Audio Engineering Society

Conference Paper 25

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

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Establishing a Virtual Listener Panel for audio characterisation

Michelle Herlufsen[1], Niels Asp Fuglsang[1], Benjamin Pedersen[1], Søren Vase Legarth[1]

[1] *SenseLab, FORCE Technology, 2970 Hørsholm, Denmark*

Correspondence should be addressed to Michelle Herlufsen (MCHE@forcetechnology.com)

ABSTRACT

Listening tests provide valuable insight into the sound quality of audio products, but the barriers to conducting such tests are high. They require specialized knowledge, facilities, equipment, time, and access to a panel of trained assessors. The cost alone of maintaining a panel of expert assessors is significant. This study introduces an AI model, the "Virtual Listener Panel", trained using a Convolutional Neural Network (CNN) to overcome these challenges. The purpose of the model is to predict expert ratings for audio characterization. While this paper briefly describes the architecture of the proposed model, the primary focus is on the data collection approach developed to obtain premium quality training data. The performance of the Virtual Listener Panel is evaluated using unseen test data from a listening test in which 20 expert assessors characterized several different earbuds. Prediction performance of five audio characterization attributes from the timbre domain were evaluated: bass strength, bass depth, midrange strength, treble strength and brilliance. Despite the modest dataset size, the model achieved Pearson correlation coefficients between 0.80 and 0.97. The Virtual Listener Panel offers a scalable solution for fast and accurate audio characterization, reducing the need for extensive listening tests and providing valuable insights for all stages of product development.

1 Introduction

Perceptual audio evaluation is crucial for understanding how listeners perceive and appreciate audio products. However, conducting listening tests to collect perceptual ratings is both costly and time-consuming. As a result, objective metrics that predict these ratings are increasingly being used and evolving rapidly. For instance, an open dataset of audio signals with corresponding subjective quality scores is available to accelerate research on perceived audio quality predictions[11]. The International Telecommunication Union (ITU) has standardized ITU-R Rec. P.863 "POLQA" [2] and ITU-R Rec. BS.1387 "PEAQ" [1]. Additionally, several other perceptually driven audio metrics exist, such as 3QUEST [17], VISQOL [3], DPAM [4], CDPAM [5]. These metrics are intrusive methods, meaning they require a clean reference signal. However, this requirement is often impractical in real-world scenarios where reference recordings are unavailable or difficult to obtain. Consequently, non-intrusive metrics—which do not require a reference signal—have been developed using different deep learning architectures, such as Convolutional Neural Network (CNN) [12]-[14],[16],[20]. As a result, state-of-the-art approaches are increasingly relying on deep learning techniques.

All the objective metrics mentioned above focus on speech degradation and mean of opinion scores. While these metrics provides valuable information about consumer preference, they do not provide insight of which characteristics are the *drivers of liking*. This additional insight is crucial for improving

sound quality in a product and would typically require listening test data from trained assessors. In this study we utilize our panel of expert listeners to train a CNN model, the Virtual Listener Panel (VLP), to characterise audio. The process of sensory assessor development to obtain expert assessors [8] requires significant resources and is therefore less common within the audio field. Consequently, the use of expert data in deep learning modelling is limited. Experts are highly trained assessors who are usually used for perceptual objective tests. This is where the human serves as the measuring instrument. The characteristics of the perceived stimuli are evaluated objectively whereas preference and other affective measures are subjective [10]. As an example, we use a wide range of sensory descriptors from the Sound Wheel[7][15], such as clarity, brilliance, bass strength or different spatial cues to characterise audio products. At SenseLab, an expert panel consisting of more than 30 assessors, understands and is repeatedly trained in these attributes. This means that they contribute with high-quality data, and the panel is potentially large enough to deliver sufficient data for a CNN model in this unique area of audio characterisation. A neural network model does not give knowledge of the internal workings, often referred to as a "black box". Developing a model that offers insight to objective perceptual characterisation, which can be used for preference mapping[1], can mitigate the black box effect. Especially, timbre is found to be essential to audio preference. Rumsey et al. [23] found that approximately 70% of the perceptual differences between loudspeakers were related to timbre, for surround sound stimuli downmixed to other loudspeaker configurations. The high importance of timbre leading to preference is also observed for headphones [21][22].

While the collection of training data is the focus of this paper, the VLP model's training architecture will be briefly described first.
Finally, to validate the model predictions, results from a listening test with expert listeners will be compared to the VLP's output.

2 Model framework

Inspired by Manocha et al. [13], that proposed an approach called No-Reference Speech Quality Assessment via Pairwise Preference (SQAPP), we propose a CNN that learns its own prediction scale for audio attribute assessment. The model is trained using 4-second snippets of variable-Q transform power spectra from audio recordings. Essentially, we present a simple architecture that, during training, learns by comparing pairs of recordings, but during inference, it requires only single recordings to make predictions.

The VLP model consists of three main components: the *Backbone CNN*, *Attribute Encoder*, and *Rating Predictor*. The training architecture is illustrated in Figure 1.

Backbone CNN: To extract features from spectrograms of audio recordings, we use a pretrained ResNet-18 [24] as the model's backbone. The backbone choice is flexible and can be replaced to adjust the model's complexity.

Attribute Encoder: A multilayer perceptron (MLP) that converts a one-hot-encoded variable into a feature vector.

Rating Predictor: An MLP that maps the combined features from the backbone and attribute encoder to a scalar rating.

With the proposed architecture, we can train the model on various attributes, enabling it to discover non-linear relationships between them across different music genres - all without relying on a reference recording.

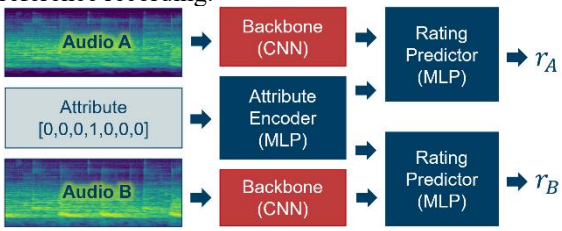


Figure 1) The training architecture of the VLP model using pairwise comparison.

3 Data collection

Data collection is a crucial component in constructing a reliable neural network model. The quantity and the quality of the data are balanced critical factors in generating accurate deep learning outputs. The high quality and volume of data were ensured using

[1] Preference mapping links consumer preference to the sensory characteristics of products. This approach provides a reliable basis for adapting or creating products to meet consumer expectations and optimize preference.

SenseLab's expert listener panel [8][9]. They undergo annually testing with a passing criterion for discrimination and reproducibility abilities to ensure they contribute with reliable data. Furthermore, they receive training in different test types, which includes audio characterization using the sensory descriptors from the Sound Wheel [7]. This results in agreement among the panel in understanding and rating the attributes. Every second year, high-frequency audiograms are measured to confirm normal hearing.

3.1 Experimental setup

Each expert listener was provided with a home-kit containing high-resolution, low-noise sound cards (FiiO K3) and high-fidelity headphones (Sennheiser HD650). Assessors were instructed to perform the listening tests only when feeling alert and when background noise at home was minimal. Each listening test was estimated to last approximately two hours, including breaks. A written test instruction was always included.

All participants rated attributes on a horizontal continues scale with verbal anchors in a multiple stimulus test method (cf. Figure 2), using SenseLabOnline – a web-based service for administering, performing and analyzing listening tests [6]. The corresponding attribute definition was present on the same screen.

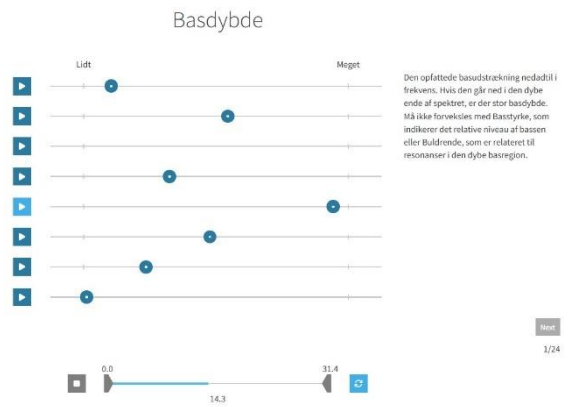


Figure 2) Screenshot example of the data collection process: Multiple stimuli test where experts provided their attribute ratings.

All stimuli in the listening test were loudness-aligned according to ITU-R BS.1770-4 [18]. Participants were instructed to adjust the level to the most comfortable using the sound card in the beginning of the session and to maintain that level throughout the test.

For the five attributes listed below, approximately 38,000 ratings comparing various hybrid and augmented stimuli were collected from experts. These ratings were subsequently transformed into approximately 159,000 pairwise comparisons for training the VLP model.

The Attributes and their corresponding definitions were selected from the existing Sound Wheel. The initial set of attributes used to train the model included:

1) Bass depth
2) Bass strength
3) Midrange strength
4) Treble strength
5) Brilliance

For each attribute, music samples were carefully selected to ensure suitability for evaluating the specific attribute. Consequently, the samples varied according to the specific attribute being assessed. However, pink noise was consistently included as one of the samples in all tests.

The samples used in the test were excerpts from original music tracks, typically 20-30 seconds in duration.

Although being classified as expert listeners, assessors were required to pass a gating test before participating in the actual data collection. To pass, they had to rate 4-6 systems in the correct order twice, with increasing difficulty. If they failed, they received immediate feedback and were given the opportunity to try again.

To ensure the highest data quality, additional evaluations were used for post-screening after the full dataset was collected. If the responses in these trials were incorrect (given in a wrong order), the corresponding dataset for that expert and attribute was excluded.
The test sequence followed this order:

- Gating (with immediate feedback)
- Post-screening
- Actual data collection for VLP
- Post-screening repeated

3.2 Stimuli generation

To optimize the variation in training material for the model, the variations were not entirely random. Measured impulse responses from real headphone

devices were convolved with the selected music excerpts. Additionally, various filters were applied to introduce a broad range of reasonable frequency response variations. Additional filter variations known to cause perceptual changes in the specific attribute were applied to achieve diverse ratings for the training material.
Based on our experience, having an assessor rate a single attribute for an extended period can lead to confusion and reduced rating accuracy. To mitigate this, anchors were frequently included throughout the test to stabilize scale usage.

Figure 3 shows an example of final frequency response variations used to generate stimuli for the data collection of *bass depth*. It is important to note that all experts had unique tests—no two assessors rated the same stimuli (except for the gating and post-screening). This approach maximized the variation observed by the model.

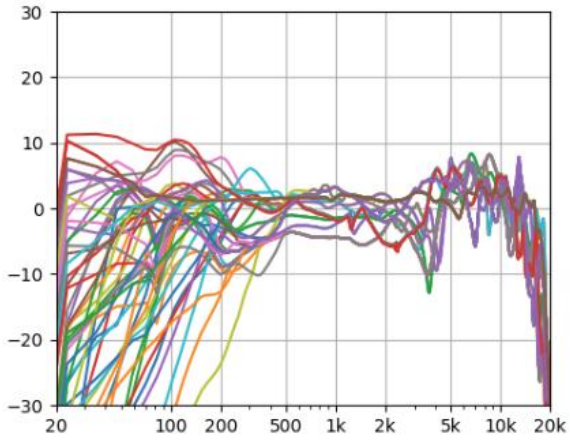


Figure 3) Example of filters that were used to process the music samples to generate stimuli for an assessor. The horizontal axis is frequency in Hz and the vertical axis is gain in dB.

4 Results

In this section, a separate listening test of earbuds is utilized as a validation dataset, to test the reliability of the VLP model.
The listening test was conducted by 20 assessors from SenseLab's expert panel in sound-insulated IAC listening booths using Sennheiser IE 900 in-ear headphones.
A total of seven midrange earbuds were recorded on a B&K 5128 HATS with five music samples of 20-30 sec length:

1. Argon Audio, IE20
2. boAt, Nirvana Ivy
3. Jlab, JBuds ANC 3
4. Samsung, Galaxy Buds FE
5. SkullCandy, Rall ANC
6. Soundcore, Liberty 4 NC
7. Urbanista, Palo Alto

None of the recordings have been used to train the VLP model. However, up to two of the five music excerpts were similar to those included in the VLP training data for certain attributes.
In the listening test a "reference" and "anchor" were included to encourage the experts to use most of the scale. The anchor was created by applying a bandpass filter with the passband frequency range of 100 to 7000 Hz. While the reference and anchor are included in the results section, it is important to note that the VLP remains a non-intrusive method. The model's predictions are unaffected by the presence or absence of a reference or anchor and yield identical results either way.

The results for each attribute are plotted, as shown in Figure 4-8. To ensure easy visual comparability, a min-max normalisation method is applied to both assessor ratings and VLP predictions. The output of the VLP is an arbitrary scale which enables relative comparisons (ranking and perceptual distance). While many models aim for prediction of the absolute rating, most or all data in perceptual audio evaluation is relative, although efforts to make scales used in a stable absolute sense has been made [25]. Because the output is relative, normalizing such that the ratings and prediction have the same min and max is needed.

The VLP generates scores in 4-second time slices at fixed intervals (time delta), e.g., every 1 second, resulting in multiple predictions across each sample. The error bars shown in the figures for the VLP predictions represent the standard deviation of the predictions across time and across all five music

excerpts, using a fixed interval (time delta) of 2 seconds.

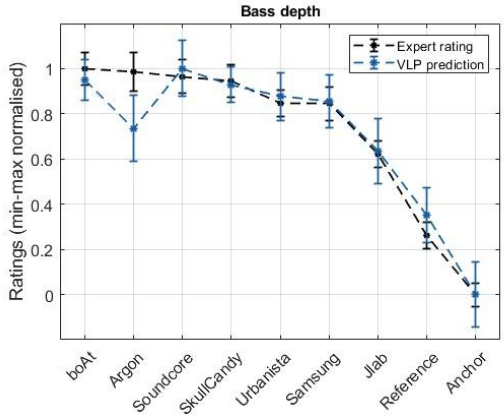


Figure 4) Bass depth evaluations. They are ordered according to the average expert ratings.

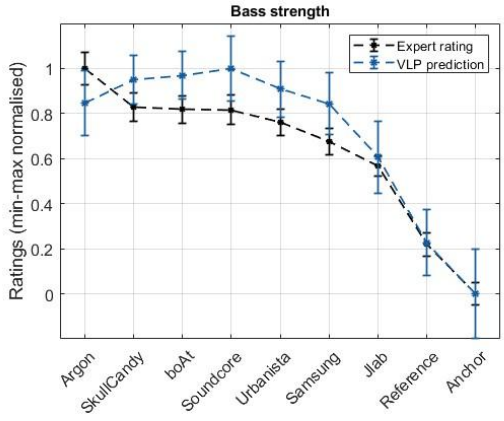


Figure 5) Bass strength evaluations. They are ordered according to the average expert ratings.

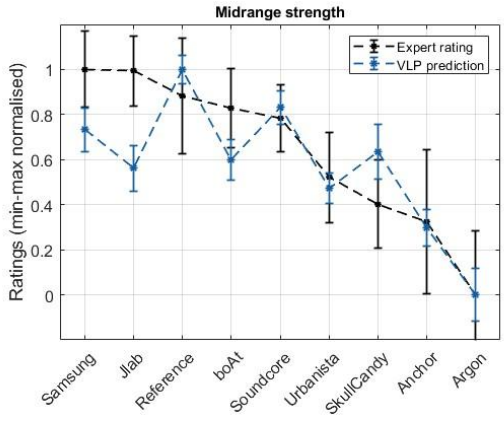


Figure 6) Midrange strength evaluations. They are ordered according to the average expert ratings.

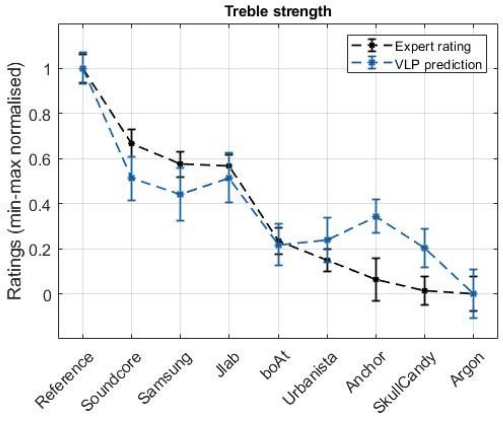


Figure 7) Treble strength evaluations. They are ordered according to the average expert ratings.

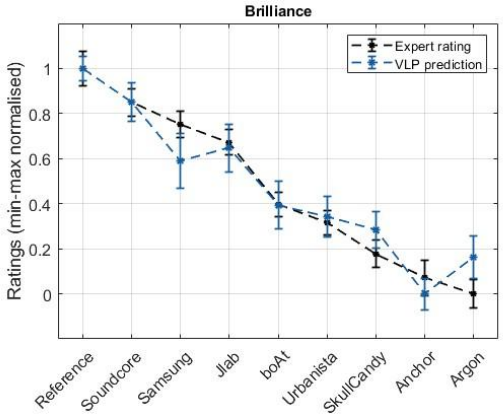


Figure 8) Brilliance strength evaluations. They are ordered according to the average expert ratings.

	Pearson	Spearman	RMSE	RMSE*
Bass depth	0.96	0.80	0.098	0.029
Bass strength	0.95	0.77	0.135	0.030
Brilliance	0.97	0.97	0.094	0.020
Treble strength	0.92	0.88	0.145	0.041
Midrange strength	0.80	0.63	0.219	0.042

Table 1. Results of the predictive VLP model with root mean square error (RMSE) and a modified version (RMSE*), Pearson-, and Spearman correlation coefficient.

In Table 1, the accuracy of the prediction model is evaluated using correlation coefficients and the root mean square (RMSE) evaluation metric. Furthermore, modified epsilon-insensitive rmse

(RMSE*) values, which take the uncertainty of the listening test scores into account [19], are included. To calculate the RMSE and RMSE* values, the rating and prediction scores are min-max normalised.

5 Discussion

The purpose of this study was to investigate whether it is possible to train a neural network model for the use of audio characterisation. Controlled laboratory listening test and maintenance of an expert panel for such characterisations are both costly and time consuming. In SenseLab we are in the unique position to have a trained listener panel. The CNN model's performance is illustrated in Figure 4-8 and summarized in Tabel 1: The comparative analysis demonstrates a high validity and performance of the prediction model. The prediction model's validity for midrange strength evaluations is inferior compared to the other attribute predictions. However, the experts' ratings from the listening test, also exhibit greater difficulty in reaching consensus on midrange strength ratings, as seen by the relatively wider confidence intervals compared to the other attribute evaluations. Hence, the RMSE* value reduces significantly compared to the traditional RMSE value, due to the uncertainty from the expert ratings itself. Midrange strength can be a difficult attribute to rate compared to bass- and treble strength. So, the larger disagreement with expert ratings could be due to difficultly in rating this attribute, both from the listening test itself but also from the training material collected for the VLP model.

While predictions of some attribute ratings can certainly be improved, the VLP is already useful at this performance level. One example is for selection of music samples for listening test; Here the VLP can predict which will give the largest differentiating between products and where the excerpt is stable enough to provide similar ratings across assessors focusing on different parts of the excerpts.

In machine learning in general, data augmentation is an important tool, because data collection can be expensive. However, within audio, even small changes to the stimuli might lead to audible differences that would lead to a different rating. The lack of published knowledge on Just Noticeable Differences of complex stimuli makes these decisions harder. Using the trick of presenting stimuli in a multiple stimulus method but converting them to paired comparison ratings is a way of multiplying the collected data with a significant factor. Although increasing the number of ratings rather than the number of stimuli, it is still very valuable. When combined with the pooled attribute data model architecture, it makes it possible to train an attribute prediction model based on one (large) test, rather than dozens or hundreds of tests, assuming excellent data quality. With each new attribute we collect data on, predictions for other attributes are likely to improve as well, as the model continuously refines its ability to identify non-linear relations – mirroring the auditory system. Currently, the model has been trained on three more attributes—timbral balance, nasal and boxy—which needs to be validated with expert listening test data.

6 Concluding remarks

In this paper, we propose a method for efficient collection of high-quality data for an audio characteristic prediction model. Starting with five timbral aspects, known to be an important domain in terms of drivers of liking. The Pearson correlation ranged from 0.80-0.97, when tested on a dataset from an independent expert listening test. To maximize the data quality, we implemented the following measures:

- Used trained listeners (experts) to provide the training material for the audio characterisation prediction model, VLP.
- Conducted data collection with known and identical hi-fi equipment.
- Required passing a gating test and post-screening.
- Included reference stimuli to anchor scale usage.
- Used carefully selected music excerpts matched to each attribute.
- Created relevant, audible and varied variations in stimuli.

The scope of the model in this study was tested only with expert data from recordings of earbuds. It is still unknown how well the model performs with recordings of other audio devices, such as loudspeakers.

References

[1] A. W. Rix, J. G. Beerends, et al., “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in ICASSP, 2001, vol. 2, pp. 749–752.

[2] J. G. Beerends, C. Schmidmer, et al., “Perceptual objective listening quality assessment (POLQA), the third generation itut standard for end-to-end speech quality measurement part i—temporal alignment,” Journal of the AES, no. 6, 2013

[3] A. Hines, J. Skoglund, et al., “ViSQOL: An objective speech quality model,” EURASIP, vol. 2015, no. 1, 2015.

[4] Manocha, A. Finkelstein, et al., “A differentiable perceptual audio metric learned from just noticeable differences,” Interspeech, 2020.

[5] P. Manocha, Z. Jin, et al., “CDPAM: Contrastive learning for perceptual audio similarity,” ICASSP 2021.

[6] “SenseLabOnline” (version 5.0) [Online]. Available: https://forcetechnology.com/en/services/acoustics-noise-sound-quality/senselabonline

[7] SenseLab’s “Sound Wheel” [Online]. Available: https://forcetechnology.com/en/services/acoustics-noise-sound-quality/senselab-sound-wheel

[8] S. V. Legarth and D. N. Zacharov, "Assessor Selection Process for Multisensory applications," in *Audio Engineering Society*, München, Germany, 2009.

[9] ISO 8586-2, “Sensory analysis — General guidelines for the selection, training and monitoring of selected assessors – Part 2: Experts”. International Standard ISO

[10] T.H. Pedersen, “The Semantic Space of Sounds”, 2008.

[11] M. Torcoli et al., “ODAQ: Open Dataset Of Audio Quality,” Dec. 2023, doi: 10.5281/zenodo.10405774.

[12] Mittag, Gabriel & Chehadi, Assmaa & Möller, Sebastian. (2021). NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets.2127-2131.10.21437/Interspeech.2021-299.

[13] P. Manocha, Z. Jin, and A. Finkelstein, “SQAPP: No-Reference Speech Quality Assessment Via Pairwise Preference.” [Online]. Available: https://pixl.cs.princeton.edu/pubs/Manocha

[14] A. H. Andersen, J. M. de Haan, Z. -H. Tan and J. Jensen, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," in *IEEE/ACM*, vol. 26, no. 10, pp. 1925-1939, Oct. 2018.

[15] Report ITU-R B.2399-0, “Methods for selecting and describing attributes and terms, in the preparation of subjective tests”, ITU 2017.

[16] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, “Non-intrusive Speech Quality Assessment Using Neural Networks,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., May 2019, pp. 631–635.

[17] ETSI EG 202 396-3, “Speech and multimedia Transmission Quality (STQ); Speech Quality performance in the presence of background noise; Part 3: Background noise transmission – Objective test methods”, 2018.

[18] Report ITU-R B.1770-4, “Algorithms to measure audio programme loudness and true peak level”, ITU 2015.

[19] Recommendation ITU-T P.1401, “Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models”, ITU 2020

[20] P. Manocha, B. Xu, and A. Kumar, “NORESQA: A Framework for Speech Quality Assessment using Non-Matching References,” Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.08125

[21] S. E. Olive and T. Welti, “The Relationship between perception and Measurement of Headphone Sound Quality,” 2012, [Online]. Available: www.aes.org

[22] G. Ravizza, J. Villegas, C. P. Volk and Tore Stegenborg-Andersen, “An Over-Ear Headphone Target Curve for Brüel & Kjær Head And Torso Simulator Type 5128 measurements” in 155th Convention Express Paper, 2023.

[23] F. Rumsey, S. Zieliński, R. Kassier, S. Bech, “On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality”, 2005, The Journal of the Acoustical Society of America, vol. 118.

[24] K. He, X. Zhang, S. Ren & J. Sun, “Deep Residual Learning for Image Recognition”, Microsoft Research, 2015, [Online] Available: https://arxiv.org/abs/1512.03385

[25] Lim J, Wood A, Green BG, “Derivation and evaluation of a labeled hedonic scale”, 2019, Chem Senses, 34, 739-751.