Metrol. Meas. Syst., Vol. 27 (2020) No. 4, pp. 721–733
DOI: 10.24425/mms.2020.134849

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229
www.metrology.wat.edu.pl

M&MS

SINGLE-ENDED QUALITY MEASUREMENT OF A MUSIC CONTENT VIA CONVOLUTIONAL RECURRENT NEURAL NETWORKS

Kamila Organiściak, Józef Borkowski

Wrocław University of Science and Technology, Chair of Electronic and Photonic Metrology, B. Prusa 53/55, 50-317 Wrocław, Poland (✉ *kamila.organisciak@pwr.edu.pl*, +48 713 206 416, *jozef.borkowski@pwr.edu.pl*)

Abstract

The paper examines the usage of Convolutional Bidirectional Recurrent Neural Network (CBRNN) for a problem of quality measurement in a music content. The key contribution in this approach, compared to the existing research, is that the examined model is evaluated in terms of detecting acoustic anomalies without the requirement to provide a reference (clean) signal. Since real music content may include some modes of instrumental sounds, speech and singing voice or different audio effects, it is more complex to analyze than clean speech or artificial signals, especially without a comparison to the known reference content. The presented results might be treated as a proof of concept, since some specific types of artefacts are covered in this paper (examples of quantization defect, missing sound, distortion of gain characteristics, extra noise sound). However, the described model can be easily expanded to detect other impairments or used as a pre-trained model for other transfer learning processes. To examine the model efficiency several experiments have been performed and reported in the paper. The raw audio samples were transformed into Mel-scaled spectrograms and transferred as input to the model, first independently, then along with additional features (Zero Crossing Rate, Spectral Contrast). According to the obtained results, there is a significant increase in overall accuracy (by 10.1%), if Spectral Contrast information is provided together with Mel-scaled spectrograms. The paper examines also the influence of recursive layers on effectiveness of the artefact classification task.

Keywords: audio data analysis, artefacts detection, convolutional neural networks, recurrent neural networks, classification model.

1. Introduction

Digital audio broadcasting services and internet streaming media providers continuously improve their encoding and delivery methods to minimize processing time and complexity while maintaining audio quality at the same time. Each modification in the broadcast chain, from content creation to a particular hardware setup on the end user side may introduce some unexpected issues to the transferred audio signal. Also, numerous factors of propagation channels and digital

standards can affect broadcast signals [1]. The traditional method of validating the quality of a real audio content is to perform subjective listening tests. However, the substantial amount of encoded contents makes it impractical to perform listening tests for each one of them. Reducing the scope of subjective tests does not resolve the problem, since the created contents differ significantly to provide a better custom user experience.

The evaluation of audio quality can be made through intrusive and nonintrusive metrics [2]. The first type of metrics provide audio quality information by comparing a degraded test signal with its reference (original unprocessed signal). One of the main problems with this technique can be limited access to the reference file. As the reference signal is usually the input for an encoder or another processing product, this approach seems to be correct only for a channel-based content. However, the channel-based approach is not sufficient anymore to encompass an immersive and interactive experience, primarily because of limited combinations of channels. Because of that, object and scene-based formats were introduced [3]. This new approach is much more extensible and efficient, however, the input signal cannot be compared directly to the encoder output without applying additional metadata and/or rendering the content first. For such cases non-intrusive (non-reference) metrics can be used which is able to predict audio quality just using the in-service signal. They can also provide continuous quality monitoring of an audio signal delivered to the end customer or regression tests for a particular node in an end-to-end broadcast ecosystem.

Two types of metrics can be used for audio quality assessment: subjective or objective [2]. While subjective metrics are more complex to analyze and more time consuming to prepare, objective metrics are more difficult to create, as their goal is to reflect the human perception as much as possible. In general, most methods are based on intrusive metrics and require a comparison with the reference signal. This paper examines a *Convolutional Bidirectional Recurrent Neural Network* (CBRNN) model as a new objective method which could be used as a step in automatic audio quality evaluation with no need to provide a reference signal, as it is natural to human perception – the humans are capable of assessing if there are any impairments in an audio content, even without a comparison with the reference (clean) signal.

2. The artefacts detection in audio signals – state of the art

In general, an artefact detection task is a part of the audio quality assessment and is performed by subjective tests where listeners rate the overall quality of a test signal against its reference, following the standard described in ITU-RBS.1284-2 [4]. Since manual tests are an expensive and time-consuming process, there was a need to develop some other fully automatic methods. Currently, the most commonly used objective audio quality predictors, PEAQ (*Perceptual Evaluation of Audio Quality*) [5] (or PESQ Perceptual Evaluation of Speech Quality), POLQA (Perceptual Objective Listening Quality Analysis) [6], PEMO-Q (*Perception Model for Quality Assessment*) [7], STOI (Short-Time Objective Intelligibility) [8], VISQOL (*Virtual Speech Quality Objective Listener*) [9], SNR (*Signal-to-Noise Ratio*) [10] require the reference signals and/or specific types of noise that may degrade audio quality [11]. Existing non-reference solutions [12] are mainly focused on quality measurements of speech (ANIQUE (*Auditory Non-Intrusive Quality Estimation*) [13], HASQI (*Hearing Aid Speech Quality Index*) [14], POSQE (*Perceptual Output-based Speech Quality Evaluation*) [15], SRMR (*Standardized Root Mean Square*) [16] and others [17, 18]), synthetic audio signals [19], image [20] or video [19, 21, 22]. To the best of the authors' knowledge, there is no recommended non-reference objective method for a real music quality assessment.

Metrol. Meas. Syst., Vol. 27 (2020), No. 4, pp. 721–733
DOI: 10.24425/mms.2020.134849

In the context of this paper artefact detection is the task to find abnormal events in a music content. This detection task is associated with several issues: first, it is impractical to reconstruct all possible audio artefacts, as well as collect all possible music tracks. Second, even during manual listening tests there are some rare cases where a listener is not able to tell if the sound event should be treated as an artefact or rather as an intended music content, in case no reference signal is provided. In this paper, the main focus is to classify artefacts clearly detectable by listening tests (which might be assessed as "very annoying" by unipolar discrete five-grade scale used for subjective assessment of impairment [4]), and examine what is the effectiveness of methods used widely for other sound event detection problems (*e.g.* classification of environmental sounds coming from various sources, such as cars, people or buildings [23]). The problem of environmental sound classification seems to be a little bit different, since the model in that case is trained to recognize the specific environmental patterns and backgrounds sound, which is more repetitive and predictable comparing to the music content. In the case of music, it seems ineffective to teach the model how the music pattern should look like, but rather which events should not be present and if they occur in the signal, how they should be classified. There has been no systematic investigation of convolutional and recurrent neural networks effectiveness for an artefact detection task in a music content. This paper is a study of this problem.

2.1. Statement of the problem

The Recommendation ITU-R BS.1284-2 [4] for assessment of impairments of audio signal specifies 11 categories which can be used for analysing and classifying the kind of artefact in digital coding or transmission techniques. These also include the typical ones for multichannel audio (like distortion of spatial image quality or correlation effect – *crosstalk*). However, in this study, only mono signals were examined within specific artefacts categories limited to: *quantization defect* (associated with insufficient digital resolution), *distortion of gain characteristics* (changes in the level or dynamic range of source signals, level jumps), *extra sound* (spurious sounds not related to the source material, such as clicks, pops and noise), and *missing sound* (loss of sound components of the source material, glitches).

At the high level, the problem of detecting artefact category can be viewed as a multinomial classification problem, where the classifier function is parametrized by the introduced neural network.

Formally, we assume that $x \in R_+^{d \times t}$ is the input spectrogram (where t is the length of the spectrogram and d is the dimension of each frame, *i.e.* number of frequency bins in the spectrogram) and $y \in \{1, \ldots, k\}$ is the corresponding signal category where k is the number of classes. Given a training set $D = \{(x_i, y_i)\}_{i=1}^{n}$ of n pairs of the spectrogram and its corresponding label, the problem of artefacts detection can be formalized as finding a model $h\colon R_+^{d \times t} \to \{1, \ldots, k\}$ which produces class predictions for all instances x. The classification model is a probabilistic classifier which assigns to each instance x and class y a probability estimate $P(y \mid x)$ of instance x belonging to class y. We use the general maximum-probability rule to generate class labels [24]:

$$h(x) = \arg\max_{y \in C} P\left(y \mid x\right). \tag{1}$$

We then solve the classification task by generating a classification model based on the supplied training set and a cost-sensitive classification algorithm.

3. Materials and methods

The examined model (Fig. 1) is based on convolutional and recurrent neural networks. These types of networks have become popular, due to their high effectiveness, in a general signal processing, especially in an acoustics event detection [25, 26, 27]. Comparing to the traditional neural network, usage of convolutional layers allows to store fewer parameters, because of so-called sparse interactions, accomplished by making a kernel smaller than the input. This reduces the memory requirements and improves the model's statistical efficiency [28].

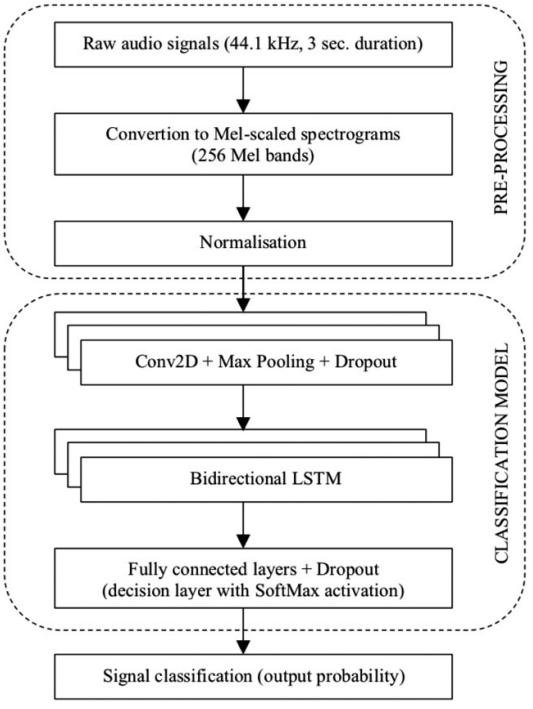

Fig. 1. A diagram of the examined classification algorithm.

The use of recurrent layers supports modelling sequences, especially popular in many natural language processing tasks. The recurrence embedded in an intelligent modelling pipeline provides a way to extend deep learning to sequential data. This allows to reconstruct time domain dependences, valid for sequential patterns, similarly to a regression analysis.

The model was first trained with Mel-scaled spectrograms, extracted from raw audio samples. Next, we examined, how the efficiency of the model could be improved depending on the number of features and the extraction technique.

3.1. Experimental dataset

To the best of the authors' knowledge, there is no existing official database representing a real music content with different types of artefacts. We decided to reuse the latest benchmark music dataset MUSDB18 [29], available upon request. MUSDB18 is a set of real music tracks which includes ~10h duration of different genres along with their isolated drums, bass, vocals and others.

It is provided in a raw uncompressed format with a sampling rate of 44,100 Hz which results in better audio quality compared to the other existing datasets. For example, GTZAN (2002) [30] with recording samples at 22,050 Hz contains a significant fraction of corrupted files and repeated clips [31], the "Million Song Dataset" (2011) [32] acquired with the same sampling rate 22,050 Hz contains audio features and metadata only, whereas FMA (2017) [33] audio samples are already compressed to MP3 format.

Based on the MUSDB18 dataset, five fully labelled sub-sets were created: one set of clean signals and four sets corresponding to the artefact categories. Four artefact categories were made by modifying the original signals, *i.e.* selected types of distortions were added to the base samples – see Fig. 2 and description below:

– *gain distortion*: randomly changing the dynamic range in parts of the signal, each level jump takes at least 20 ms (Fig. 2b presents an example of gain distortion at timestamp 1 sec. and duration ~250 ms);
– *missing sound*: samples created by inserting glitches with low-level random noise (up to –50 dBFS); simulating dropped frames and repeating the last valid frame with a variable duration in range 20–100 ms (Fig. 2c presents an example with a single missing frame with duration of 20 ms at timestamp 240 ms);
– *quantisation defect*: converting to bit depth lower than 16 (Fig. 2d shows an example of insufficient bit resolution, sample size significantly reduced (to 8 bits));

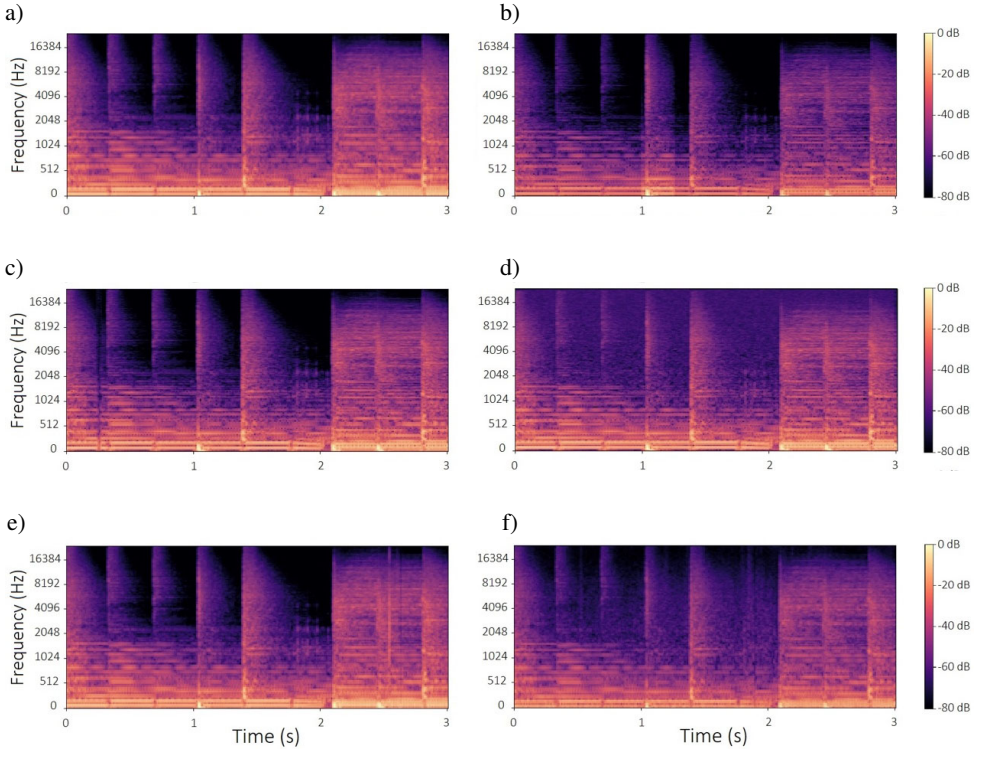

Fig. 2. Example model inputs: Mel-scaled spectrograms extracted from audio signals. The first dimension represents time, the second one – frequency and each value corresponds to its power in dB scale. The first spectrogram contains a 3 sec part of "Summerghost" by Leaf, original clean signal from the MUSDB18 dataset (a). The others present selected types of artificially added artefacts: gain distortion (b), missing frame (c), quantisation defect (d), clicks (e) and noise (f).

– *extra sound (noise)*: mixing a clean signal with generated samples of coloured noise (white, pink, blue, brown, violet) with *Signal-to-Noise Ratio* up to 20 dB; modifying random bits in the signals; combining with real distorted audio samples (additive and burst noise, clicks and pops) from Freesound database [34] (Fig. 2e, f).

The dataset was divided into non-overlapped training, test and validation sets, 60%, 20% and 20% respectively. Currently, there are no distinguished dependencies on a music genre, however, for the future work it would be worth extending the database and analyze the model effectiveness for each music type, separately.

3.2. Data pre-processing

Since the implemented model takes a fixed-size input, the raw audio was divided into 3-seconds chunks. If the last frame was shorter than 3 seconds, it was padded with zeros. According to [23, 26, 35], a different time window is used for audio classification or audio quality measurement task, mostly in range 1–10 seconds. The decision to use 3-seconds chunks is a compromise between reducing the input data complexity and capturing enough audio content and its distortions. In this particular case, signals are long enough to detect an artefact through manual listening tests. The pre-processing stage consists of two phases: spectrograms generation and normalization. First, short-time Fourier transform was performed for each audio sample with a 75% overlap and window size equal to 2048 samples. The spectrograms were created from average power of each band across each signal frame. The power of the obtained spectrum was mapped onto the Mel scale, using triangular overlapping windows with the maximum frequency equal to $fs/2$. All input signals have the same length in the time and frequency domain and were transformed into 256 Mel bands. Each Mel-scaled spectrogram was normalized into $\langle -1, 1\rangle$ [dB] range.

3.3. Methodology

The evaluated model is purely data-driven, it does not make any assumptions about the signal content. The model consists of three main parts. The architecture of the examined model is as follows (Table 1): first, three convolutional layers with 256 filters are used, each with kernel size 3 [23]. To reduce computational complexity each convolutional layer is followed by max pooling layer and dropout. Following the CNN component, three bidirectional recurrent layers of an LSTM are used, with 128 hidden units, and one dense layer with 64 neurons and drop out. The output layer comprises of 5 nodes (equal to the number of recognized categories). To handle number of classes k, the *SoftMax* activation function was used:

$$\sigma(z_i) = \frac{\exp(z_i)}{\sum_{p=1}^{K} \exp(z_p)} \quad (i = 1, \ldots, k), \tag{2}$$

where $\sigma(z_i)$ is the activation function on the output nodes. Since $\sigma(z_i)$ are always positive and their sum is 1, they can be viewed as probabilities, while output nodes with an inserted activation function can be used for probability estimation [36].

The network was fed with 2D feature maps extracted from input signals (Mel-scaled spectrograms where the first dimension represents time, the second dimension represents frequency and each value corresponds to its power on the dB scale). A single-label multiclass classification

Metrol. Meas. Syst., Vol. 27 (2020), No. 4, pp. 721–733
DOI: 10.24425/mms.2020.134849

with one-hot encoding vector was used with a categorical cross-entropy loss [27]:

$$L = -\frac{1}{N_T}\sum_{n=1}^{N_T}\sum_{i=0}^{2}(t_n)_i \ln \sigma(z_{in}), \tag{3}$$

where N_T is a training dataset size and t_n is a label vector associated with an n-th sequence in the training set.

The number of epochs was fixed to 30 with a scheduled learning rate starting from 0.001 to 0.0001. To prevent the model from overfitting, the best model weights were saved for each cycle based on its accuracy on a validation set. The model was evaluated using the Keras framework [37].

Table 1. The examined model architecture.

Layer	# of filters	Kernel / pool size	Stride	Activation function
Conv2D	256	(3, 3)	(1, 2)	ReLu
Maxpooling 2D	N/A	(3, 2)	N/A	N/A
Dropout	0.3			
Conv2D	256	(3, 3)	N/A	ReLu
Maxpooling 2D	N/A	(1,2)	N/A	N/A
Dropout	0.3			
Conv2D	256	(3, 3)	N/A	ReLu
Maxpooling 2D	N/A	(1, 4)	N/A	N/A
Dropout	0.3			
Reshaping	(86, 4 * 256)			
B-LSTM	128 nodes			
B-LSTM	128 nodes			
B-LSTM	128 nodes			
Dense	64 nodes			ReLu
Dropout	0.3			
Dense	5 nodes			softmax

4. Results

The first comparison was performed for two different model architectures: *Convolutional Neural Network* (CNN) and *Convolutional Bidirectional Recurrent Neural Network* (CBRNN). The comparison was performed based on the following metrics: specificity, precision, recall, F1-score and (overall) accuracy [38]:

$$Specificity\ (TNR) = \frac{TN}{TN + FP}, \tag{4}$$

$$Precision = \frac{TP}{TP + FP}, \tag{5}$$

$$Recall = \frac{TP}{TP + FN}, \tag{6}$$

$$Fscore = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{8}$$

where: TP – true positive (the number of artefacts correctly classified as artefacts); FP – false positive (the number of samples incorrectly classified as artefacts); TN – true negative (the number of samples which were not in the selected artefact category, classified correctly as not this artefact type); FN – false negative (the number of samples from different categories classified incorrectly as the selected artefact type). As regards accuracy, the overall result is reported (Table 2).

Table 2. Classification performance measured for various model architectures used in studies.

Category	**CNN**				**CBRNN**			
	precision	recall	F1	TNR	precision	recall	F1	TNR
Clean signals	0.542	0.703	0.612	0.851	0.635	0.748	0.687	0.893
Quantisation defect	0.997	0.900	0.946	0.999	0.999	0.899	0.946	0.999
Gain distortions	0.657	0.611	0.633	0.920	0.764	0.744	0.754	0.943
Extra sound (noise)	0.984	0.729	0.838	0.997	0.990	0.784	0.875	0.998
Missing sound	0.657	0.611	0.876	0.951	0.841	0.973	0.902	0.954
Overall accuracy [%]	77.5				83.0			

The comparison was performed using the same test dataset and the same feature extraction technique – spectrograms scaled to 256 Mel bands. The results (Table 2) show that the usage of bidirectional LSTM layers increases values of almost all metrics. Based on F1-score, the CBRNN model was significantly better for four categories of signals, one of them (*quantisation defect*) remained unchanged. The highest improvement was achieved for *gain distortion* (value increased by 0.121). For other categories, *clean signals, extra sound* (*noise*), *missing sound*, the difference is also noticeable (0.075, 0.037 and 0.026 respectively). The overall accuracy score was raised from 77.5% to 83%.

The next experiments examined model efficiency in terms of the number of input features and their contents. Namely, for sound event classification, usage of 40 Mel bands per frame results in an appropriate model evaluation [23]. However, in the case of analysis of music, it is not enough to get the proper amount of information. Comparing 40, 128 and 256 Mel bands per frame, only the last value resulted in a sufficiently good model evaluation. To improve the effectiveness of the model, two additional input features were examined (Table 3): *Zero Crossing Rate* (ZCR)

Table 3. Performance of CBRNN results depending on the extracted input features.

Category	**Mel-scaled spectrogram + Zero crossing rate**				**Mel-scaled spectrogram + Spectral contrast**			
	precision	recall	F1	TNR	precision	recall	F1	TNR
Clean signals	0.637	0.777	0.700	0.889	0.775	0.969	0.861	0.930
Quantisation defect	0.999	0.899	0.946	0.999	0.999	0.899	0.946	0.999
Gain distortions	0.765	0.752	0.758	0.942	0.949	0.935	0.942	0.988
Extra sound (noise)	0.989	0.796	0.882	0.998	0.987	0.800	0.884	0.997
Missing sound	0.886	0.967	0.924	0.969	0.918	0.967	0.942	0.978
Overall accuracy [%]	83.8				91.4			

Metrol. Meas. Syst., Vol. 27 (2020), No. 4, pp. 721–733
DOI: 10.24425/mms.2020.134849

and *Spectral Contrast*. The first one represents the noisiness of sound – a higher value means more noise in the signal [39]. The second feature – *Spectral Contrast* – widely used in a music classification, finds the difference between spectral peaks and spectral valleys for each sub-band (6 in this case) [40]. Comparing to the previous results, when only spectrograms with 256 Mel bands were used (Table 2), the addition of the first feature improved the overall accuracy only by 0.8 percentage points, while the *Spectral Contrast* information increased it by 8.4 percentage points. In this case, the achieved increase of F1-score is as follows: 0.188 for *gain distortion*, 0.174 for *clean signals*, 0.040 and 0.009 for *extra noise* and *missing sound* respectively.

To examine the impact of these additional features, a transfer learning was involved, where the existing CBRNN model was used as an integrated feature extractor. The pre-trained CBRNN model (without the last classification layer) was integrated into two new models. The first one (Table 4) uses a single Bidirectional LSTM layer with 128 units to process ZCR data. The output of this layer is concatenated to features extracted by the pre-trained CBRNN from Mel-scaled spectrograms. Similarly, to the main model, there is one following dense layer with 64 neurons and a decision layer which consists of five nodes. The input layer of the next model (Table 5) consists of a single convolutional layer followed by max pooling, dropout and also a single Bidirectional LSTM, to process *Spectral Contrast* inputs. Layers and weights of the pre-trained model were frozen during the second training.

Table 4. Model integrating an additional Zero Crossing Rate feature extractor and the pre-trained CBRNN.

Layer	**Nodes**	**Activation function**	**Pre-trained CBRNN (features extractor)**
B-LSTM	128 nodes		
Concatenate			
Dense	64 nodes	ReLu	
Dropout	0.3		
Dense	5 nodes	softmax	

Table 5. Model integrating an additional Spectral Contrast feature extractor and the pre-trained CBRNN.

Layer	**# of filters**	**Kernel / pool size**	**Stride**	**Activation function**	**Pre-trained CBRNN (features extractor)**
Conv2D	256	(3, 3)	(1, 2)	ReLu	
Maxpooling 2D	N/A	(3, 2)	N/A	N/A	
Dropout	0.3				
Reshaping	(86, 7 * 256)				
B-LSTM	128 nodes				
Concatenate					
Dense	64 nodes			ReLu	
Dropout	0.3				
Dense	5 nodes			softmax	

4.1. Analysis of misclassified samples

A correct decision if a signal contains an artefact or not is the most critical from the monitoring and regression tests perspective. A classification which kind of artefact was detected is a matter

of secondary importance, however, this categorization helps us predict which artefacts are most problematic to detect in the designed methodology. The highest false-positive rates occur for clean signals (0.1074) and gain distortions (0.0573). Usage of the *Spectral Contrast* feature reduces these values to 0.0703 and 0.0124 respectively. The highest false-negative rate occurs for gain distortion (0.2559), but also clean signals and signals with extra sound (noise) are significantly affected by this error (0.2525 and 0.2164 respectively). These are also reduced by the additional feature to (0.2482, 0.2233 and 0.2040).

However, they still need to be improved in the feature work on the final product. In the ideal scenario, we would like to achieve the error value close to zero. The algorithm seems to work well for the other basic artefacts examined in this paper. In the light of this, our future work will focus on improving the effectiveness of the model for clean signals, gain distortions and extra sound (noise). The possible solutions for this misclassification would be: 1) extracting additional features; 2) adding multi-label classification; 3) extending the database to include more music samples, especially with an electronic and synthetic content.

5. Conclusions and summary

The goal of this work was to develop a prototype model for the artefact detection in a real-world music content. The provided performance measurements can constitute a basis for further research. The described topic is often perceived to be identical with a standard acoustic event detection problem (like environmental sound classification), however, the results show that in this case more detailed features are required to effectively evaluate the model. The analysis has also shown that the presented CBRNN method with a transfer learning (additional features) results in considerably better performance than the other objective benchmark method presented in this paper. Comparing to the standard CNN, the overall classification accuracy is 7.1% higher. Also, the results show that the current algorithm and selected hyperparameters perform best using Mel-scaled spectrograms with at least 256 filters and addition of *Spatial Contrast* information as input features.

The current implementation covers only selected audio artefacts. The future work will focus on extending the database to reduce false-positive and false-negative errors (especially for clean signals) and recognize more distortions. The actual database was prepared based on the MUSDB18 which includes ~10 h duration of music content. A similar amount of data is referred to in some existing publications regarding audio classification, *e.g.* [26]. This seems to be sufficient on the prototype examination level, especially for a very limited number of kinds of artefacts. However, the prospective goal is to provide a much more extended dataset. To exploit the full capability of the latest deep learning techniques and improve the model effectiveness of classification, a dataset exceeding 5000 hours would be preferable [41]. This may increase the model accuracy and reliability, especially for more complex scenarios. Since the main focus of this work was to classify very basic artefacts, clearly detectable in listening tests (assessed as “very annoying” by unipolar discrete five-grade scale used for subjective assessment of impairment [4]), the other future goal would be to extend the model and perform experiments for more subtle, real impairments and to improve the robustness for such content.

References

[1] Gilski, P., & Stefański, J. (2017). Transmission Quality Measurements in DAB+ Broadcast System. *Metrology and Measurement Systems, 24*(4), 675–683. https://doi.org/10.1515/mms-2017-0050

[2] Rix, A. W., Beerends, J. G., Kim, D. S., Kroon, P., & Ghitza, O. (2006). Objective assessment of speech and audio quality – technology and applications. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(6), 1890–1901. https://doi.org/10.1109/TASL.2006.883260

[3] International Telecommunication Union. (2017). *Audio Definition Model* (Recommendation ITU-R BS.2076-1). https://www.itu.int/rec/R-REC-BS.2076-1-201706-S/en

[4] International Telecommunication Union. (2019). *General methods for the subjective assessment of sound quality* (Recommendation ITU-R BS.1284-2). https://www.itu.int/rec/R-REC-BS.1284-2-201901-I/en

[5] Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., & Colomes, C. (2000). PEAQ-The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society, 48*(1–2), 3–29.

[6] International Telecommunication Union. (2011). *Perceptual Objective Listening Quality Assessment* (Recommendation ITU-T P.863). https://www.itu.int/rec/T-REC-P.863-201101-S/en.

[7] Sloan, C., Harte, N., Kelly, D., Kokaram, A. C., & Hines, A. (2017). Objective assessment of perceptual audio quality using ViSQOLAudio. *IEEE Transactions on Broadcasting, 63*(4), 693–705. https://doi.org/10.1109/TBC.2017.2704421

[8] Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(7), 2125–2136. https://doi.org/10.1109/TASL.2011.2114881

[9] Hines, A., Gillen, E., Kelly, D., Skoglund, J., Kokaram, A., & Harte, N. (2015). ViSQOLAudio: An objective audio quality metric for low bitrate codecs. *The Journal of the Acoustical Society of America, 137*(6), EL449–EL455. https://doi.org/10.1121/1.4921674

[10] Plapous, C., Marro, C., & Scalart, P. (2006). Improved Signal-to-Noise Ratio Estimation for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(5), 2098–2108. https://doi.org/10.1109/TASL.2006.872621

[11] Li, Z., Wang, J. C., Cai, J., Duan, Z., Wang, H. M., & Wang, Y. (2013). Non-reference audio quality assessment for online live music recordings. *Proceedings of the 21st ACM international conference on Multimedia*, Spain, 63–72. https://doi.org/10.1145/2502081.2502106

[12] Akhtar, Z., & Falk, T. H. (2017). Audio-visual multimedia quality assessment: A comprehensive survey. *IEEE Access, 5*, 21090–21117. https://doi.org/10.1109/ACCESS.2017.2750918

[13] Doh-Suk, K. (2005). ANIQUE: An auditory model for single-ended speech quality estimation. Speech and Audio Processing. *IEEE Transactions, 13*(5), 821–831. https://doi.org/10.1109/TSA.2005.851924

[14] Kates, J. M., & Arehart, K. H. (2010). The hearing-aid speech quality index (HASQI). *Journal of the Audio Engineering Society, 58*(5), 363–381. http://www.aes.org/e-lib/browse.cfm?elib=15451.

[15] Mahdi, E. A., & Picovici, D. (2010). New single-ended objective measure for non-intrusive speech quality evaluation. *Signal, Image Video Process, 4*, 23–38. https://doi.org/10.1007/s11760-008-0092-1

[16] Falk, T. H., Zheng, C., & Chan, W. Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, Language Processing, 18*(7), 1766–1774. https://doi.org/10.1109/TASL.2010.2052247

[17] Orcik, L., Voznak, M., Rozhon, J., Rezac, F., Slachta, J., Toral-Cruz, H., & Lin, J. C. W. (2017). Prediction of speech quality based on resilient backpropagation artificial neural network. *Wireless Personal Communications, 96*(4), 5375–5389. https://doi.org/10.1007/s11277-016-3746-2

[18] Falk, T. H., & Chan, W. Y. (2006). Single-ended speech quality measurement using machine learning methods. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(6), 1935–1947. https://doi.org/10.1109/TASL.2006.883253

[19] Babić, D., Pul, M., Vranješ, M., & Peković, V. (2017, October). Real-time audio and video artifacts detection tool. *International Conference on Smart Systems and Technologies (SST)*, Croatia, 251–256. https://doi.org/10.1109/SST.2017.8188704

[20] Shen, J., Li, Q., & Erlebacher, G. (2011). Hybrid no-reference natural image quality assessment of noisy, blurry, JPEG2000, and JPEG images. *IEEE Transactions on Image Processing, 20*(8), 2089–2098. https://doi.org/10.1109/TIP.2011.2108661

[21] Li, Y., Po, L. M., Cheung, C. H., Xu, X., Feng, L., Yuan, F., & Cheung, K. W. (2015). No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology, 26*(6), 1044–1057. https://doi.org/10.1109/TCSVT.2015.2430711

[22] Chen, C., Izadi, M., & Kokaram, A. (2016, October). A perceptual quality metric for videos distorted by spatially correlated noise. *Proceedings of the 24th ACM international conference on Multimedia*, Netherlands, 1277–1285. https://doi.org/10.1145/2964284.2964302

[23] Jung, S., Park, J., & Lee, S. (2019). Polyphonic sound event detection using convolutional bidirectional lstm and synthetic data-based transfer learning. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, United Kingdom, 885-889. https://doi.org/10.1109/ICASSP.2019.8682909

[24] Cichosz, P. (2015). *Data Mining Algorithms: Explained Using R*. John Wiley & Sons.

[25] Zhao, H., Zarar, S., Tashev, I., & Lee, C. H. (2018). Convolutional-Recurrent Neural Networks for Speech Enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Canada, 2401–2405. https://doi.org/10.1109/ICASSP.2018.8462155

[26] Sang, J., Park, S., & Lee, J. (2018). Convolutional recurrent neural networks for urban sound classification using raw waveforms. *26th European Signal Processing Conference (EUSIPCO)*, Italy, 2444–2448. https://doi.org/10.23919/EUSIPCO.2018.8553247

[27] Portsev, R. J., & Makarenko, A. V. (2018). Convolutional Neural Networks for Noise Signal Recognition. *IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, Denmark. https://doi.org/10.1109/MLSP.2018.8516920

[28] Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep Learning. London*. The MIT Press.

[29] Rafii, Z., Liutkus, A., Stöter, F. R., Mimilakis, S. I., Bittner, R. (2018). *The MUSDB18 corpus for music separation*. https://doi.org/10.5281/zenodo.1117372

[30] Sturm, B. L. (2014). The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of New Music Research, 43*(2), 147–172. https://doi.org/10.1080/09298215.2014.894533

[31] Lu, Y. C., Wu, C. W., Lu, C. T., & Lerch, A. (2016, July). An unsupervised approach to anomaly detection in music datasets. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, United States, 749–752. https://doi.org/10.1145/2911451.2914700

[32] Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & Lamere, P. (2011). The Million Song Dataset. *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, United States. https://doi.org/10.7916/D8NZ8J07

[33] Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2017). FMA: A Dataset For Music Analysis. *18th International Society for Music Information Retrieval Conference (ISMIR)*, China. https://arxiv.org/abs/1612.01840.

[34] Font, F., Roma, G., & Serra, X. (2013). Freesound technical demo. *Proceedings of the 21st ACM international conference on Multimedia (MM '13)*, New York, 411–412. https://doi.org/10.1145/2502081.2502245

[35] Min, X., Zhai, G., Zhou, J., Farias, M. C., & Bovik, A. C. (2020). Study of Subjective and Objective Quality Assessment of Audio-Visual Signals. *IEEE Transactions on Image Processing, 29*, 6054–6068. https://doi.org/10.1109/TIP.2020.2988148

[36] Camastra, F., & Vinciarelli, A. (2015). *Machine Learning for Audio, Image and Video Analysis: Theory and Applications*. Springer-Verlag London. https://doi.org/10.1007/978-1-4471-6735-8

[37] Chollet, F. *et al.* (2015). Keras. GitHub. https://github.com/fchollet/keras

[38] Mesaros, A., Heittola, T., & Virtanen, T. (2016). Metrics for polyphonic sound event detection. *Applied Sciences, 6*(6), 162. https://doi.org/10.3390/app6060162

[39] Gulhane, S. R., Badhe, S. S., & Shirbahadurkar, S. D. (2018). Cepstral (MFCC) feature and spectral (Timbral) features analysis for musical instrument sounds. *2018 IEEE global conference on wireless computing and networking (GCWCN)*, India, 109–113. https://doi.org/10.1109/GCWCN.2018.8668628

[40] Khonglah, B. K., & Prasanna, S. M. (2017). Clean speech/speech with background music classification using HNGD spectrum. *International Journal of Speech Technology, 20*(4), 1023–1036. https://doi.org/10.1007/s10772-017-9464-7

[41] Wu, Y., & Lee, T. (2018, April). Reducing model complexity for DNN based large-scale audio classification. *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Canada, 331–335. https://doi.org/10.1109/ICASSP.2018.8462168

Józef Borkowski received his Ph.D. degree (1997) and D.Sc. degree (2012) from Wrocław University of Science and Technology, Wrocław, Poland in the field of electronics. He is currently Associate Professor at Wrocław University of Science and Technology, the Chair of Electronic and Photonic Metrology.
His main research areas are precise spectrum analysis methods (IpDFT, LIDFT), data processing in measurements using DSP methods, and methods of analysis and data processing in renewable energy systems.

Kamila Organiściak received the M.Sc. degree in electronic engineering from Wrocław University of Science and Technology, Wrocław, Poland, in 2017, where she is currently working toward her Ph.D. degree at the Chair of Electronic and Photonic Metrology. Her current research interests include audio signals processing and machine learning algorithms.