Audio Engineering Society

Convention Paper 10494

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

Predicting Audio Quality for different assessor types using machine learning

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 Nick Zacharov (nvz@force.dk)

ABSTRACT

In this paper we study how sound quality is evaluated by different groups of assessors, with different levels of hearing loss. Formal listening tests using the Basic Audio Quality scale were designed using 22 headphones spanning a wide range of qualities and sound quality characteristics. The tests were performed with two formally selected listening panels with normal hearing (NH), and mild (N2) or moderate (N3) hearing loss characteristics. It is shown that not only do the two panels evaluate the sound quality consistently within each panel, but also that there are systematic changes in the manner in which hearing loss impacts the evaluation and ranking of the devices under study. Using this data we successfully train machine learning algorithms to predict the sound quality for the two assessor type panels. The prediction performance for each panel is NH: RMSE = 7.1 ± 3.0, PCC = 0.91 ± 0.13; HI: RMSE = 8.7 ± 2.4, PCC = 0.91 ± 0.12. Whilst it may not be practical to run listening tests with multiple panels of assessors, we demonstrate here that machine learning based models can be practically and cost effectively employed to predict the perception of multiple assessor groups rapidly and simultaneously.

1 Introduction

In the field of audio there are different perspectives on how sound quality is perceived. In listening tests we use different assessor groups for different purposes. For example in telecommunications ITU-T based listening tests (e.g. [1, 2]), panels of consumers are employed, to be representative of end-users. Alternatively, in the domain of audio codec listening tests, as recommended by the ITU-R (e.g. [3, 4]), experienced or expert assessors are selected, to represent the most acute members of the population. The general categorisation of assessors is defined in ISO 8586-1 [5], ISO 8586-2 [6] and discussed in detail in [7, 8]. Furthermore, in the field of hearing aids and audiology, hearing aid devices are individually tuned for the needs of each user based on the nature of their personal hearing loss. In this field it is well understood that hearing loss has a significant impact on the perception of sound, sound quality and intelligibility. Prior to 2012, the evaluation of hearing aids was primarily performed on an individual basis. Due to the very individual nature of hearing loss, assessor panels were rarely employed. In recent years, assessors with similar hearing losses have been grouped in panels to take on the challenge of designing devices for hearing impaired users. This will be discussed further in the next section.

From these observations and other experiences, it is clear that at a panel level, we can expect different responses from listening tests due to the hearing acuity or expertise of the assessors. In many audio applications it would be interesting to know how well audio technologies perform from the perspective of multiple groups of assessors. For example, it would be interesting to know what the optimal bit rate would be for naive consumers and for experts, as potentially the latter group is more stringent than the former. However, due to the cost and complexity, it is rare for listening tests to be run with more than one group of assessors and thus this comparison is seldom possible.

With the continued development in smart wireless headphones, the advent of hearables and the continued rapid development of hearing aids, we are seeing an increasing overlap of these product categories - e.g. hearing aids are being tuned to be able to reproduce music and audio, while smart headphones and hearables are developed to help the mildly hearing impaired. There is thus a growing interest to be able to evaluate the audio quality perceived by different assessor groups and understand whether this is the same for all groups or not.

With the current development in the field of machine learning, computing power and tools, it is becoming increasingly accessible to develop predictive models. In this paper we thus study how audio quality is perceived by a number of different panels and evaluate how well the performance from these can be predicted using machine learning on a modest size dataset.

2 Background

It can be demonstrated that certain aspects of sound quality are perceived in a commonly agreed manner, more or less by all people. For example the perception of loudness will generally be ranked in a similar manner. These aspects are objective in the sense, that they are only affected by perception and not subjective weighting by factors such as cultural or personal connotations or degree of liking. Ideally, these aspects can be described in terms of one-dimensional attributes as illustrated in the sound wheel for reproduced sound [9, 10, 11]. However, when evaluating sound quality overall on a single scale, it is less evident whether all assessors will weight and integrate all the sound quality characteristics in a similar manner. For example, for normal hearing individuals many aspects of sound quality are important including the key attribute families such as *loudness, timbre, artefacts, spaciousness, dynamics, intelligibility,* etc. However, as hearing loss increases, certain characteristics become more important, such as loudness and intelligibility, whilst other aspects have less importance, such as *distortion* or *spaciousness*. Further discussion of the nature of sound quality attributes employed by hearing impaired panels, can be found in Chapter 9 of [8].

In order to be able to robustly and repeatably study how people perceive sound, we commonly employ groups of assessors or panels, with similar characteristics. For example expert assessors are screened for normal hearing characteristics and also their basic aptitude for performing sound quality evaluation. Once selected such assessors are then trained and assessed for their skill in performing in different types of perceptual evaluation tasks.

As hearing loss is very individual in nature, it is more complex to find groups of assessors with similar hearing loss characteristics. Hearing loss can be conductive and / or sensorineural in nature and can be characterised by pure tone audibility thresholds (audiograms) or other more specific traits (loudness recruitment, fine structure, spectral smearing, etc.). In order to study the sound perception of different hearing loss groups, we have set about developing different panel of expert assessors. The IEC 60118-15 standard [12] provides one way to group assessors based on the simple audiogram. In this standard several different categories of hearing loss are defined ranging from mild to severe hearing losses, as illustrated in Figure 1.

In order to evaluate headphones, hearables and hearing aids, it is of interest to span a range of hearing losses. For this study we considered including three types of assessors with normal hearing (NH), a mild hearing loss (N2) and with a moderate (N3) hearing loss. In the research leading to the IEC 60118-15 standard reported by Bisgaard et al. [13], the N3 group was the largest of 28244 individuals tests with an average hearing loss (HL) of 46 dB. The N2 group was the third largest group with a hearing loss of 31 dB.

3 Experimental setup

In order to study how different panels of assessors evaluate the sound quality of a range of products, we selected a range of 20 headphones and two "anchor" headphones for evaluation. These were recorded on a high resolution head and torso simulator. The anchor

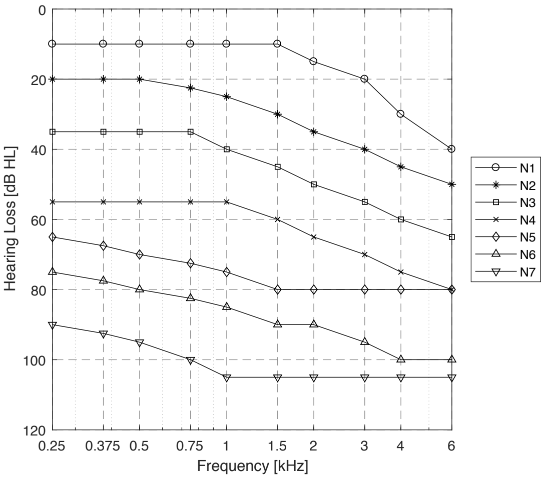

Fig. 1: Example of standard audiograms for the flat and moderately sloping group found in IEC 60118-15 [12].

headphones recordings were manually altered to degrade the sound quality. A novel design of experiment were utilised whereby 4 clips (15-35s) were selected from each of 15 samples (music track). The specific experimental design is used to ensure that all assessors evaluate all headphones with one clip from each sample to increase the number of audio files without increasing the duratio or complexity of the listening tests - this approach is called pre-augmentation and is described in detail in [14], with further details of the tested headphones, samples, etc.In the current study a second anchor stimuli mid anchor was removed from the test with the hearing impaired panel, as it was rated too similar to the low anchor. Furthermore, the hearing impaired panel (being older than the normal hearing panel), was spared listening to three samples: Metallica, Jay-Z and Rage Against the Machine. Also to give them a bit more time per response within the same test duration. The recorded clips for each headphone were equalised for reproduction over reference headphones. These stimuli were presented to each assessor using a double blind multiple stimulus presentation test paradigm and evaluated using a 100-point basic audio quality (BAQ) continuous quality scale (CQS). Presentation order was randomised for each assessor and the tests were performed as independent experiments for the normal hearing- and the hearing impaired assessors.

4 Assessors

For these two experiments we employed a number of selected and trained expert assessors panels. Since 2008 we have maintained a panel of $\sim 25-35$ trained expert assessors with normal hearing (NH) characteristics, selected based on the principles outlined in [15]. Additionally, we also have a panel of hearing impaired assessors following the N3 profile, according to [12] the selection of which is described in [16]. The panel consists of ~ 20 selected and trained assessors with an average age of ~ 70.

Additionally, an N2 panel was developed with an average symmetrical hearing loss of 38 dB.

For this study the following details summarise the nature of each panel employed:

Normal hearing (NH): n = 20, median age = 36.9, min = 19.5, max = 58.8; 18 men and 2 women.

Hearing loss group (N2+N3): n = 20 (7 N2 & 13 N3), average HL = 42.1 dB, median age = 73.4, min = 64.4, max = 86.2; 11 men and 9 women.

The audiograms of the assessors included in this study are shown in Figure 2.

5 Results

Data from the two experiments were analysed to study the differences, if any, between the two panels. For the sake of comparison alone, the normal hearing results are selected as a point of reference.

Prior to analysis, the data from each experiment is normalised using a two-step z-transform introduced in [17] and examined for this data set in the pre-augmentation paper [14]. This transform reduces the variability of individual assessor differences with regards to difference in scale usage related to multiple factors.

5.1 Normal hearing (NH)

The normal hearing data is taken as the baseline for comparison. The average scores for all 20 assessors for all clip/sample combinations are plotted in Figures 3 illustrating the mean and 95 % confidence intervals. Note, that the data quality here is very high, leading to confidence intervals smaller than the dot presenting the mean. The data extends a bit beyond the BAQ scale as

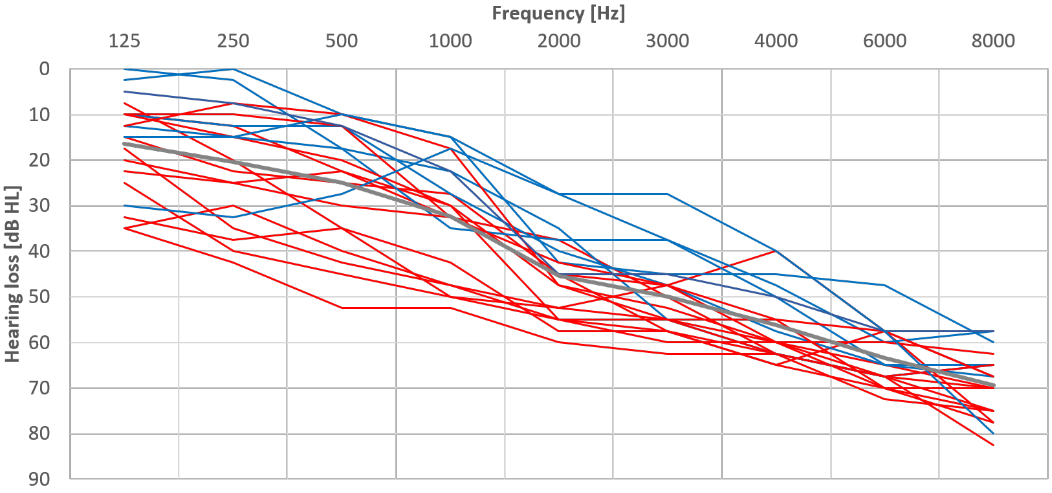

Fig. 2: Audiograms of the hearing impaired panel with seven N2 assessors (blue) and thirteen N3 assessors (red) and the average in grey.

an artefact of the normalisation[1]. From the data we can see that the selected headphones span the entire range of audio quality. The headphones are sorted by BAQ scores to assist in later comparisons.

5.2 Mild and moderate hearing loss (N2 & N3)

Data for the HI panel are shown in Figure 4. Firstly, we see that the HI assessors are able to identify the reference (highest scoring) and the low anchor (lowest scoring) and that the overall trends look similar to that of the NH panel, while clear differences in ranking are also seen. The confidence intervals are small allowing statistically significant separation of headphones and an indication of assessor agreement overall.

5.3 Data comparison

To compare the data collected from each panel summary statistics were calculated and are presented in Tables 1-2 and Figures 3-4.

Taking the normal hearing data as a point of reference, the HI panel rates the stimuli in a significantly different

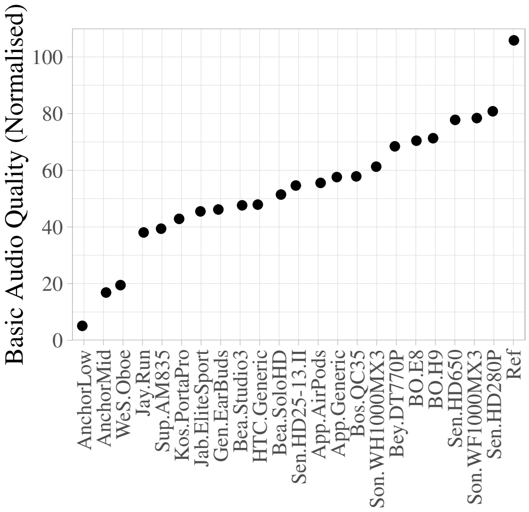

Fig. 3: Mean values and 95% confidence intervals of 20 normal-hearing (NH) assessors. Averaged over assessors, samples, and clips. Note that the error bars for confidence intervals are too small to show.

[1] The data normalisation proposed in [17] seeks to maintain the original scale. However for certain data, such as these, the transformed scale extends beyond the original 100 points, due to an assumption of normal distribution, which for smaller subsets of data are known to follow a t-distribution with heavier tails, e.g. higher probabilities at the extremes.

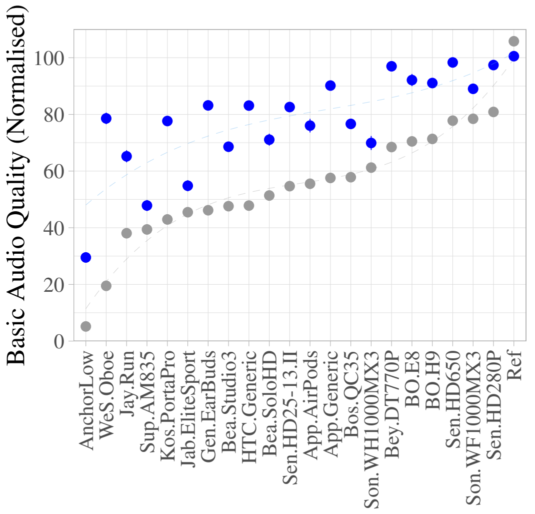

Fig. 4: Mean values and 95% confidence intervals of 20 hearing impaired (blue) assessors. Averaged over assessors, samples, and clips. Sorted by normal hearing (grey) mean values. Lines are 3. order polynomial fits.

manner, viewed from several perspectives. We see that the span of the data from the HI panel is compressed compared to the NH panel. This suggests that timbral and distortion artefacts are less significant to hearing impaired assessors, as for example the low anchor and similar stimuli are rated higher by the hearing impaired panel compared to the NH panel. This upwards compression of the scale with hearing loss, is systematic across system averages.

Additionally, from Figure 4 we can observe that while the overall trend is similar between the two panels the manner in which the hearing impaired panel rates the stimuli still differs significantly . This is further confirmed by the correlations between the panels in Table 2. Whilst we do not have data to fully explain the cause for these differences, it might be hypothesised that there is a weighting of different perceptual characteristics or attributes for each panel. This would be logical as it is generally known that with hearing loss the importance of sound quality characteristics shifts towards an emphasis upon fundamental characteristics e.g. *intelligibility, loudness*, etc. However, this would require further data to verify, which is beyond the scope of this study and data.

Panel	Min. score (points)	Max. score (points)	Data span (points)	Avg. 95% CI (points)
NH	5.2	105.8	100.6	2.3
HI	29.5	100.5	71.0	3.3

Table 1: Summary statistics of the data ranges for each assessor panel.

Comparison	Pearson	Spearman	MAD	RMS
NH vs. HI	0.79	0.79	23.1	26.1

Table 2: Statistical comparison of differences between the panel data in terms of Pearson correlation coefficient, Spearman rank correlation coefficient, mean absolute difference (MAD) and root-mean square (RMS).

Nonetheless, we can confirm that there are statistically significant differences in the ratings between the three panels, which are non-trivial and are worthy of model modelling and prediction.

6 Machine learning

For establishing a machine learning prediction model of these two data sets it was decided to use previously successful metrics combined in regression models. This choice was made as training a deep-learning model would require more data to perform well as was established in our pre-augmentation paper using the same normal hearing data set as in this one [14]. The metrics with the best performance are described in the next section. Besides the ones described here, PEAQ [18, 19] (in the GstPEAQ implementation [20]) and CPAM [21] were tested, but not found to improve the model.

6.1 Metrics

ViSQOLv3 [22] is an open source implementation that merges the speech quality model proposed in VISQOL with the audio quality model proposed in VISQOLAudio [23]. It constructs a Gammatone spectrogram of each audio input, finds spectrogram patches with the best alignments and computes a difference between them using Neurogram Similarity Index Measure (NSIM) [24]. For the audio quality mode, the NSIM is mapped to a Mean Opinion Score (MOS) using a Support Vector Regression model, while for speech quality a polynomial mapping is used.

CDPAM from 2021 by Manocha et al. [25] expands upon CPAM [21], which is a deep learning model using a Convolutional Neural Network on raw audio waveform. The dataset used for training was constructed using synthetic distortions that were evaluated by a crowd sourced listening test. The assessors were asked to determine if two audio clips are exactly equal or not, and the dataset is referred to as Just Noticeable Differences (JND). The model outputs a distance metric, where 0.0 indicates no perceptual difference. CDPAM extends CPAM by including a 3-stage learning process. The first stage uses contrastive learning inspired by SimCLR[26] to learn a vector representation of each audio input, the second stage uses JND training like CPAM, and a third stage performs fine tuning on a new dataset of triplets. The output is a distance metric like CPAM, which outperforms CPAM, especially on large distortions well outside the just noticeable difference region.

A metric derived only from the frequency response was also included. This is inspired by works showing that loudspeaker and headphone preference can be modelled effectively using simple models based solely on frequency response deviations (see e.g. [27]). The overall frequency response of each headphone was calculated by subtracting the spectrogram of the processed audio by the spectrogram of the corresponding reference audio. The spectrogram type used was 32-bins log-mel-spectrogram. The overall frequency response difference from the reference track was summarised using Mean Squared Error (MSE). We dub this metric FreqERR.

6.2 Modelling

Regression models (linear and non-linear) were created using the three mentioned metrics to predict the listening tests results. For each audio file (one clip) the output from the metrics and the normalised response (averaged across assessors) is used as input during training. The predictions of all audio files from the same system is then averaged to become the estimate of Basic Audio Quality for that system. The model performance metrics used are root mean square error (RMSE), Pearson regression coefficient, and Spearman rank coefficient. The evaluation was done with 20-fold cross-validation. The folds were stratified on system (headphone), such that samples from the same system does not appear in both training- and test split. In each fold, samples from 4 systems were used as test data ($\sim 25\%$). The number of folds was selected such that every systems appears at least once in the test subset. Results from two types of regression models are reported: Linear regression using FreqErr, CDPAM and ViSQOLv3 (referred to as *LN*) and Gradient Boosting Regressor using all metrics (referred to as *GBT*). A few other model- and metrics combinations were also tested, without improvements.

Model	Group	RMSE	Pearson	Spearman
LN	NH	9.7 ± 4.1	0.86 ± 0.22	0.81 ± 0.22
	HI	8.7 ± 2.4	0.91 ± 0.12	0.89 ± 0.12
GBT	NH	7.1 ± 3.0	0.91 ± 0.13	0.84 ± 0.22
	HI	9.9 ± 4.4	0.85 ± 0.21	0.80 ± 0.27

Table 3: Results of predictive models for each assessor group and using either linear regression (LN) or Gradient Boosting Regression (GBT). Performance metrics are root-mean square error (RMSE), Pearson correlation coefficient and Spearman rank correlation coefficient.

6.3 Modelling results

Using the model structure described in the previous section, the two assessor groups, NH and HI, were modelled separately. The results are summarised in Table 3 as system averaged predictions on the combined test sets across all folds.

Figure 5 compares the model prediction per system with the normalised responses.

Two of the largest prediction errors in the NH model is that of the low anchor, AnchorLow, and the reference, Ref. These are particular of the test methodology in which the reference must be correctly identified and rated at 100 (before Z-normalisation, in a similar manner to the user of references and anchors in standardised methodologies. See Sec. 5.1) and the low anchor is artificially generated and easily identified by our assessors as the artificial system that must be scored low on the scale. Consequently, the true perceived quality of these two systems might differ from the collected ratings. That the model is unable to predict these systems well, might actually be a good sign and the prediction error of Gen.EarBuds of greater concern. Due to the position of AnchorLow and Ref at the extremes on scale, they were still beneficial to keep in the dataset and improved the overall modelling.

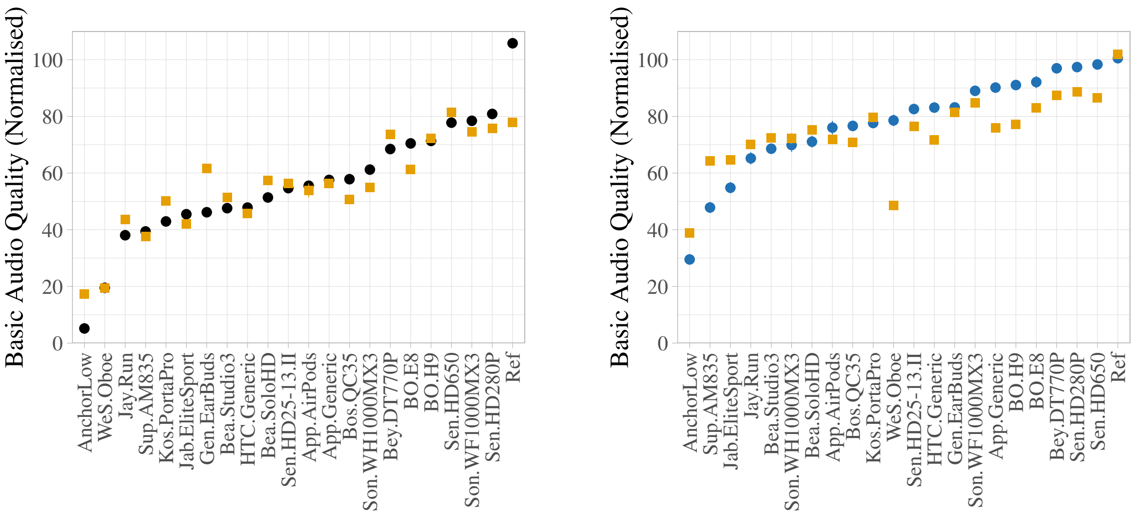

Fig. 5: Normalised listening test responses (black/blue circles) vs. model predictions (orange squares), with 95%-confidence intervals. Left: Results for normal hearing panel, using GBT model. Right: Results for hearing impaired panels, using LN model.

7 Discussion

Using the pre-augmentation approach described in Volk et al. [14], an identical experiment was run with a panel of hearing impaired assessors with either mild (N2) or moderate (N3) hearing loss (see Figure 2). Although it would have been interesting to model the N2 group and the N3 group separately, we had insufficient data and statistical power to allow for a complete analysis of each group and as a result the N2 and N3 data were combined for analysis. While responses do not differ too much between N2 and N3 assessors in this test, we do not believe this to be the case in general.

The key observation from the two experiments is that normal hearing and assessors with hearing loss score the performance of the systems under test very differently. These differences are summarised by two key characteristics for these systems see in Figure 4. Firstly, we see that the hearing impaired panel compressed the BAQ scale upwards, illustrating a potential tolerance to sound quality aberrations. Secondly, from this figure we can also see a difference in the rank of the systems under test compared to the NH panel. Thirdly, due to the very tight confidence intervals, we can see a a clear pattern of results for both the NH and HI panels. Based on these significant results, it makes sense to develop separate predictive models for the NH and HI panels.

With average RMSE values of the sizes in Table 3, it is possible to predict system performance for both assessor groups with sufficient performance to be of significant value even with the relatively small data sets collected for this study, due to utilisation of relevant metrics based on previously published models. However, the few systems with large errors are a concern, which require further data or model optimisation, but it is certainly a good step along the way. With regards to the RMSE metric, we aim to switch to a version better suited for target data with uncertainty in our model optimisation process, namely RMSEpsilon ITU-T [28], which does not penalise predictions within the confidence intervals of the target data.

8 Conclusions

In this paper we illustrate how assessor groups with systematically different hearing thresholds yield different audio quality ratings on the Basic Audio Quality (BAQ) scale. The differences found with a hearing impaired panel consisting of assessors with mild (N2) and moderate (N3) hearing losses are significant compared to a normal hearing panel.

Using the combined data collected from listening tests for normal hearing and hearing impaired listening panels predictive models were successfully created to predict the basic audio quality scores from unseen audio files. Whilst this is work in progress, this *proof-of-concept* level modelling inspires us to understand that machine learning can be used to predict overall sound quality for different hearing impaired groups and also potentially other groups of assessors such experts and consumers, etc. Even common subsets of consumers such as the *bass-lovers* and the *naturalness seekers*. In future efforts, this approach could enable the optimisation of product performance from multiple user perspectives.

9 Future work

This paper shows the basic principle of how audio quality varies with hearing loss. For certain this element should be studied in further depth to understand the role of hearing loss on audio quality at different hearing loss levels, or using other approaches to cluster assessors. For this to be of greater interest, this would need to be tested on more complex audio stimuli, e.g. with hearables, hearing aids, that also employ digital signal process audio enhancement for a wide range of audio scene types.

We can also see that different assessor groups weigh audio quality characteristics differently. Exploring this phenomenon further and understanding the underlying mechanisms for this would also be valuable.

Exploring the applicability of modelling to different assessor group or clusters (e.g. different hearing impairment levels, consumers, experts, audio professionals, etc.) would also be of interest to both the scientific and industrial community.

9.1 Acknowledgements

This work was partially funded by The Danish Council for Technology and Innovation. We also thank the members of the listening panels for their contributions.

References

[1] ITU-T Rec. P.800, “Methods for subjective determination of transmission quality,” Recommendation ITU-T Rec. P.800, International Telecommunication Union, Geneva, Switzerland, 1996.

[2] ITU-T Rec. P.835, “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm,” Technical Report ITU-T Rec. P.835, International Telecommunication Union, Geneva, Switzerland, 2003.

[3] ITU-R BS.1534-3, “Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems,” Technical Report ITU-R BS.1534-3, International Telecommunication Union, Geneva, Switzerland, 2015.

[4] ITU-R Rec. BS.1116-3, “Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems,” Technical Report ITU-R Rec. BS.1116-3, International Telecommunication Union, Geneva, Switzerland, 2015.

[5] ISO 8586-1, “Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 1: Selected assessors,” Technical Report ISO 8586-1, International Standards Organization, Geneva, Switzerland, 1993.

[6] ISO 8586-2, “Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 2: Expert sensory assessors,” Technical Report ISO 8586-2, International Standards Organization, Geneva, Switzerland, 2008.

[7] 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.

[8] Zacharov, N., *Sensory Evaluation of Sound*, CRC, New York, first edition, 2019.

[9] Pedersen, T. H. and Zacharov, N., “The Development of a Sound Wheel for Reproduced Sound,” in *Audio Engineering Society Convention 138*, 2015.

[10] Zacharov, N., Pedersen, T. H., and Pike, C., "A common lexicon for spatial sound quality assessment - latest developments," in *Proceedings of 8th International Conference on Quality of Multimedia Experience (QoMEX 2016)*, 2016.

[11] ITU-R Rep. BS.2399, "Methods for selecting and describing attributes and terms in the preparation of subjective tests," Technical Report ITU-R Rep. BS.2399, International Telecommunication Union, Geneva, Switzerland, 2017.

[12] International Electrotechnical Commission, "Electroacoustics - Hearing aids - Part 15: Methods for characterising signal processing in hearing aids with a speech-like signal," Technical Report IEC 60118-15, IEC, Geneva, Switzerland, 2012.

[13] Bisgaard, N., Vlaming, M. S., and Dahlquist, M., "Standard Audiograms for the IEC 60118-15 Measurement Procedure," *Trends in Amplification*, 14(2), pp. 113–120, 2010, ISSN 10847138, doi:10.1177/1084713810379609.

[14] Volk, C. P., Nordby, J., Stegenborg-Andersen, T., and Zacharov, N., "Efficient data collection pipeline for audio machine learning," in *Proc. of Audio Engineering Society Convention 150*, p. 10, Audio Engineering Society, Copenhagen, Denmark, 2021.

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

[16] Legarth, S. V., Simonsen, C. S., Dyrlund, O., Bramsløw, L., and Jespersen, C. T., "Establishing and qualifying a hearing impaired expert listener panel," in *International Hearing Aid Research Conference*, 2012.

[17] 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.

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

[19] Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., and Colomes, C., "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc*, 48(1/2), pp. 3–29, 2000.

[20] Holters, M. and Zã, U., "GstPEAQ – an Open Source Implementation of the PEAQ Algorithm," in *Proc. of the 18th Conference on Digital Audio Effects (DAFx-15)*, p. 4, Trondheim, Norway, 2015.

[21] 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.

[22] Chinen, M., Lim, F. S. C., Skoglund, J., Gureev, N., O'Gorman, F., and Hines, A., "ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric," *arXiv:2004.09584 [cs, eess]*, 2020.

[23] Hines, A., Gillen, E., Kelly, D., Skoglund, J., Kokaram, A., and Harte, N., "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *The Journal of the Acoustical Society of America*, 137, pp. EL449–EL455, 2015, doi: 10.1121/1.4921674.

[24] Hines, A. and Harte, N., "Speech intelligibility prediction using a Neurogram Similarity Index Measure," *Speech Communication*, 54(2), pp. 306–320, 2012, ISSN 0167-6393, doi:10.1016/j.specom.2011.09.004.

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

[26] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G., "A Simple Framework for Contrastive Learning of Visual Representations," *arXiv:2002.05709 [cs, stat]*, 2020, version: 3.

[27] Olive, S., Welti, T., and Khonsaripour, O., "A Statistical Model that Predicts Listeners' Preference Ratings of Around-Ear and On-Ear Headphones," in *Proc. of the 144th Audio Engineering Society Convention*, Audio Engineering Society, 2018.

[28] ITU-T, “Recommendation ITU-T P.1401 - Methods, metrics and procedures for statistical evaluation, qualification.pdf,” Recommendation ITU-T P.1401, ITU Telecommunication Standardization Sector (ITU-T), Geneva, Switzerland, 2020.