2016-6

Bitrate Classification of Twice-Encoded Audio using Objective Quality Features

Colm Sloan
Trinity College Dublin, sloanco@tcd.ie

Damien Kelly
Google.inc.

Naomi Harte
Trinity College Dublin

Anil C. Kokaram
Google, Inc.

Andrew Hines
Dublin Institute of Technology, andrew.hines@dit.ie

Recommended Citation

C. Sloan, N. Harte, D. Kelly, A. C. Kokaram and A. Hines, (2016) Bitrate classification of twice-encoded audio using objective quality features," *Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon*, 2016, pp. 1-6. doi: 10.1109/QoMEX.2016.7498956

Bitrate Classification of Twice-Encoded Audio using Objective Quality Features

Colm Sloan*, Naomi Harte*, Damien Kelly†, Anil C. Kokaram† and Andrew Hines*‡
*Sigmedia, Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland
†Google, Inc., Mountain View, CA, USA
‡School of Computing, Dublin Institute of Technology, Dublin, Ireland
Email: sloanco@tcd.ie, andrew.hines@dit.ie

***Abstract*—When a user uploads audio files to a music streaming service, these files are subsequently re-encoded to lower bitrates to target different devices, e.g. low bitrate for mobile. To save time and bandwidth uploading files, some users encode their original files using a lossy codec. The metadata for these files cannot always be trusted as users might have encoded their files more than once. Determining the lowest bitrate of the files allows the streaming service to skip the process of encoding the files to bitrates higher than that of the uploaded files, saving on processing and storage space. This paper presents a model that uses quality predictions from ViSQOLAudio, a full reference objective audio quality metric, as features in combination with a multi-class support vector machine classifier. An experiment on twice-encoded files found that low bitrate codecs could be classified using audio quality features. The experiment also provides insights into the implications of multiple transcodes from a quality perspective.**

I. Introduction

Streaming services such as Google Play Music and SoundCloud handle terabytes of audio data every week. These services aim to encode audio with a balance between quality of experience (QoE) [1] for the end user, the size of the encoded audio files, and the processing cost of the encoding. Users may upload files to a streaming service that have already been encoded because the user wants to reduce file size to decrease upload time. The same audio encoded as a 3 MB uncompressed WAV, a 510 KB 256kb/s AAC-LC, or a 250 KB 128 kb/s Opus all seem similar in quality to expert listeners [2]. Streaming services encode audio to a number of bitrates and formats to provide the best experience for users of different devices. For example, mobile users may prefer to compromise quality to limit bandwidth consumption. Services do not encode to bitrates higher than that of the uploaded files as there will be no increase in quality. Determining the lowest bitrate of the files allows the streaming service to forgo encoding the files to bitrates higher than that of the uploaded files, saving on processing and storage space.

A number of approaches have been taken to correctly classify the codec and bitrate used to encode audio that has been encoded multiple times. We will refer to the combination of codec and bitrate as the *treatment*. Deep learning has been used to detect if Adaptive Multi-Rate files have been re-encoded, to alert for potential tampering [3]. Frame offsets have been used to detect MP3 files that have been re-encoded to a bitrate higher than the first encoding to advertise a quality they do not deliver [4]. Power spectral density was shown to identify lossy AAC-LC files that have been converted to lossless ALAC or FLAC files [5]. Another strategy measured the similarity between the modified discrete cosine transform histograms of once-encoded and twice-encoded MP3 files to determine the bitrate of the first pass for detecting tampering of the twice-encoded MP3 [6].

978-1-5090-0354-9/$31.00

The two techniques most similar to ours are described in [7]. The first method uses a number of quality measures to train support vector machines (SVMs) to classify the bitrate of once-encoded audio files. The second method uses an analysis of an audio bit stream to train SVMs that predict the lowest encoded bitrate of twice-encoded MP3 files where the first encoded bitrate is higher than the second.

For twice encoded audio, it is more difficult to predict the lower bitrate used when the bitrate of the second-pass treatment is lower than the first-pass. Our experiments use audio quality measures with Multi-Class Support Vector Machines (MC-SVMs) for predicting the lowest encoded bitrate of twice-encoded files where the codec used of the first and second pass are different, and where the bitrates of the second encodings can be higher, the same, or lower than the bitrate of the first encoding. Additionally, we have used a wider collection of codecs and bitrates in our experimental evaluation than has been previously reported in the literature.

The technique proposed in this paper relies on treatments applied to multiple short samples and is seen as a first step toward an more robust model with fewer constraints.

This paper is structured as follows: Section II reflects on the impact of sample selection on audio quality and describes the audio dataset and experimental encoding scenarios used in this paper. Section III outlines the notation used in the paper before Section IV details the experiment to validate and then evaluate our approach. Section V discusses the results before we present conclusions in Section VI.

II. Content-Quality Relationship for Codecs

Experiments have shown that the perceived quality for a given treatment (bitrate and codec) can vary depending on the content of the audio sample [2]. This phenomena has also been observed using objective metrics [8] along with the fact that treatments with imperceptible quality differences are similarly ranked in terms of their predicted quality. As a result, only treatments that are perceptually different from the perspective of the quality model will produce quality estimates with a

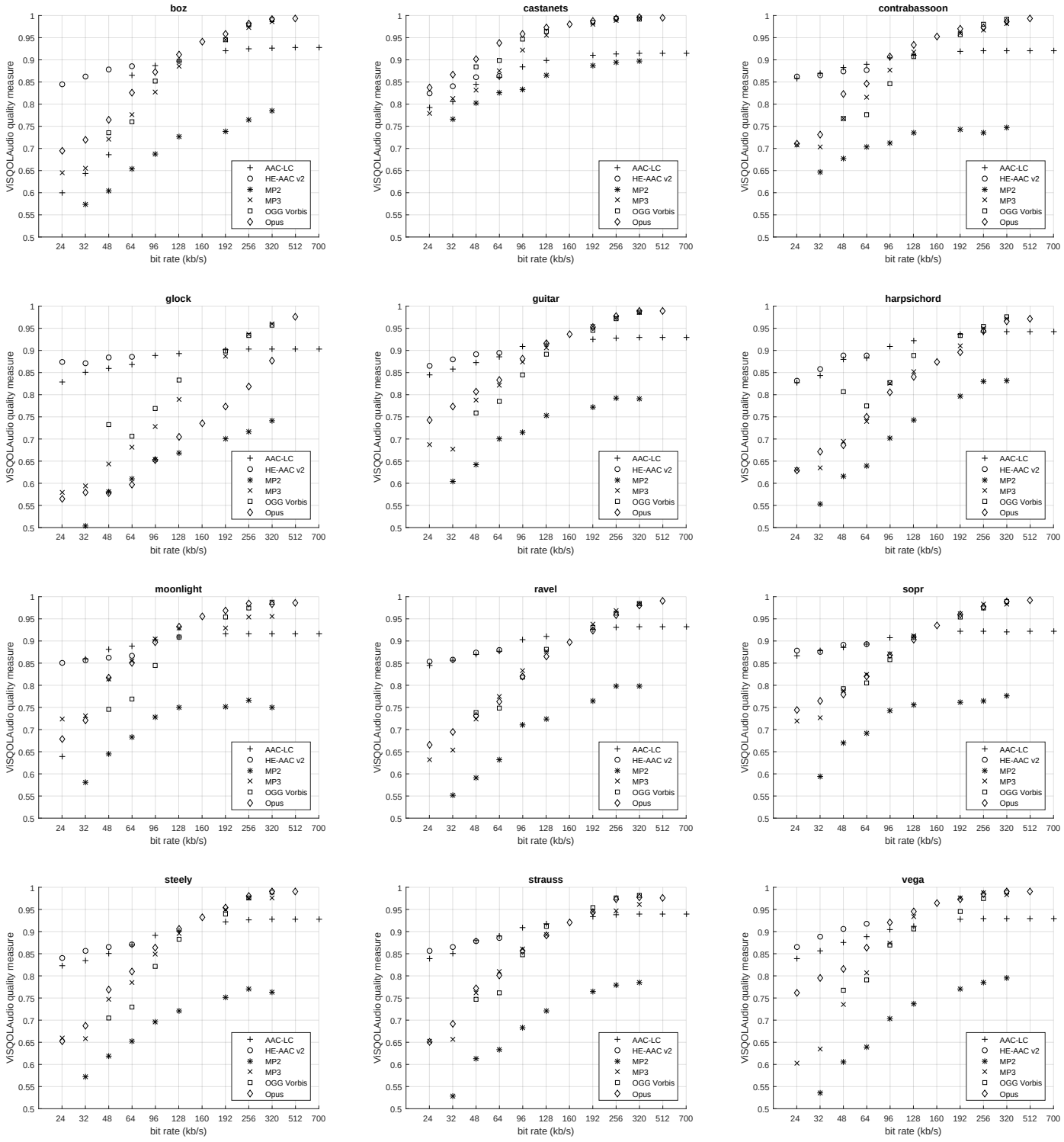


Fig. 1: ViSQOLAudio quality predictions. Demonstrates the variation in quality for treatment across different samples.

TABLE I: Treatments used to encode the 12 uncompressed audio file samples to create the set of single-pass samples D^1. Each ✓ indicates a treatment that was used to train the MC-SVMs in the experiment in Section IV, where the set of ticked treatments is T^1. Each ✗ indicates treatments analysed in Section II that were found unsuitable for the experiment.

codec	FFMPEG encoder	bitrates(kb/s)											
		24	32	48	64	96	128	160	192	256	320	512	700
AAC-LC	libfdk_aac	✓	✓	✓	✓	✓	✓		✓	✗	✗	✗	✗
HE-ACC-v2	libfdk_aac	✗	✗	✗	✗								
MP2	libtwolame		✓	✓	✓	✓	✓		✓	✓	✗		
MP3	libmp3lame	✓	✓	✓	✓	✓	✓		✓	✓	✗		
Ogg Vorbis	libvorbis			✓	✓	✓	✓		✓	✓	✓		
Opus	libopen	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗

variation that can be used to develop treatment classification features.

A number of models exist that can be used to estimate the objective quality of audio and the QoE of a user listening to the audio [9]. ViSQOLAudio is a full-reference signal-based objective metric for measuring the quality of audio, where the quality measure is how similar a human subject would rate a reference audio file against a degraded audio file [8]. ViSQOLAudio was selected for use in this research because it has been shown to be accurate at estimating the quality of degraded audio files [8], though we could equally have used another objective metric in this proof-of-concept.

The uncompressed audio samples used in this paper are the same as those used in [2], and are described in Table II. These samples are used as references by ViSQOLAudio when calculating the quality of encoded (degraded) samples. The reference audio is uncompressed PCM WAV and the degraded audio is the reference audio encoded to some codec and at a lower bitrate. The uncompressed reference files consist of stereo music samples of 7 to 15 seconds in duration covering a variety of musical sounds. The clips were selected to cover a range of genres and instruments. The audio files in this set were sourced from CDs and the EBU music database [10] and were all originally 48 or 44.1 kHz, 16-bit stereo. PCM WAV files were created at 48 kHz for all files. These uncompressed samples were then encoded using FFMPEG using the treatments shown in Table I, which were selected because these treatments are typical of current digital audio broadcasting systems [11]. One second portions of audio for each sample were used as inputs to ViSQOLAudio. These one second samples were found to give similar quality estimates as to when the entire signal was used as input but required less processing time.

ViSQOLAudio predicts perceived audio quality by comparing a reference signal to a test signal. It has a three stage process: preprocessing, alignment, and comparison. It uses a Neurogram Similarity Index Measure (NSIM) [12] to compare spectrograms across critical frequency bands from 50 Hz to 16000 Hz and patches consisting of 30 0.016 s time frames and outputs a quality score between 0 and 1, with 1 signifying perfect quality. Although the samples are stereo, when predicting quality with ViSQOLAudio, only the audio in the left channel of the samples is considered.

Figure 1 is illustrative of the quality predictions from ViSQOLAudio using twelve samples. The bitrates are shown on the X-axis and the quality measure of a sample with a given treatment are shown along the Y-axis. A wide variation in quality across bitrates is evident. Generally, samples encoded at bitrates below 128 kb/s had a much wider variation in

TABLE II: Audio Samples

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

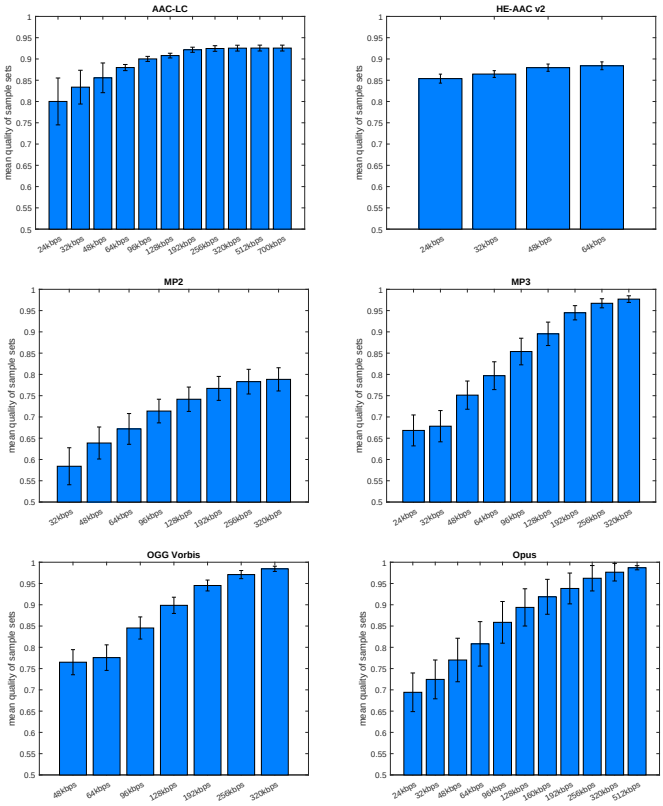
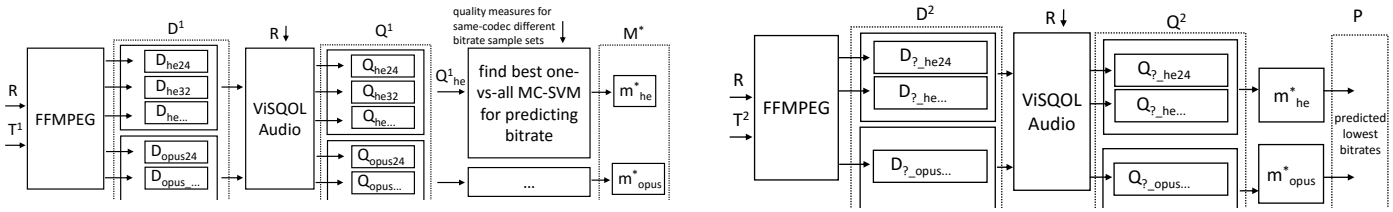


Fig. 2: The mean sample set quality per treatment, grouped by codec.



(a) Training the MC-SVMs.

(b) Testing the MC-SVMs.

Fig. 3: Training and testing MC-SVMs to classify lowest encoded bitrate of two-pass samples.

predicted quality than those above 128 kb/s. Samples encoded using AAC-LC steadily increase in quality up to 128 kb/s but then make only marginal improvements in quality for higher bitrates. HE-AAC-v2 often had the highest quality across all bitrates from 24 kb/s to 64 kb/s, though the difference between quality scores for 24 kb/s and 64 kb/s were small. MP2 quality is lower across bitrates for the illustrated samples. MP3, Opus and Vorbis exhibit similar quality trends, with a steady improvement in predicted quality as bitrate increased, though Opus generally has the highest quality measures above 128 kb/s.

More generally, for each codec in Figure 2 the mean quality scores per bitrate is shown with 95% confidence interval error bars. The scores were calculated using 12 samples for a given treatment. As in Figure 1, there is a trend of quality increasing as treatment bitrate increases. The confidence interval also decreases as bitrates increase for all codecs except MP2 and HE-AAC-v2. It has been shown that expert human listeners cannot distinguish degraded audio quality from uncompressed when encoded at high bitrates [2]. This is reflected in the ViSQOLAudio quality predictions where the difference in mean quality for a treatment (Q_t) becomes very small at high bitrates for some codecs, e.g. MP3 at bitrates greater than 256 kb/s. As ViSQOLAudio cannot distinguish these high bitrate quality measures, it would not be possible for our proposed model to distinguish them based on ViSQOLAudio's objective quality predictions.

For this reason, the following treatments are excluded from our classification experiments A and B: MP2, MP3 where treatment bitrate $t_b = 320$ kb/s, AAC-LC where $t_b \geq 256$ kb/s, Opus where $t_b = 700$ kb/s, and all HE-AAC-v2. The subset treatments T^1 used in these experiments are marked with a ✓in Table I.

III. Methodology

In this section, we will describe how the degraded samples are created, rated by ViSQOLAudio, and then used to train and test MC-SVMs that predict the lowest encoded bitrate of the degraded test samples. MC-SVM were used during the experiments because tests revealed MC-SVMs to be more accurate than regressions and other machine learning approaches.

Figure 3 presents two block diagrams that illustrate the training and testing setup for our model. The notation used in the figure and throughout the paper are introduced below. Superscript is used to denote if something belongs to once or twice-encoded information. For example, D^1 is the set of all degraded samples that have been encoded once, and D^2 for sample sets encoded twice.

Degraded samples are created by passing a set of uncompressed reference samples R (described in Table II) into FFMPEG 2.3.6 with a number of treatments. For example, the reference samples with the treatment he48 will produce a sample set of each of the reference samples encoded with HE-AAC-v2 at 48 kb/s. The one-pass treatments are shown as tick marks in Table I and are used to create the one-pass degraded sample sets D^1, which will be used to train the MC-SVMs.

The two-pass treatments T^2 are illustrative combinations of treatments from T^1, where the first pass codec and second pass codec are different. The treatments in T^2 are: AAC-LC to 24 kb/s MP3; AAC-LC to 128 kb/s Opus; MP3 to AAC-LC 256 kb/s; MP3 to Opus 64 kb/s; Opus to AAC-LC 64 kb/s; and Opus to MP3 48 kb/s, where all first-pass treatments were done at 24, 48, 64, 128 and 256 kb/s. These treatments are passed to FFMPEG with the reference sample set to create the set of two-pass samples D^2, which will be used to test the MC-SVMs.

Every sample in a set has its quality rated using ViSQOLAudio. A quality measure for any encoded sample is generated by presenting ViSQOLAudio with the sample for comparison to the relevant original reference from the set R. The output is a quality measure from 0 to 1, e.g. $\mathsf{ViSQOLAudio}(boz_{\mathsf{ref}}, boz_{\mathsf{he24}}) \rightarrow 0.8$. For each degraded sample set, a vector of 12 quality measures is generated, one for each of the 12 samples in the sample set. This will later be used as a feature vector for the MC-SVMs. The set of quality measure vectors is Q^1 for all sample sets in D^1, and Q^2 for all sample sets in D^2.

One MC-SVM is created per codec to classify the bitrate of samples. Each MC-SVM model consists of a collection of SVMs, one for each bitrate we classify with. For example, in the HE-AAC-v2 MC-SVM, there are four SVMs to classify bitrates: 24, 32, 48, and 64 kb/s. The label for an observation is 1 if the quality vector was generated from a sample set that the SVM is trained to recognise, and 0 otherwise. For example, if the SVM is trained to recognise a sample set encoded with he24, only a quality measure vector generated from the sample set encoded as he24 will be labelled with 1. The SVMs use MATLAB (R2015b) default parameters with a radial kernel.

To find the best MC-SVM for classifying sample sets encoded with a particular codec, an MC-SVM is trained for every combination of quality measures in the quality measure vectors of Q^1. For example, one MC-SVM is trained using only Opus *boz* quality measures and another using only Opus *castanets* and *steely* quality measures.

To find which MC-SVMs work best for classifying the

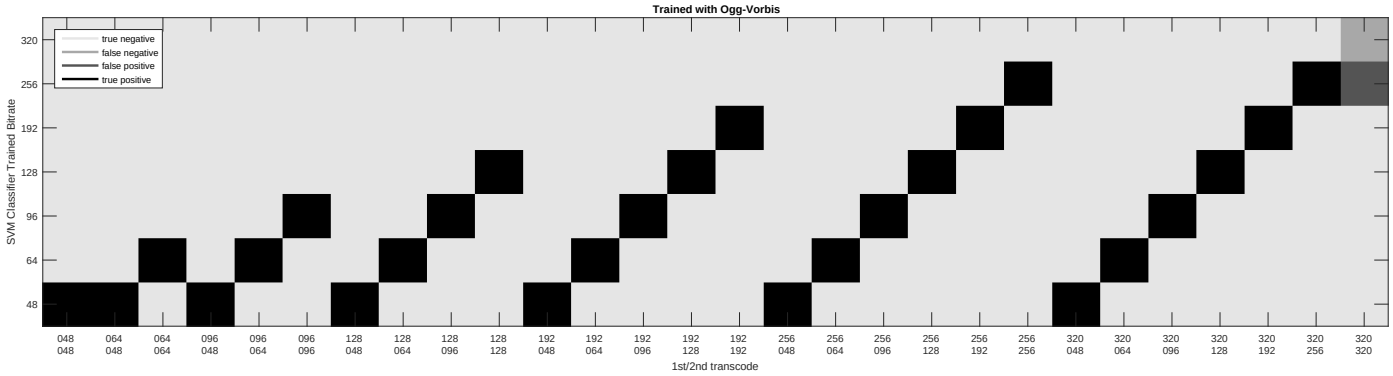


Fig. 4: Predicted lowest bitrate for the most accurate Ogg-Vorbis MC-SVM $m*_{\mathsf{vorbis}}$. Misclassifications only occur at high bitrates.

lowest encoded bitrate of two-pass samples, a set of two-pass sample sets is created from the reference set where the first pass treatment is a treatment in T^1 and the second pass treatment has the same codec as the first treatment but at a different bitrate. These two-pass same-codec-different-bitrate sample sets are input to ViSQOLAudio which outputs their quality measures, which are used to test the MC-SVMs. The MC-SVMs that are most accurate at classifying for a particular codec are stored as $M^* = \{m*_{\mathsf{aaclc}}, ..., m*_{\mathsf{opus}}\}$.

The visualisation of the confusion matrix of the predicted versus the actual lowest encoded bitrates for the most accurate Ogg-Vorbis MC-SVM $m*_{\mathsf{vorbis}}$ is shown in Figure 4. The bitrates that the model is trained to classify are shown on the Y-axis and the two bitrate encodings for the tested sample sets is shown on the X-axis. The figure shows that even the best classifiers misclassified high bitrate encodings due to similarities in quality to other high bitrate encodings, as illustrated in the Ogg-Vorbis misclassification of the sample set twice passed at 320 kb/s.

IV. Experiment

In this experiment, M^* is used to predict the lowest encoded bitrate for the two-pass samples D^2 using their quality measures Q^2. The treatment of the second pass is known, as it is available from the audio files metadata, but the treatment of the first pass (both codec and bitrate) is unknown.

The treatments used in the experiments are: AAC-LC to 24 kb/s MP3; AAC-LC to 128 kb/s Opus; MP3 to AAC-LC 256 kb/s; MP3 to Opus 64 kb/s; Opus to AAC-LC 64 kb/s; and Opus to MP3 48 kb/s, where all first-pass treatments were done at 24, 48, 64, 128 and 256 kb/s. As an exhaustive search was infeasible, these treatments were selected because they cover a wide range of bit rates and codecs.

TABLE III: Accuracy for most accurate MC-SVMs for each sample set when tested on Q^2. First-pass treatment was performed for all bitrates tested for that codec.

treatment	MC-SVM classified with	MC-SVM accuracy (%)
AAC-LC to MP3 24k	MP3	100
AAC-LC to Opus 128k	Opus	40
MP3 to AAC-LC 256k	AAC-LC	80
MP3 to Opus 64k	Opus	40
Opus to AAC-LC 64k	AAC-LC	60
Opus to MP3 48k	MP3	100

The visualisation of the confusion matrix of the predicted versus the actual lowest encoded bitrates in the experiment is shown in Figure 5b. Although the accuracies from Table III are mixed, the behaviour of the predictions is consistent. Of the predictions, only 2 out of the 25 tested were wrong by more than one bitrate range, shown in Figure 5b. For all but two of the misclassifications (Figures 5d and 5e), the predicted lowest bitrate was lower than the actual lowest bitrate. Predictions where the lowest encoded bitrate was 48 kb/s or lower were more accurate than higher bitrate predictions because difference in quality between 24 and 48, and 48 and 96 kb/s is larger than the difference between higher bitrates.

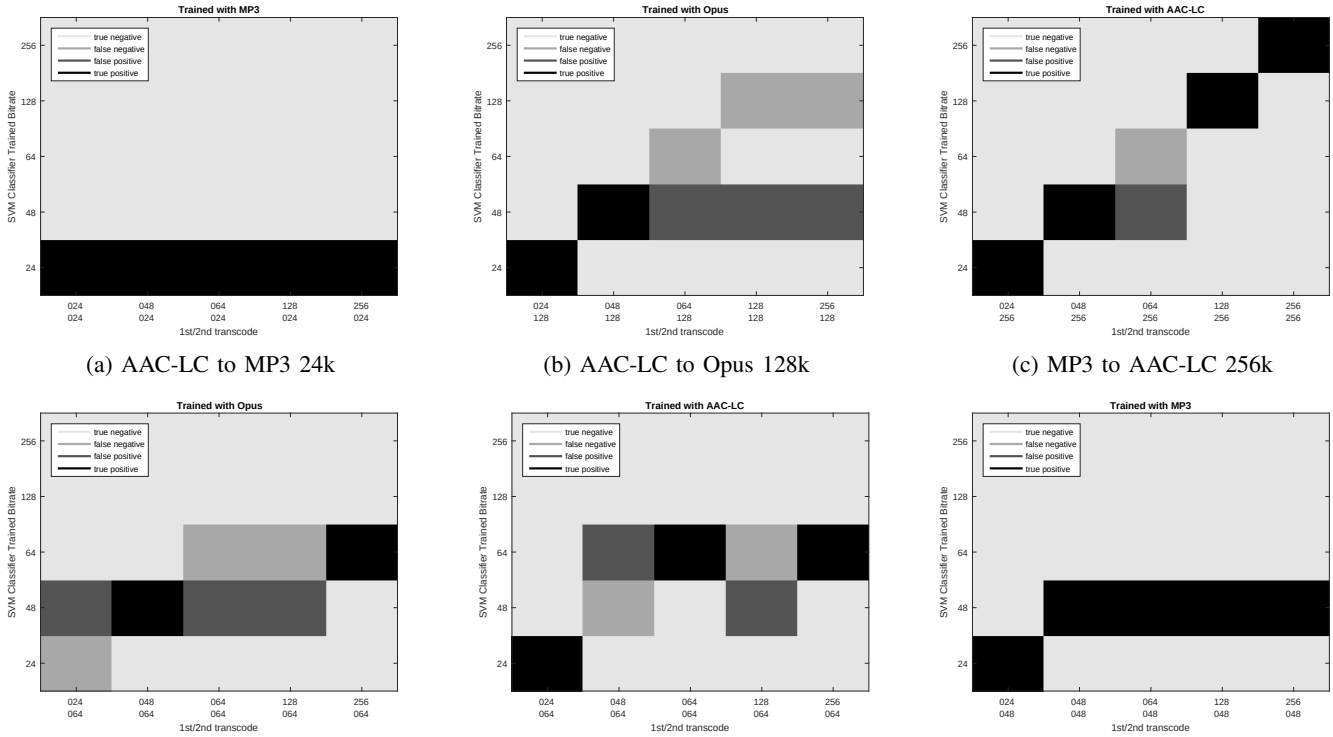
V. Discussion

Though it is well known that quality changes with encoded bitrate and codec, this paper has highlighted that quality varies considerably with the samples used (as illustrated in Figure 1). We also observed a quality threshold for certain codecs after which an increase bitrate had no effect on quality. The threshold limits the use of a previously derived quality metric as a predictor of treatment and suggests that this metric alone should not be used to differentiate bitrates for treatments.

It also suggests that to optimise storage at the expense of additional pre-processing on upload, audio content could be analysed to optimise quality in codec selection, e.g. a speech codec for speech audio.

The experiment showed that our model is able to identify the lowest encoded bitrate of sample sets, usually to within one range of bitrate, for two-pass samples sets across a range of bitrates where the first pass used a codec different to that of the second pass. Almost all of the misclassifications observed in the experiment underestimated the actual bitrate. This suggests that enough quality is lost during cross-codec encoding that an expert listener would believe that the encoded bitrate is lower than it actually is. To compensate for this, rather than using the quality measure of single-pass samples to train the MC-SVMs, we could use the same quality measures but with their values reduced to reflect the loss of quality that would occur during a second encoding.

The experiment considered a range of 5 different bitrates with levels of quality that are not always very different from that of the bitrates adjacent them in the range. The results of the experiment suggest that our model would perform better for a service with a smaller range of bitrates .



(a) AAC-LC to MP3 24k (b) AAC-LC to Opus 128k (c) MP3 to AAC-LC 256k

(d) MP3 to Opus 64k (e) Opus to AAC-LC 64k (f) Opus to MP3

Fig. 5: Lowest bitrate predictions for cross-codec encoded samples sets.

VI. Conclusions and Future Work

Streaming services aim to encode audio with a balance between QoE for the end user, the size of the encoded audio files, and the processing cost of the encoding. Services do not encode to bitrates higher than that of the uploaded files as there will be no increase in quality. Determining the lowest bitrate of the user-uploaded files allows the streaming service to skip encoding the files to bitrates higher than that of the uploaded files, saving on processing and storage space. The proposed system uses objective audio quality scores to predict the lowest encoded bitrate of twice-encoded audio, allowing streaming service providers to validate that the QoE for a given treatment meets the expected QoE. For experiments considering five bitrates classifications from 24 – 256 kb/s, the system can predict the lowest encoded bitrate to within one class of the actual bitrate for 23 of the 25 tested treatments. The analysis of the dataset in Section II revealed that the system is limited to bitrates no greater than 256 kb/s for some codecs, as neither humans nor objective metrics can differentiate samples at these bitrates. Future work will focus on creating a no-reference model that can classify using a single sample.

Acknowledgement

This publication has emanated from research conducted in the CONNECT research centre in collaboration with Google, Inc. with the financial support of Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2077.

References

[1] P. Le Callet, S. Möller, and A. Perkis, eds., "Qualinet white paper on definitions of quality of experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)," 2013.

[2] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "Perceived audio quality for streaming stereo music," in *Proceedings of the ACM International Conference on Multimedia*, 2014.

[3] D. Luo, R. Yang, and J. Huang, "Detecting double compressed AMR audio using deep learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

[4] R. Yang, Z. Qu, and J. Huang, "Exposing MP3 audio forgeries using frame offsets," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2012.

[5] J. Lacroix, Y. Prime, A. Remy, and O. Derrien, "Lossless audio checker: A software for the detection of upscaling, upsampling, and transcoding in lossless musical tracks," in *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015.

[6] T. Bianchi, A. De Rosa, M. Fontani, G. Rocciolo, and A. Piva, "Detection and localization of double compression in MP3 audio tracks," *EURASIP Journal on Information Security*, 2014.

[7] S. Hicsonmez, E. Uzun, and H. T. Sencar, "Methods for identifying traces of compression in audio," in *Communications, Signal Processing, and their Applications (ICCSPA), 2013*. IEEE.

[8] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *The Journal of the Acoustical Society of America*, 2015.

[9] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality - technology and applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1890–1901, 2006.

[10] G. Waters, "Sound quality assessment material recordings for subjective tests: Users handbook for the EBU–SQUAM compact disk," *European Broadcasting Union (EBU), Tech. Rep 3253-E*, 1988.

[11] P. Pocta and J. G. Beerends, "Subjective and objective assessment of perceived audio quality of current digital audio broadcasting systems and web-casting applications," *Broadcasting, IEEE Transactions on*, 2015.

[12] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Communication*, 2012.