameters could enhance performance and mitigate some of the observed limitations. However, this is expected to offer minor improvements when compared to the other more significant developments such as a larger and more diverse dataset, greater and more effective data augmentation methods, and more binaural colouration-specific feature extraction methods.

5 Summary

This paper has presented an exploratory study into the use of shallow neural networks in predicting the colouration between binaural signals, utilising feature embeddings obtained from a VGGish pre-trained convolutional neural network for large-scale audio classification. Seven versions of the proposed model with different data augmentation methods have been tested and compared to three signal processing methods for predicting binaural colouration. Results show that the MLP networks offer comparable, and in some cases improved, performance to the simple signal processing methods BSD and LSD, but fall short of the more complex PBC-2.

Although the results of this study were not close to the state-of-the-art PBC-2 model, this approach shows promise. A number of avenues can be explored to improve the results. First and foremost is the need for a larger dataset with more diverse signals in areas such as: complexity of scene (number of sources), reverberance, colouration type, sound source content (transient and periodic, wide-band and narrow-band), among others. Secondly, other features could be explored: while VGGish does work, using features specific to colouration may be more appropriate, such as those produced by the PBC-2 model's peripheral processing stages. Additional data augmentation and hyperparameter optimisation may serve to improve model performance with the given data. Finally, future work to reduce the overfitting will only serve to improve the overall generalisability of the models.

6 Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/X032914/1, project Challenges in Immersive Audio Technology.

References

[1] Rumsey, F., Zieliński, S., Kassier, R., and Bech, S., "On the Relative Importance of Spatial and Timbral Fidelities in Judgments of Degraded Multichannel Audio Quality," *J. Acoust. Soc. Am.*, 118(2), pp. 968–976, 2005, doi:10.1121/1.1945368.

[2] Ono, K., Pulkki, V., and Karjalainen, M., "Binaural Modeling of Multiple Sound Source Perception: Methodology and Coloration Experiments," *111th AES Convention*, 2001.

[3] Ono, K., Pulkki, V., and Karjalainen, M., "Binaural Modeling of Multiple Sound Source Perception: Coloration of Wideband Sound," in *112th AES Convention*, 2002.

[4] McKenzie, T., Armstrong, C., Ward, L., Murphy, D. T., and Kearney, G., "Predicting the Colouration Between Binaural Signals," *Appl. Sci.*, 12(5), pp. 1–15, 2022, doi:10.3390/app12052441.

[5] McKenzie, T. and Brinkmann, F., "Toward an Improved Auditory Model for Predicting Binaural Coloration," *J. Audio Eng. Soc.*, 73(3), 2025, doi:10.17743/jaes.2022.0192.

[6] Lladó, P., Hyvärinen, P., and Pulkki, V., "Auditory Model-Based Estimation of the Effect of Head-Worn Devices on Frontal Horizontal Localisation," *Acta Acustica*, 6(1), 2022, doi: 10.1051/aacus/2021056.

[7] May, T., Van De Par, S., and Kohlrausch, A., "A Pobabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE Trans. on Audio, Speech and Lang. Proc.*, 19(1), pp. 1–13, 2010, doi:10.1109/TASL.2010.2042128.

[8] Francl, A. and McDermott, J. H., "Deep Neural Network Models of Sound Localization Reveal how Perception is Adapted to Real-World Environments," *Nature Human Behaviour*, 6(1), pp. 111–133, 2022.

[9] Saddler, M. R. and McDermott, J. H., "Models Optimized for Real-World Tasks Reveal the Task-Dependent Necessity of Precise Temporal Coding in Hearing," *Nature Communications*, 15(1), pp. 1–29, 2024.

[10] Baby, D., Van Den Broucke, A., and Verhulst, S., "A Convolutional Neural Network Model of Human Cochlear Mechanics and Filter Tuning for Real-Time Applications," *Nature Machine Intelligence*, 3(2), pp. 134–143, 2021.

[11] Leer, P., Jensen, J., Carney, L. H., Tan, Z.-H., Østergaard, J., and Bramsløw, L., "Hearing-Loss Compensation Using Deep Neural Networks: A Framework and Results From a Listening Test," *IEEE Trans. on Audio, Speech and Lang. Proc.*, pp. 828–841, 2025, doi:10.1109/TASLPRO.2025.3536183.

[12] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al., "CNN Architectures for Large-Scale Audio Classification," in *IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, pp. 131–135, 2017.

[13] McKenzie, T., Murphy, D. T., and Kearney, G. C., "Diffuse-Field Equalisation of Binaural Ambisonic Rendering," *Appl. Sci.*, 8(10), 2018, doi:10.3390/app8101956.

[14] McKenzie, T., Murphy, D. T., and Kearney, G., "Towards a Perceptually Optimal Bias Factor for Directional Bias Equalisation of Binaural Ambisonic Rendering," in *EAA Spatial Audio Sig. Proc. Symp.*, 2019, doi:10.25836/sasp.2019.08.

[15] McKenzie, T., Murphy, D. T., and Kearney, G., "An Evaluation of Pre-Processing Techniques for Virtual Loudspeaker Binaural Ambisonic Rendering," in *EAA Spatial Audio Sig. Proc. Symp.*, pp. 149–154, 2019, doi:10.25836/sasp.2019.09.

[16] Lladó, P., McKenzie, T., Meyer-Kahlen, N., and Schlecht, S. J., "Predicting Perceptual Transparency of Head-Worn Devices," *J. Audio Eng. Soc.*, 70(7/8), pp. 585–600, 2022, doi:10.17743/jaes.2022.0024.

[17] ITU-R BS.1534-3, *Methods for the Subjective Assessment of Intermediate Quality Level of Audio Systems*, International Telecommunication Union, Geneva, Switzerland, 2015.

[18] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S., "Youtube-8M: A Large-Scale Video Classification Benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[19] Singh, N. K., Chanu, Y. J., and Pangsatabam, H., "A Study of Various Audio Augmentation Methods and Their Impact on Automatic Speech Recognition," in *Lecture Notes in Electrical Engineering*, volume 1071 LNEE, pp. 481–491, Springer Science and Business Media Deutschland GmbH, 2024, doi:10.1007/978-981-99-4713-3_46.