INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 2, PP. 331-339
Manuscript received February 14, 2024; revised June 2024. doi: 10.24425/ijet.2024.149549

Non-intrusive method for audio quality assessment of lossy-compressed music recordings using convolutional neural networks

Aleksandra Kasperuk, and Sławomir Krzysztof Zieliński

***Abstract*—Most of the existing algorithms for the objective audio quality assessment are intrusive, as they require access both to an unimpaired reference recording and an evaluated signal. This feature excludes them from many practical applications. In this paper, we introduce a non-intrusive audio quality assessment method. The proposed method is intended to account for audio artefacts arising from the lossy compression of music signals. During its development, 250 high-quality uncompressed music recordings were collated. They were subsequently processed using the selection of five popular audio codecs, resulting in the repository of 13,000 audio excerpts representing various levels of audio quality. The proposed non-intrusive method was trained with the data obtained employing a well-established intrusive model (ViSQOL v3). Next, the performance of the trained model was evaluated utilizing the quality scores obtained in the subjective listening tests undertaken remotely over the Internet. The listening tests were carried out in compliance with the MUSHRA recommendation (ITU-R BS.1534-3). In this study, the following three convolutional neural networks were compared: (1) a model employing 1D convolutional filters, (2) an Inception-based model, and (3) a VGG-based model. The last-mentioned model outperformed the model employing 1D convolutional filters in terms of predicting the scores from the listening tests, reaching a correlation value of 0.893. The performance of the Inception-based model was similar to that of the VGG-based model. Moreover, the VGG-based model outperformed the method employing a stacked gated-recurrent-unit-based deep learning framework, recently introduced by Mumtaz *et al.* (2022).**

***Keywords*—objective audio quality assessment; non-intrusive audio quality evaluation, convolutional neural networks**

I. Introduction

SUBJECTIVE scores obtained in audio quality listening tests are regarded as reference data. Therefore, they are commonly utilized during the development and optimization of products and services. However, obtaining such data is time-consuming, expensive, and complicated, often requiring researchers to follow rigorous experimental protocols [1],[2]. Hence, some scientists and engineers prefer using objective methods of audio quality assessment as they are less expensive, faster, and relatively easy to apply. However, most of the developed objective quality assessment methods so far are intrusive since the two signals are required at their inputs: (1) an unimpaired reference audio signal, and (2) a signal under test. The above requirement precludes such methods from many real-life applications. The intrusive quality assessment methods are also referred to as 'double-ended' techniques [3].

In this paper, we introduce a non-intrusive (single-ended) audio quality assessment method employing convolutional neural networks. No reference signal is required by the proposed technique. During the development of the method, three convolutional neural networks were adapted for our purposes and their performance compared, namely: (1) a model employing one-dimensional (1D) convolutional filters [4], (2) an Inception-based model [5], and (3) a VGG-based model [6]. The proposed method is intended to account for audio artefacts arising from the lossy compression of music signals. This type of artefacts can be regarded as one of the most common types of distortions encountered in modern audio delivery systems such as Internet-based music streaming services. In this study, a repository of 13,000 music audio excerpts was employed. The proposed method was trained with the data obtained using a well-established intrusive model (VISQOL v3) [7]. Finally, its performance was evaluated by means of the quality scores acquired in the subjective listening tests undertaken remotely over the Internet. The proposed method, after further optimization, could be applied for real-time audio quality monitoring of music recordings streamed over the Internet.

We make the following contributions: (1) We introduce a non-intrusive method for the audio quality assessment of lossy-compressed music recordings, employing convolutional neural networks. (2) We demonstrate that our method satisfactorily matches the scores obtained from the subjective listening tests, outperforming the state-of-the-art technique recently introduced in the literature.

The next section overviews the related work in the area of the objective audio quality assessment. The methodology applied in this work is described in Sec. III. The obtained results are discussed in Sec. IV. The conclusions are provided in the last section of the paper.

II. Related Work

Objective modelling of audio quality perception can be traced back to the early work of Karjalainen [8] who in 1985

This work was supported by the grant from Białystok University of Technology (WZ/WI-IIT/5/2023) and funded with resources for research by the Ministry of Science and Higher Education in Poland.

A. Kasperuk and S. K. Zieliński are with Faculty of Computer Science, Białystok University of Technology, Poland (e-mail: aleksandra.kasperuk.105570@student.pb.edu.pl, s.zielinski@pb.edu.pl).

demonstrated that it is possible to computationally predict audio quality scores obtained by human listeners. In the years 1998−2000, joint efforts of several research groups culminated in the development of a method for Perceptual Evaluation of Audio Quality (PEAQ) [9], which was subsequently standardized by ITU [10]. Since then, several competitive methods have been developed, most notably PEMO-Q [11], HAAQI [12], InSE-NET [13], and ViSQOL [3],[7]. They exhibited superior performance compared to the PEAQ technique under some conditions. Despite its outdated neural network, the PEAQ method is still used by some engineers [14]. It constitutes the only international standard for the objective assessment of audio quality [10].

The performance of the objective audio quality methods is widely regarded as satisfactory [9]-[14]. However, since these methods are predominantly intrusive, requiring access to a reference signal, the scope of their real-world applications is significantly reduced. Therefore, several non-intrusive techniques have been developed. The state-of-the-art non-intrusive methods comprise such techniques as MOSA-Net [15], DNSMOS [16], WAWEnets [17], Quality-Net [18], NISQA [19], and NIC-STOI [20]. Nevertheless, all the above-mentioned non-intrusive methods have been developed for the quality assessment of 'speech' signals. Hence, they are not suitable for the evaluation of audio recordings, as demonstrated in [21].

To the best of the authors' knowledge, the literature provides only one study with the non-intrusive method intended for the quality assessment of audio recordings, including music. Namely, it is the technique proposed in 2022 by Mumtaz *et al.* [21]. It was developed for the audio quality assessment of multimedia content generated by the users of popular video-sharing services. The scope of their method was limited to the quality assessment of the two types of distortions, namely, low-bitrate coding distortions and background noise. It was designed by combining a traditional feature extraction procedure with a deep learning approach based on a stacked gated-recurrent-unit framework. The correlation coefficient between the quality scores obtained using their method and those acquired in the listening tests was equal to 0.834.

While Organiściak and Borkowski [22] also published a paper on '*single-ended quality measurement of a music content*', a closer examination of their report revealed that, contrary to the paper's title, its authors developed a method for the 'classification' of audio distortion types rather than for the quality assessment.

III. METHOD

This section provides the experimental details regarding the development and evaluation of the proposed method. For clarity, Fig. 1a shows a block diagram demonstrating how the intrusive method employing the ViSQOL technique [7] can be employed to assess the quality of low-bitrate compressed audio signals. This approach was used in this work to generate the training data, which will be described in more detail below. Note, that the ViSQOL algorithm takes two signals at its input: an original reference recording and a signal to be assessed. The latter one is subject to a low-bitrate lossy compression algorithm. The ViSQOL technique yields an estimated audio quality value at its output.

A non-intrusive method introduced in this study is illustrated in Fig. 1b. In the proposed approach, low-bitrate audio signals are initially converted to spectrograms and then directed to the convolutional neural network. The network estimates the audio quality scores. In contrast to the intrusive approach, the proposed algorithm undertakes the quality assessment task "blindly," solely based on low-bitrate compressed signals.

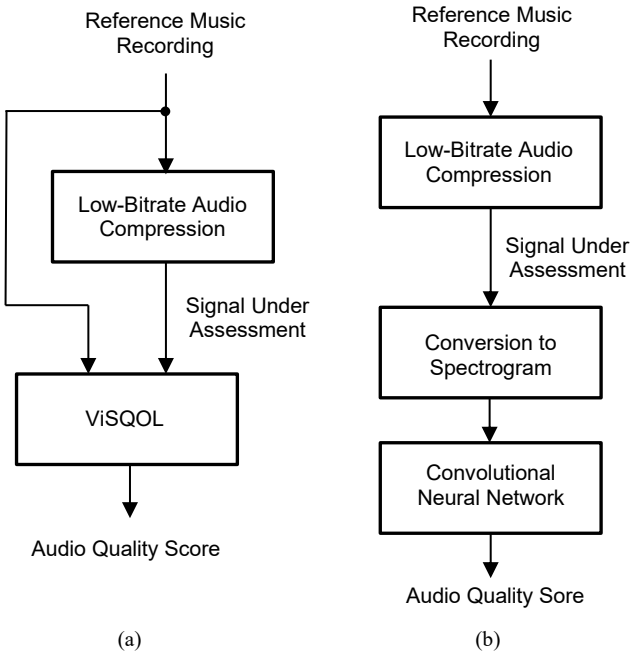


Fig. 1. Audio quality assessment techniques: (a) intrusive method employing ViSQOL algorithm [7], (b) non-intrusive method proposed in this study

A. Selecting Reference Music Recordings

For the purpose of this study, 250 uncompressed reference music recordings were selected. They were acquired from private CD collections as well as from publicly available Internet-based repositories. The selected reference recordings represented a broad range of music genres, including rock, heavy metal, progressive rock, jazz, hip-hop, pop, country, reggae, disco, blues, classical music, and opera. The recordings were trimmed to a duration of 10 seconds each. They were stored in an uncompressed monophonic WAV format at a 48 kHz sample rate with a 16-bit resolution.

B. Introducing Audio Quality Degradations

The audio quality of the selected 250 reference recordings was degraded using two processes: lossy low-bitrate audio coding and low-pass filtering. To this end, the five following audio codecs were utilized: mp2 (MPEG-1 Audio Layer 2), mp3 (MPEG-2 Audio Layer 3), ADTS, ogg (Ogg Vorbis), and Opus. The bitrate values considered for each codec along with associated process identification numbers (IDs) are presented in Table I. In addition to lossy low-bitrate audio coding, the audio quality of the reference recordings was also degraded by means of a low-pass filter, with the cut-off frequencies and associated process IDs outlined in Table II. To this end, a 4^{th}-order infinite impulse response (IIR) filter was used. Note, that in total there were 45 processes, out of which 41 accounted for the audio coding conditions, whereas the remaining 4 processes represented the low-pass filtering conditions.

TABLE I
DEGRADATION PROCESSES USING LOSSY LOW-BITRATE CODECS

Codec	Process ID	Bitrate (kb/s)
mp2	1–8	32, 48, 64, 96, 128, 192, 256, 320
mp3	9–6	32, 48, 64, 96, 128, 192, 256, 320
adts	17–27	24, 32, 48, 64, 96, 128, 192, 256, 320, 512, 700
ogg	28–32	48, 64, 96, 128, 192
opus	33–41	24, 32, 48, 64, 96, 128, 160, 192, 256

TABLE II
DEGRADATION PROCESSES EMPLOYING A LOW-PASS FILTER

Process ID	Cut-off Frequency (Hz)
42	3500
43	5000
44	7500
45	9000

The audio quality of the processed recordings was objectively assessed using the state-of-the-art intrusive algorithm, namely ViSQOL v3 [7]. Initial examination of the objective quality scores obtained using the above 45 processes revealed that their distribution was highly skewed towards high-quality scores. In order to obtain a more uniform distribution of the quality scores, the following algorithm was applied:

1. For all the 250 reference recordings, processes 1−45 were applied, yielding 11,250 audio excerpts (250×45).
2. The excerpts obtained using processes 6−8, 14−16, 23−27, 32, and 39−41 were excluded from the repository, as they were deemed redundant in terms of the resultant quality levels.
3. The above repository was supplemented by the excerpts obtained using 'cascaded' processing, whereby reference recordings were sequentially processed by two randomly selected processes from the list of all 45 processes described in Tables I and II.
4. The previous step was repeated until 64,960 excerpts were generated in total.
5. The obtained excerpts were trimmed to 10 seconds in duration and loudness equalized to -23 LUFS [23].
6. The quality of the excerpts was objectively assessed using the ViSQOL v3 method.
7. The final repository of 13,000 excerpts was obtained by randomly drawing the excerpts in such a way that the quality distribution of the resultant repository was relatively uniform.

The audio quality scores generated by the ViSQOL v3 algorithm used in step 6 originally ranged from 1 (the worst) to 5 (the best). In our study, they were scaled to 0−1 range.

C. Calculating Spectrograms

The final repository of 13,000 audio excerpts was converted to Mel-spectrograms. They were subsequently used as monochromatic 'images' at the inputs of the convolutional neural networks. During the calculation of the spectrograms, 256 Mel-frequency bands were used. A moving window of FFT analysis was employed. Its length was equal to 2048 samples with an overlap of 256 samples. Recall, that the sample rate of the audio excerpts was equal to 48 kHz. An analysed signal was weighted using a Hann window. Finally, the spectrogram values were normalized to the range of 0−1. The spectrograms were stored as two-dimensional tensors of the size of 1876×256, where the first dimension represented time and the second one signified frequency.

TABLE III
TOPOLOGY OF THE NEURAL NETWORK EMPLOYING 1D CONVOLUTIONAL FILTERS

Layer Acronym	Description	Output Tensor Shape
Input	Input layer with shape: 1876×256	1876×256
Conv1D	1D convolutional layer (number of filters: 64, length: 11)	1866×64
MaxPooling1D	Max pooling layer (stride: 2)	933×64
Dropout	Dropout layer (rate: 0.01)	933×64
Conv1D	1D convolutional layer (number of filters: 128, length: 7)	927×128
MaxPooling1D	Max pooling layer (stride: 2)	463×128
Dropout	Dropout layer (rate: 0.01)	463×128
Conv1D	1D convolutional layer (number of filters: 256, length: 3)	461×256
MaxPooling1D	Max pooling layer (stride: 2)	230×256
Flatten	Flattening layer	58880
Dense	Dense layer	256
Dropout	Dropout layer (rate: 0.01)	256
Dense	Dense layer	128
Dense	Dense layer	1

TABLE IV
INCEPTION-BASED NETWORK ARCHITECTURE

Layer Acronym	Description	Output Tensor Shape
Input	Input layer with shape: 1876×256	1876×256
Reshape	Reshape layer	1876×256×1
Conv2D	2D convolutional layer (number of filters: 3, size: 11×11)	1866×246×3
Resizing	Resizing layer	224×224×3
Inception	Inception v3 model	1000
Dense	Dense layer	256
Dense	Dense layer	128
Dense	Dense layer	1

D. Convolutional Neural Networks

Three types of convolutional neural networks (CNN) were adapted for our purposes. The first one involved 1D convolutional filters [4]. It was employed in our work due to the promising results of the pilot study. Its topology is presented in Table III. It can be seen in the table that the network takes spectrograms of size 1876×256 at its input. It consists of three 1D convolutional layers, with an increasing number of filters in each consecutive layer (64, 128, 256) and a decreasing length of the filters in each successive layer (11, 7, 3). Max pooling and dropout layers were used after each convolutional layer, with the max pooling stride and dropout rate being equal to 2 and 0.01, respectively. After converting two-dimensional tensors to one-dimensional ones in the flattening layer, the data were processed using three fully connected layers (dense layers), intertwined with one dropout layer. Rectified Linear Unit (ReLU) activation function was used in all the convolutional layers, whereas a linear activation function was utilized in the dense layers (the dense layers performed a linear regression).

The second network used in this study was based on the Inception v3 model [5], whereas the third one employed the VGG19 model [6]. The adapted topology of the Inception-based model is presented in Table IV. The architecture of

the VGG19 model was almost identical. The only difference was the fifth layer, where instead of the Inception model, the VGG19 algorithm was exploited. Similar to the first model, both the Inception-based and VGG-based models accepted spectrograms of size 1876×256 at their inputs. Their topology was mutually identical in terms of the four input layers (Input, Reshape, Conv2D, and Resizing) as well as the three dense layers. The core difference regarded the fifth layer. It contained either an Inception or a VGG19 model. The presented topologies were designed heuristically during the pilot tests. Similar to the first network, they both employed the ReLU activation function in the convolutional layers, whereas a linear activation function was utilized in the dense layers.

The VGG19 and Inception models were used with the weights pre-trained on the ImageNet database [24]. These weights were kept intact during the experiments. The weights of the remaining layers in the proposed topologies of the network were subject to optimization.

The initial inspection of the audio quality scores produced at the output of the above three networks revealed that their values occasionally exceeded unity, representing a notional top quality. Therefore, all the output scores extending beyond the top-quality score were truncated to 1.0. The networks were implemented in Python using the TensorFlow library.

E. Training and Validation

As mentioned above, in total, 13,000 music excerpts were converted to spectrograms, out of which approximately 75% were used for the training purposes, whereas the remaining 25% served for the validation procedures. The training and validation sets were unique in terms of music recordings. Out of 250 reference music recordings exploited in this study, 190 were employed in the training set, whereas 60 reference recordings were utilized in the validation set.

All three networks shared the same optimization approach. However, they differed in terms of the batch size (see below). They were all optimized using the Adam optimization algorithm [25]. Mean square error was employed as a loss function, whereas root mean squared error was utilized as a regression metric. For all three networks, the initial learning rate value was set to 0.001. Subsequently, its value was reduced by a factor of 2 when a validation loss was not improving for ten consecutive epochs. For the 1D convolution-based model, the batch size was set to 64. For the Inception-based model and for the VGG19-based model, the batch size was adjusted to 64 and 32, respectively. The above-mentioned optimization values were identified heuristically during the pilot tests.

The maximum number of training epochs was limited to 300. However, the networks were trained until an overfitting effect was observed based on the validation loss curve. To this end, an early stopping procedure was employed. The training process was terminated when no improvement in the validation loss was observed for 50 consecutive epochs. After the termination of the training procedure, the model parameters were reversed to those exhibiting the minimum validation loss.

The music excerpts exploited in this study are not publicly available due to copyright restrictions. However, the trained network models along with the source code developed in this work have been made publicly available at GitHub [26].

F. Testing with Objective Data

To test the trained models using objective data, eleven new high-quality reference music recordings were acquired. They represented various music genres. The selected music recordings were not exploited earlier during the training and validation of the models.

The selected new recordings were degraded in audio quality in the same way as described above in Sec. III B. As a result, 750 test excerpts were generated. These new excerpts were evaluated in audio quality using the three networks implemented in our study. The above procedure represented a non-intrusive approach. Moreover, the new test excerpts were also assessed by means of the intrusive ViSQOL v3 [7] technique. Subsequently, the quality scores obtained with non-intrusive and intrusive approaches were compared.

G. Testing with Subjective Data

The developed models were also tested using the subjective data, obtained from the two listening tests undertaken in conformance with the MUSHRA recommendation (ITU-R BS.1534-3 [2]). Both experiments were carried out remotely, over the Internet, using the WebMUSHRA software [27]. The reason for undertaking two listening tests instead of one was the insufficiency of data obtained from the first listening test, preventing these authors from reaching reliable conclusions. Since the methodology applied in both listening tests was almost identical, the first test will be described in detail, whereas only the methodological differences will be discussed with regard to the second test.

1) Audio Stimuli

Eight new reference music recordings were selected for the first listening test. They were acquired from the publicly available repository [28]. The first recording served for the listeners' training purposes, whereas the remaining seven recordings were used in the listening tests.

Each recording was degraded in the audio quality using three selected processes employing the low-bitrate audio coding. Moreover, every recording was also low-pass filtered to produce the so-called anchor recordings in conformance with the MUSHRA recommendation [2]. To this end, an IIR (Chebyshev Type I) filter of the 10^{th} order was employed. According to the standard, the cut-off frequency of the filter was set to 3.5kHz and 7kHz, respectively, to produce two distinct quality levels of the low-pass filtered anchors. As a result, 40 excerpts (8×5) were generated. They represented different levels of quality. Out of 40 degraded excerpts, five were used during the initial training session, whereas the remaining 35 recordings were employed in the listening test. The duration of the excerpts under assessment ranged from 5 to 9 sec. (The exact trimming points of the excerpts were adjusted in a musically aesthetic way, as assessed by the first author). The excerpts were looped during the listening test. It must be emphasized that these excerpts were not employed formerly in the training or validation of the models.

2) Listening Test Procedure

As mentioned above, the listening tests were undertaken remotely over the Internet. During the tests, the participants were asked to assess the basic audio quality of the audio excerpts in comparison to the reference recordings. Prior to undertaking the above task, the listeners were initially provided with a set of instructions, explaining the assessment

methodology. They were requested to perform the test using headphones. Moreover, they were asked to adjust the playback volume to a comfortable level. Furthermore, they were instructed that at least one of the evaluated items must be assigned the maximum score, as reference recording was included in the set of evaluated excerpts. Before the test, the listeners completed a training session, similar to the assessment task performed during the actual test. This procedure allowed them to get acquainted with the interface and familiarize themselves with the 100-point audio quality scale recommended by the MUSHRA standard [2].

The stimuli were presented to each listener in random order. To assess the listeners' repeatability, five excerpts were presented twice. The listening test was typically completed by each participant within approximately 30 minutes. After finishing the test, the participants were requested to confirm whether they had used headphones.

3) Participants and Data Screening

In total, 24 participants took part in the first listening test. They were recruited from the population of students at Białystok University of Technology. The data from the three participants were rejected as they had not used the headphones (based on the self-reports acquired after the test). Recall, that employing headphones constituted one of the requirements provided to the listeners in the instructions. Moreover, the data from four listeners were removed due to the abnormalities observed in their assessment scores. These listeners assigned the top score to all the excerpts, evaluated the 3.5kHz anchor using the top score, or assessed hidden reference recordings with mid-scale quality values. Furthermore, the data from two other listeners were rejected based on their poor repeatability. For these two listeners, the discrepancy between the scores obtained for the replicated stimuli exceeded 50% of the grading scale. The data from the remaining 15 listeners were retained. In line with the typical practice [1],[2], for each evaluated audio excerpt, the retained data were summarized by calculating mean opinion scores (MOS).

4) Methodological Differences in the Second Listening Test

Regarding the second listening test, the methodological differences were as follows. Seven new reference music recordings were gathered. They were degraded in quality using the same procedure as before. They were acquired from a different repository [29]. In total, 24 participants were recruited for the second listening test, out of which the data from six participants were rejected due to the abnormalities in their assessment scores. These six listeners assessed hidden reference recordings with mid-scale quality values, rated the 3.5kHz anchor as much better than the 7kHz anchor, or evaluated the 3.5kHz anchor as better than the hidden reference recording. Moreover, the data from one participant were removed because of their poor repeatability (assessing the replicated excerpts with the discrepancy exceeding 40% of the grading scale). The data from the remaining 17 participants were kept. Subsequently, the retained data were aggregated by calculating MOS values for each evaluated audio excerpt. Finally, the MOS values from both listening tests were merged and used to test the accuracy of the developed models.

IV. Results

This section provides the results of the evaluation of the developed models using objective and subjective data. For consistency with the subjective data obtained employing a 100-point scale recommended by the MUSHRA standard [2], the scores calculated by the models were scaled to a 0−100 range.

Many metrics could be used to assess the goodness of fit of the proposed method, including mean absolute error, mean square error, root mean square error, Pearson's correlation coefficient, Spearman's rank-order correlation, Kendall's rank-order correlation, or the coefficient of determination. In this study, the two following metrics were employed, namely: root mean square error (RMSE) and Pearson's correlation coefficient. They were selected because they are commonly applied by researchers in the area of objective speech and audio quality modelling [12],[17],[19],[21]. RMSE is regarded as an adequate descriptor of the model's accuracy [16],[19] whereas Pearson's correlation coefficient is considered to be a suitable measure of the strength of a linear association between the target and predicted quality scores [3],[9].

A. Test Results Using Objective Data

Figure 2a shows the audio quality scores achieved using the proposed non-intrusive method employing 1D convolutional filters (vertical axis). They are plotted against the data calculated with the ViSQOL v3 intrusive algorithm (horizontal axis). It can be seen that the results obtained using the proposed method match the scores obtained with the ViSQOL technique relatively well, given that the non-intrusive algorithm worked without access to the reference recordings. In this example, Pearson's correlation coefficient between the compared scores was equal to 0.823, whereas root mean square error (RMSE) amounted to 15.65 points relative to a 100-point scale. Moreover, a saturation effect could be observed, as some data for reference recording No. 8 (grey circles) were limited to the top of the scale. Recall, that during the development of the networks, their output values were deliberately 'clipped' to the maximum permissible value of the grading scale (Sec. III D).

The results of the quality scores predicted using the Inception-based model, plotted against the scores from the ViSQOL v3 algorithm, are presented in Fig. 2b. Overall, the degree of match between the scores is better than in the previous case, with no saturation effect observed. For this example, the correlation coefficient was equal to 0.902, whereas the RMSE value amounted to 10.42 points, indicating a considerable improvement compared to the previous model.

The results obtained using the VGG-based model, illustrated in Fig. 2c, were similar to those achieved with the Inception-based model discussed above. However, occasional 'clipping' of the scores occurred at the top end of the grading scale. For this example, the correlation coefficient between the scores was equal to 0.921, whereas the RMSE value amounted to 9.53 points.

TABLE V
COMPARISON OF THE NON-INTRUSIVE MODELS TESTED USING DATA OBTAINED WITH THE VISQOL ALGORITHM

	1D Convolution	Inception v3	VGG19
Correlation Coefficient	0.701 (0.075)	0.847 0.076)	0.875 (0.038)
RMSE	18.80 (2.29)	12.97 (3.09)	11.75 (1.66)

The test results of the developed models using the objective data are summarized in Table V. The presented data were acquired by repeating the experiments 15 times. The table reports the mean values and standard deviations of the correlation coefficients as well as the RMSE values, calculated across all the experimental repetitions. The VGG-19 model exhibited the best performance, with the correlation coefficient equal to 0.875 (SD 0.038). While the correlation coefficient obtained for the Inception-based model was slightly lower, amounting to 0.847 (SD 0.076), the difference was not statistically significant according to the *t*-test ($p > 0.05$). This outcome implies that the performance of the VGG19-based and Inception-based models was the same in a statistical sense. The performance of the model employing 1D convolution was the worst, yielding the correlation coefficient of 0.701 (SD 0.075). The difference between the best-performing model (VGG19) and the worst-performing technique (1D convolution), in terms of the correlation coefficients, was statistically significant at $p = 9.7\times10^{-9}$ level.

Considering the RMSE values, the observations that can be made regarding the performance of the models are the same as above. Namely, the VGG19 model exhibited the best performance, while the technique employing 1D convolutional filters showed the worst operation. For these models, RMSE values were equal to 11.75 (SD 1.66) and 18.80 (SD 2.29), respectively. This difference was statistically significant at $p = 2.2\times10^{-10}$ level. However, the difference between the RMSE values obtained using the Inception-based and VGG19-based models was statistically not significant. This result indicates that the performance level of these two models was similar. The difference between the RMSE values reached by the Inception-based model and that employing 1D convolutional filters was statistically significant at $p = 2.6\times10^{-6}$ level.

B. Test Results Using Subjective Data

Figure 3 illustrates the example test results of the three developed models using the subjective data. The performance of the model employing 1D convolutional filters is presented at the top pane of that figure (Fig. 3a). It can be seen that it exhibits a mediocre level of performance. Moreover, it tends to systematically underestimate the scores (bias effect) compared to the reference data obtained from the listening tests. In this example, the correlation coefficient between the objective and subjective scores equals 0.902, whereas the RMSE value amounts to as much as 20.79 points relative to a 100-point scale. A slightly better performance is shown by the Inception and VGG-based models, with their data presented in Figs. 3b and 3c, respectively. Observe a better match between the subjective data and the scores predicted by the two models, with most of the data points scattered in the vicinity of a diagonal line. Nevertheless, a small level of an underestimation bias is also present in the data, indicating the need for further improvements of the proposed techniques. In contrast to the test results with the objective data discussed in the previous section, none of the models exhibited a saturation effect when evaluated with the subjective scores.

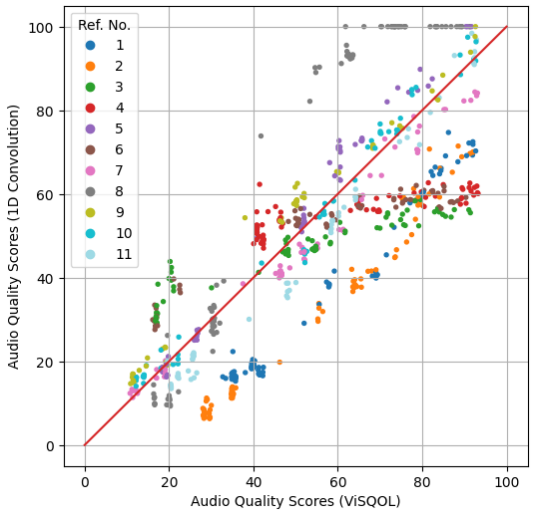
The results obtained in the tests employing the subjective data are summarised in Table VI. It presents the correlation coefficients as well as the RMSE values. Like the previously discussed table above, it provides the mean values and standard deviations calculated with the results obtained in 15 experimental repetitions.

Parallel to the results described in the previous section, the VGG-based model proved to be the best-performing network, reaching the mean correlation coefficient of 0.893 (SD 0.036), whereas the model employing 1D convolutional filters exhibited the worst performance, with the average correlation coefficient of 0.814 (SD 0.056). The difference between these correlation coefficients was statistically significant at $p = 9.3\times10^{-5}$ level. In terms of its performance, the Inception-based model ranked in the middle position, between the above two models, attaining the correlation coefficient of 0.857 (SD 0.098). The differences between the correlation coefficients obtained for the Inception-based model and those achieved by the remaining two models were statistically not significant.

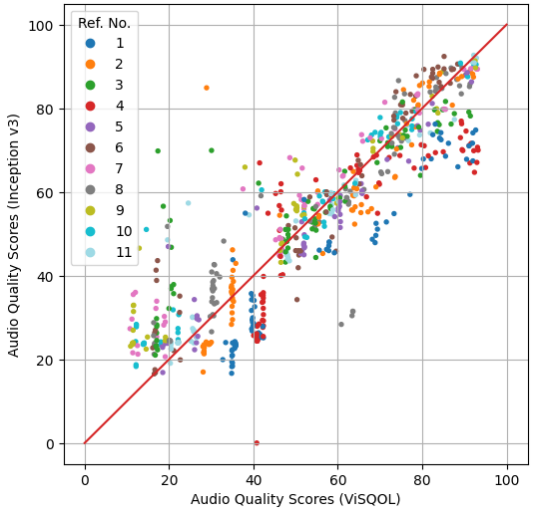
Similar observations can be made by inspecting the RMSE values in Table VI. The presented results confirm the superiority of the VGG-based model compared to the model based on the 1D convolutional filters. The mean RMSE value exhibited by the VGG-based model was equal to 16.96 (SD 1.83). This could be considered as an acceptable level of the prediction error, given its non-intrusive topology. In contrast, the RMSE value of the worst-performing algorithm was greater, amounting to 20.69 (SD 2.00). In terms of its prediction error, the Inception-based model was also ranked in the middle, between the above two models, yielding the RMSE value of 18.60 (SD 4.26). The difference between the RMSE values obtained by the best and the worst-performing models was statistically significant at $p = 1.1\times10^{-5}$ level. However, the differences between the mean RMSE values achieved by the Inception-based model and the remaining two models were statistically not significant.

C. Discussion

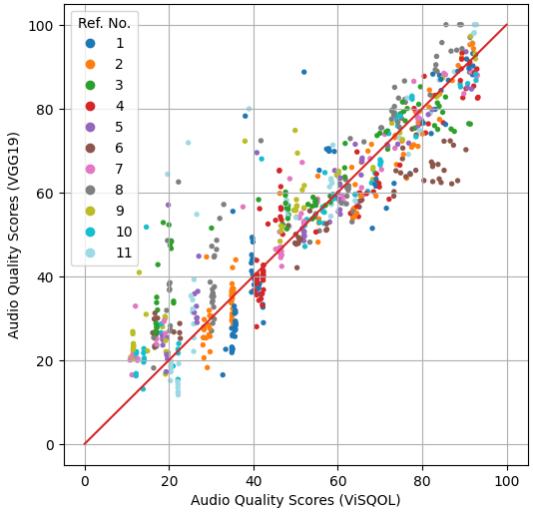
The non-intrusive method proposed in this study was trained with the data obtained using the intrusive algorithm (ViSQOL v3), inheriting not only its 'knowledge' but also its potential biases. The risk of perpetuating bias effects in objective models is discussed in [30]. Ideally, the proposed algorithm should have been trained using the subjective data obtained with the listening tests. However, due to the scarcity of such data available in the public domain and considering the high cost of performing large-scale listening tests required to train deep learning algorithms, the above approach could be justified.



(a)

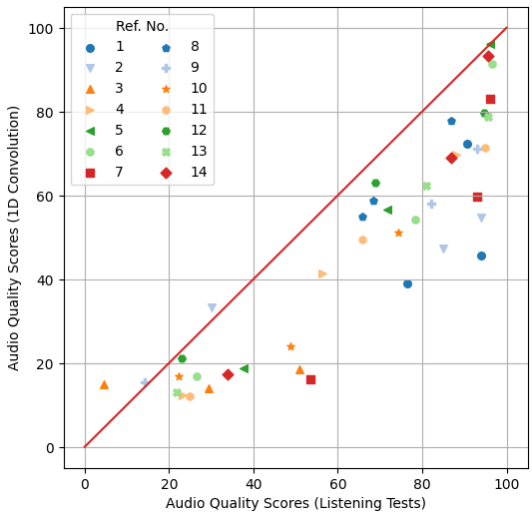


(b)

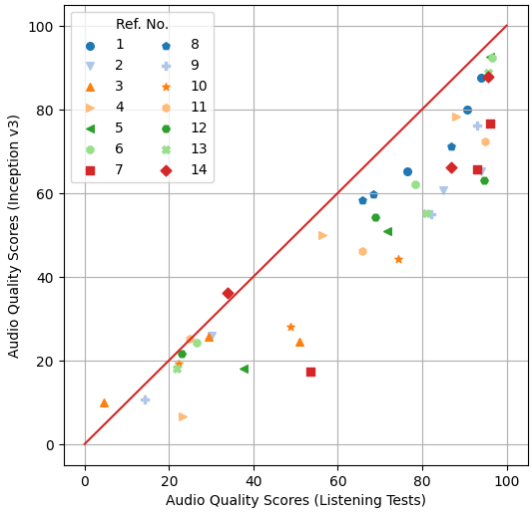


(c)

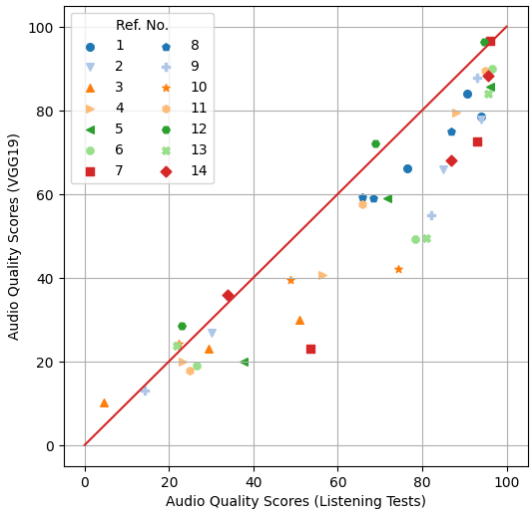
Fig. 2. Comparison of the audio quality scores obtained using the intrusive ViSQOL method with the scores calculated by the non-intrusive algorithms: (a) 1D convolution model (b) Inception v3 model (c) VGG19 model



(a)



(b)



(c)

Fig. 3. Comparison of the audio quality scores obtained using the listening tests with the scores calculated by the non-intrusive algorithms: (a) 1D convolution model (b) Inception v3 model (c) VGG19 model

TABLE VI
COMPARISON OF THE NON-INTRUSIVE MODELS EVALUATED WITH THE DATA FROM THE LISTENING TESTS

	1D Convolution	Inception v3	VGG19
Correlation Coefficient	0.814 (0.056)	0.857 (0.098)	0.893 (0.036)
RMSE	20.69 (2.00)	18.60 (4.26)	16.96 (1.83)

The results of the evaluation of the proposed method using the objective data, presented in Sec. IV A, demonstrate that our models successfully learned the knowledge from the ViSQOL algorithm. However, these outcomes are of little value with regard to assessing the generalization property of the developed models. In turn, the results of the evaluation with the subjective data, provided in Sec. IV B, prove that our technique is generalizable.

Out of the three models compared in this study, the VGG-based model turned out to be the best-performing neural network in terms of predicting the subjective data. It reached the correlation between the subjective and objective scores equal to 0.893 (SD 0.036). Hence, it outperformed the non-intrusive audio quality assessment method proposed by Mumtaz *et al.* [21], who reported the correlation coefficient between the subjective and objective data as being equal to 0.834. The difference between the above-quoted correlation coefficients is statistically significant at $p = 2.09\times10^{-5}$ level (according to the one-sample *t*-test). However, the conclusion regarding the superiority of the method proposed in this paper must be treated with some caution due to the difficulty in the direct comparison of the methods between the studies. The music recordings employed in our work were solely affected by low-bitrate codecs, whereas those utilized in the work of Mumtaz *et al.* were also degraded by background noise [21].

CONCLUSIONS

Most of the objective audio quality methods developed so far are intrusive, limiting the scope of their real-life applications. In this study, we introduce a non-intrusive audio quality assessment method based on convolutional neural networks. In contrast to the traditional intrusive techniques, it does not require a reference recording.

The following three convolutional neural networks were compared as candidate techniques for the non-intrusive audio quality assessment: (1) a model employing 1D convolutional filters, (2) an Inception-based model, and (3) a VGG19-based model. The last-mentioned model performed the best in terms of predicting the scores from the listening tests, yielding a correlation value of 0.893. While the model employing 1D convolutional filters exhibited significantly worse results, the performance of the Inception-based model was almost the same as that of the VGG-19-based model. Moreover, the VGG19-based model outperformed the method employing a stacked gated-recurrent-unit-based deep learning framework, recently introduced by Mumtaz *et al.* [21].

The applicability scope of the proposed non-intrusive method is limited to the assessment of the audio quality of music recordings affected by artefacts produced by lossy audio low-bitrate codecs. Extending the capability of the method to other types of distortions will be considered in future work.

ACKNOWLEDGEMENTS

We are grateful to Albert Bielak for sharing the code and audio repository developed within his undergraduate engineering project at Białystok University of Technology regarding the application of 1D convolutional networks to non-intrusive assessment of audio quality.

REFERENCES

[1] ITU-R BS. 1116-3 Recommendation. "Methods for the subjective assessment of small impairments in audio systems," International Telecommunication Union, Geneva, 2015.

[2] ITU-R BS. 1534-3 Recommendation. "Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, Geneva, 2015.

[3] C. Sloan, N. Harte, D. Kelly, A.C. Kokaram, and A. Hines, "Objective Assessment of Perceptual Audio Quality Using ViSQOLAudio," IEEE Transactions on Broadcasting, vol. 63, pp. 693–705, Dec. 2017. https://doi.org/10.1109/TBC.2017.2704421

[4] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D.J. Inman, "1D convolutional neural networks and applications: A survey," Mechanical Systems and Signal Processing, vol. 151, 107398, 2021. https://doi.org/10.1016/j.ymssp.2020.107398

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed *et al.*, "Going deeper with convolutions," in Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1−9, 2015. https://doi.org/10.1109/CVPR.2015.7298594

[6] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. International Conference on Learning Representations (ICLR), arXiv:1409.1556, 2015. https://doi.org/10.48550/arXiv.1409.1556

[7] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric," in Proc. 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), Athlone, Ireland, 2020. https://doi.org/10.1109/QoMEX48832.2020.9123150

[8] M. Karjalainen, "A new auditory model for the evaluation of sound quality of audio systems," in Proc. ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing, Tampa, FL, USA, 1985. https://doi.org/10.1109/ICASSP.1985.1168376

[9] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ—the ITU standard for objective measurement of perceived audio quality," J. Audio Eng. Soc., vol. 48, pp. 3−29, 2000. http://www.aes.org/e-lib/browse.cfm?elib=12078

[10] ITU-R BS. 1387-2 Recommendation. "Method for objective measurements of perceived audio quality," International Telecommunication Union, Geneva, 2023.

[11] R. Huber and B. Kollmeier, "PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, pp. 1902−1911, 2006. https://doi.org/10.1109/TASL.2006.883259

[12] J. M. Kates and K. H. Arehart, "The Hearing-Aid Audio Quality Index (HAAQI)," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, pp. 354−365, 2016. https://doi.org/10.1109/TASLP.2015.2507858

[13] G. Jiang, A. Biswas, C. Bergler, and A. Maier, "InSE-NET: A Perceptually Coded Audio Quality Model based on CNN," in Proc. 151st Audio Engineering Society Convention, Online, 2021. http://www.aes.org/e-lib/browse.cfm?elib=21478

[14] P. M. Delgado and J. Herre, "Can We Still Use PEAQ? A Performance Analysis of the ITU Standard for the Objective Assessment of Perceived Audio Quality," in Proc. Twelfth International Conference on Quality of Multimedia Experience (QoMEX), Athlone, Ireland, 2020. https://doi.org/10.1109/QoMEX48832.2020.9123105

[15] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep Learning-Based Non-Intrusive Multi-Objective Speech

Assessment Model With Cross-Domain Features," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 54−70, 2023. https://doi.org/10.1109/TASLP.2022.3205757

[16] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022. https://doi.org/10.1109/ICASSP43922.2022.9746108

[17] A. A. Catellier and S. D. Voran, "Wawenets: A No-Reference Convolutional Waveform-Based Approach to Estimating Narrowband and Wideband Speech Quality," in Proc. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020. https://doi.org/10.1109/ICASSP40776.2020.9054204

[18] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in Proc. Interspeech, Hyderabad. India, pp. 1873−1877, 2018. https://doi.org/10.48550/arXiv.1808.05344

[19] G. Mittag, B. Naderi, A. Chehadi, and Sebastian Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in Proc. Interspeech, Brno, Czechia, pp. 2127−2131, 2021. https://doi.org/10.21437/Interspeech.2021-299

[20] C. Sørensen, J. B. Boldt, and M. G. Christensen, "Validation of the Non-Intrusive Codebook-based Short Time Objective Intelligibility Metric for Processed Speech," in Proc. Interspeech, Graz, Austria, pp. 4270−4274, 2019. https://doi.org/10.21437/Interspeech.2019-1625

[21] D. Mumtaz, V. Jakhetiya, K. Nathwani, B. N. Subudhi, and S. C. Guntuku, "Nonintrusive Perceptual Audio Quality Assessment for User-Generated Content Using Deep Learning," IEEE Transactions on Industrial Informatics, vol. 18, pp. 7780−7789, 2022. https://doi.org/10.1109/TII.2021.3139010

[22] K Organiściak and J. Borkowski, "Single-ended quality measurement of a music content via convolutional recurrent neural networks," Metrology and Measurement Systems, vol. 27, pp. 721−733, 2020. https://doi.org/10.24425/mms.2020.134849

[23] EBU R.128 Recommendation, "Loudness normalization and permitted maximum level of audio signals," European Broadcasting Union, Geneva, 2020.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge." International Journal of Computer Vision, vol. 115, pp. 211–252, 2015. https://doi.org/10.1007/s11263-015-0816-y

[25] D. P. Kingma and J. L. Ba, "ADAM: a method for stochastic optimization," in Proc. 3rd International Conference on Learning Representations (ICLR 2015), San Diego, pp. 1–15, 2015. https://doi.org/10.48550/arXiv.1412.6980

[26] A. Kasperuk, "Software repository. Nonintrusive audio quality assessment ISSET2023," GitHub, https://github.com/WaitWhatSon/nonintrusive_audio_quality_assessment_isset2023 (accessed on August 18, 2023).

[27] M. Schoeffler, F. Stöter, B. Edler, and J. Herre, "Towards the Next Generation of Web-based Experiments: A Case Study Assessing Basic Audio Quality Following the ITU-R Recommendation BS.1534 (MUSHRA)," in Proc. 1st Web Audio Conference, Paris, France, 2015.

[28] "The 'Mixing Secrets' Free Multitrack Download Library," Cambridge Music Technology, https://cambridge-mt.com/ms/mtk/ (accessed on June 10, 2023).

[29] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research," in Proc. 15th International Society for Music Information Retrieval Conference, Taipei, Taiwan, 2014.

[30] S. K. Zieliński, "On Some Biases Encountered in Modern Audio Quality Listening Tests (Part 2): Selected Graphical Examples and Discussion," J. Audio Eng. Soc., vol. 64, pp. 55−74, 2016. http://www.aes.org/e-lib/browse.cfm?elib=18105