Objective Assessment of Perceptual Audio Quality Using ViSQOLAudio

Colm Sloan, Naomi Harte, Damien Kelly, Anil C. Kokaram, and Andrew Hines

***Abstract*—Digital audio broadcasting services transmit substantial amounts of data that is encoded to minimize bandwidth whilst maximizing user quality of experience. Many large service providers continually alter codecs to improve the encoding process. Performing subjective tests to validate each codec alteration would be impractical, necessitating the use of objective perceptual audio quality models. This paper evaluates the quality scores from ViSQOLAudio, an objective perceptual audio quality model, against the quality scores of PEAQ, POLQA, and PEMO-Q on three datasets containing fullband audio encoded with a variety of codecs and bitrates. The results show that ViSQOLAudio was more accurate than all other models on two of the datasets and performed well on the third, demonstrating the utility of ViSQOLAudio for predicting the perceptual audio quality for encoded music.**

***Index Terms*—Perceived audio quality, subjective audio quality assessment, objective audio quality assessment, ViSQOLAudio, ViSQOL, POLQA, PEAQ, PEMO-Q.**

I. Introduction

DIGITAL audio broadcasting systems and streaming services such as YouTube Music are popular platforms for consuming audio media. These streaming services use codecs to minimize bandwidth and maximize users' quality of experience whilst not degrading perceptual quality.

Frequent modifications are made to the codecs to fix bugs and improve efficiency. Subjective listening tests are ideally performed after each codec modification to assess changes in the perceptual audio quality of audio encoded with the modified codec. In these tests, subjects listen and assign a perceptual quality score to each clip in a set of encoded audio clips. The average score from all subjects is taken to create a mean opinion score (MOS) for each clip. The effect of the codec modification is then be assessed by comparing the MOS values before and after the modification.

Because subjective testing is time-consuming, objective perceptual audio quality models are used to predict MOS values in an automated and timely manner. A number of objective models exist that can predict the perceptual audio quality of an encoded audio clip given the encoded clip and its uncompressed equivalent as reference. PEAQ [1], POLQA [2], PEMO-Q [3] and ViSQOLAudio [4] are four such full-reference models. Each of these models have been used previously to rate the quality of encoded fullband audio [3]–[6].

One model in particular, ViSQOLAudio, will be the focus of this paper. ViSQOL [7] is a speech quality model that was later adapted to function as a perceptual audio quality model, yielding a prototype that delivered promising results [4]. This paper builds upon a proof of concept presented in [4] that showed that objective speech quality metrics such as POLQA and ViSQOL could be adapted for audio quality prediction. POLQA Music [8] demonstrated that training with audio data improved the performance of the speech metric. In this paper, ViSQOLAudio introduces a number of novel additions that improve upon its predecessor to produce MOS values for compressed audio. These additions include:

- Using machine learning to create quality scores that better match those made using human perception.
- Considering information from both channels when evaluating stereo audio clips.
- Compensating for subframe misalignments of the reference and degraded signals caused by encoder padding.
- Using a filter bank more suitable for fullband (music) content.
- Outputting MOS values rather than similarity scores, making the output more intuitive to humans.

In this paper, ViSQOLAudio is evaluated against POLQA, PEMO-Q, and PEAQ on three datasets of music content encoded with an assortment of bitrates and codecs used for popular digital broadcasting and streaming services. This is done to determine which objective models are suitable for assessing the perceptual quality of encoded music. Models are evaluated by the accuracy, consistency and linearity (defined in Section V-A) of their objective perceptual quality scores. ViSQOLAudio is, to the authors' knowledge, this is the first totally free and open source audio quality metric with accuracy comparable to models used in industry when tested upon compressed audio.

Manuscript received October 11, 2016; revised January 17, 2017; accepted March 28, 2017. Date of publication June 6, 2017; date of current version December 8, 2017. This work was supported in part by the CONNECT Research Centre through YouTube, Google Inc., Science Foundation Ireland, and in part by the European Regional Development Fund under Grant 13/RC/2077. *(Corresponding author: Colm Sloan.)*

C. Sloan and N. Harte are with the Sigmedia, Department of Electronic and Electrical Engineering, Trinity College Dublin, 2 Dublin, Ireland (e-mail: sloanco@tcd.ie).

D. Kelly and A. C. Kokaram are with Google, Inc., Mountain View, CA 94043 USA.

A. Hines is with the School of Computing, Dublin Institute of Technology, Dublin 8, Ireland, and also with Sigmedia, Department of Electronic and Electrical Engineering, Trinity College Dublin, 2 Dublin, Ireland (e-mail: andrew.hines@dit.ie).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2017.2704421

This paper has the following structure. Section II describes the objective quality models, focusing on POLQA, PEMO-Q and PEAQ, and explaining why they were selected for comparison to ViSQOLAudio. Section III describes improvements made to ViSQOLAudio. Section IV gives details on the datasets used to compare the objective quality models. Section V describes and justifies the evaluation metrics for comparing the objective models. Section VI gives the results of the experiment comparing the models, leading to a discussion of the results in Section VII. Section VIII then closes with conclusions.

II. Background

This section presents a number of objective quality models. The PEAQ [1], POLQA [2] and PEMO-Q [3] models are given particular attention because they will be part of the experiments described in Section VI.

Objective models for predicting perceptual audio quality can be classified as being in two categories: parameter-based or signal-based. Parameter-based models such as ITU-T G.107 [9] predict quality by modeling characteristics of a transmission channel of audio, such as packet loss rate and delay jitter.

Signal-based models estimate quality based on information taken from signals rather than the medium of their transmission. Signal-based models are subcategorized into no-reference models (also known as single-ended and non-intrusive) and full-reference models (also known as comparison-based and intrusive). No-reference models only analyze a degraded signal when predicting the quality of that signal. No-reference models such as ITU-T P.563 [10] analyze speech and clipping among other techniques to estimate signal quality.

Full-reference models predict quality by comparing features from a perfect quality reference signal to a degraded version of that signal. This category is the focus of our research, where the reference signal is the uncompressed audio uploaded by a user to a streaming service, and the degraded signal is an encoding of that uncompressed audio.

Early full-reference models, such as ITU-T P.861 (PSQM) [11] were focused on speech and predicted quality within a narrow frequency band (300 – 3400 Hz). PSQM was made obsolete when VoIP introduced problems such as larger signal distortions and variable delays between the reference and degraded signals [12]. ITU-T P.862 (PESQ) [13] fixed many weaknesses in PSQM and widened the frequency band of audio it could evaluate. However, PESQ had issues with loudness loss, echoes and sidetone. These were addressed by the successor to PESQ, ITU-T P.863 (POLQA) [2].

POLQA [2] was designed to predict the quality of speech from narrow up to super-wideband (50 – 14000 Hz). The POLQA quality score prediction process begins by creating a psychophysical representation of the reference and degraded signals. This process includes time alignment, level alignment, time-frequency mapping, frequency warping and compressive loudness scaling. The reference signal then undergoes an "idealization" process which adjusts timbre if the signal is noisy. POLQA then eliminates reference signal noise and suppresses the degraded signal noise. These modified signals are passed into a cognitive model that computes quality indicators such as a frequency response indicator and a noise indicator, and are combined to give a MOS value. A MOS-LQO (mean opinion score - listening quality objective, the objective MOS) value is the objective equivalent to a subjective MOS (MOS-LQS) value.

Unlike POLQA, which was designed to predict perceptual speech quality, ITU-T BS.1387 (PEAQ) [1] was designed for encoded audio. There are two versions of PEAQ: PEAQ-Basic with a lower complexity model for fast quality score predictions, and PEAQ-Advanced with a high complexity model that takes longer to calculate. Our evaluation will focus on PEAQ-Advanced as this is considered the most accurate version by the developers of PEAQ [1] and because, although PEAQ-Basic has been shown to perform better than PEAQ-Advanced when targeting degraded audio [14], PEAQ-Advanced performs better on the datasets used as part of this work.

The PEAQ-Advanced quality prediction process begins by passing the reference and degraded signals into an ear model that segments the signals into auditory filter bands that, among other steps, are passed into weighted transfer functions representing the different parts of the ear. A process then identifies excitation patterns in loudness and modulation. These patterns are used to calculate several psycho-acoustically based model output variables, such as average linear distortions, that quantify differences between the reference and degraded signals. These model output variables and a set of coefficients are inputted to an artificial neural net which outputs a distortion index that is mapped to an Objective Difference Grade (ODG). An ODG is analogous to the Subjective Difference Grade, defined as $SDG = grade_{degraded} - grade_{reference}$, where the grade is the ITU-T BS.562 [15] impairment scale from 1 (very annoying) to 5 (imperceptible).

Another model for predicting perceptual audio quality, PEMO-Q [3], was shown by its authors to predict quality more accurately than PEAQ [3]. PEMO-Q predicts quality using time-aligned reference and degraded signals that are level aligned before deleting silence from the signals. The signals are input to a psychoacoustically motivated model that transforms the signals into a three dimensional representation, where the dimensions represent activity patterns in time, frequency and modulation-frequency. The correlations between the reference and degraded patterns are used to create error estimations that are divided into target distortion, interference and artifact components. Each component is weighted for salience and the weights are input to a trained non-linear mapping that produces an Overall Perception Score (OPS) value ranging from 0 (bad quality) to 100 (excellent quality).

III. ViSQOLAudio

This section will present ViSQOLAudio, the full-reference objective quality model that is the focus of this paper. An overview of ViSQOLAudio is shown in Figure 1. The quality score prediction process of ViSQOLAudio has

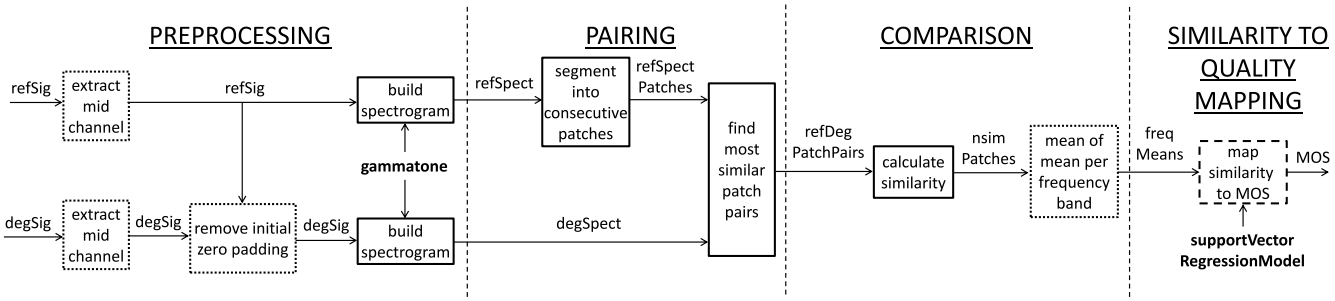

Fig. 1. A high-level representation of ViSQOLAudio. The dotted line boxes represent processes added to since the previous version of ViSQOLAudio [4]. The dashed line boxes represent processes modified since the previous version. Bold text denotes inputs modified since the previous version.

four phases: preprocessing, pairing, comparison, and similarity to quality mapping. A high level explanation of each phase will be given followed by detailed explanations of the processes of each phase.

In the *preprocessing* stage, the mid channel is extracted from the reference and degraded signals to consider information from both channels (described in more detail in Section III-A). An alignment process is then performed on the reference and degraded signals, compensating for subframe misalignments caused by encoder padding (Section III-B). A spectrogram of the reference and degraded signals is then built using a Gammatone filter (Section III-C).

The *pairing* phase first segments the reference spectrogram into patches of 30 frames. These patches are used as input into a robust alignment process that matches each reference spectrogram patch with the most similar patch from the degraded spectrogram, creating a set of most similar reference-degraded patch pairs (Section III-D). This alignment process helps to correct drift and warping in the degraded signal.

In the *comparison* stage, the similarity of each most similar patch pair is measured (Section III-E), outputting similarity patches representing the similarity of each of the pairs. For each of these similarity patches, the similarity across each frequency band is measured. This allows each of the similarity of each frequency band in the degraded signal to be considered separately, allowing a machine learning model to find relationships between similarities across frequency bands that are used to make more accurate quality score predictions.

The *similarity to quality mapping* phase inputs the mean frequency band similarity scores of each similarity patch into a support vector regression (SVR) model that outputs a MOS-LQO value (Section III-F).

A. Channel Selection

Subjective studies have shown that perceptual quality of audio output from codecs is not uniform for all musical sounds [5]. In audio where the one channel contains more of one instrument than the other channel, an objective model taking information only from one channel may analyze an input signal unrepresentative of the signal heard by the subject. Furthermore, non-expert audio users may upload a stereo audio file containing only one audio signal.

One of our goals was to extend ViSQOLAudio to consider information from both channels of stereo signals. A number of approaches were attempted. ViSQOLAudio was used to evaluate the left and right channel signals separately and combine the quality scores to form a single score representative of the stereo quality. The use of the mid and side channel signals were also considered, where $mid(y) = (y_{left} + y_{right})/2$ and $side(y) = y_{left} - y_{right}$ where y is a left-right stereo input signal.

Tests revealed that considering the signals from two channels gave more accurate scores than only considering one channel. The two most accurate model stereo configurations came from taking the maximum predicted quality of the left and right channel signals, and the maximum predicted quality of the mid and side channel signals. Further analysis showed that the maximum predicted quality of the mid-side channel pairs almost always came entirely from the mid channel as the side channel contained little information, which meant that it was not necessary to consider the side channel.

A repeated-measures ANOVA test with a significance $p < 0.05$ was performed to confirm that there was no significant difference between the quality scores produced by considering both the left and right channel signals or just considering the mid channel signal. Besides requiring half of the computational power, using the mid channel signal also alleviated the problems of having different instruments in different audio channels and the problem where users may upload a stereo audio file containing only one audio signal. These results led to the incorporation of the use of the mid channel signal by ViSQOLAudio.

B. Removing Initial Zero Padding

When audio is encoded, many popular encoders add a buffer of zero signal samples to the beginning of the degraded (encoded) signal during the encoding windowing process [16], [17]. The number of samples added can be over 4000 for some codecs. These additional samples at the beginning of the degraded signal causes misalignment with the uncompressed signal it was encoded from.

In most tested cases, the patch alignment process of ViSQOLAudio (Section III-D) was enough to compensate for this misalignment. However, some encodings were more affected than others, particularly MP3. This misalignment is caused by the difference in ViSQOLAudio window size and codec window size, which resulted in a subframe misalignment. ViSQOLAudio compensates for this misalignment using

a frequency-domain cross-correlation on the Hilbert transformed envelope of the reference and degraded signals to find the correct sample number offset for the degraded signal.

C. Building the Spectrograms

Prior to generating the spectrograms, the power level of the degraded signal is scaled to match that of the reference signal. Following this scaling, a short-time Fourier transform is performed with a 32 band Gammatone filter bank with a minimum frequency of 50 Hz and a 50% overlap with a window of 1536 samples (16 ms). The average power of each band across each frame is used to create spectrograms for both the reference and degraded signals. The spectrograms are floored to the minimum value of the reference spectrogram to level the signals with a 0 dB reference.

D. Aligning Spectrogram Patches

The reference spectrogram is segmented into an ordered set of grids, each 30 frames wide, and with a height equal to the number of filter bank frequency bands, as shown in Figure 2. Each segment is referred to as a *patch*. The patch alignment process enables compensation for local time misalignments. The goal of the process is to match each reference patch with its most similar corresponding degraded patch, forming a patch pair. A patch pair is denoted (P_i^r, P_j^d), where i is the reference patch index, j is the degraded patch index, P_i^r is a patch from the reference spectrogram and P_j^d is a patch from the degraded spectrogram.

The set of all degraded patches P^d in the degraded spectrogram consists of all possible 30 consecutive frames in the degraded spectrogram. To find the degraded patch most similar to a reference patch, the process iterates through each possible degraded patch and compares it to the reference patch using the Neurogram Similarity Index Measure (NSIM) [18] (described in Section III-E). The degraded patch with the highest similarity measure is selected as the degraded part of the reference-degraded patch pair and added to the set of the best patch pairs. The process of finding the most similar degraded patch for a reference patch is described as:

$$s = \underset{d \in P^d}{\operatorname{argmax}} \overline{NSIM(P_i^r, d)} \tag{1}$$

where P^d is the set of all degraded patches, P_i^r is the reference patch being paired and the overbar is the *mean* operation. This process is performed for all reference patches, yielding a set of the most similar reference-degraded spectrogram patch pairs, *bestPatchPairs*, that will be used during the mapping from patch-pair similarity scores to a MOS-LQO quality score. Before that, we discuss how the similarity scores used to pair reference and degraded patches are generated.

E. Similarity Comparison

Structural Similarity (SSIM) [19] was originally developed to measure the degradation of compressed JPEG images by comparing the weighted luminescence, contrast and structure of the uncompressed reference image and degraded (compressed) image. The NSIM is a similarity measure specialized

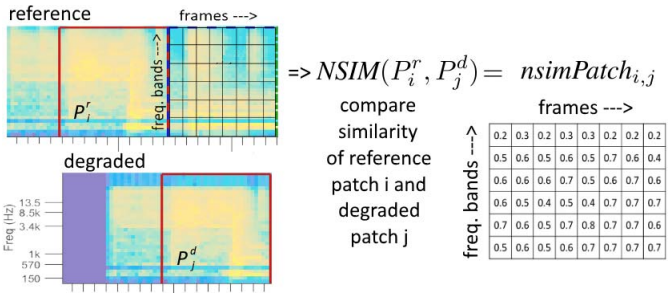

Fig. 2. The process of creating an NSIM patch by comparing the similarity of a reference and degraded patch pair.

for comparing spectrograms. NSIM has been shown to give more accurate similarity measures than SSIM when comparing spectrograms for speech audio [18].

Figure 2 shows part of a reference and degraded spectrogram being compared for similarity. The NSIM of a reference-degraded patch pair, $NSIM(P_i^r, P_j^d)$, is calculated the same way as SSIM index is calculated in [19], but where the luminescence weight $\alpha = 1$, the contrast weight $\beta = 0$, the structural weight $\gamma = 1$, and the regularization constant regularization constants c_1 and c_3 are 0.01 and 0.03 respectively (the constants recommended in [19]). Using the windowing method described in [19], a 3x3 Gaussian window with a radius of 0.5 is used when weighting pixels in the area of interest. The NSIM of a reference and degraded patch is described as:

$$NSIM\left(P_i^r, P_j^d\right) = l\left(P_i^r, P_j^d, c_1\right) \cdot s\left(P_i^r, P_j^d, c_3\right) \tag{2}$$

where P_i^r is the reference patch, P_i^d is the degraded patch, l is luminosity and s is structure. Each NSIM value is placed into its respective cell, forming an NSIM patch where a cell represents the similarity between the reference and degraded signals for a given frame and within a given frequency band. As such, patch columns (frames) represent information over time and patch rows (frequency bands) represent information over frequencies.

F. Mapping Similarity to Quality

The ViSQOLAudio process of generating a MOS-LQO (objective quality score) from similarity patch pairs is shown in Figure 3 and described as:

$$q = SVR\left(\frac{1}{M}\sum_{i=1}^{M}\Omega_i\right) \tag{3}$$

where q is a MOS-LQO value from 1 to 5, M is the number of patches in *bestPatchPairs*, Ω is the row (similarity scores across frequency bands) sums of the set of most similar reference-degraded spectrogram patch pairs, and *SVR* is the support vector regression (SVR) mapping function.

As shown in Figure 3, the row means Ω over all M patches gives a set of vectors $\mathbf{f}$, where each $\mathbf{f}_i$ is a vector of similarity scores (one for each frequency band). The mean of $\mathbf{f}$ is calculated $\bar{\mathbf{f}}$ which is input to the *SVR* mapping function that takes a frequency similarity vector as input and outputs a MOS-LQO.

The *SVR* mapping function is an SVR model. The model is a ν-SVR with a radial kernel, where the $\nu = 0.6$, cost $= 0.4$,

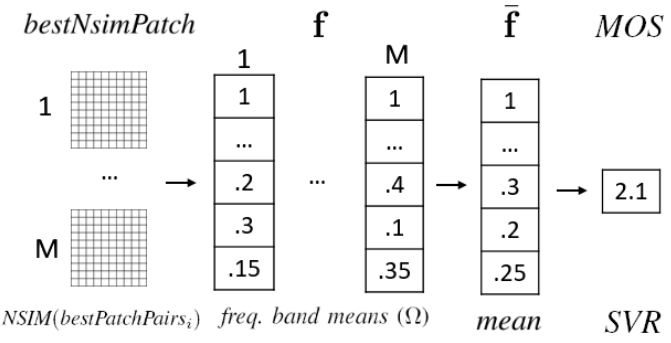

Fig. 3. The process of generating a MOS-LQO from NSIM patches. NSIM patches have their mean similarity across frequency bands calculated, the mean of which is input to an SVR that outputs a MOS-LQO. Variable names are shown above their visual representation and the function that produced the variable is shown below.

TABLE I
AUDIO SAMPLE TREATMENTS IN THE *TCDAudio14* DATASET

Codec	Format	Bitrate (kb/s)
N/A	3.5 kHz narrowband	256
N/A	7 kHz wideband	512
libfdk_aac	HE-AAC v2	24
libfdk_aac	HE-AAC v2	48
libfdk_aac	HE-AAC v2	64
libmp3lame	MP3	96
libfdk_aac	AAC-LC	128
libopus	Opus	128
libfdk_aac	AAC-LC	256

and the remaining values are the LIBSVM [20] defaults. The SVR is trained using $\bar{\mathbf{f}}$ as an observation and the degraded audio clip MOS-LQS as the target. More details of the SVR training is described in Section V-D.

IV. DATASETS

This section presents the datasets used to evaluate the performance of the objective models in the experiment in Section VI. These include the content of the datasets and the conditions (treatments) under which the datasets were created. These datasets include the *TCDAudio14* [5], *AACvOpus15*, and *CoreSV14* [21] datasets. Each dataset was created at different locations using different subjects and prior to the development of this model. Each dataset differs in methodology because each was created by different teams. Subject pool sizes for each dataset conform to the required standard [22].

A. TCDAudio14

The *TCDAudio14* dataset was created to assess the quality of several popular formats at a variety of bitrates commonly used by streaming services. The full list of treatments is shown in Table I (all at constant bitrates) and includes the treatments of 3.5 kHz (lowpass-filtered) narrowband and 7 kHz (lowpass-filtered) wideband as a low and mid quality anchors (as recommended in ITU-R BS.1534 [22]). The samples tested, shown in Table II, were selected to capture a variety of different audio types and were taken from CDs and the EBU music database [23]. With nine treatments and 12 samples, the dataset contains a total of 108 audio clips.

TABLE II
AUDIO SAMPLES IN THE *TCDAudio14* DATASET

Label	Music Type	Source
boz	Rock/R&B (Boz Scaggs)	CD
castanets	Castanets	EBU
contrabassoon	Arpeggio / Melodious Phrase	EBU
glock	Glockenspiel	EBU
guitar	Larry Coryell	EBU
harpsichord	Arpeggio / Melodious Phrase	EBU
moonlight	Piano (Moonlight Sonata)	CD
ravel	Tzigane	EBU
sopr	Soprano singer	EBU
steely	Soft Rock (Steely Dan)	CD
strauss	R. Strauss (Orchestra)	EBU
vega	Vocals (Suzanne Vega)	CD

TABLE III
AUDIO SAMPLE TREATMENTS IN THE *AACvOpus15* DATASET

Codec	Format	Bitrate (kb/s)
N/A	3.5 kHz narrowband	256
libfdk_aac	HE-AAC v2	48
libopus	Opus	48
libfdk_aac	HE-AAC v2	64
libopus	Opus	64
libfdk_aac	AAC-LC	128
libopus	Opus	128
libfdk_aac	AAC-LC	160
libopus	Opus	160

The subjective tests were fully compliant with the ITU-R BS.1534 [22] standard. The subjective scores were given using the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) format, as recommended for audio with the intermediate quality like that in this dataset. Ten expert assessors [24], also trained according to standard [22], wore high quality Sennheiser HD headphones and assigned quality scores ranging from 0 (bad) to 100 (excellent) for every audio clip in the dataset. The duration of the tests were within the limits of [22] and all tests took place in a sound proof room in Dublin, Ireland in 2014. Further details on this dataset are found in [5].

B. AACvOpus15

The *AACvOpus15* dataset was created to access the quality of the AAC format against the Opus format at a variety of bitrates commonly used by streaming services. The full list of treatments is shown in Table III (all at constant bitrates) and includes 3.5 kHz narrowband as low quality anchor. The samples tested, shown in Table IV, were selected to capture a variety of different audio types including the kind of audio that might be found on YouTube. All audio clips had a frame size of 20 and were resampled to 48 kb/s. With nine treatments and ten samples, the dataset contains a total of 90 audio clips.

The subjective tests were based on ITU-R BS.1534 [22] standard with MUSHRA. A total of 19 expert assessors, also trained according to standard [22], wore high quality AKG K550 headphones. The tests were run in a quiet room where users gave quality scores from 0 to 100 using a Nexus 7 running the HTML-based MUSHRA program described in [5]. The duration of the tests were within the limits of [22] and

TABLE IV
AUDIO SAMPLES IN THE *AACvOpus15* DATASET

Label	Music Type	Source
boz	Rock/R&B (Boz Scaggs)	CD
castanets	Castanets	EBU
contrabassoon	Arpeggio / Melodious Phrase	EBU
glock	Glockenspiel	EBU
guitar	Larry Coryell	EBU
harpsichord	Arpeggio / Melodious Phrase	EBU
hero	Violin and Arrows	DVD
limp	Vocals and Electric Guitar	CD
strauss	R. Strauss (Orchestra)	EBU
vega	Vocals (Suzanne Vega)	CD

TABLE V
AUDIO SAMPLE TREATMENTS IN THE *CoreSV14* DATASET

Codec	Format	Bit rate
libfaac	AAC-LC	48 (VBR)
libfaac	AAC-LC	96 (VBR)
qaac	AAC-LC	96 (CVBR)
libopus	Opus	96 (VBR)
venc603	Ogg Vorbis	96 (VBR)
libmp3lame	MP3	128 (ABR)

all tests took place in a sound proof room in San Francisco, USA in 2015.

C. CoreSV14

The *CoreSV14* dataset [21] was created to access the quality of the Opus format at 96 kb/s compared to AAC and Ogg Vorbis at 96 kb/s and MP3 at 128 kb/s at a variety of bitrates. The full list of treatments is shown in Table V. The libfaac at 48 kb/s and 96 kb/s are used as the low and mid quality anchors, respectively. There are 40 different samples in total including 5 speech samples and 35 music samples. The music samples contain several solos but mostly excerpts from popular songs across many genres. With six treatments and 40 samples, the dataset contains a total of 240 audio clips.

The subjective tests used the ABC/Hidden Reference (ABC/HR) methodology, a hidden reference variation of the ABX methodology [25], where subjects played an uncompressed reference and then rated two files: one being the hidden reference and the other being the compressed audio. Ratings were scored on a continuous impairment scale from 1 (very annoying) to 5 (imperceptible), as described ITU-T BS.562 [15]. Tests were crowd sourced from 30 music enthusiasts of unknown assessor ability. The tests took place in homes of each subject using a variety of sound setups and not in a controlled environment. This dataset is included in the experiments because it covers a wide range of samples and treatments. Further details on the dataset can be found at [21].

V. EXPERIMENT METHODOLOGY

The experiment in Section VI sees the subjective scores in each dataset compared to the objective scores predicted by PEAQ, POLQA, PEMO-Q and ViSQOLAudio. These objective models were selected as each model has been shown to accurately predict perceptual quality for musical audio [3]–[6]. Although POLQA was designed for use with speech rather than music, tests have shown POLQA performs well when predicting the quality of encoded audio [4]. Although a version of POLQA designed specifically for predicting music quality exists [8], this version is currently unavailable for commercial use and so is not used in our experiment.

This section explains the experiment evaluation metrics, the configuration of the objective models, the post-screening process performed on the datasets, and the describes how the SVR in ViSQOLAudio is trained.

A. Evaluation Metrics

It is recommend that objective models should be assessed at least in terms of their linearity, accuracy, and consistency [26]. This section defines the metrics used to determine each of these model properties.

Two fittings are performed as per the recommendation in ITU-T P.1401 [26]. The two fittings are first and third order polynomial regressions of the raw objective quality scores to the MOS-LQSs. Monotonically increasing polynomials for the first and third order fits are found using the Hawkins algorithm [27]. Regression is employed with these monotonic fittings to map objective scores to minimize the RMSE and compensate for biases within the subjective data without changing the rank order of the objective scores [26]. The unmapped and mapped objective quality scores of each model will be compared to the MOS-LQSs using the evaluation metrics described in Section V-A for each treatment in each dataset (as recommended in [28] and [29]). The evaluation metrics are defined as follows.

1) Linearity - R: Pearson's correlation coefficient (R) is used to measure the linear relationship between a sequence of objective and subjective quality scores. R is calculated:

$$R = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \tag{4}$$

where X_i is the MOS-LQS for audio clip i, Y_i is the MOS-LQO (objective score) for audio clip i, $\bar{X}$ is the mean MOS-LQS, $\bar{Y}$ is the mean objective score, and N is the number of audio clips in the dataset.

2) Accuracy - ε-RMSE: The root-mean-square error (RMSE) can be used to describe the absolute prediction error between a sequence of MOS-LQS and objective score values. MOS-LQS values are an average of subjective scores and do not represent variance. The epsilon insensitive root-mean-square ($\boldsymbol{\varepsilon}$-RMSE) can be used to describe the prediction error between a sequence of MOS-LQS and objective score values that accounts for variance in the subjective scores [26]. To consider variance, $\boldsymbol{\varepsilon}$ is first set to the (one-sided) 95% confidence interval of the subjective scores that compose a MOS-LQS. An $\boldsymbol{\varepsilon}$ insensitive prediction error can then be calculated by first predicting an objective score for an audio clip and testing if the score falls within the range of the MOS-LQS $\pm$ $\boldsymbol{\varepsilon}$. If it does, the error for that MOS-LQO prediction is set to 0. $\boldsymbol{\varepsilon}$-RMSE is defined as:

$$\boldsymbol{\varepsilon}\text{-RMSE} = \sqrt{\frac{1}{N-d}\sum_{i=1}^{N}\max(0, |X_i - Y_i| - ci_i)^2} \tag{5}$$

where d is the degree of the polynomial fit and where ci is the 95% confidence interval of the subjective scores for audio clip i. Determining the confidence interval per audio clip is defined:

$$ci_i = t(0.05, M_i)\frac{\sigma_i}{\sqrt{M_i}} \tag{6}$$

where t is the Student's t-distribution, σ_i is the standard deviation of the subjective scores for the audio clip i, and M_i is the number of subjective scores for the audio clip i.

3) Consistency - Outlier Ratio: Prediction score consistency is calculated using the outlier ratio (OR), as recommend in [26], where an outlier is defined for an objective score on an audio clip as:

$$o_i = \begin{cases} 1, & \text{if } |X_i - Y_i| > 2\sigma_i \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $2\sigma_i$ is twice the standard deviation of the subjective scores given for the audio clip i, X_i is the MOS-LQS and Y_i is the objective score for audio clip i. The outlier ratio is therefore given as:

$$OR = \frac{\sum_{i=1}^{N} o_i}{N} \tag{8}$$

where N is the number of audio clips in the dataset.

B. Objective Models Configuration

ViSQOLAudio uses the configuration described in Section III, sampling from the mid channel of the stereo audio and using a ν-SVR with a radial kernel. The Matlab code for ViSQOLAudio can be found at [30]. PEAQ-Advanced (Opera 3.5 distribution) was tested using the default settings given in the example batch files that come with PEAQ: static gain on, the DC filter on, and automatically inverting the test signal. PEMO-Q was tested with the default settings supplied in version 2.0 of the PEASS Toolkit [31], as optimized in [32]. POLQA (version 2.4, with Opticom's POLQA OEM Library for 64-bit Linux, version 1.22) is run in super-wide band mode.

C. Post-Screening of Assessors

It is recommended to remove subjective and objective data prior to analysis if there is reason to believe the data is invalid. The subjective data from the datasets specified in Section IV is removed from the experiment if it meets any of the following criteria specified in ITU-T P.1401 [26] and ITU-R BS.1534 [22]:

1) Data from any subject that did not understand the instructions.
2) Data from a subject that rates the hidden reference condition for more than 15% of the test items less than a score of 90% of the maximum possible score.
3) Data from a subject that rates the mid-range anchor for more than 15% of the test items greater than a score of 90% of the maximum possible score.
4) Data from a sample that more than 25% of the subjects rate the mid-range anchor greater than a score of 90% of the maximum possible score.
5) Data relating to a reference that scored below 4 MOS-LQS.

The following number of subject results were excluded for satisfying one of the post-screening criteria follow: three subjects from *TCDAudio14* for satisfying criteria 3, five subjects from *ACCvOpus15* for satisfying criteria 2, and one subject from *CoreSV14* for satisfying criteria 3. The reason the *CoreSV14* dataset has so few exclusions given its large size is because this dataset had already been screened according to criteria detailed in [21].

PEAQ and other models have been trained on some of the samples in the *TCDAudio14* and *AACvOpus15* datasets. Therefore, none of the models are tested on these samples and no results from these samples are be considered during evaluation. These samples are: *castanets*, *glockenspiel*, *harpsichord* and *vega* from the *TCDAudio14* and *AACvOpus15* datasets. Also, as music is the use-case of interest, the samples *speech English male*, *speech German male* and *speech Korean male* are excluded from the *CoreSV14* dataset.

D. Training and Testing

Each of the models has been trained on a dataset to map signal derived attributes to an objective score. Each test should be performed with the same mapping function. A mapping function is usually trained on several datasets and tested on another. When evaluating the audio clips in the *CoreSV14* dataset, the mapping function for ViSQOLAudio (an SVR described in Section III-F) was trained on frequency band similarity scores (observations) and MOS-LQS values (targets) from the samples in the *TCDAudio14* and *AACvOpus15* datasets.

However, the mapping function for ViSQOLAudio when predicting quality for audio clips in the *TCDAudio14* and *AACvOpus15* datasets is different due to a scarcity of subjective score datasets. When rating a clip in the *TCDAudio14* and *AACvOpus15* datasets, the mapping function will have been trained on all samples in those datasets except for the sample currently being tested, e.g., if predicting the quality of a *boz* audio clip, the mapping function is trained on all clips except for the *boz* clips. This cross-validation approach, necessitated by the scarcity of datasets, is taken to make the mapping function as similar as possible to the one used to test *CoreSV14* while not testing on the same data the model was trained on. By not testing on anything the SVR has been trained on and given that codecs encode different sounds and instruments with different qualities and characteristics [5], we consider it fair to compare the performance of ViSQOLAudio to other objective models for the *TCDAudio14* and *AACvOpus15* datasets.

VI. Experiment

This section presents the results of applying ViSQOLAudio, PEAQ, POLQA and PEMO-Q to each dataset. The quality predictions, both unmapped and mapped with polynomial regression, are compared to the subjective scores to determine the accuracy, consistency and linearity of each model. Mapped quality predictions are then aggregated by group to discuss how models performs on each audio treatment.

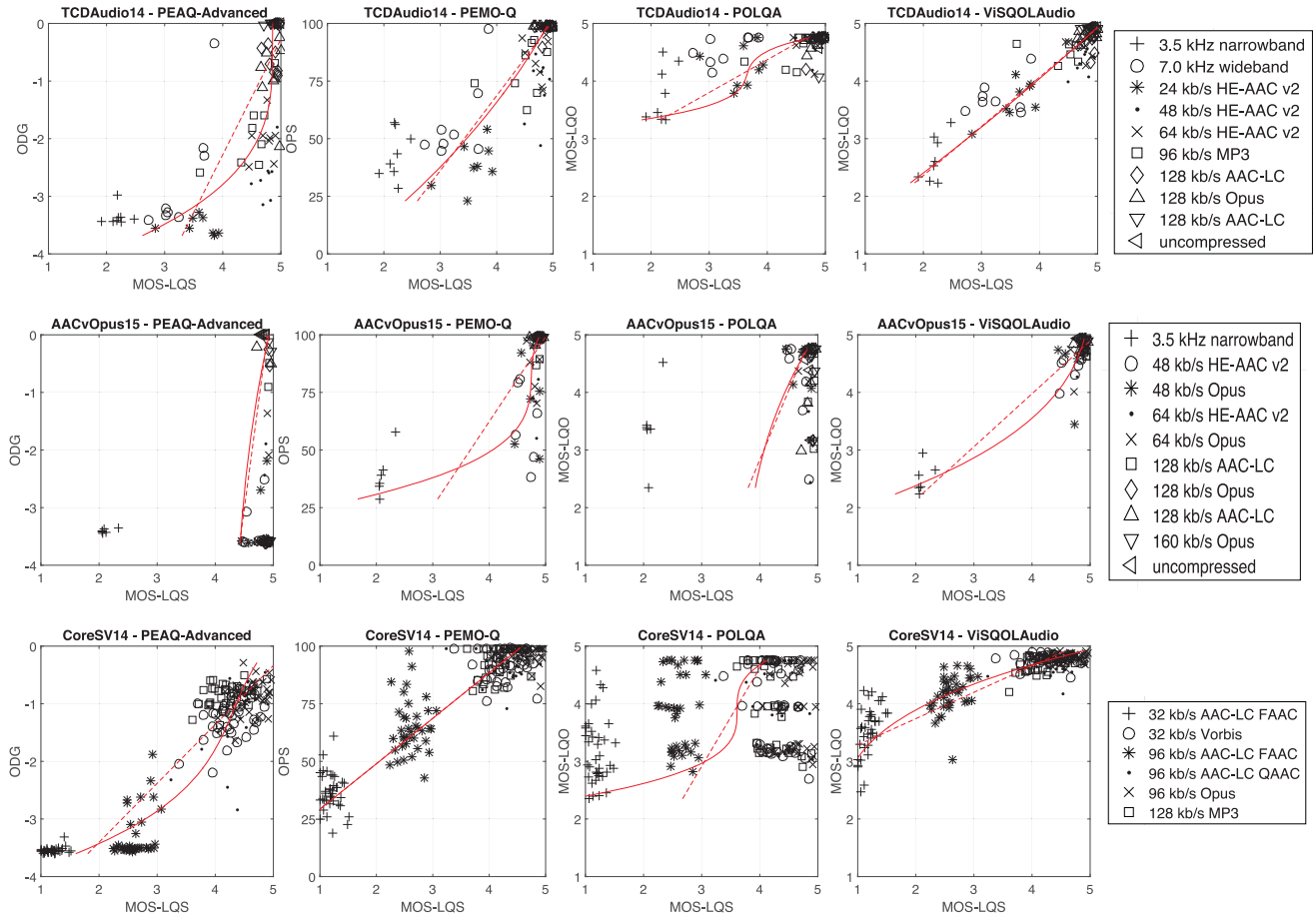

Fig. 4. Subjective versus objective quality scores. First and third order polynomial fits are shown as dashed and solid lines, respectively.

A. Results

Scatter plots show the objective versus subjective scores for each model in Figure 4, demonstrating how well each model performed. The x-axis of each scatter plot is the MOS-LQS of an audio clip and the y-axis is the objective quality prediction for the same audio clip. Each solid line is a third order polynomial fit and each dashed line is a first order fit where the polynomial values were found using the Hawkins algorithm [27].

These scatter plots show that ViSQOLAudio and PEMO-Q fit well to the subjective scores for each dataset. PEAQ consistently underestimates the quality of medium quality audio. PEAQ performs particularly badly on the *AACvOpus15* dataset, where even a third order polynomial cannot reasonably account for the predictions. POLQA has fair performance on *TCDAudio14*, and acceptable performance on *AACvOpus15* with the exception of the predictions for the low-quality anchor clips, but struggles on *CoreSV14* where poor thresholding reduces prediction accuracy.

Table VI presents the $\boldsymbol{\varepsilon}$-RMSE, OR (outlier ratio) and R (Pearson correlation), respectively representing the accuracy, consistency and linearity of model predictions, for the unmapped predictions and the predictions regressed with the first and third order polynomials shown in Figure 4. The unmapped objective predictions are scaled to MOS-LQS (using linear interpolation) to allow a comparison of the unmapped predictions of each model. The results for PEAQ-Basic are included for completeness though discussion of the results will refer exclusively to PEAQ-Advanced as it performed better.

The unmapped results in Table VI show that ViSQOLAudio scores correlate strongly with the subjective scores across all datasets and that ViSQOLAudio has the lowest $\boldsymbol{\varepsilon}$-RMSE in two of the three datasets. The OR for PEAQ is high for the *AACvOpus15* dataset because of the accurately predicted low pass (narrow and wideband) audio clips being poorly fitted, and because of the poorly predicted medium quality audio. PEAQ also has the lowest $\boldsymbol{\varepsilon}$-RMSE for the *CoreSV14* dataset. PEAQ benefits the most from the polynomial fits, with $\boldsymbol{\varepsilon}$-RMSE dropping substantially for the *TCDAudio14* and *AACvOpus15* datasets. PEMO-Q performs well on all datasets, always near to the best in each dataset. The unmapped predictions for POLQA are reasonably accurate for the high quality audio clips in *TCDAudio14* and *AACvOpus15* but inaccurate for low quality clips. The predictions for *CoreSV14* are particularly poor compared to the other models, where POLQA seems unable to distinguish high and low quality clips. However, emphasis should not be placed on the results of the unmapped evaluation metrics as they do not account for potential bias in the

TABLE VI
EVALUATION METRICS FOR OBJECTIVE MODELS ON ALL TREATMENTS

		unmapped scaled to MOS-LQS			first order polynomial mapped			third order polynomial mapped		
dataset	model	R	OR	ε-RMSE	R	OR	ε-RMSE	R	OR	ε-RMSE
TCDAudio14	PEAQ-Adv.	0.73	0.25	1.15	0.74	**0.00**	0.40	0.81	**0.00**	0.32
TCDAudio14	PEAQ-Basic	0.50	0.34	1.32	0.50	0.03	0.55	0.52	0.01	0.55
TCDAudio14	PEMO-Q	0.82	0.01	0.40	0.82	**0.00**	0.31	0.82	**0.00**	0.31
TCDAudio14	POLQA	0.72	0.04	0.50	0.72	0.01	0.43	0.74	**0.00**	0.42
TCDAudio14	ViSQOLAudio	**0.93**	**0.00**	**0.18**	**0.93**	**0.00**	**0.17**	**0.93**	**0.00**	**0.17**
AACvOpus15	PEAQ Adv.	0.22	0.65	2.47	0.22	0.10	0.70	0.23	0.10	0.71
AACvOpus15	PEAQ-Basic	0.09	0.58	2.65	0.09	0.10	0.72	0.22	0.10	0.71
AACvOpus15	PEMO-Q	0.71	0.05	0.58	0.71	**0.00**	0.46	0.81	0.03	0.40
AACvOpus15	POLQA	0.35	0.17	0.85	0.35	0.10	0.67	0.36	0.10	0.68
AACvOpus15	ViSQOLAudio	**0.94**	**0.00**	**0.19**	**0.94**	**0.00**	**0.18**	**0.96**	**0.00**	**0.14**
CoreSV14	PEAQ-Adv.	0.92	**0.00**	**0.36**	0.92	**0.00**	**0.28**	**0.93**	**0.00**	**0.23**
CoreSV14	PEAQ-Basic	0.90	0.01	0.43	0.90	**0.00**	0.37	0.91	**0.00**	0.36
CoreSV14	PEMO-Q	**0.93**	**0.00**	0.57	**0.93**	**0.00**	**0.28**	**0.93**	**0.00**	0.28
CoreSV14	POLQA	0.33	0.04	1.00	0.33	0.01	0.94	0.43	0.03	0.90
CoreSV14	ViSQOLAudio	0.89	0.05	0.99	0.91	**0.00**	0.34	0.91	**0.00**	0.29

datasets, which the polynomial fittings are meant to compensate for.

For mapped data, the first order polynomial regressed data follows the same pattern as the third order regressed results. PEAQ benefits most from the data regressions, substantially reducing its associated ε-RMSE and OR. For third order mapped data, the OR values for all models but PEAQ drop to negligible values. ViSQOLAudio continues to have to lowest ε-RMSE values in two of the three datasets. PEAQ continues to be the most accurate for *CoreSV14*. POLQA does not benefit as much from the mappings as the other models because of the poor correlation between its objective score and the subjective scores.

Table VII shows the evaluation metrics for all score predictions but without the anchors (3.5 and 7.0 kHz treatments for *TCDAudio14*, 3.5 kHz for *AACvOpus15*, and both 32 kb/s treatments for *CoreSV14*). For the unmapped data, ViSQOLAudio has the best ε-RMSE for all datasets. OR values are high for PEAQ and POLQA due to the large variation in prediction quality for high quality audio clips. Correlations are generally poor across all datasets and models because, without anchors, the results cluster only around the MOS-LQS 4 to 5 area, giving little direction to the mappings. This is particularly true for POLQA, where having a negative correlation made it impossible to find a monotonically increasing fit using the fitting algorithm, which is why the mapped data for POLQA for the *AACvOpus15* and *CoreSV14* datasets have been excluded.

For the no-anchor third order polynomial regressed data, only the metrics from *TCDAudio14* and *CoreSV14* should be considered reliable as there is a reasonably large spread in quality among the audio clips. However, for *AACvOpus15*, without anchors, the fittings for all models is a nearly vertical line (as opposed to the fittings with anchors shown in the *AACvOpus15* plots in Figure 4) because there is too little difference in the subjective scores of non-anchor audio clips. This puts almost all predictions within the epsilon of the RMSE and results in unreliable ε-RMSE values. In *TCDAudio14* and *CoreSV14*, PEAQ performs best by nearly all metrics. However, as seen in Figure 4 for PEAQ *TCDAudio14*, with the anchors, the fitting is a steep curve, pushing the inaccurate high quality audio clip quality predictions up to the range where the majority of subjective scores are. PEMO-Q experiences the same problem to a lesser extent.

Table VIII shows the evaluation metrics for ViSQOLAudio and ViSQOLAudio as it was in 2015 [4]. Across all datasets, mappings and metrics, ViSQOLAudio is as good or substantially better than its predecessor. This affirms the benefit of the changes to the ViSQOLAudio model.

A breakdown of the accuracy of each model by treatment is shown in Figure 5. These box plots compare the subjective scores to the third order polynomial regressed objective scores. The error bars represent 95% confidence intervals.

The subjective scores in Figure 5 highlight that all treatments with a bitrate above 48 kb/s for all but AAC-LC FAAC audio clips have a score near 4.5 MOS-LQS. The figure also shows that objective score accuracy increases with perceptual audio quality suggesting that all models are generally reliable for high quality audio. In all datasets, ViSQOLAudio MOS-LQO mean values are always less than 0.5 from the MOS-LQS mean values.

For all datasets, models were least accurate and had the highest variation when scoring low quality treatments, such as anchors and files with bitrates of 48 kb/s and lower.

The results from the tested datasets indicate that PEAQ is inaccurate for predicting low-bitrate audio quality, especially for the *TCDAudio14* and *AACvOpus15* anchors. PEAQ also exhibits an unusual pattern of predicting large differences in quality for clips with the same treatment, as clearly demonstrated in the *AACvOpus15* dataset by the large PEAQ error bars, even at high bitrates. This large variation in quality prediction suggests that PEAQ is quite sensitive to different kinds of sample content, e.g., *guitar* samples are predicted correctly but *contrabassoon* samples are predicted poorly.

PEMO-Q is accurate on all but the low quality anchors of *TCDAudio14* and *AACvOpus15*. The variation in PEMO-Q scores is large at low bitrates but reduces to more acceptable levels at greater than 4 MOS-LQS values. This variation suggests that PEMO-Q becomes less sensitive to different kinds of sample content as perceptual audio quality increases.

POLQA is inaccurate when predicting the quality of anchors for each of the datasets. POLQA also has a consistently

TABLE VII
EVALUATION METRICS FOR OBJECTIVE MODELS ON ALL TREATMENTS EXCLUDING THE ANCHOR TREATMENTS

		unmapped scaled to MOS-LQS			first order polynomial mapped			third order polynomial mapped		
dataset	model	R	OR	ε-RMSE	R	OR	ε-RMSE	R	OR	ε-RMSE
TCDAudio14	PEAQ-Adv.	0.71	0.50	1.25	0.72	0.02	0.20	0.84	**0.00**	**0.14**
TCDAudio14	PEAQ-Basic	0.67	0.52	1.48	0.67	0.02	0.21	0.82	0.02	0.17
TCDAudio14	PEMO-Q	**0.83**	0.16	0.43	**0.83**	**0.00**	**0.15**	**0.85**	0.02	**0.14**
TCDAudio14	POLQA	0.60	0.03	0.24	0.60	0.05	0.24	0.65	0.03	0.22
TCDAudio14	ViSQOLAudio	0.82	**0.02**	**0.17**	0.82	0.02	**0.15**	**0.85**	0.02	**0.14**
AACvOpus15	PEAQ-Adv.	0.28	0.83	2.60	0.28	0.07	**0.02**	0.29	0.07	**0.02**
AACvOpus15	PEAQ-Basic	0.27	0.80	2.79	0.27	0.07	**0.02**	0.27	0.07	**0.02**
AACvOpus15	PEMO-Q	0.45	0.37	0.60	0.45	0.09	**0.02**	**0.50**	**0.04**	**0.02**
AACvOpus15	POLQA	-0.15	0.52	0.81	N/A	N/A	N/A	N/A	N/A	N/A
AACvOpus15	ViSQOLAudio	**0.46**	**0.22**	**0.18**	**0.46**	**0.06**	**0.02**	**0.50**	**0.04**	**0.02**
CoreSV14	PEAQ-Adv.	**0.40**	0.20	0.33	**0.40**	**0.02**	**0.11**	0.40	**0.01**	0.11
CoreSV14	PEAQ-Basic	0.36	**0.19**	0.48	0.36	**0.02**	0.13	**0.50**	**0.01**	**0.10**
CoreSV14	PEMO-Q	0.21	0.20	0.21	0.21	0.03	0.15	0.21	0.03	0.15
CoreSV14	POLQA	-0.20	0.48	0.62	N/A	N/A	N/A	N/A	N/A	N/A
CoreSV14	ViSQOLAudio	0.29	**0.19**	**0.13**	0.29	**0.02**	0.13	0.29	0.02	0.13

TABLE VIII
EVALUATION METRICS FOR THE OLD AND NEW VERSION OF VISQOLAUDIO FOR ALL TREATMENTS

		unmapped scaled to MOS-LQS			first order polynomial mapped			third order polynomial mapped		
dataset	model	R	OR	ε-RMSE	R	OR	ε-RMSE	R	OR	ε-RMSE
TCDAudio14	ViSQOLAudio 2015	0.81	**0.00**	0.34	0.82	**0.00**	0.32	0.83	**0.00**	0.31
TCDAudio14	ViSQOLAudio	**0.93**	**0.00**	**0.18**	**0.93**	**0.00**	**0.17**	**0.93**	**0.00**	**0.17**
AACvOpus15	ViSQOLAudio 2015	0.86	**0.00**	0.37	0.90	**0.00**	0.26	0.94	**0.00**	0.20
AACvOpus15	ViSQOLAudio	**0.94**	**0.00**	**0.19**	**0.94**	**0.00**	**0.18**	**0.96**	**0.00**	**0.14**
CoreSV14	ViSQOLAudio 2015	0.79	0.13	1.16	0.79	**0.00**	0.54	0.80	**0.00**	0.54
CoreSV14	ViSQOLAudio	**0.89**	**0.05**	**0.99**	**0.91**	**0.00**	**0.34**	**0.91**	**0.00**	**0.29**

large variation in its quality predictions for samples with the same treatment. The mean POLQA quality predictions for all treatments with a bitrate of 48 kb/s or more is lower than the MOS-LQS values, with the exception of FAAC 96 kb/s AAC-LC. This is likely due to the music being interpreted as noise because the instruments have few of the characteristics of a voice.

VII. DISCUSSION

Before describing the findings from the objective scores, we will first take a moment to describe the difference in the subjective scores across the datasets as these have a large impact on the evaluation metrics.

Although the subjective tests for the *AACvOpus15* dataset were based on ITU-R BS.1534 [19], it deviated from the standard by having subjects told an uncompressed reference was among each set of degradations and that the subjects must assign a score of 100 to at least one of audio clips per audio clip set. This deviation did not greatly impact the subjective scores as the scores in *AACvOpus15* are close to those in *TCDAudio14* for clips with the same treatment, where subjects for the *TCDAudio14* could vote without condition.

The *CoreSV14* dataset used the ITU-T BS.562 [15] impairment scale from 1 (very annoying) to 5 (imperceptible) whereas the *TCDAudio14* and *AACvOpus15* datasets used the ITU-R recommendation BS. 1534-3 [33] scale of 0 (bad) to 100 (excellent). The wording of low quality ratings may have affected the scores, explaining why the low quality anchor in *CoreSV14* has a MOS-LQS mean around 1.2 while the low quality anchors in *TCDAudio14* and *AACvOpus15* have a MOS-LQS mean around 2. The anchor scores are made even more puzzling given that, in the opinion of the authors, the (3.5 kHz narrowband) low quality anchor audio in *TCDAudio14* and *AACvOpus15* are perceptually lower in quality than the (32 kb/s AAC-LC) low quality anchor audio in *CoreSV14*. Moreover, the low quality anchor scores are consistent across the *TCDAudio14* and *AACvOpus15* datasets despite using different subjects, and the low quality anchor scores in *CoreSV14* are very consistent around 1.2 MOS-LQS. However, this kind of inconsistency is simply the nature of subjective tests.

When considering all treatments, MOS-LQS means are generally lower in the *CoreSV14* dataset than the other datasets. This is illustrated by the MOS-LQS means of Opus and MP3 treated audio clips present in both datasets. The MOS-LQS mean per treatment is lower in *CoreSV14* for Opus and MP3 even at bitrates higher than those tested in the *TCDAudio14* and *AACvOpus15* datasets. The low subjective scores in *CoreSV14* help explain why the models consistently underestimate the quality of *CoreSV14* audio clips.

For mapped and unmapped objective scores, with respect to the tested data, PEAQ was inaccurate on *TCDAudio14* and *AACvOpus15* when compared to PEMO-Q and ViSQOLAudio (Table VI). We believe this may be because the bulk of PEAQ's training has been performed on a different scale and with very different low quality anchors to the narrowband anchors used in *TCDAudio14* and *AACvOpus15*. This makes sense when considering the accuracy of PEAQ on *CoreSV14*, which had a low quality anchor with quality much higher than that of the *TCDAudio14* and *AACvOpus15* low quality anchor. The large variation in PEAQ quality predictions among high quality audio shows an undesirably strong sensitivity to sample content.

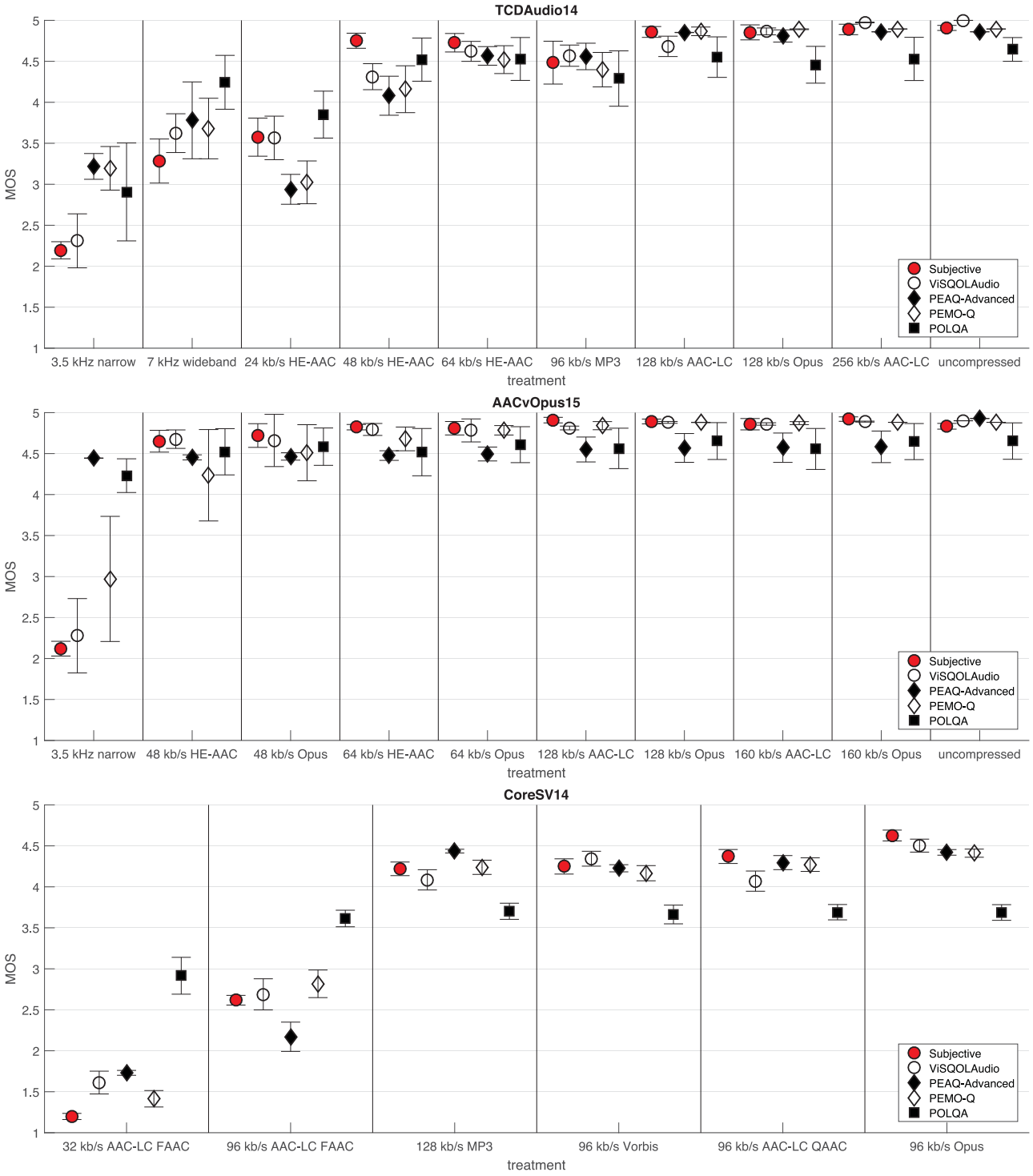

Fig. 5. Box plot of third order polynomial mapped data by treatment with 95% confidence intervals.

PEMO-Q performed well overall, with reasonable correlation to the subjective scores for all sets on mapped and unmapped data. PEMO-Q predictions became increasingly accurate and had a lower variation across samples for perceptually higher quality samples. PEMO-Q predictions varied greatly for low bitrate audio, as seen in *TCDAudio14*. Overall, PEMO-Q performs well in linearity, consistency and accuracy across all datasets.

POLQA performed well for high quality audio clips in *TCDAudio14* and *AACvOpus15* but not *CoreSV14*. POLQA performed poorly for all other treatments, when compared to the predictions of the other models. This is likely because POLQA is specialized to identify and extract voice data from audio and does not translate well to a musical domain, unlike POLQA Music [8], which is not released at the time of this publication.

ViSQOLAudio performed well on the *TCDAudio14* and *AACvOpus15* datasets, having the best linearity and accuracy on third order polynomial mapped data with anchors. ViSQOLAudio was able to give accurate predictions for both low and high quality audio. ViSQOLAudio also gave the most accurate quality predictions for all but one of the anchor treatments.

Ideally, all models would be trained and tested using common data, however datasets used to develop PEAQ and PEMO-Q were not available to the authors. The *CoreSV14* tests with an ITU-T BS.562 [15] scale rather than MUSHRA scale data that were used to train ViSQOLAudio's mapping scale revealed the robustness of ViSQOLAudio's performance. It can be seen in Table VIII that the improvements in correlation statistics between the proof of concept ViSQOL model adaptation [4] and the newly presented ViSQOLAudio model is largest for the CoreSV14 dataset (i.e., the dataset not used during training). This highlights that the improvements for all datasets come from the other improvements to the model and are not simply as a result of training to map from a similarity measure to a MOS value. This reassured the authors that the leave-one-out training model has not given ViSQOLAudio an unfair advantage when evaluating performance with the *TCDAudio14* and *AACvOpus15* datasets.

ViSQOLAudio was greatly enhanced by the culmination of a number of improvements (as demonstrated in Table VIII). Apart from being more accurate, ViSQOLAudio also has greater utility than its predecessor as it now outputs an easily interpreted MOS-LQO value rather than just a similarity score.

ViSQOLAudio overestimates scores in *CoreSV14* as a result of its training data. MOS-LQS values of the low quality anchors in the datasets that ViSQOLAudio was trained on (*TCDAudio14* and *AACvOpus15*) were higher than the MOS-LQS values for low quality anchor in *CoreSV14*, though the *CoreSV14* audio was of perceptually higher quality. As the *TCDAudio14* low quality anchors scored around 2 MOS-LQS, the *CoreSV14* low quality anchor could not be given a score lower as they were perceptual higher in quality than the *TCDAudio14* low quality anchor. Upon removing the anchors from each of the datasets, ViSQOLAudio then had the highest accuracy for unmapped data across all datasets (Table VII), including *CoreSV14*.

As well as limitations to the machine learning approach given the current data, there are weaknesses to the use of mid channel data to consider information from both channels of a stereo signal. For example, consider a stereo file where the left channel signal is 180 degrees out of phase with the right channel signal, resulting in a silent mid channel. However, no such issue was found in the tested music domain.

VIII. Conclusion

The goal of this paper was to determine the viability of objective perceptual audio quality models as a tools for codec regression testing. This was done by evaluating the accuracy, linearity and consistency of perceptual quality predictions from ViSQOLAudio, PEAQ, POLQA and PEMO-Q compared to the subjective quality scores. The evaluation was performed on encoded musical audio with a variety of samples and treatments. The results showed that ViSQOLAudio performed best on all metrics for two of the three datasets and just short of the best accuracy for the third dataset. These results demonstrate that ViSQOLAudio, a free and open source objective metric, is a powerful alternative to PEAQ, POLQA, PEMO-Q when evaluating perceptual audio quality at a variety of bitrates making it suitable for codec regression testing. Future work on ViSQOLAudio will focus on finding a more robust method for handling stereo audio and investigating wavelet transforms in place of the Gammatone filterbank.

References

[1] "ITU-R Rec. BS.1387: Method for objective measurements of perceived audio quality," Int. Telecomm. Union, Geneva, Switzerland, 2001.
[2] "ITU-T Rec. P.863: Perceptual objective listening quality assessment," Int. Telecomm. Union, Geneva, Switzerland, 2014.
[3] R. Huber and B. Kollmeier, "PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception," *IEEE Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.
[4] A. Hines *et al.*, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *J. Acoust. Soc. Amer.*, vol. 137, no. 6, pp. EL449–EL455, 2015.
[5] A. Hines *et al.*, "Perceived audio quality for streaming stereo music," in *Proc. 22nd ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 1173–1176.
[6] T. Thiede *et al.*, "PEAQ—The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, nos. 1–2, pp. 3–29, 2000.
[7] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Visqol: An objective speech quality model," *EURASIP J. Audio Speech Music Process.*, vol. 2015, no. 1, p. 13, 2015. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-015-0054-9
[8] P. Počta and J. G. Beerends, "Subjective and objective assessment of perceived audio quality of current digital audio broadcasting systems and Web-casting applications," *IEEE Trans. Broadcast.*, vol. 61, no. 3, pp. 407–415, Sep. 2015.
[9] "ITU-T Rec. G.107—The E-model, a computational model for use in transmission planning," Int. Telecomm. Union, Geneva, Switzerland, 2003.
[10] "ITU-T Rec. P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications," Int. Telecomm. Union, Geneva, Switzerland, 2004.
[11] "ITU-T Rec. P.861: Objective quality measurement of telephone-band (300-3400 Hz) speech codecs," Int. Telecomm. Union, Geneva, Switzerland, 1996.
[12] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 2. Salt Lake City, UT, USA, 2001, pp. 749–752.
[13] "ITU-T Rec. P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Int. Telecomm. Union, Geneva, Switzerland, 2001.
[14] C. D. Creusere, K. D. Kallakuri, and R. Vanam, "An objective metric of human subjective audio quality optimized for a wide range of audio fidelities," *IEEE Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 129–136, Jan. 2008.
[15] "ITU-R Rec. BS.562: Subjective assessment of sound quality," Int. Telecomm. Union, Geneva, Switzerland, 1990.
[16] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the opus audio codec," Internet Eng. Task Force (IETF), Reston, VA, USA, Tech. Rep. RFC 6716, 2012.
[17] M. Taylor. (2000). *LAME Technical FAQ*. [Online]. Available: http://lame.sourceforge.net/tech-FAQ.txt
[18] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Commun.*, vol. 54, no. 2, pp. 306–320, 2012.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[20] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[21] CoreSV Team. (2014). *CoreSV Listening Test*. [Online]. Available: http://listening-test.coresv.net/results.htm
[22] "ITU-R Rec. BS.1534: Subjective assessment of sound quality," Int. Telecomm. Union, Geneva, Switzerland, 2015.
[23] G. Waters, "Sound quality assessment material recordings for subjective tests: Users handbook for the EBU–SQUAM compact disk," Eur. Broadcast. Union (EBU), Geneva, Switzerland, Tech. Rep 3253-E, 1988.
[24] *8586. Sensory Analysis—General Guidelines for the Selection, Training and Monitoring of Selected Assessors and Expert Sensory Assessors*, Int. Organ. Standardization (ISO), Geneva, Switzerland, 2012.
[25] W. Munson and M. B. Gardner, "Standardizing auditory tests," *J. Acoust. Soc. Amer.*, vol. 22, no. 5, p. 675, 1950.
[26] "ITU-T Rec. P.1401: Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," Int. Telecomm. Union, Geneva, Switzerland, 2012.
[27] K. Murray, S. Müller, and B. A. Turlach, "Fast and flexible methods for monotone polynomial fitting," *J. Stat. Comput. Simulat.*, vol. 86, no. 15, pp. 2946–2966, 2016.
[28] S. Möller *et al.*, "Speech quality estimation: Models and trends," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 18–28, Nov. 2011.
[29] T. H. Falk *et al.*, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.
[30] (2016). *ViSQOLAudio [Computer Program]*. Accessed on Sep. 9, 2016. [Online]. Available: http://www.sigmedia.tv/tools
[31] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.
[32] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*. Tel Aviv, Israel, 2012, pp. 430–437.
[33] "ITU-R Rec. BS.1534-3: Method for the subjective assessment of intermediate quality levels of coding systems," Int. Telecomm. Union, Geneva, Switzerland, 2015.

Colm Sloan received the Ph.D. degree from the Artificial Intelligence Research Centre, Dublin Institute of Technology. Since then, he has been a Post-Doctorate Researcher with the CONNECT Research Centre, Trinity College Dublin.

Naomi Harte is an Associate Professor in Digital Media Systems with the School of Engineering. In 2015, she was a Visiting Professor with ISCI in Berkeley, CA, USA. Her work research interests focus on human-to-human and human-to-machine speech communication, specifically speech quality, audio visual speech processing, speaker verification for biometrics, and emotion in speech.

Damien Kelly received the B.A./B.A.I. degree in computer and electronic engineering from Trinity College Dublin, Dublin, Ireland, in 2005, and the Ph.D. degree in 2010. Since then, he has been a Research Fellow with the Media Processing Group (www.sigmedia.tv), Department of EEE, Trinity College Dublin and for Green Parrot Pictures Ltd. developing software tools for video enhancement. In 2011, he joined the Chrome Media Team with Google. In 2013, he joined the Video Infrastructure Team with YouTube, where he currently works on VR audio, audio/video transcoding and quality enhancement. His research interests include perceptual audio/video quality metrics, spatial audio, and audio-visual tracking.

Anil C. Kokaram is a Professor with Trinity College Dublin, Ireland and leads the www.sigmedia.tv research group. His main expertise is in the broad areas of DSP for Video Processing, Bayesian inference, and motion estimation. In 2007, he was a recipient of the Science and Engineering Academy Award for his work in video processing for post-production applications. He is a former Associate Editor of the IEEE TRANSACTIONS ON VIDEO TECHNOLOGY and the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Andrew Hines is a Lecturer with the School of Computing, Dublin Institute of Technology, Ireland, and an Adjunct Assistant Professor with Trinity College Dublin. His primary research interests are in speech, audio, and video signal processing. He has develop metrics for predicting speech intelligibility for people with hearing impairments, speech and audio quality for Voice over IP (VoIP), and audio codec compression degradations. His broader research interests include using signal processing and machine learning for data driven quality of experience prediction for a variety of domains.