2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)

Audio Impairment Recognition using a Correlation-Based Feature Representation

Alessandro Ragano[1,2,3,4], Emmanouil Benetos[3,4], and Andrew Hines[1,2]
[1] *School of Computer Science, University College Dublin, Ireland* [2] *Insight Centre for Data Analytics, Ireland*
[3] *School of EECS, Queen Mary University of London, UK* [4] *The Alan Turing Institute, UK*
alessandro.ragano@ucdconnect.ie, emmanouil.benetos@qmul.ac.uk, andrew.hines@ucd.ie

***Abstract*—Audio impairment recognition is based on finding noise in audio files and categorising the impairment type. Recently, significant performance improvement has been obtained thanks to the usage of advanced deep learning models. However, feature robustness is still an unresolved issue and it is one of the main reasons why we need powerful deep learning architectures. In the presence of a variety of musical styles, hand-crafted features are less efficient in capturing audio degradation characteristics and they are prone to failure when recognising audio impairments and could mistakenly learn musical concepts rather than impairment types. In this paper, we propose a new representation of hand-crafted features that is based on the correlation of feature pairs. We experimentally compare the proposed correlation-based feature representation with a typical raw feature representation used in machine learning and we show superior performance in terms of compact feature dimensionality and improved computational speed in the test stage whilst achieving comparable accuracy.**

***Index Terms*—audio impairments, feature representation, feature dimensionality, feature robustness, convolutional neural networks**

I. Introduction

Audio classification tasks range from general classification of audio content e.g., music, speech or noise [1], to more specific classification of musical content e.g., music genre recognition or music annotation [2]. Identification and recognition of audio impairments is needed in several applications such as audio restoration of sound archives [3], [4], voice activity detection [5] and audio quality assessment [6]. Particularly, recognition of audio impairments can be useful in non-intrusive audio quality assessment i.e., when quality is predicted without the usage of the clean signal as opposed to intrusive quality metrics that make use of both clean and degraded signals. Non-intrusive quality assessment is more difficult to tackle and usually performs worse than the intrusive setting. However, non-intrusive metrics are necessary when assessing quality in real-time applications such as VoIP calls [7], [8] or when the clean signal cannot be available and sound is inherently noisy, for example with audio archives [9].

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 17/RC-PhD/3483 and 17/RC/2289_P2 and was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. EB is supported by RAEng Research Fellowship RF/128 and a Turing Fellowship.

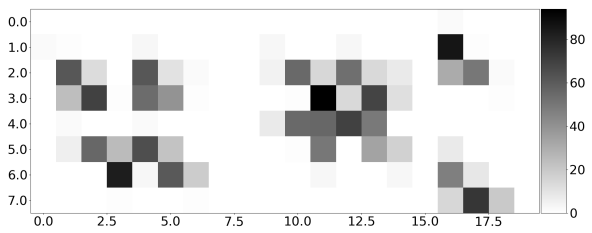
Improving performance in the above-mentioned applications is critical for multimedia service providers in order to provide better quality of experience (QoE) for the end-users.

In a typical machine learning scenario, researchers employ a two-stage approach: 1) extracting features, 2) designing the classifier. Feature extraction refers to the selection of salient and discriminatory measured values or the computation of secondary values from the measured values. Where the features are carefully selected or chosen for the task at hand they are often referred to as *hand-crafted*. This selection process requires expertise or knowledge about the task and signal type under evaluation.

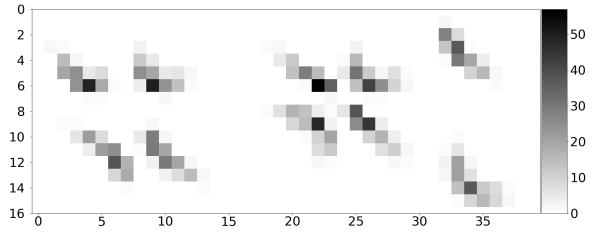
For music classification, hand-crafted features include low-level features and high-level features. Low-level features such as spectral centroid, spectral bandwidth and zero crossing rate [10] are typically extracted with a frame duration between 10 ms and 100 ms [2]. These features do not represent a human-level understanding of musical events, in contrast to high-level features such as pitch and beat which are closer to the perception of musical events.

Hand-crafted features have been employed in audio classification tasks for noise identification [4], audio impairment recognition [4], [6], audio anomaly detection [11] and general classification tasks in music informatics [12]. However, the problem of lack of robustness of those features is still an unresolved issue. Improved performance in music informatics tasks is mainly due to advanced deep learning models which partly compensate the lack of robustness in hand-crafted features as explained by Humphrey et al. [12]. This problem is emphasized in audio impairment recognition and noise identification scenarios where hand-crafted features could fail in recognising impairment types given the presence of different styles of musical content.

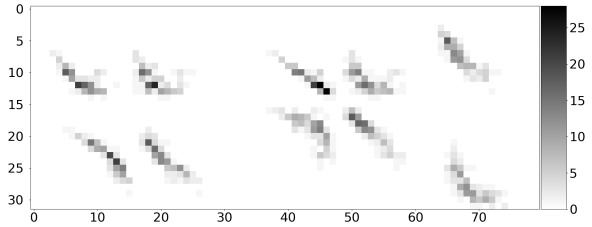
In this paper, we propose a new representation of hand-crafted features based on the correlation between pairs of features for the tasks of audio impairment recognition and noise identification. Rather than feeding the classifier with raw feature values we use a correlation-based feature representation as input of the classifier. We carry out experiments on noise identification and audio impairment recognition. To demonstrate the wide applicability of the proposed method we also investigate the music genre recognition task which is widely explored in music informatics. Insights from our study show that the correlation-based feature representation achieves



(a) Combination of 10 two-dimensional histograms using 4 bins.



(b) Combination of 10 two-dimensional histograms using 8 bins.



(c) Combination of 10 two-dimensional histograms using 16 bins.

SP BW / SP CT	SP BW / SP FL	SP BW / SP RO	SP BW / ZCR	SP CT / SP FL
SP CT / SP RO	SP CT / ZCR	SP FL / SP RO	SP FL / ZCR	SP RO / ZCR

(d) Respective names of each feature pair.

Fig. 1. An example of the correlation-based representation of one training sample belonging to the first 4 seconds of "The Slump" by Tony Williams mixed with a vinyl hiss at 30 dB. 4 bins representation (a), 8 bins representation (b), and 16 bins representation (c). The corresponding positions of the feature pairs are shown in (d).

the same accuracy of hand-crafted raw features and results in reduced feature dimensionality and higher computational speed when testing the model.

II. Related Work and Motivations

The motivation to explore the effect of using the correlation between features as input of the classifier originated while exploring the problem of feature robustness in the presence of noise for audio quality assessment applications [9], [13]. Classifying the impairment type can be useful when predicting audio quality in non-intrusive scenarios as it can give information on the expected perceived audio distortion [13].

We assume that a more compact representation of features could be achieved by representing correlation between each pair of features rather than using raw feature values. In this way, our expectation is that we provide more informative features to the classifier by attenuating redundant information in hand-crafted features. New directions in music informatics propose the use of feature learning instead of using hand-crafted features [12]. However, the same does not apply for audio impairment recognition where hand-crafted features and spectrograms are still employed [6]. For this reason, we decided to use hand-crafted features for our task. Designing robust features for general music informatics tasks has been analysed by Humphrey et al. [12]. The authors discover that insufficient non-linearity, poorly tuned parameters, and the inherent problems with short-time analysis are the main problems of hand-crafted features. One of the main conclusions from their analysis is that the lack of robustness in features can be partly compensated by more powerful classifiers which are based on deep learning. The same problem is found in audio impairment recognition [3]–[6]. The authors have shown that hand-crafted features are not robust in various contexts. They conclude that more powerful classifiers such as deep convolutional neural networks (CNNs) or features based on psychoacoustic models are needed to reach high performance. A similar scenario is also found in audio anomaly detection where detecting noise in the audio data requires advanced deep learning models to compensate for the lack of feature robustness [11], [14], [15]. Modifying feature representation has been successful when detecting non-speech audio events where methods such as principal component analysis (PCA) have been used for reducing dimensionality and improve feature robustness at the same time [16]. Motivated by the above-mentioned issues, in this paper we create a correlation-based representation to tackle the above-mentioned problem of feature robustness. Our assumption is that the correlation as a two-dimensional representation with spatial relationships could be more robust for detecting noise hidden in a dataset characterised by many different musical styles given that it could eliminate redundant information that belongs to the music content rather than the impairment type. The goal of the paper is to contrast two feature representations in the presence of noise. We carry out 3 experiments: noise identification, audio impairment recognition, and music genre recognition. We assess the differences in accuracy, feature dimensionality, and running time between the correlation-based feature representation and the typical usage of raw feature values. Contrary with our assumptions, we show that there is no improvement in terms of feature robustness where both feature representations perform similarly. However, results show improvement in terms of feature dimensionality reduction and running time which support the usage of the proposed method.

III. Representation of Feature Correlation using Two-Dimensional Histograms

The correlation-based feature representation can be obtained by generating the equivalent of the individual scatter plot used

TABLE I

CNN ARCHITECTURE USED IN AUDIO IMPAIRMENT RECOGNITION (AIR) AND NOISE IDENTIFICATION (NI) BY USING THE CORRELATION AND THE RAW REPRESENTATION. WE SHOW THE NUMBER OF CHANNELS (c), THE KERNEL SIZE (k), AND THE PADDING FACTOR (p) FOR CONVOLUTIONAL LAYERS, THE POOLING SIZE (ps) FOR THE MAX POOLING LAYERS AND THE NUMBER OF NODES (n) FOR THE FULLY CONNECTED LAYERS. $F = 5$ IS THE NUMBER OF EMPLOYED FEATURES.

	Audio Impairment Recognition		Noise Identification	
Operation	AIR-Correlation	AIR-Raw	NI-Correlation	NI-Raw
Conv	$c = 16$, $k = (7, 7)$, $p = same$	$c = 16$, $k = (4, F)$	$c = 8$, $k = (3, 3)$, $p = same$	$c = 8$, $k = (4, F)$
Conv	$c = 32$, $k = (5, 5)$	$c = 32$, $k = (4, 1)$	$c = 16$, $k = (3, 3)$	$c = 16$, $k = (4, 1)$
MaxPool[Dropout(50)]	$ps = (2, 2)$	$ps = (PrevOutput, 1)$	$ps = (2, 2)$	$ps = (PrevOutput, 1)$
Dense[Dropout(50]	$n = 200$	$n = 200$	$n = 128$	$n = 128$
Dense	$n = 4$	$n = 4$	$n = 2$	$n = 2$

for analysing the correlation between every pair of features. To represent the correlation we generate an individual scatter plot by computing the two-dimensional histograms of each pair of features employed in the experiment.

The correlation-based representation is generated as follows:

1) We split the whole audio track into frames of 4 seconds with 50% of overlap as done in [17], [18].
2) Given an audio frame we compute hand-crafted features. We compute two groups of features F as defined in [10]. The first group, the short-time (ST) features, includes spectral roll-off (SP RO), spectral centroid (SP CT), zero crossing rate (ZCR), spectral bandwidth (SP BW), and spectral flatness (SP FL). The second group is characterised by the first 5 mel-frequency cepstral coefficients (MFCCs).
3) For each pair of features we compute a 2D histogram specifying the number of histogram bins B and the histogram range. The number of histogram bins is a crucial parameter as we will show in the results section.
4) The histogram range is given by the maximum/minimum absolute value of each feature with respect to the whole dataset. In this way we avoid eventual similarity between different pairs of features due to the automatic scaling of each feature when combining different individual 2D histograms.
5) We combine each individual 2D histogram of each pair of features in a bigger matrix which represents a training sample. We concatenate along the rows or the columns depending on the number of features employed.

Given that we use $F = 5$ features we need $\binom{F}{2} = 10$ individual two-dimensional histograms to be concatenated with each other. Each individual histogram has a size equal to $B \times B$. Therefore the size of one training sample is given by:

$$(2 \times B) \times (5 \times B)$$

where $(2 \times B)$ is the number of rows and $(5 \times B)$ is the number of columns.

In Figure 1 we visualise one training sample from our dataset which has been created by mixing tracks from the GTZAN dataset [19] with impairments taken from the Freesound database [20]. The track is "The Slump" by Tony Williams mixed with a vinyl hiss at 30 dB. We show the correlation-based feature representation using 4 bins, 8 bins, and 16 bins indicating the respective feature pairs. The number of histogram bins B affects both resolution and dynamic range of the proposed feature representation. A small value of B results in higher dynamic range but lower resolution. This is due to the fact that with using fewer histogram bins we associate more different points to the same square as shown in Figure 1a that consequently increases the magnitude at the expense of lower resolution.

It should be noted that we experimented with using images of the scatter plot matrix generated from data analysis libraries such as Pandas as alternative to the proposed histogram-based method but informal tests yielded inferior results compared to the histogram method.

IV. EXPERIMENTAL SETUP

In this section we describe the experimental setup which includes a description of the dataset and model architectures.

A. Dataset

We employ the GTZAN dataset [19], widely used for music genre recognition [21], because it allows us to use a good variety of musical styles avoiding any potential genre-specific bias. The dataset includes 1000 tracks equally divided in 10 different genres, each with 30 seconds duration. For audio impairment recognition and noise identification we extract 120 tracks from the dataset by excluding those containing repetitions, mislabelings and distortions as discussed by Sturm [22]. We mix the 120 tracks with different audio impairments as discussed below.

B. Model Architectures

In each experiment we use a CNN as classifier. The motivation to use CNNs is due to their success in computer vision applications where the input of the architecture is a two dimensional matrix with spatial relationships as in the case of the proposed method. In Table I we show the architectures used in the two tasks for each group of features. In both tasks we try to keep a very simple architecture to minimise the capacity of the classifier. Regarding raw features we use a domain-knowledge approach [23] to design the kernels. By using 5 columns in the kernel of the first convolutional layer, the model learns the combination of features. Then we

TABLE II

PERFORMANCE EVALUATION OF AUDIO IMPAIRMENT RECOGNITION AND NOISE IDENTIFICATION WITH MFFCS AND ST FEATURES. WE COMPARE THE ACCURACY, THE SIZE, AND RUNNING TIME AT 3 DIFFERENT HISTOGRAM BINS B BETWEEN THE PROPOSED METHOD CORR(B=4,B=8,B=16) AND THE RAW FEATURE REPRESENTATION. BOLD TEXT INDICATES THE CASES WHERE THE PROPOSED METHOD SHOWS SUPERIOR PERFORMANCE.

	Audio Impairment Recognition - MFCCs				Audio Impairment Recognition - ST			
	Corr(B=4)	Corr(B=8)	Corr(B=16)	Raw	Corr(B=4)	Corr(B=8)	Corr(B=16)	Raw
Accuracy	57.49	**63.57**	**63.45**	63.05	54.64	**59.35**	**61.78**	59.16
Size	**160**	**640**	2560	865	**160**	**640**	2560	865
Running Time (*secs*)	**2.21**	**2.71**	6.35	4.71	**2.05**	**3.44**	6.69	5.22

	Noise Identification - MFCCs				Noise Identification - ST			
	Corr(B=4)	Corr(B=8)	Corr(B=16)	Raw	Corr(B=4)	Corr(B=8)	Corr(B=16)	Raw
Accuracy	73.15	76.15	77.08	77.91	73.57	73.79	76.54	77.38
Size	**160**	**640**	2560	865	**160**	**640**	2560	865
Running Time (*secs*)	**2.40**	**3.09**	6.35	4.90	**2.30**	**3.36**	7.09	4.70

design the second convolutional layer to model the temporal dependencies as adopted similarly in [17]. We have noticed a more robust learning when using this approach.

V. RESULTS

The goal of these experiments is to compare the proposed feature representation with a typical feature representation that uses raw values. We evaluate performance by varying the histogram bins B in the proposed method to assess the accuracy, the size of the training dataset, and running time for the test stage. The size is expressed in terms of feature dimensionality and represents the number of matrix entries in the training dataset. We expect that the running time is shorter when using a more compact representation. It should be noted that the dataset is created in both methods with the same amount of audio data and the size difference in the training data depends on the employed representation. The running time is the amount of time that the model takes to be tested and it is computed on a MacBook Pro with 6-Core Intel i9 2.9GHz. One training/test sample captures 4 seconds of audio with 2 seconds of overlap in both raw and correlation representations. This aggregation of features over time has been used to improve the classification accuracy as discussed in [18]. We split the dataset into training, validation, and test subsets. The training and test subsets are partitioned using cross-validation while the validation set represents 15% of the training dataset obtained after cross-validation. We use the validation set to find the best model i.e., the one with the lowest categorical cross-entropy loss after 800 epochs. The best model is selected for evaluation with the test set. It should be noted that the split is made on a track-level instead of splitting on a frame-level to avoid any possible repetitive use of the data between the training set, the test set, and the validation set.

A. Audio Impairment Recognition

We perform a stratified 6-fold cross validation to test our model which guarantees a balanced trade-off between bias and variance and equal partition of the data with respect to the ground truth. We classify 4 types of noise that are typically found in archive recordings: vinyl hiss, tape noise, gramophone noise, and white noise. The first 3 real-world noises have been obtained from the Freesound database [20]. We create mixtures at different SNR levels from 0 dB to 30 dB with 5 dB increments as done by Reddy et al. [6]. The goal is to classify the 4 different impairments. The accuracy and runtime are averaged over the different test partitions while the size is not averaged since it is the same in every experiment. The results of the accuracy, the size, and running time are shown in Table II. The correlation-based feature representation with 8 bins allows to use a dataset $\approx 30\%$ smaller than the one obtained from the raw representation. Regarding running time, the correlation representation allows to have a model which is $\approx 53\%$ faster when using MFCCs and $\approx 41\%$ faster with ST features. Shorter running time is due to the compact feature representation as expected. The accuracy of the classifier has no significant difference between the correlation representation and the raw values. The usage of other bin values is discouraged due to lower accuracy ($B = 4$), larger dataset ($B = 16$) and longer running time ($B = 16$). The difference in the results between MFCCs and ST features is negligible when comparing the correlation representations with the raw values. However, it is significant in terms of absolute performance where MFCCs outperform ST features. The obtained accuracy scores show that the proposed method does not address the hypothesis of feature robustness that we discussed above. However, reduction in both size and running time support the usage of the proposed method for reducing feature dimensionality.

TABLE III
MUSIC GENRE RECOGNITION ACCURACY.

Method	Features	Classifier	Accuracy
[19]	{8 ST+12 MFCCs}×MuVar+beat+pitch	GMM	61% ± 4%
[17]	5 ST+2 Tonality+ ST Energy	CNN+RB	89.6% ±2.4%
Raw Features	5 ST	CNN	72.20% ± 4.26%
Correlation Features	5 ST	CNN	68.90% ± 4.54%
Late Fusion	5 ST Raw and Corr.	CNN	74.70% ± 5.24%

B. Noise Identification

As with the previous task, we perform a stratified 6-fold cross-validation to test the proposed model in several equally split partitions. Here we have a binary classification problem. Half the dataset contains clean recordings while the other half contains mixtures at 30, 15, and 0 dB SNR levels. The types of noise employed are the same as above. The mixtures with SNR levels lower than 30 dB are considered as noisy recordings. Using different SNR levels guarantees more robustness during training and ensures that the classifier learns from the noise instead of the different mixture levels. The goal is to distinguish noisy recordings from the clean ones regardless of the noise type. We decided to include different types of noise to explore a harder task compared to the case with only one class of noise. The results are shown in Table II. We obtain results similar to the audio impairment recognition task. The correlation-based feature representation results in a dataset $\approx 30\%$ smaller and in a model runtime $\approx 43\%$ faster with MFCCs and $\approx 33\%$ faster with ST features despite a comparable accuracy. Again, results support the usage of the proposed method for reducing feature dimensionality instead of improving feature robustness as we assumed in the motivations section above.

C. Music Genre Recognition

In these experiments we apply a 10-fold cross validation on the whole GTZAN dataset as used in [2], [17], [19]. Unlike the previous tasks, we evaluate only the accuracy by setting the histogram bins equal to 8 as we are interested in understanding the capability of the method in a different task and using a much larger dataset. Therefore, we have a dataset that $\approx 30\%$ smaller than the raw feature representation.

The goal of this task is to classify 10 genres as labelled in the GTZAN dataset. Results are shown in Table III. The proposed method shows a slightly lower accuracy to the one using the raw feature representation with the advantage of $\approx 30\%$ reduction in feature dimensionality. It should be noted that this method does not outperform the best model [17] which uses CNNs with residual blocks (RB) and a combination between ST features and time-frequency energy. This result is aligned with the previous tasks where we show that the advantages of the proposed method are due to feature dimensionality and running time instead of higher accuracy. We also explore the combination of the raw feature values and the correlation after training the two models separately. The accuracy of 74.70% suggests that the two methods do not learn exactly the same concepts.

VI. DISCUSSION

In this paper we have shown that using the correlation between features using two dimensional histograms allows us to use a reduced feature dimensionality and a shorter running time whilst achieving comparable accuracy. The performance of the proposed method depends on the choice of the histogram bins B up to a limit of 16 bins (Figure 1c). Testing with higher values for B had no positive effect on the accuracy. For $B = 4$, accuracy is reduced and the correlation structure seen for $B = 8$ or 16 is not reproduced in Figure 1a. This means that higher resolution of feature correlation, given by a bigger value of B, is more informative than higher dynamic range, which is obtained for smaller values e.g., $B = 4$.

It could be assumed that the comparable accuracy between the raw and correlation trained models is explained by the capacity of the CNN to learn the feature correlation. However, the result shown in Section V-C for late fusion suggests that the two representations are not exactly the same.

Results do not confirm our assumption on feature robustness. However, our proposed method contributes in terms of feature dimensionality. Classifying impairment types in audio quality assessment application can benefit from having a reduced feature dimensionality for saving bandwidth in real time monitoring and storage in case of large volume of data. Given that the motivation of using the proposed method is related to feature robustness instead of feature dimensionality reduction, a comparison with other methods for the latter task is not given in this paper. Nevertheless, research studies on feature dimensionality reduction methods such as PCA and factor analysis, suggest that these methods require a large ratio between number of samples and number of features in order to preserve consistency in the data [24]–[26]. This is crucial as the dataset that we used has an approximate ratio observations-features that is slightly lower than 2:1 which is not the recommended ratio for using PCA or factor analysis. Therefore, as the proposed method is not dependent from the ratio observation-features, it could be preferred to the above-mentioned methods on datasets that show this peculiarity.

VII. CONCLUSIONS AND FUTURE WORK

In this paper we presented a new way to represent hand-crafted features for audio impairment recognition which is based on the correlation between features. We obtained significant performance improvement in terms of feature dimensionality and running time while maintaining accuracy levels similar to those obtained using raw feature representation. We also obtained a reduced feature dimensionality for music genre recognition i.e., $\approx 30\%$ smaller, at the expense of a slightly lower accuracy.

In the future we want to explore different features to see if the method is robust in more scenarios. In particular, we want to go beyond the usage of low-level features by exploring the correlation-based representation on high-level features. We also intend to explore methods that further decrease the feature dimensionality and the runtime by compressing the sparse matrices obtained from the proposed representation. Finally,

we believe that a comparison with other feature dimensionality reduction methods is needed. In particular, we want to compare the proposed method with other feature dimensionality reduction methods such as factor analysis and PCA using a dataset with higher ratio observations-features and with other feature selection methods such as correlation-based elimination and backward feature elimination.

REFERENCES

[1] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, "Classification of audio signals using AANN and GMM," *Applied Soft Computing*, vol. 11, no. 1, pp. 716–723, 2010.

[2] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.

[3] C. F. Stallmann and A. P. Engelbrecht, "Gramophone Noise Detection and Reconstruction Using Time Delay Aritificial Neural Networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 6, pp. 893–905, 2016.

[4] M. Brandt, S. Doclo, T. Gerkmann, and J. Bitzer, "Impulsive disturbances in audio archives: Signal classification for automatic restoration," *Journal of the Audio Engineering Society*, vol. 65, no. 10, pp. 826–840, 2017.

[5] J. Saeedi, S. M. Ahadi, and K. Faez, "Robust voice activity detection directed by noise classification," *Signal, Image and Video Processing*, vol. 9, no. 3, pp. 561–572, 2015.

[6] C. K. Reddy, R. Cutler, and J. Gehrke, "Supervised Classifiers for Audio Impairments with Noisy Labels," in *Interspeech 2019*, 2019, pp. 2573–2577. [Online]. Available: https://arxiv.org/pdf/1907.01742.pdf

[7] A. Hines, E. Gillen, and N. Harte, "Measuring and monitoring speech quality for voice over ip with polqa, visqol and p. 563," in *Proc. Interspeech 2015*, 2015.

[8] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Monitoring the effects of temporal clipping on voip speech quality." in *Proc. Interspeech 2013*, 2013, pp. 1188–1192.

[9] A. Ragano, E. Benetos, and A. Hines, "Adapting the quality of experience framework for audio archive evaluation," *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), IEEE*, 2019.

[10] A. Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley-IEEE Press, 2012.

[11] E. Marchi, F. Vesperini, S. Squartini, and B. Schuller, "Deep Recurrent Neural network-based Autoencoders for Acoustic Novelty Detection," *Computational Intelligence and Neuroscience*, 2017.

[12] E. J. Humphrey, J. P. Bello, and Y. Lecun, "Feature learning and deep architectures: New directions for music informatics," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 461–481, 2013.

[13] Z. Akhtar and T. H. Falk, "Audio-Visual Multimedia Quality Assessment: A Comprehensive Survey," *IEEE Access*, vol. 5, pp. 21 090–21 117, 2017.

[14] E. Rushe and B. Mac Namee, "Anomaly Detection in Raw Audio Using Deep Autoregressive Networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3597–3601.

[15] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous sound detection using deep audio representation and a blstm network for audio surveillance of roads," *IEEE Access*, vol. 6, pp. 58 043–58 055, 2018.

[16] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, "Non-speech audio event detection," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 1973–1976.

[17] C. Senac, T. Pellegrini, F. Mouret, and J. Pinquier, "Music feature maps with convolutional neural networks for music genre classification," in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*. ACM, 2017, p. 19.

[18] W. Zhang, W. Lei, X. Xu, and X. Xing, "Improved music genre classification with convolutional neural networks," *Interspeech 2016*, pp. 3304–3308, 2016.

[19] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[20] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proocedings of the 21st ACM international conference on Multimedia*, 2013, pp. 411–412.

[21] B. L. Sturm, "A survey of evaluation in music genre recognition," in *International Workshop on Adaptive Multimedia Retrieval*. Springer, 2012, pp. 29–66.

[22] ——, "The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[23] J. Pons, O. Slizovskaia, R. Gong, E. Gómez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," in *25th European Signal Processing Conference, EUSIPCO 2017*, 2017, pp. 2744–2748.

[24] J. W. Osborne and A. B. Costello, "Sample size and subject to item ratio in principal components analysis," *Practical Assessment, Research, and Evaluation*, vol. 9, no. 1, p. 11, 2004.

[25] D. J. Mundfrom, D. G. Shaw, and T. L. Ke, "Minimum sample size recommendations for conducting factor analyses," *International Journal of Testing*, vol. 5, no. 2, pp. 159–168, 2005.

[26] S. S. Shaukat, T. A. Rao, and M. A. Khan, "Impact of sample size on principal component analysis ordination of an environmental data set: effects on eigenstructure," *Ekológia (Bratislava)*, vol. 35, no. 2, pp. 173–190, 2016.