Audio Engineering Society

Convention Paper 10488

Presented at the 150th Convention
2021 May 25–28, Online

Efficient data collection pipeline for machine learning of audio quality

Christer P. Volk[1], Jon Nordby[2], Tore Stegenborg-Andersen[1], and Nick Zacharov[1]

[1]*FORCE Technology, SenseLab, 2970 Hørsholm, Denmark*
[2]*Soundsensing, 0349 Oslo, Norway*

Correspondence should be addressed to Christer P. Volk (cvo@force.dk)

ABSTRACT

In this paper we study the matter of perceptual evaluation data collection for the purposes of machine learning. Well established listening test methods have been developed and standardised in the audio community over many years. This papers looks at the specific needs for machine learning and seeks to establish efficient data collection methods, that address the requirements of machine learning, whilst also providing robust and repeatable perceptual evaluation results. Following a short review of efficient data collection techniques, including the concept of data augmentation and introduce the new concept of *pre-augmentation* as an alternative efficient data collection approach. Multiple stimulus presentation style listening tests are then presented for the evaluation of a wide range of audio quality devices (headphones) evaluated by a panel of trained expert assessors. Two tests are presented using a traditional full factorial design and a pre-augmented design to enable the performance comparison of these two approaches. The two approaches are statistically analysed and discussed. Finally, the performance of the two approaches for building machine learning models are reviewed, comparing the performance of a range of baseline models.

1 Introduction

The evaluation of audio quality is an important element of product design, often studied both objectively and subjectively. Audio quality can be estimated without listening tests, by the use of objective metrics for specific application areas. The International Telecommunication Union (ITU) has standardized ITU-R Rec. BS.1387 "PEAQ" [1] and ITU-T Rec. P.863 "POLQA" [2]. Recently, approaches have used more advanced modelling techniques, such as deep learning to estimate audio quality based on spectrogram or audio waveforms [3, 4, 5].

In our field, there is a long tradition of performing full factorial within-subjects (FFWS) designs for listening tests, an overview of which can be found in [6]. The data from such FFWS experiments is often considered to yield some form of *ground truth* performance of the technologies under study. In such experiments it is common to select a few clips of samples or sound tracks, typically ∼10-30s in duration, and process them through the systems under evaluation to be subsequently perceptually evaluated. This results is some 5-20 clips being used to generalise system performance. For machine learning this is considered a very small amount of audio training data, which easily leads to over-training of the model specifically for the traits of these samples and thus lacking the desired generalised applicability to the broader range of audio. Training ML algorithms benefit from a large and varied

set of audio samples. But how can this data hunger be catered for without significantly increasing the size, complexity and cost of such listening test?

To be better equipped to train ML models, we would need a perceptual evaluation protocol that when compared to traditional listening tests ideally: 1) Yields similar results, 2) Is no larger, 3) Cost no more, 4) Provides more (audio) data.

This paper introduces a proposed machine learning pipeline for recording-based listening tests and includes the novel concept of *pre-augmentation*. It covers investigations of the influence of this approach on the statistical analysis in comparison to a traditional full-factorial listening test design and, finally, the impact on machine learning performance on the basis of these two approaches is investigated and discussed.

2 Efficient data collection approaches

Moving beyond full factorial design of experiments, there are several approached to more efficiently gather data from listening tests.

The field experimental design or efficient *design of experiment* (DoE) is well established on certain field study due to measurement constrains or the high number of test conditions. These methods are however rarely employed in the audio field (outside telecommunication), potentially due to the lack of necessity.

The statistics of design of experiment originated largely from the work of Sir Ronald Fisher in the 1920's and 30's Fisher [7]. The full factorial design provides for all conditions (combinations of experimental factors) in an experiment to be *measured* and as such is a sure manner to ensure a *good* experimental design, assuming there is sufficient statistical power. In our domain a full factorial within-subjects design would comprise processing all test samples through all system under test and presenting them to all assessors in a balanced and randomised manner.

By comparison any design of experiment (see e.g. Montgomery [8]) aims to reduce the size of the experiment in some manner such that we can either run a larger experiment, i.e. with more test conditions, or that the test can be run with less time and effort. In either case, a partial set of conditions are tested or presented to each assessor. Thus when employing DoE, we are seeking to gain the same level of statistical analysis and interpretability of the data, but with less effort. This is only possible by making some compromises that need to be understood. With traditional DoE structures such as balanced incomplete block designs or D-optimal designs [9] the aim is to maintain the statistical power of the independent variables of interest in the experiment. The trade-off is made by compromising the statistical power of higher order interactions of independent variables or to confound them, such that they cannot be analysed with any certainty. Audio examples of efficient DoE using the *response surface method* (RSM) can be found in Lorho [10] and other examples of using IV-optimal design to handle very large numbers of test conditions can be found in Zacharov and Schevciw [11, 12].

More recently, we have seen the development of alternative methods for efficient data gathering, most often tied to machine learning applications. The concepts of *adaptive sampling* or *active learning* are efficient methods for gaining as much information from a stimulus data set without needing to test all of the conditions. These techniques have been successfully employed with the paired comparison paradigm to interactively select the next pair of stimuli based on the the prior and current stimuli (see e.g. [13, 14, 15]). The aim of this approach is to gain the maximum amount of information from the data, in as few trials as possible, achieved through adaptive interactive modelling of the data during the sampling/measurement. Adaptive sampling has been successfully applied to the evaluation and tuning of audio quality in hearing aids by Nielsen et al. [13]. Active learning has also been extensively researched and successfully applied to video quality testing [14, 15].

3 Pre-augmentation concept

When performing listening tests, it is typical that we want to span the range of stimuli to stress and test products as completely as possible to be able to identify differences and key areas of weakness. Also such tests are meant to be a generalisable representation of "all audio material". This is quite a tall order and furthermore needs to be done efficiently, e.g. within a 2-hour listening test. To do this, usually 10-20 audio clips of short duration ($\sim$15 s) are selected from different audio tracks, as illustrated on left the of Figure 1. The results from such listening test are considered to yield the ground truth performance for the devices under test,

even with less than 5 minutes of original audio material. From the perspective of machine learning, this is a very small and limited amount of audio data to train on, and does not allow for generically robust models to be build. Furthermore, deep-learning (DL) types of approaches are even more data hungry. In application to DL models, methods such as data-augmentation [16] are used to ensure the robust and generic applicability of models.

Data augmentation (DA) comprises of slightly modifying the audio files, in a manner that would leave the listening test scores unaffected and thus creating a different new version of the same audio file for training DL models. DA can take many forms, for example time-shifting, re-sampling, etc. and it has generally been found that 16x data-augmentation can be a valuable way of making DL models more generically robust, by feeding the learning process with more samples. The challenge for regression DL problems is that augmentations must be large enough to matter and small enough to assume that the change would not lead to a difference in rating. Consequently, changes might be too small to improve performance/robustness of a regression model. Our concept sought to address the question of how to get more audio samples into a listening test without the size and complexity of experiments growing excessively.

We thus developed and introduce the concept of *pre augmentation* as a way of using more audio files in a listening tests, without increasing the size and duration of the listening test. The idea is to select several clips from each audio track, as illustrated on the right of Figure 1 thus increasing the amount of audio data available. We hypothesis that many of the sound quality characteristics of a track remain largely constant across the track and thus across several selected clips. To a large extent the spectral balance, spatial nature, loudness and dynamics of samples are tied to the nature of the track and its production. Additionally, by sampling across several places in a track, the finer nuances are also evaluated, providing a more generalisable perspective.

In a traditional full factorial, within-subjects design, as traditionally employed in listening test, all samples are processed though all test systems and these resulting conditions are then presented to all assessors in the listening test. With sample augmentation, using a full factorial approach would increase the experiment size by the number of clips, leading to a manageable experiment size. The alternative approach we selected and sought to test was to use a factorial design, whereby one set of clips is presented to a subset of assessors, in a balanced incomplete block design (BIBD). Samples and clips are distributed using a Latin square design approach to each of the assessor groups to yield a balanced design, as show in Figure 1.

4 Audio capture

For maximising machine learning data input a audio capture protocol was established. This was done to ensure that input audio would be suitable for machine learning purposes. Input audio for machine learning and listening test was captured using a B&K head-and-torso-simulator (HATS 5128-C) in an effort to capture the same stimuli as was heard by listeners to maximise the direct link between stimuli and listener responses [17]. Using a HATS recording technique furthermore allows capture of most audio product types in the same audio format, which has many benefits including pre-training of models across products types and using transfer learning [18, 19, 5] techniques to allow further fine-tuning to specific product types with smaller data sets.

The recording process was handled by a custom Python script, *BenREC*, which combines all audio clips in an input folder into one .wav file, plays and records it simultaneously, and splits each recording into separate .wav files (for each original audio clip) based on either automatically added pre-stimuli makers (e.g. a short 10 kHz burst or a short sweep) or cross-correlation analysis of the original audio clip and the recording. Using an automated script allowed efficient unmanned capture of large amounts of audio.

5 Experimental setup

5.1 Product selection

Headphones was selected as the product category for evaluation. The headphones were selected to generate a broad spectrum of stimuli both in terms of overall audio quality and in terms of various perceptual differences (i.e. distortion, frequency characteristics etc.). In-ear, on-ear, and over-the-ear headphones; the latter two in both open-back and closed-back versions were included in the study. An overview of the different product types is shown in Table 1.

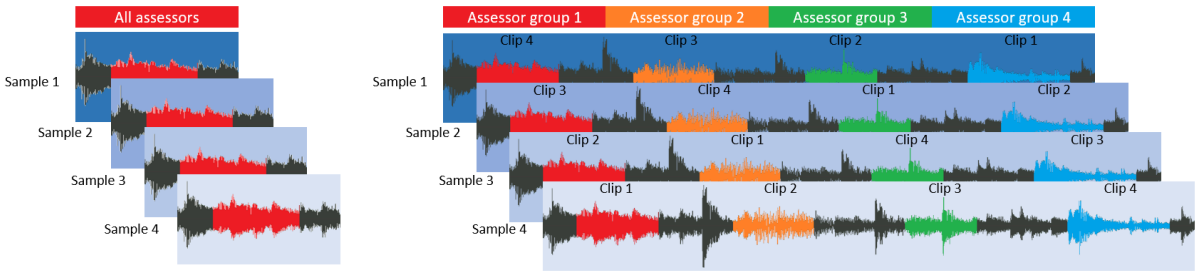

Fig. 1: Illustration of the usage of sound clips per sample for. *Left*: one clip per sample as commonly employed. *Right*: multiple clips per sample, as per pre-augmentation method.

Type	Number
On-ear	4
In ear	8
Over-the-ear open back	1
Over-the-ear closed back	7

Table 1: Summary of employed headphone types.

A total of 20 headphones plus two anchor systems were included - mid and low anchor. The mid anchor system was generated using recordings of a poorly performing headphone, which were degraded further using dynamic compression, band-pass filtering and by adding two resonances to create the low anchor.

5.2 Sample selection & clip extraction

Samples were selected to cover a wide range of musical genres and production styles, as well as to excite as many perceptual differences as possible in the devices under test. A total of 15 samples were included, covering female and male voices, contemporary pop, heavy rock, classical, jazz, electronic, hip-hop, instrumental music and known critical musical samples.

Extracted clips from samples typically vary in length between 15-35s. For these experiments 15s clips were carefully manually selected, to be representative, similar, perceptually stationary and loopable.

5.3 Assessors

Twenty of SenseLab's trained expert assessors participated in the study [20]. Each assessor attended two sessions of two hours, including instructions, familiarisation, breaks etc. Two assessors were female, 18 were male. The median age was 36.9 years, the minimum and maximum age were 19.5 years and 58.8 years respectively.

All assessors were normal hearing and regularly tested as part of the routine of being employed in our trained assessor panel.

5.4 Experimental setup

The test method and test design was configured in SenseLabOnline which also handled all assessor invitations, randomisation of stimuli presentation order, audio playback, and data collection.

Assessors listened through Sennheiser HD650 headphones, connected to RME ADI-2 DAC sound cards. The test took place in IAC listening booths with background noise levels $\leq$ NR15.

5.4.1 Post-processing of recordings

Recordings were processed 1) to remove the frequency colouring of the HATS' ear canal, 2) the influence of the playback headphones used in the listening test, and 3) to loudness align across headphones. The inverse filter applied to compensate for the influence of the HATS's ear canal as well as the playback headphones was made using the AKtools Matlab toolbox by Brinkmann and Weinzierl [21]. A 16384-sample Minimum-phase FIR filter with automatic time-alignment and 1/6-octave regularisation of both input- and output frequency response. The filter was designed to compensate within the frequency range 16 Hz - 16 kHz and based on multiple and repositioning of multiple models of our Sennheiser HD650 headphones.

All recordings were analysed using ITU-R Rec. BS.1770-3 [22] and adjusted to the same loudness level

and lastly all audio files were renamed using the naming scheme *condition_sample_system* for automatic upload in SenseLabOnline.

6 Full factorial experiment

A traditional full factorial test was designed and build comprising: 23 systems x 15 samples = 345 unique wav files. Due to the blocking of the experiment, we had (20+4*3) systems * 15 samples * 2 groups/repetitions = 960 as input to the machine learning model. Each clip was rated independently by two groups of assessors.

7 Sample augmentation design

For our sample augmentation experiment, we carefully selected four clips from each of the 15 selected critical audio samples. Each clip was recorded twice with a headphone repositioning in-between to end up with eight conditions per sample. The 20 products were divided into four blocks and the panel of assessors divided into four groups; which evaluated either different clips or clips in different orders. Altogether this led to the fractional factorial design illustrated in Table 2. Note that this design deviated slightly from the description as clip 5 was replaced with clip 1 in Group 2 and Group 3 in order to allow for comparison of assessor performance between groups for this particular experiment. This deviation is not part of the proposed pre-augmentation method in general. The advantage over the traditional full-factorial design is that this design includes eight times the number of audio files without an increase in test duration and with the potential of being able to analyse data as if it was a full-factorial design by considering clips from the same sample as one.

	Group 1 (n= 5)	**Group 2 (n= 5)**	**Group 3 (n= 5)**	**Group 4 (n= 5)**
Block 1: Products 1-5	Clip 1	Clip 6	Clip 8	Clip 4
Block 2: Products 6-10	Clip 2	Clip 1	Clip 7	Clip 3
Block 3: Products 11-15	Clip 3	Clip 7	Clip 1	Clip 2
Block 4: Products 16-20	Clip 4	Clip 8	Clip 6	Clip 1

Table 2: Example fractional design for sample augmentation design of experiment. *n* refers to the number of assessors.

8 Results

8.1 Assessor normalisation

Assessors typically use rating scales differently depending on their expectations and experience. This is well-established and includes rating-effects related to level, range, variability, etc. [23]. Due to the fractional-factorial experimental design, assessors did not evaluate the same stimuli between assessor groups, which may increase the size of these effects. This is something that is seen in the data (see Figure 2 (Left)). Consequently, a rating normalisation was utilised.

Specifically the method described by Athar et al. [24], which normalises data based on the original scale usage range of each individual assessor and the scale usage range of each set of headphones across all assessors. The method assumes normal distributions. It does not exactly normalise back into the original scale, but the output range is similar. For this particular data set the normalisation reduces the 95% confidence intervals significantly, as depicted in Figure 2, and thereby improves the statistical power of further analyses.

8.2 Overall results

The overall results are summarised in Figure 3 as normalised mean Basic Audio Quality ratings of the headphones with 95%-confidence intervals. The headphone means span the majority of the scale (which isn't min-max normalised), making it suited for machine learning. The majority of headphones have overlapping mean ratings across the three data sets: PreAug-set1, PreAug-set2, and Full (-factorial), suggesting that effect of pre-augment on ratings is minor, although the differences in mean ratings are large enough to affect system ranking.

The artificial *AnchorMid* system designed to anchor scale usage at the mid section of the scale is degraded too much - rated lower than any of the headphones and could have been left out or replaced by either Beats Studio 3 or Apple generic concha EarPods, which are at approximately 50 and without influence from the difference in clips between data sets.

8.3 Pre-augmentation influence on ratings

The ideal sought after state would be that the pre-augmented data provides identical ratings to the full

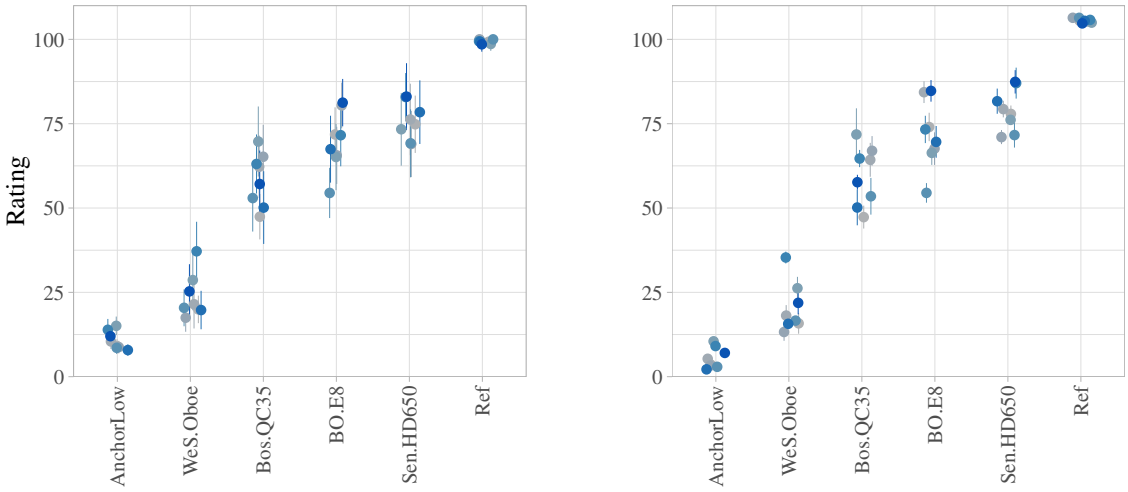

Fig. 2: Mean values with 95% confidence intervals of a representative subset of headphones and samples. Average across assessors and clips. Colours represent the sample (a subset of eight selected here and separated using random jitter). *Left*: Before normalisation. *Right*: After normalisation.

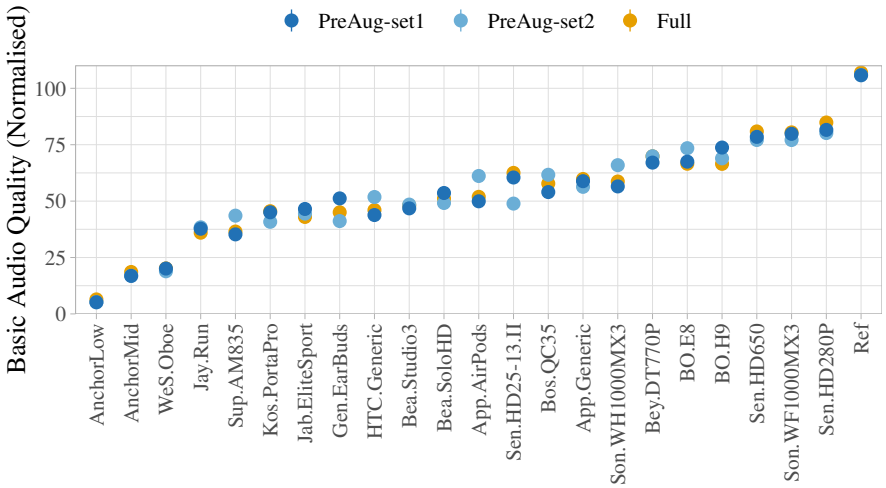

Fig. 3: Mean ratings with 95% confidence intervals (barely visible) for the full-factorial data set ("Full") and the pre-augmentation data set. The pre-augmentation data set was split into two subsets; Each comparable to the full factorial data set.

	Df	Mean Sq.	F-value	p-value
System	22	411895	8718.75	<2e-16
Sample	14	5804	122.85	<2e-16
System:Sample	308	770	16.29	<2e-16
Sample:Clip	90	483	10.23	<2e-16
Residuals	9146	47		

Table 3: Mixed-effect ANOVA with Assessor as a random effect and Clip nested within Sample.

factorial design, when averaged across the panel. i.e. the inference on the mean should be identical between clips, to enable us to analyse all clips as if they were one, i.e. most or all of the clips' confidence intervals should overlap with the CIs of *RefClip1*. When we study Figure 3 we can observe that overall there is close similarity between the pre-augmented data and the full factorial data, but they are not identical. This difference can be further studied in Figure 4. The statistical influence of clips was investigated in a mixed model ANOVA with assessor as a random effect and Clip nested within Sample. The result, shown in Table 3, illustrates that Clips nested interaction with Sample is highly significant, but that it's influence (F-value) is very minor in comparison to the main effects for System and Sample. Whilst this result deviates from the expected, it remains to be seen whether this is an issue in general. To confidently conclude upon this matte, a larger experiment would be needed with more assessor per group.

9 Machine learning

The main purpose of using pre-augmentation is to increase the number of audio files available for training of a machine learning model. To investigate the influence of this approach, machine learning models were trained on all three data sets to test the hypothesis that an increased amount of audio files will improve the performance of a machine learning model.

The machine learning models were of the type *full-reference/intrusive*, which utilise that the listening test included comparison to reference, i.e. all models have both a reference clip and a corresponding stimuli (headphone recording) as input. The models were trained with supervised learning, using the normalised ratings from the listening test as targets.

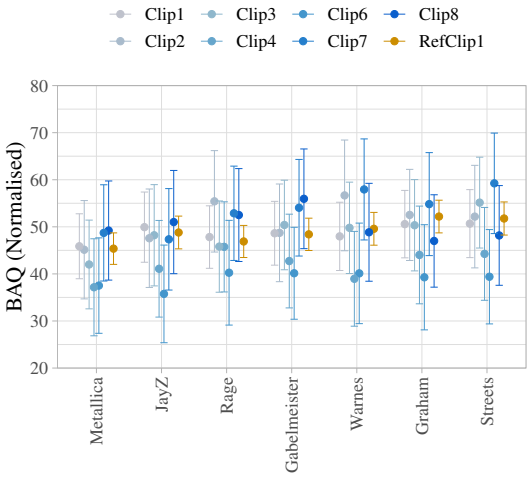

Fig. 4: Comparison of the averaged ratings across systems for seven of the fifteen samples. *RefClip1* is the average from the full factorial test and the remaining clips are from PreAug-set1.

In the following subsections a deep-learning model is described and its performance compared for the full-factorial data set and the pre-aug data sets. A low-complexity machine learning model was chosen to accommodate the modest-sized data sets. This limits overall performance to a point where other machine learning approaches can outperform the model used, but was found well-suited for evaluating the potential of the pre-augmentation approach.

9.0.1 Common machine learning pipeline & model

For each audio file, a log-mel-spectrogram is computed with 32 mel-bands, a window length of 2048 samples (46.44 *ms*), and hop length of 512 samples (11.61 *ms*) between frames. The *residual* of the spectrogram was computed by subtracting the spectrogram of the reference audio. This approach works when the data is well time-aligned, which is a standard criteria for multiple stimulus test types to allow instantaneous switching between stimuli.

The machine learning models were implemented in Python using the user-friendly Keras Deep-learning API on top of TensorFlow. A model type of structure

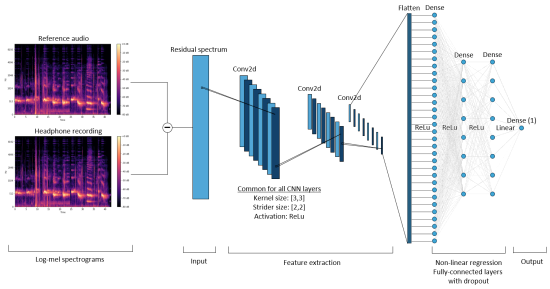

Fig. 5: Overview of the convolutional neural network (CNN) employed

Model	Eval.	RMSE	Pearson	Spearman
Full	Clip	19.5 ± 4.3	0.52 ± 0.26	0.48 ± 0.25
Full	Sys	14.6 ± 5.1	0.61 ± 0.39	0.52 ± 0.39
Pre-aug 1	Clip	16.2 ± 4.9	0.66 ± 0.24	0.64 ± 0.24
Pre-aug 1	Sys	11.8 ± 4.8	0.73 ± 0.31	0.62 ± 0.41
Pre-aug 2	Clip	15.4 ± 4.3	0.71 ± 0.21	0.65 ± 0.22
Pre-aug 2	Sys	$\mathbf{11.6 \pm 5.0}$	$\mathbf{0.80 \pm 0.24}$	$\mathbf{0.69 \pm 0.30}$

Table 4: Results of predictive models with root mean square error (RMSE), Pearson-, and Spearman rank correlation coefficients with standard deviations.

was chosen on the basis of approaches applied successfully in the literature to similar audio tasks [25, 26, 27], i.e. using the residual spectrogram as input to a Convolutional Neural Network (CNN), as illustrated in Figure 5. The CNN consisted of three convolutions blocks, followed by three fully-connected layers. Each convolutional layer has eight channels and uses rectified linear unit (ReLu) non-linearity activation functions, while the last layer is a fully-connected layer with linear activation.

The model performance evaluation was done with 20-fold cross-validation, where folds were stratified on system (headphone), such that stimuli from the same system does not appear in both training- and test split. In each fold, approximately 25% of the systems were used as test data.

9.1 Performance comparison

The performance of the models trained on either the full-factorial data set ("Full") or the pre-augmentation data sets is presented in Table 4. All three data sets have the same total number of perceptual evaluations and the same systems and samples. "Pre-aug 1" was evaluated by the same assessors as "Full", while "Pre-aug 2" was evaluated by another group of assessors without overlap. Pre-augmentation provides an average **18%** improvement of prediction accuracy compare to the full factorial experiment.

10 Discussion

The purpose of this study was to investigate how a traditionally designed perceptual audio evaluation might be modified to better accommodate the data requirements of machine learning. Controlled laboratory listening tests are costly and thus reducing cost and complexity is desirable.An important target with the pre-augmentation concept was to improve the value of the collected data for the machine learning purposes without introducing extra complexity in the statistical analysis. The machine learning performance of input from pre-augmentation vs a full-factorial design, is summarised in Table 4: It indicates that an improvement is obtained by using pre-augmentation for all performance metrics. Thus using the same size of perceptual evaluation data with pre-augmentation, it is possible to increase the prediction performance using machine learning. Even though the overall prediction performance is insufficient for our application the improvement is noteworthy and makes pre-augmentation worth exploring further in future studies. The generally large standard deviations is a consequence of the system-stratified split into training and test subsets. With only 22 systems in total, the split can have a large effect on performance, depending on which systems are missing in the training set, when assigned to the test set. This variation does not seem to be reduced in the two data sets with pre-augmentation, but a large performance variation in standard deviation per model re-run is observed making it difficult to conclude on without additional data.

As discussed in section 8.3 pre-augmentation experiments can be performed at the same size as traditional listening tests, with n-times the number of audio files. This yields a very significant improvement in machine learning performance, discussed later in this section. The presently proposed pre-augmentation design would require further refinement to yield more stable and re-

peatable results compared to traditional full factorial experiments. For example, increasing the number of assessors per clip, more broadly distributing the clips across assessor groups and invoking a more elaborate design of experiment might yield more robust results, without a major cost/complexity penalty.

Furthermore, from the data we can see that not all clips are equivalent. In some respects this is a benefit for machine learning, but leads to less similar assessor data compared to the full factorial, single clip design. On the one hand we may seek to select more similar clips, to strive towards closer similarity to the traditional full factorial design results. Alternatively, we might consider that selecting different clip types from a sample beneficial to bring a more varied selection of audio into the listening tests and for machine learning. This begs the question of *what is the ground truth data from a listening test?*

11 Conclusion

This study proposes a method for collecting perceptual audio evaluation data requiring little extra effort in the workflow of product evaluation, while optimising the suitability of the data set for machine learning. It proposes a well-defined method of recording stimuli suited for most devices (e.g. headphones, hearables, hearing aids, loudspeakers, advanced sound systems, etc.) and conditions, while resulting in stimuli with identical binaural audio files. Furthermore, the novel concept of *pre-augmentation* is introduced, which entails using multiple clips from each included sample in a fractional-factional experimental design, which avoids the risks involved in traditional data augmentation for regression problems. A workflow with an automatic recording process is described, allowing recording and post-processing of a large set of stimuli efficiently to ensure a cost-effective method. An assumed benefit of pre-augmentation of statistical analysis is that data can be analysed as if all assessors evaluated the same clip from each sample. This assumption is tested and discussed and finally a simple machine learning model is described and its modelling performance with and without pre-augmentation presented. A small decrease in modelling error (RMSE) and a larger increase in modelling robustness was found as shown in Table 4.

11.1 Acknowledgements

This work was partially funded by The Danish Council for Technology and Innovation. The authors thank the SenseLab expert assessor panel for their efforts in this work. Thor Bundgaard Nielsen is thanked for his contributions to the project, especially his involvement in establishing the recording protocol.

References

[1] ITU-R, "Method for objective measurements of perceived audio quality," Recommendation ITU-R BS.1387-1, International Telecommunication Union Radiocommunication Assembly (ITU-R), Switzerland, 2001.

[2] ITU-T, "Perceptual objective listening quality assessment," Recommendation ITU-T P.863, ITU Telecommunication Standardization Sector (ITU-T), United States, 2011.

[3] Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M., "Fr\'echet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms," *arXiv:1812.08466 [cs, eess]*, 2019.

[4] Manocha, P., Finkelstein, A., Zhang, R., Bryan, N. J., Mysore, G. J., and Jin, Z., "A Differentiable Perceptual Audio Metric Learned from Just Noticeable Differences," *Interspeech 2020*, 2020, doi:10.21437/interspeech.2020-1191.

[5] Manocha, P., Jin, Z., Zhang, R., and Finkelstein, A., "CDPAM: Contrastive learning for perceptual audio similarity," 2021, _eprint: 2102.05109.

[6] Bech, S. and Zacharov, N., *Perceptual Audio Evaluation - Theory, Method and Application*, John Wiley & Sons, Ltd, 1st edition, 2006, ISBN 978-0-470-86923-9, doi:10.1002/9780470869253.ch1.

[7] Fisher, R. A., *Design of Experiments*, Oliver and Boyd, 1935.

[8] Montgomery, D. C., *Montgomery: Design and Analysis of Experiments*, Wiley, 2017.

[9] Ozdemir, A., *Development of the D-Optimality-Based Coordinate-Exchange Algorithm for an Irregular Design Space and the Mixed-Integer Nonlinear Robust Parameter Design Optimization*, Phd thesis, Clemson University, 2017.

[10] Lorho, G., "Subjective evaluation of headphone target frequency responses," in *126th Audio Engineering Society Convention 2009*, volume 3, pp. 1575–1594, 2009, ISBN 9781615671663.

[11] Zacharov, N. and Schevciw, A., "Procedure for identification of optimal handset receive mask for coded speech," Contribution S4-140069, 3rd Generation Partnership Project (3GPP), 2014.

[12] Zacharov, N. and Schevciw, A., "Frequency Response masks for EVS," Contribution S4-141209, 3rd Generation Partnership Project (3GPP), 2014.

[13] Nielsen, J. B. B., Nielsen, J., and Larsen, J., "Perception-based personalization of hearing aids using gaussian processes and active learning," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 23(1), pp. 162–173, 2015, ISSN 23299290, doi:10.1109/TASLP.2014.2377581.

[14] Jiang, Y., Xu, Q., Zhang, W., and Huang, Q., "Active Sampling for Subjective Video Quality Assessment," in *2018 IEEE 4th International Conference on Multimedia Big Data (BigMM 2018)*, 61672514, pp. 11–15, IEEE, 2018, ISBN 9781538653210, doi:10.1109/BigMM.2018.8499064.

[15] Li, J., Mantiuk, R. K., Wang, J., Ling, S., and Le Callet, P., "Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation," in *32nd Conference on Neural Information Processing Systems (NIPS)*, Canada, 2018.

[16] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-September, pp. 2613–2617, 2019, doi:10.21437/Interspeech.2019-2680.

[17] Lavandier, M., Herzog, P., and Meunier, S., "Comparative measurements of loudspeakers in a listening situation," *J. Acoust. Soc. Am.*, 123(1), pp. 77–87, 2008, ISSN 00014966, 15208524.

[18] Cramer, J., Wu, H., Salamon, J., and Bello, J. P., "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3852–3856, 2019, doi:10.1109/ICASSP.2019.8682475, ISSN: 2379-190X.

[19] Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D., "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *arXiv:1912.10211 [cs, eess]*, 2020.

[20] Legarth, S. V. and Zacharov, N., "Assessor Selection Process for Multisensory Applications," in *Audio Engineering Society Convention 126*, 2009.

[21] Brinkmann, F. and Weinzierl, S., "AKtools – An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics," in *Proc. of Audio Engineering Society Convention 142*, e-Brief, Audio Engineering Society, Berlin, Germany, 2017.

[22] ITU-R, "Algorithms to measure audio programme loudness and true-peak audio level," Recommendation ITU-R BS.1770-3, International Telecommunication Union Radiocommunication Assembly (ITU-R), United States, 2012.

[23] Brockhoff, P. B., Schlich, P., and Skovgaard, I., "Taking individual scaling differences into account by analyzing profile data with the Mixed Assessor Model," *Food Quality and Preference*, 39, pp. 156–166, 2015, ISSN 09503293, doi:10.1016/j.foodqual.2014.07.005.

[24] Athar, S., Costa, T., Zeng, K., and Wang, Z., "Perceptual Quality Assessment of UHD-HDR-WCG Videos," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1740–1744, 2019, doi:10.1109/ICIP.2019.8803179.

[25] Salamon, J. and Bello, J. P., "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, 24(3), pp. 279–283, 2017, ISSN 1558-2361, doi:10.1109/LSP.2017.2657381, conference Name: IEEE Signal Processing Letters.

[26] Zhang, Y., Suda, N., Lai, L., and Chandra, V., "Hello Edge: Keyword Spotting on Microcontrollers," *arXiv:1711.07128 [cs, eess]*, 2018.

[27] Gamper, H., Reddy, C., Cutler, R., Tashev, I., and Gehrke, J., "Intrusive and Non-Intrusive Perceptual Speech Quality Assessment Using a Convolutional Neural Network," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.